THE UNIVERSITY OF CHICAGO

NEURAL SIGNATURES OF STORAGE ISOLATE CORE MECHANISMS OF WORKING MEMORY

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

BY

HENRY MORROW JONES

CHICAGO, ILLINOIS

AUGUST 2026

Abstract

To successfully navigate through the world, we must be able to actively hold information in mind. From remembering your dual-factor authentication code to the name of someone you just met, working memory is a core component of everyday cognition, and individual differences in working memory have been associated with broad measures of intellectual function (e.g., IQ, scholastic achievement, and psychological well-being). This work investigates a recently identified neural signature that tracks the number of items in working memory, using it as a lens into the mechanisms underlying this critical process, and offers two major results. First, voluntary control over what gets stored in working memory is distinct from voluntarily directing attention towards specific regions of space. Even when observers direct spatial attention in the absence of external cues -- an act that requires sustained maintenance of information -- neural signatures of storage in working memory are not necessarily observed. This observation motivates a revision of standard definitions of working memory storage that focus only on information retention. Instead, I propose that WM storage is best defined as the item-based, capacity-limited storage system that drives the individual differences in intellectual ability described above. Second, WM storage mechanisms are applied across vastly different types of information and track behavioral capacity limits. This supports theories suggesting that WM storage relies on the abstract binding of items to a representation of the current event to enable selective maintenance and access to these items. I introduce an initial test of this theory, and close with a discussion of possible implementations of this theory in the brain. In summary, this work offers a framework for understanding the core components of working memory and the role they play in our ability to flexibly navigate everyday life.

For my family, and Anna. I love you, and this work would not exist with you.

Contents

List of Figures

List of Tables

Acknowledgments

Thank you to Edward Awh for being the nicest person in science, and an incredible mentor to boot.

Thank you to Edward Vogel & Monica Rosenberg for effectively serving as additional advisors for the duration of my PhD.

Thank you to David Freedman for being interested in bridging human and animal research to build a deeper understanding of the mind and brain.

Thank you to William Thyer, William Ngiam, Gisella Diaz and the rest of the Awh Vogel lab for sharing so much of your time to teach me how to be a scientist.

Thank you to my parents for always challening your sons to follow their passions; thank you to my brothers for not shying away from that challenge and giving me the space to do the same.

Thank you to Anna for wanting to spend every day with me, now and for the rest of our lives.

Introduction

How much information can we hold onto in a given moment? How is this information represented in the brain so that we can use it to form new associations in memory, solve problems, make decisions, or perform other abstract cognitive processes? These are central questions in cognitive neuroscience, as they tackle core mechanisms behind our ability to behave flexibly and adaptively in new and dynamic situations. One critical cognitive system underlying this flexibility is working memory (WM), an active, capacity-limited storage system designed to keep a privileged set of information accessible for a short amount of time (on the order of seconds) for further manipulation. The effective amount of information that an individual can retain in WM is highly predictive of a variety of positive life outcomes, including educational and career attainment as well as both physical and psychological well-being (Mashburn et al., 2023). Thus, learning how information is encoded into and maintained in WM is important for understanding flexible and adaptive behavior both in theory and in practice. The current proposal seeks to investigate the neural underpinnings of WM through the use of electroencephalography (EEG) to advance and interrogate theories of WM storage.

Multiple theories of WM storage have been developed over time based on a combination of behavioral and neural data. Studies of visual working memory (VWM) often rely on the change detection task, for which many variants exist. In the standard change detection paradigm, participants briefly view an array of objects (e.g., colored squares), maintain them in VWM across a brief delay, and then must judge whether a probed item has changed (e.g., a red square becoming a green square) (Luck & Vogel, 1997). Data from the change detection task and its

variants have revealed a sharp capacity limit of ~3-4 items (Cowan, 2001; Balaban, Fukuda, & Luria, 2019). This limited capacity, and the pattern of errors people make when they are asked to remember more information than they have capacity for, are two critical data points that theories of WM storage need to explain. For example, an early (and ongoing) debate in the field of VWM was whether this limit reflects the discrete maintenance of a subset of items in the display (so-called "slot" models), or whether it is driven by a more continuous resource which is increasingly spread thin with increasing numbers of items to remember (Alvarez & Cavenaugh, 2004; Awh, Barton & Vogel, 2007, Zhang & Luck, 2008; Bays, Catalao, & Husain 2009, Brady, Stormer & Alvarez, 2011, Adam, Vogel & Awh, 2017, Bays et al., 2024).

The examination of neural data has inspired mechanistic theories of VWM storage and capacity limits. Historically, research has focused on sustained patterns of neural activity that track the specific contents maintained in VWM (Fuster & Alexander, 1971; Fuster & Jervey, 1981; Funahashi et al., 1989; Goldman-Rakic, 1995; Harrison and Tong, 2009; Serences et al., 2009; D'Esposito and Postle, 2015; Rademaker et al., 2019). One of the primary results of these investigations is that content information in VWM is actively represented in the same sensory cortices that originally processed it. This has given rise to the sensory recruitment hypothesis, which states that WM content is stored in these same sensory cortices (for a recent review, see Adam et al., 2025). The sensory recruitment hypothesis is often linked to interference- and noise-based theories of WM capacity limits, which assume that the neural representations in primary sensory cortex interfere with one another, reducing the fidelity of representations, lowering the overall signal-to-noise ratio and leading to the behavioral patterns describes above (Bayes, 2014; Oberauer & Lin, 2017; Bouchacourt & Buschman, 2019; Schurgin, Wixted, & Brady, 2020).

A separate branch of investigations has focused on neural signals that track the amount of information in WM (WM load), rather than the specific content. These signals have been studied using functional magnetic resonance imaging (fMRI; Todd & Marois, 2004; Xu & Chun, 2009), as well as EEG (Vogel & Machizawa, 2004; Luria et al., 2016; Adam et al., 2020; Thyer et al., 2022). The most commonly studied neural signature is an event-related response from EEG, termed the contralateral delay activity, or CDA (Vogel & Machizawa, 2004). The CDA can be measured using a variant of the change detection task in which sets of items are presented in both the left and right hemifields, but participants are cued to attend to one side. The CDA is a lateralized negativity at the posterior electrodes that is contralateral to the attended hemifield and scales with the number of items participants are holding in WM. In addition to the CDA, more recent work relies on multivariate classifiers, which are statistical models that can be trained to detect and discriminate between patterns in neural data, such as how neural activity differs when one vs two items are held in WM (Adam et al., 2020; Thyer et al., 2022). These multivariate techniques are useful because they can be trained using all electrodes, rather than just the posterior set, and avoid lateralized designs, making them sensitive to processes that aren't lateralized in the brain. In addition, they are sensitive enough to accurately classify WM load above chance at the level of individual trials (Adam et al., 2020). By varying the set size, or the number of items participants need to remember, from trial to trial, studies of the CDA and multivariate classification find signals that increase monotonically until an individual's behavioral capacity limit, ~3 items, and then either plateau or even slightly decrease at supracapacity set sizes (Vogel & Machizawa, 2004; Fukuda, Woodman & Vogel, 2015). In addition, these WM load signals appear to be independent of the specific content being maintained, as they are consistent across different types of remembered features (e.g., colored

squares, oriented lines, as well as colored orientations; Woodman & Vogel, 2008; Thyer et al., 2022). These results have been used to argue against purely content-based models of WM, such as the interference-based accounts above.

A final class of theories proposes that a major component of WM is its ability to bind items to the current context as a means of maintaining selective access to those items. The process of binding items to the current context is a computational mechanism that has been proposed in two distinct fields of cognitive neuroscience: long-term memory formation and attentional tracking of dynamic scenes. Within long-term memory formation, research has focused on how the hippocampus may be used to form item-context associations that underlie episodic memory (Eacott and Gaffan, 2005; Davachi, 2006; Ranganath, 2010; Knierim et al., 2006; Libby et al., 2020). For example, Libby et al. (2020) found that hippocampal activity discriminated between trials with either different contexts or items, but showed consistent codings of trials with both similar items and similar contexts, suggesting that the hippocampus can both reduce interference in memory between events with overlapping components and enable generalization and abstraction across highly similar events. Within the attentional tracking literature, researchers have proposed an indexing operation that enables the online tracking of items through time and space (Kahneman et al., 1992; Pylyshyn, 1989). The key idea is that item representations must be bound to a point in spacetime to maintain a coherent sense of the object's representation, even as low-level features (colors, brightness, size, shape, etc) consistently change throughout the ongoing scene. This spacetime index allows people to consistently update the object's low-level feature values without breaking the continuity of that object's identity. The index has been compared to demonstratives in language, like "this" or "that", which refer to an entity with clear informational content, despite lacking any content

information themselves. Indexing mechanisms have been incorporated into theories of WM as a means of maintaining and selectively accessing items in WM (e.g., Oberauer & Lin, 2017; Hedayati et al., 2022; Awh & Vogel, 2025). These theories posit that items need to be bound to context to be effectively accessed and propose that the capacity limit of WM is largely driven by an individual's ability to effectively allocate these indices, often referred to as pointers. In line with this, it has been proposed that the previously identified general load signals reflect the allocation of pointers (Thyer et al., 2022), rather than the specific content being pointed to.

The pointer theory is an attractive explanation of WM storage for a number of reasons. First, it offers an explanation of the seemingly content-independent WM load activity described above. Second, it explains previous findings that the CDA also appears to scale with the number of target objects in attentional tracking tasks (Drew & Vogel, 2008). Third, it still allows for item-level interference and differences in fidelity at the level of the content representations being maintained in sensory cortex, explaining dissociations between estimates of individuals' capacity and their memory precision (Awh, Barton & Vogel, 2007; Fukuda et al., 2010). Finally, it introduces a theoretical bridge between attentional control and long-term memory, helping to explain previous findings linking WM capacity to the rate of learning (e.g., Fukuda & Vogel, 2019). However, the theory is largely untested, and many questions remain unanswered. For example, can other content-independent processes, such as attention to spatial locations, explain the seemingly general load signal? Does it extend to different types of remembered information? Does this load signal, and the process underlying it, actually mediate WM performance? Can these load signals be explicitly linked to the process of binding items to context and spatiotemporal indexing?

The following chapters characterize the content-independent load signal through the questions outlined above. The main analytical technique is multivariate classification, though this is supplemented by other multivariate techniques. In Chapter 1, I assess whether the WM mechanism reflected by the load signal is distinct from the allocation of spatial attention, a common confound in studies of WM load signals. By independently manipulating set size and relevant spatial area, I show that both signals are present in the EEG data, but are dissociable. In Chapter 2, I assess the generality of the load signal, which has previously only been studied using visual features that are jointly coded in early visual cortices. Using features with neurally distinct representations, I show that the neural load signal is content-independent across both distinct visual features and distinct sensory modalities. In Chapter 3, I examine whether this WM load signal is behaviorally relevant. That is, whether it tracks the capacity limits identified in previous behavioral and neural work. To do this, I introduce a new analytical technique to isolate changes in the WM load signal from changes in spatial attention signals across set sizes. In doing so, I find that the WM load signal plateaus around set size 2 or 3, consistent with previous estimates of WM capacity. Notably, I also find evidence that spatial attention can be allocated, even in the absence of external cues, in cases where nothing appears to be stored in WM. The fact that the pattern of activity related to spatial attention matches standard definitions of working memory storage - the active maintenance of information to support cognitive processes - prompts us to refine our definition of working memory to focus on phenomena specific to it, including its capacity-limited nature. In Chapter 4, I investigate whether this load signal can be directly linked to spatiotemporal indexing. I compare the load signal for a standard WM task to that of a spatiotemporal tracking task that relies on illusions of motion to isolate the process of spatiotemporal indexing, divorced from the maintenance of any content. We find that the WM

load signal is not consistent with the tracking load signal, suggesting that distinct mechanisms are recruited for these different paradigms. I conclude by theorizing on productive definitions of working memory and possible mechanisms underlying this load signal, and future directions that can be taken to understand this core component of working memory.

Chapter 1: Electroencephalogram decoding reveals distinct processes for directing spatial attention and encoding into working memory

Abstract

Past work reveals a tight relationship between spatial attention and storage in visual working memory. But is spatially attending an item tantamount to working memory encoding? Here, we tracked electroencephalography (EEG) signatures of spatial attention and working memory encoding while independently manipulating the number of memory items and the spatial extent of attention in two studies of adults (N_1=39; N_2=33). Neural measures of spatial attention tracked the position and size of the attended area independent of the number of individuated items encoded into working memory. At the same time, multivariate decoding of the number of items stored in working memory was insensitive to variations in the breadth and position of spatial attention. Finally, representational similarity analyses provided converging evidence for a pure load signal that is insensitive to the spatial extent of the stored items. Thus, although spatial attention is a persistent partner of visual working memory, it is functionally dissociable from the selection and maintenance of individuated representations in working memory.

Introduction

There has been longstanding interest in our ability to exert voluntary control over attention to direct limited processing resources toward the most relevant aspects of the environment. But is voluntary attentional control a unitary process, or is it better understood as a constellation of distinct forms of control? We examined this question in the context of two well-known examples

of goal-driven selective attention. First, covert spatial attention can be voluntarily deployed to relevant regions of space and can modulate some of the earliest stages of sensory processing (Hillyard et al., 1998; Martinez et al., 1999). Second, people may selectively control which items enter working memory, an online memory system that allows us to store, manipulate, and rapidly access information (Cowan, 1999; Panichello & Buschman, 2021). Importantly, working memory gating operates during relatively late stages of processing; items can be excluded from working memory storage even after full perceptual and semantic processing (e.g., Chun & Potter, 1995; Luck et al., 1996; Vogel et al., 1998, Vogel & Luck, 2002). Although it has been established that attention can operate during both early and late stages of processing, the possibility remains that a single control process mediates both types of selection. For example, directing covert attention toward an object may simultaneously modulate sensory processing (Martínez et al., 2006) and encode that item into working memory. In line with this possibility, spatial attention is sustained at the location of items encoded into working memory even when location is irrelevant to the memory task (Foster et al., 2017), and disruption of spatial attention also disrupts working memory performance (Awh et al., 1998; Williams et al., 2013). This close partnership between spatial attention and visual working memory has motivated the perspective that storage in visual working memory is best understood as internally directed visual attention (Chun, 2011). Here, we present evidence that challenges this unitary model.

To study the relationship between spatial attention and working memory gating, we examined the neural signals that track each of these forms of selection. With scalp electroencephalogram (EEG) recordings, spatial attention can be tracked via measures of oscillatory activity in the alpha frequency band (8–12 Hz; see, e.g., Foster et al., 2017; Woodman et al., 2022), whereas the number of items encoded into working memory has been tracked via

univariate and multivariate analyses of raw voltage (Adam et al., 2020; Luria et al., 2016; Thyer et al., 2022; Vogel & Machizawa, 2004). Recent work has identified plausible dissociations among these signals. They explain distinct variance in individual differences in working memory performance (Fukuda et al., 2015). They respond differently to distractors, and with different time courses (Hakim et al., 2021). Finally, they respond differently to manipulations of the number of relevant locations, compared to the number of individuated items occupying those locations (Diaz et al., 2021; Hakim et al., 2019; Thyer et al., 2022). These dissociations indicate that spatial attention may be controlled separately from encoding into working memory, but these findings are limited in two ways. First, almost every study dissociating spatial attention and working memory load signals conflates the number of relevant items with the number of relevant locations (but see Diaz et al., 2021). In addition, the alpha measures in these studies were limited to univariate measures of power or laterality within posterior electrodes, precluding a precise link between those alpha oscillations and the deployment of covert attention to specific regions of space. Thus, we employed refined EEG measures of covert spatial attention that track both the position and breadth of the spatially attended regions, and we designed a task that deconflates the number of relevant items and the breadth of the relevant locations. As discussed below, these refinements provided strong traction for examining the separability of spatial attention and WM gating.

Two experiments provided clear evidence that distinct neural signals tracked the deployment of spatial attention and the maintenance of individuated items in working memory. In Experiment 1, we independently manipulated the number of relevant items and the number of relevant locations by sequentially presenting stimuli within overlapping or unique locations. We found that posterior alpha power was sensitive to the number of spatial locations, and not the

number of items, whereas a load-decoding model tracking the number of individuated items was insensitive to the number of locations occupied by those items. In Experiment 2, we used novel dot cloud stimuli that varied strongly in spatial extent and that could overlap with minimal perceptual interference. These stimuli enabled independent manipulation of the number of individuated items and the spatial extent of the relevant regions. Using inverted encoding models to decode both the location and precision of spatial attention (Foster et al., 2017), we found that spatial attention tracked the breadth of the spatial area occupied by the memorized items but was minimally impacted by the number of separate clouds in the display. Simultaneously, multi-variate decoding of the number of items in the display (mvLoad) precisely tracked the number of dot clouds held in working memory, despite strong variations in the spatial extent of those stimuli. Finally, we used representational similarity analysis (RSA) to assess the contributions of multiple independent factors that could influence our load decoding results. RSA provided converging evidence for a pure load signal that is uniquely determined by the number of individuated objects that are stored in working memory, as well as unique variance in EEG activity that tracked the spatial extent of the stored items. Thus, spatial attention and visual working memory storage are intertwined but distinct aspects of voluntary attentional control.

Experiments 1a and 1b

Materials & Methods

Participants.

Participants were recruited from the University of Chicago and the surrounding community. In total, 12 (9 female; mean age = 25 years, SD = 3.7) and 27 (18 female; mean age = 25 years, SD = 4.0) participants were used in Experiments 1a and 1b, respectively. We excluded data from 2

participants in Experiment 1a and 7 participants in Experiment 1b because of excessive EEG artifacts (< 200 trials remaining per condition in Experiment 1a and < 140 in Experiment 1b). For Experiment 1a, the intended sample size was 12 participants, because previous research had shown that this is a sufficient number of participants for observation of set-size effects on parieto-occipital alpha power (see Experiment 1 of Diaz et al., 2021). The 2 participants who were excluded from our analyses because of an insufficient number of trials were not replaced in favor of collecting data using a modified task design that balanced visual stimulation across displays (i.e., Experiment 1b). For Experiment 1b our intended sample size was 20 participants, because previous work had demonstrated that this is an appropriate number of participants for observation of more nuanced set-size effects on alpha power (see Experiment 2 of Diaz et al., 2021); we anticipated such effects when we introduced placeholders to balance visual stimulation across displays, because we were concerned that some participants might inadvertently attend the placeholders. Participants with excessive EEG artifacts were replaced until we completed our intended sample size.

Experimental procedures were approved by the Institutional Review Board at the University of Chicago. All participants gave informed consent and were compensated for their participation at a rate of $15 per hour. Participants reported normal color vision and normal or corrected-to-normal visual acuity.

Apparatus.

Participants were tested in a dimly lit, electrically shielded chamber. Stimuli were generated using MATLAB (The MathWorks, Natick, MA) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). Stimuli were presented on a 24-in. LCD monitor (refresh rate: 120 Hz,

resolution: 1080 x 1920 pixels) at a viewing distance of approximately 75 cm and against a dark-gray background.

EEG acquisition.

We recorded EEG activity using 30 active Ag/AgCl electrodes mounted in an elastic cap (actiCHamp, Brain Products, Munich, Germany). We recorded from International 10–20 sites Fp1, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, C3, Cz, C4, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, PO7, PO3, PO4, PO8, O1, Oz, and O2. Two additional electrodes were placed on the left and right mastoids, and a ground electrode was placed at position Fpz. All sites were recorded with a right-mastoid reference and were rereferenced offline to the algebraic average of the left and right mastoids. We recorded electrooculograms (EOG) using passive electrodes with a ground electrode placed on the left cheek. Horizontal EOG was recorded with a bipolar pair of electrodes placed 1 cm from the external canthus of each eye, and vertical EOG with a bipolar pair of electrodes placed above and below the right eye. Data were filtered online (low cutoff = 0.01 Hz, high cutoff = 80 Hz, slope from low-to-high cutoff = 12 dB/octave) and were digitized at 500 Hz using BrainVision Recorder (Brain Products, Munich, Germany) running on a PC. During preparation, impedances were set to be below 10 kΩ.

Eye tracking.

We recorded gaze position using a desk-mounted infrared eye-tracking system (EyeLink 1000 Plus, SR Research, Ottawa, Ontario, Canada). Gaze position was sampled at 1000 Hz. Stable head position was maintained during the task using a chin rest. The eye tracker was recalibrated

as needed throughout the session, including whenever participants removed their chin from the chin rest.

Artifact rejection.

For artifact rejection, each trial was segmented into 400 ms pretrial and 1,550 ms poststimulus array onset epochs. We used an automated procedure to flag trials that were contaminated by ocular or EEG artifacts. Next, we used this procedure as a guideline during manual visual inspection where it was ultimately determined which trials were to be rejected. Experimenters were blind to condition when inspecting the data for artifacts. Trials contaminated by artifacts were excluded from EEG analyses but not from behavioral analyses. Participants were excluded from the final sample if they had fewer than 200 artifact-free trials per condition in Experiment 1a (average trials per condition = ~277 out of 320) and fewer than 140 artifact-free trials per condition in Experiment 1b (average = ~198 out of 240).

An automated artifact-detection procedure was used to detect eye movements, blinks, and EEG artifacts. Trials were flagged as containing a saccade if the Euclidean vector between the mean gaze positions in the first and second halves of an 80-ms sliding window (advanced in 10-ms increments) was greater than 0.5° of visual angle. When eye tracking data were not available, we used horizontal EOG to detect saccades. Trials were flagged as containing a saccade if the mean voltage during the first and second halves of a 150-ms sliding window (advanced in 10-ms steps) exceeded 20 μV.

For blinks, trials were flagged as containing a blink if the eye tracker could not detect the pupil at any point during the trial. When eye-tracking data were not available, we used vertical EOG to

detect blinks. Trials were flagged as containing a blink if the mean voltage during the first and second halves of a 150-ms sliding window (advanced in 10-ms steps) exceeded 30 μV.

For EEG artifacts, we flagged trials as containing voltage drifts (e.g., skin potentials) if the absolute change in voltage from the first quarter of the trial to the last quarter of the trial exceeded 100 V. We flagged trials as including a sudden step in voltage (which can occur when an electrode is damaged) if the mean volt-age during the first and second halves of a 250-ms sliding window (advanced in 20-ms increments) differed by more than 100 μV. We marked trials as containing high-frequency noise (e.g., muscle artifacts) if any electrode had a peak-to-peak amplitude greater than 150 μV within a 15-ms sliding window (advanced in 50-ms increments). Finally, we flagged trials as containing amplifier saturation if any electrode had 60 time points within a 200-ms sliding window (advanced in 50-ms increments) that were within 1 μV of each other.

Experiment 1a procedure.

Participants performed a change-detection task (Fig. 1). The trial began with a black fixation dot (diameter = 0.20º) presented at the center of a dark-gray background for a randomly determined duration between 600 ms and 1,500 ms. The fixation dot remained visible throughout the trial. A stimuli array followed consisting of one or two colored squares (set size two or four, respectively; length = 2º) presented for 150 ms. The stimuli were presented within a predetermined area (Experiment 1a: 9.90° x 9.90°; Experiment 1b: 12.5° x 12.5°) and at least 2.1° (Experiment 1a) or 3° (Experiment 1b) away from fixation. The predetermined area was divided into a 4 x 4 grid; each of the sections could contain a single stimulus. The positions needed for the first array were randomly chosen from the 16 possible locations without

replacement. In the same-location condition, the same locations were used for the second array. Otherwise, new locations were randomly selected from the set of possible locations remaining that were not already used in the first array. Jitter (~0.25º) was added to the actual locations occupied by stimuli.

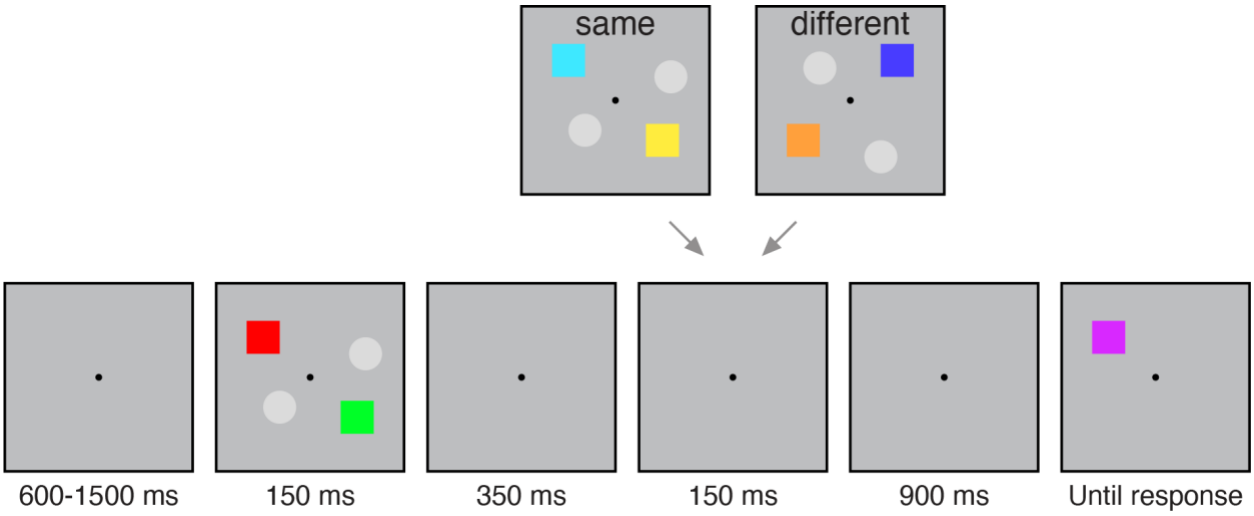

Figure 1. Example trials for sequential change detection task. Gray placeholders were not presented in Experiment 1a. For both Experiment 1a and 1b, two stimuli arrays were separated by a brief interstimulus interval. The first stimuli array contained either one or two colored squares depending on the Set Size, two or four respectively. The second stimuli array also contained one or two squares that were either presented in the same locations occupied in the first array or in different locations.

The stimuli in the first array were rendered in a color drawn randomly from nine possible colors without replacement (red, green, blue, yellow, magenta, cyan, white, black, and orange). Next, an interstimulus interval followed in which only the fixation dot remained on the screen for 350 ms. Then the second stimuli array followed, consisting of either one or two squares to

complete the set size to be maintained for that trial (set size two or four, respectively). The stimuli in this second array were either presented in the same locations that were previously occupied or in new locations (same or different locations, respectively).

The colors for the stimuli presented in the second array were determined randomly from the colors remaining and excluded the colors used in the first array. Participants were asked to remember the color and locations of the stimuli over a 900-ms blank delay interval during which only the fixation dot remained on the screen. After the delay, a single probe stimulus reappeared in one of the locations that was previously occupied and was either rendered in the same color as one of the original stimuli presented there or in a different color drawn randomly from the entire set of colors, excluding the color that was actually presented there. This meant that when the locations of the stimuli were the same between the first and second array (same-location condition), the set of possible change colors excluded both colors presented in the probed location. Participants used a keyboard button press to indicate whether this probe stimulus was the same or not. Participants pressed the "z" or "/" key to indicate whether the color of the probe stimulus was the same or different, respectively. There were no practice trials given before the formal experiment. Participants were given verbal and written task instructions with the aid of an example trial image similar to that of Figure 1.

Participants completed 20 blocks with each containing 64 trials. Within a block, half of the trials were no-change trials, and the remaining half were change trials in which the probe stimulus was rendered in a different color than the one (or any) presented there. Similarly, half of the trials were set size 2 (SS2); the first array contained one square and the second array contained an additional square. The remaining half were set size 4, with two squares presented in each of the first and second arrays. Within each block, there were also an equal number of trials

using the same location and different locations that determined whether the stimuli in the second array were presented in the same locations as the first array or not. Finally, the probe stimulus was drawn from the first and second array with equal probability.

Participants self-initiated each block by pressing the space bar key. The experiment session was scheduled to take 3 hours, but the actual duration of the session depended on the participants' pace because they initiated each block and decided when (and if) to take breaks between blocks.

Experiment 1b procedure.

The procedure was similar to that of Experiment 1a, with the following exceptions. First, an additional condition was included that simultaneously presented two or four squares in the first array, though the data from this condition are not analyzed further. Additionally, gray placeholder circles (diameter ~2.26º) were presented in both the first and second arrays so that four items were always presented. For set size 2, this meant that each array contained three placeholders, whereas set size 4 trials contained two placeholders in each array. On each trial, four positions were randomly chosen from the set of possible locations. Depending on the set size, positions were assigned to stimuli and place-holders. In same-location-condition trials, the locations of the gray placeholders (and stimuli) were the same for both arrays. The locations were switched on the different-location-condition trials so that the stimuli were placed in the placeholder locations from the first array, and the placeholders were placed in the stimuli locations from the first array.

Participants completed 20 blocks with each containing 72 trials. The experimental session was scheduled to take 3.5 hours, but the actual duration of the session depended on the

participants' pace because they initiated each block and decided when (and if) to take breaks between blocks.

Experimental design.

Both experiments used a 2 x 2 within-participants design. The factors were set size (2 or 4) and location condition (same or different locations). Behavioral data (i.e., accuracy) were analyzed using a repeated-measures analysis of variance (ANOVA), computed in JASP version 0.18.1 (Love et al., 2019). For this and all ANOVAs, we also computed a Bayesian repeated-measures ANOVA in JASP. All models were given an equal prior probability, and the change in model odds from prior to posterior (BFM) of the winning model is reported, along with the Bayes factor of the winning model against the null model (including only participant intercepts and random slopes).

Parieto-occipital alpha power analysis.

EEG signal processing was performed in MATLAB. We band-pass filtered the raw EEG data using a filter from the FieldTrip toolbox (ft_preproc_bandpassfilter.m; Oostenveld et al., 2011), and then extracted instantaneous power values for the alpha band (8–12 Hz) by applying a Hilbert transform from the Signal Processing Toolbox in MATLAB (hilbert.m) to the filtered data. We calculated alpha power for the parieto-occipital electrodes: P7, P3, Pz, P4, P8, PO7, PO3, PO4, PO8, O1, Oz, O2. For illustrative purposes in the figures, we subtracted the mean baseline (-400 ms to 0 ms) at each time point in the trial for each condition and converted to percent change from baseline.

Multivariate classification analysis.

For the mvLoad analysis, we used a logistic regression model to classify the number of items in working memory (i.e., working memory load) using baselined EEG (Thyer et al., 2022). EEG activity was calculated using a baseline from -500 ms to -100 ms relative to the onset of the stimulus array. The mean baseline amplitude was subtracted from EEG amplitude at each time point in the trial. To improve our signal-to-noise ratio, we randomly selected trials within each condition of interest to create groups of 20 trials and then averaged across the trials in each group. The classification procedure was performed by averaging voltage for each electrode across a 50-ms time window that advanced in 25-ms steps. At each time point the training data were standardized, and the testing data were standardized using the mean and standard deviation of the training data (StandardScaler in Scikit-learn; Pedregosa et al., 2011). The classifiers were trained to discriminate between our conditions of interest and then tested on a held-out set of data (StratifiedShuffleSplit in Scikit-learn). In this cross-validation procedure, the data for any given condition were split so that 80% of the data were used in training and the remaining 20% of data were used in testing. The training data were stratified by condition, and the number of trial groups per condition were equated by randomly down-sampling the condition with more trial groups. This procedure was repeated 1,000 times with results averaged across all iterations.

Statistical analysis.

Behavioral data were analyzed using a repeated-measures ANOVA as well as a Bayesian repeated-measures ANOVA. To increase the power of our EEG results, we averaged alpha and mvLoad results across the delay period for testing. To avoid bleedover from the test period in alpha, we excluded the last 100 ms; we did the same for the mvLoad results. We tested alpha

power using a repeated-measures ANOVA and a Bayesian repeated-measures ANOVA with a factor for set size (2 or 4) and a factor for location condition (same or different). We also planned to test whether the two conditions that differed in set size without differing in the number of locations—the set size 2 condition with different locations and the set size 4 condition with repeated locations—actually differed in terms of alpha power. We tested this using a repeated-measures t-test. For this and all t-tests, we also examined the Bayes factor (BF; computed with a standard Cauchy scale of .707) for evidence against the null (BF10), the reciprocal of which reflects evidence in favor of the null.

We tested decoding accuracy for each location condition separately using a repeated-measures t-test of decoding accuracy for true labels against decoding accuracy for shuffled labels in each model. In the second variation of mvLoad analysis, repeated-measures t-tests were used to test for differences in the classifiers' confidence between conditions of interest at each time window. Confidence scores refer to the signed distance of the test sample to the hyperplane in arbitrary units (decision_function in Scikit-learn).

Results

Behavior

For Experiment 1a, there was a main effect of Set Size, $F(1,9) = 61.90$, $p < .001$, $\eta_G^2 = .64$, and Location Condition, $F(1,9) = 8.80$, $p = .016$, , $\eta_G^2 = .048$, on accuracy, such that accuracy was higher for Set Size 2 (M = 0.98, SD = 0.01) than Set Size 4 (M = 0.87, SD = 0.06) and when different locations where occupied in the second array (M = 0.93, SD = 0.07) compared to the same locations (M = 0.92, SD = 0.07). There was a significant interaction between Set Size and Location Condition on accuracy, $F(1,9) = 8.17$, $p = .019$, $\eta_G^2 = .027$, such that the benefit of

appearing in different locations was greater for Set Size 4 than Set Size 2, perhaps due to a ceiling effect in Set Size 2. A Bayesian rmANOVA also supported a model with main effects of Set Size and Location, along with an interaction (BF_M=10.29, next highest =0.91; BF_{10}=8,889.37).

The pattern of behavioral results was replicated in Experiment 1b. There was a main effect of Set Size, $F(1,19) = 69.47$, $p < .001$, $\eta_G^2 = .41$, and a trending effect of Location Condition, $F(1,19) = 4.09$, $p = .057$, $\eta_G^2 = .007$, such that accuracy was higher for Set Size 2 (M = 0.95, SD = 0.06) than Set Size 4 (M = 0.82, SD = 0.10) and when different locations where occupied in the second array (M = 0.89, SD = 0.11) compared to the same locations (M = 0.88, SD = 0.10). Similar to Experiment 1a, there was a significant interaction between Set Size and Location Condition on accuracy, $F(1,19) = 9.49$, $p = .006$, , $\eta_G^2 = .005$, in which the benefit of appearing in different locations was only present for Set Size 4 compared to Set Size 2. A Bayesian rmANOVA also supported a model with main effects for Set Size and Location, along with an interaction (BF_M=11.19, next highest =0.75; BF_{10}=868,661.57).

Across both experiments, there was a reliable set size effect, such that performance was higher for Set Size 2 than Set Size 4. For Set Size 4, there was also a small benefit when the stimuli locations occupied in the second array where different than those used in the first array.

Parieto-occipital alpha power

We examined the effects of Set Size and Location Condition on alpha power at parieto-occipital electrodes (Maris & Oostenveld, 2007). Based on previous research, we predicted that alpha power would be sensitive to both factors (Diaz et al., 2021, Fukuda et al., 2015). Thus, we computed a repeated-measures ANOVA with Set Size and Location Condition as factors on the

average alpha power during the delay period (Figure 2). In Experiment 1a, there was a significant main effect of Set Size, $F(1,9) = 8.06$, $p = .019$, $\eta_G^2 = .0027$, as well as of Location Condition, $F(1,9) = 10.60$, $p = .010$, $\eta_G^2 = .0054$, and no significant interaction, $F(1,9) = .05$, $p = .82$, $\eta_G^2 = 3e\text{-}6$. A Bayesian rmANOVA also supported a model with main effects for Set Size and Location, along with an interaction (BF_M=5.03, next highest =1.06; BF_{10}=26.47). A planned t-test comparing set size 2 at different locations with set size 4 at the same location found no significant difference between the conditions, but weak evidence in favor of the null, $t(9) = 1.57$, p=.15, d=.04, BF_{10}=.787.

Experiment 1b produced similar results to Experiment 1a. There was a trend towards a main effect of Set Size, $F(1,19) = 3.5$, $p = .076$, $\eta_G^2 = .0019$, and a main effect of Location Condition, $F(1,19) = 5.09$, $p = .036$, $\eta_G^2 = .0023$, and no significant interaction, $F(1,19) = .20$, $p = .66$, $\eta_G^2 = 8e\text{-}6$. A Bayesian rmANOVA also supported a model with main effects for Set Size and Location, but no interaction, though evidence was weak (BF_M=1.66, next highest=1.21; BF_{10}=2.39). A planned t-test comparing set size 2 at different locations with set size 4 at the same location found no significant difference between the conditions, and moderate evidence in favor of the null, $t(19) = .30$, p=.76, d=.007, BF_{10}=.242.

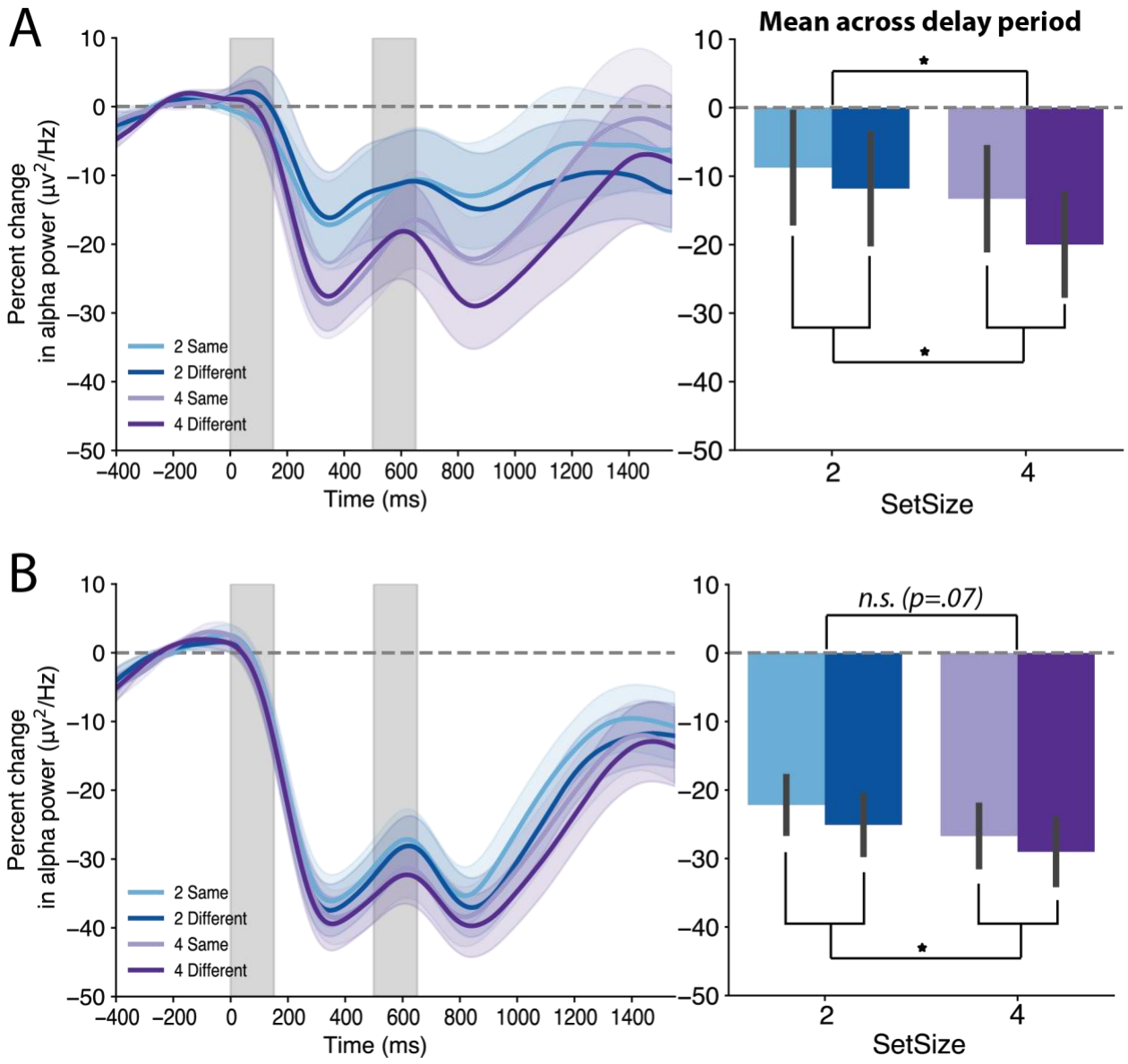

Figure 2. Averaged alpha power suppression observed at parieto-occipital electrodes in Experiment 1a (A) and 1b (B). Left, the timecourse of alpha power suppression across the trial. Shaded regions indicate duration of stimuli arrays. Right, the average suppression across the delay period. Significance markers above indicate a main effect of set size, and significance markers below indicate a main effect of location. There was no significant interaction in either experiment.

As predicted, alpha power at parieto-occipital electrodes appeared to be sensitive to Set Size. For both experiments, there was greater alpha suppression for four items than two items during the delay. However, Alpha power was also sensitive to whether items were added to the same locations or different locations. Specifically, there was greater alpha suppression when items were added to different locations for both Set Size 2 (Experiment 1) and Set Size 4 (Experiment 1 and 2). In addition, when we directly compared conditions which differed in the number of items but not the number of locations, we found weak-to-moderate evidence for no difference in alpha power, though it is possible that a set size effect is present in alpha power, and we lacked the statistical power to detect it. One possible interpretation for these findings is that alpha power tracked the shift in spatial attention required in the Different Location trials. However, we propose that the dominant signal in alpha power tracks the number of locations occupied by the memoranda. This latter interpretation is in line with previous work suggesting that alpha power at posterior electrodes also tracks the number of relevant locations when all items are presented simultaneously (Fukuda et al., 2015; Diaz et al., 2021). Thus, the effects of Set Size and Location Condition might both be explained by variations in the number of attended locations.

Multivariate analysis of voltage

Besides parieto-occipital alpha power, the number of items in working memory (i.e., WM load) can be decoded using multivariate analysis of EEG voltage, or mvLoad (Adam et al., 2020, Thyer et al., 2022). However, the extant work with this approach has always presented distinct items in unique locations, producing a confound between WM load and the number of relevant positions in the display. Thus, the central goal of Experiment 1 was to test whether the mvLoad

approach is sensitive to the total number of items stored when the confound between load and number of locations is eliminated. We accomplished this by presenting each half of the memory array sequentially, either with the second half in the same or different positions from the first half of the array. This design allowed us to manipulate the number of relevant positions while the number of stored items was held constant.

We used a logistic regression model to classify WM load using baselined EEG. For each condition of interest, trials were divided into groups of 20 and then averaged with the resulting matrix (electrodes × time points) subsequently referred to as a trial for ease of reference. The trials (-400 ms to 1550 ms) were divided into 50 ms time windows with a sliding window of 25 ms. Data were averaged within each time window, so that each trial was represented by a 30 × 80 matrix (electrodes × time windows). Finally, classification analyses were performed at each time window for each participant.

First, we investigated whether mvLoad could decode the number of items encoded into working memory. To this end, we trained and tested separate classification models to discriminate between set size 2 and 4 within the Same Location and Different Location conditions. We tested whether accuracy was above chance by averaging performance over the delay period and applying a repeated-measures t-test comparing accuracy using true labels to accuracy using shuffled labels (Figure 3). In Experiment 1a, we could classify Set Size across the delay period in both Location Conditions (Same: $t(9)=9.0$, $p=9e\text{-}6$, $d=4.1$, $BF_{10}=2370.29$; Different: $t(9)=6.67$, $p=9e\text{-}5$, $d=3.0$, $BF_{10}=308.63$). The results from Experiment 1b follow the same pattern (Same: $t(19)=6.48$, $p=3e\text{-}6$, $d=2.0$, $BF_{10}=6067.20$; Different: $t(19)=5.77$, $p=1.5e\text{-}5$, $d=1.7$, $BF_{10}=1568.96$).

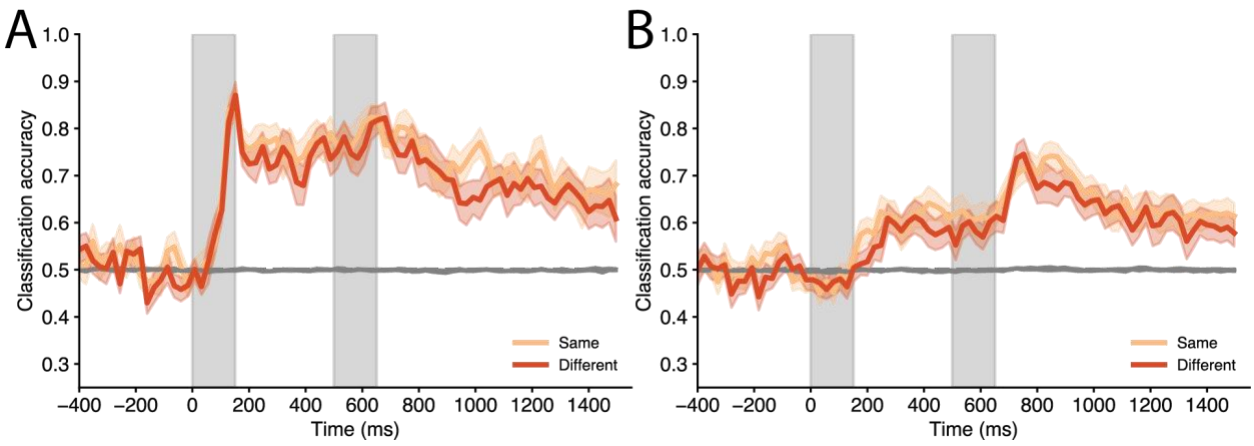

Figure 3. Classification accuracy over time for Set Size for Same Locations (light orange) and Different Locations (dark orange) in Experiment 1a (A) and 1b (B). Shaded regions indicate duration of stimuli arrays. Gray lines varying about 0.5 indicate accuracy for shuffled labels.

Although these findings are consistent with a WM load signal that does not depend on the number of relevant locations, this analysis is inconclusive because load was still confounded with the number of relevant locations. To examine whether there is a load signature that does not depend on the number of relevant positions, we tested whether the multivariate signature of load in the Same Location condition generalized to that in the Different Location condition. If the mvLoad analysis was classifying load based on the number of locations occupied by the memory items, then a classifier trained on the Same location conditions should be biased towards a *higher* set size when tested with data from the Different location condition. Likewise, a reliance on the number of occupied positions would lead a classifier trained on the Different location conditions to be biased towards a *lower* set size when it is tested on data from the Same location condition. The following results show that this was not the case.

We obtained a measure of the classifiers' confidence for each decision made about the test sample ('decision_function' in scikit-learn). This confidence score reflects the signed distance of the test sample to the hyperplane in arbitrary units. Positive scores indicated that the trial was classified as Set Size 4 (Same or Different depending on the analysis) with higher scores reflecting stronger evidence for this decision. Meanwhile, negative scores indicated a Set Size 2 (Same or Different) classification with lower scores reflecting stronger evidence. Our time window of interest was the delay period (starting at 650 ms after the initial sample onset, excluding the last 100ms to match our alpha power analyses above) given that our comparisons of interest relied on the total set size. To increase SNR, we averaged these distance values across the time window of interest, and tested for differences using a repeated measures t-test.

First, we discuss the analysis that used training data from the Different Location condition. As demonstrated above, multivariate analysis could distinguish the patterns of activity between Set Size 2 Different and Set Size 4 Different. The key question was whether or not the model trained on the Different Location condition would generalize to data from the Same Location condition. Recall that the Set Size 4 Same condition had the same number of items as the Set size 4 Different condition, but had the same number of relevant locations as the Set size 2 Different condition. Despite this, in both experiments the classifier trained on Different Location data (Figure 4) consistently classified Set Size 4 Same trials as significantly different from Set Size 2 Different (1a: $t(9)=5.57$, $p=.0003$, $d=2.85$, $BF_{10}=98.89$; 1b: $t(19)=5.75$, $p=1.5e\text{-}5$, $d=1.75$, $BF_{10}=1489.00$), but not significantly different from Set Size 4 Different, with weak-to-moderate evidence of the null (1a: $t(9)=.119$, $p=.91$, $d=.04$, $BF_{10}=.31$; 1b: $t(19)=1.62$, $p=.123$, $d=.55$, $BF_{10}=.71$). Thus, the patterns of activity between two items in two locations and four items in

two locations were discernibly and reliably different during the delay period, while the patterns of activity between four items in four locations and four items in two locations were not.

We conducted a complementary analysis after training the load classifier on data from the Same Location condition. In this case, we tested the model on data from the Set Size 2 Different Location condition, which had the same number of individuated items as Set Size 2 Same, but occupied the same number of locations as the Set Size 4 Same Location condition. Figure 5 shows that the mvLoad analysis could distinguish the patterns of activity between Set Size 2 Same and Set Size 4 Same, just as seen in the initial decoding accuracy results. Moreover, when the same load model was tested on Set Size 2 Different trials, the trials were more likely to be classified as Set Size 2 Same trials throughout the delay period in Experiment 1a, even though the number of locations in the Set Size 2 Different condition contained more locations than did the Set Size 2 Same condition (Figure 5A). The distances from the hyperplane were not significantly different between Set Size 2 Different and Set Size 2 Same, with weak-to-moderate evidence in favor of the null ($t(9)=.249$, $p=.81$, $d=.08$, $BF_{10}= .32$), suggesting that the patterns of voltage activity reflecting WM load were equivalent for two items in two locations and two items in one location (delay). In contrast, the distances from the hyperplane for Set Size 2 Different and Set Size 4 Same were significantly different ($t(9)=6.50$, $p=.0001$, $d=3.53$, $BF_{10}=260.63$). In Experiment 1b, the distances from the hyperplane for Set Size 2 Different trials were different for both Set Size 4 Same trials ($t(19)=4.76$, $p=.0001$, $d=1.67$, $BF_{10}=209.65$) and Set Size 2 Same trials ($t(19)=3.66$, $p=.0017$, $d=.96$, $BF_{10}=23.54$). From the decoding results across time (Figure 5B), it appears that Set Size 2 Different was initially similar to Set Size 2 Same through the beginning of the delay period, before moving towards the hyperplane. It is unclear whether this inconclusive finding reflects Set Size 2 Different becoming different from

both of the training conditions, or more similar to Set Size 4 Same. However, given that Set Size 2 is within capacity for most individuals, it is possible that participants inadvertently processed placeholders presented in the same location as a preceding target on Set Size 2 Different trials, resulting in an intermediate load signal. Experiment 2 controls for stimulus energy and spatial attention without placeholders, avoiding this potential confound.

Together, it appears the mvLoad analysis can discriminate the number of individuated items stored in working memory, regardless of the number of positions occupied by those items, though this decoding is not conclusive in all conditions.

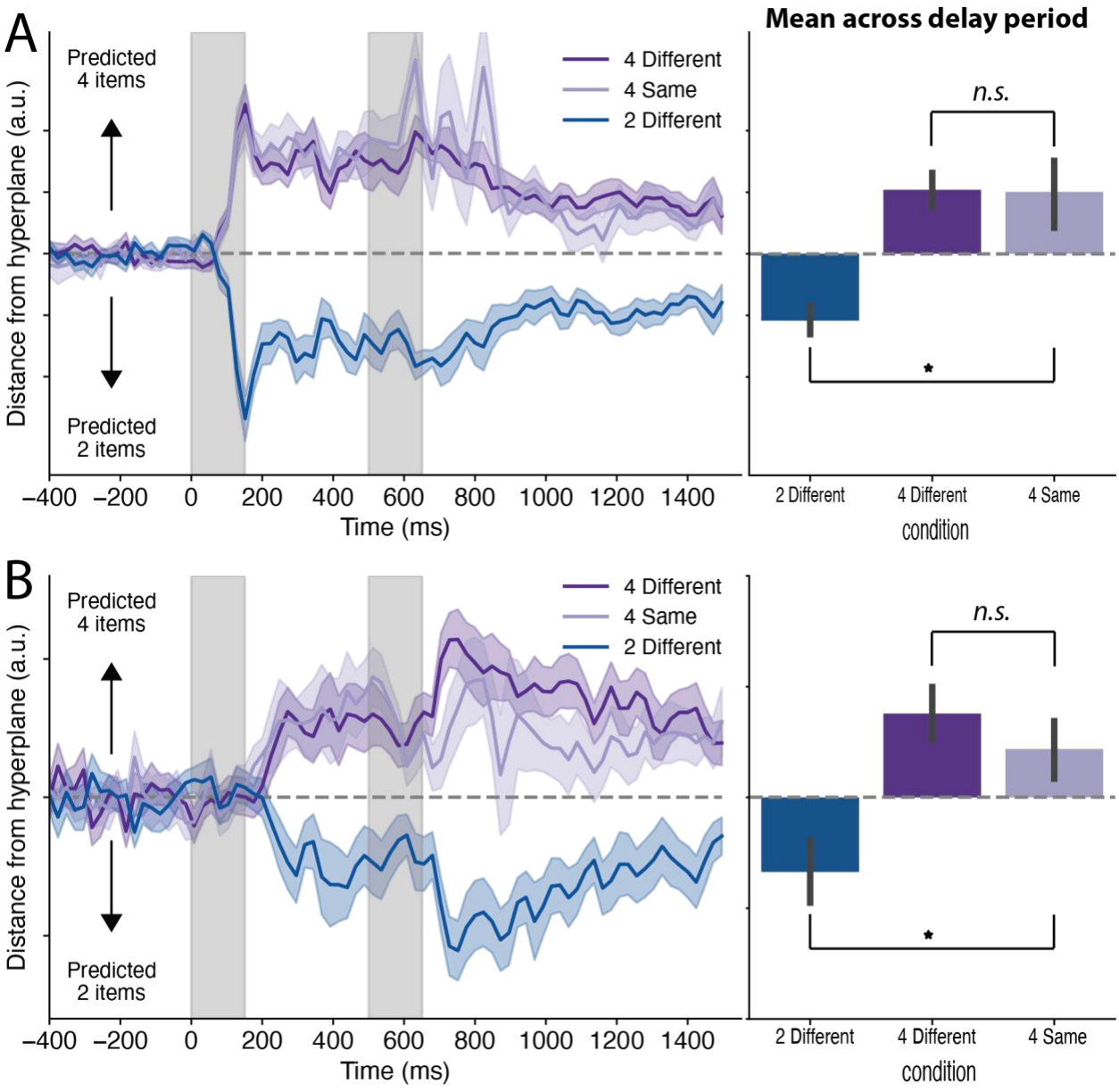

Figure 4. Distance from the classification hyperplane for Set Size 2 Different, Set Size 4 Different, and Set Size 4 Same trials across time for Experiment 1a (A) and 1b (B). Classification is trained on Set Size 2 and 4 Different trials and tested on all three conditions. Dashed gray line indicates hyperplane. Left, distances across time. Right, mean distances during the delay period, with significance markers comparing the Set Size 4 Same distances to those of the training conditions.

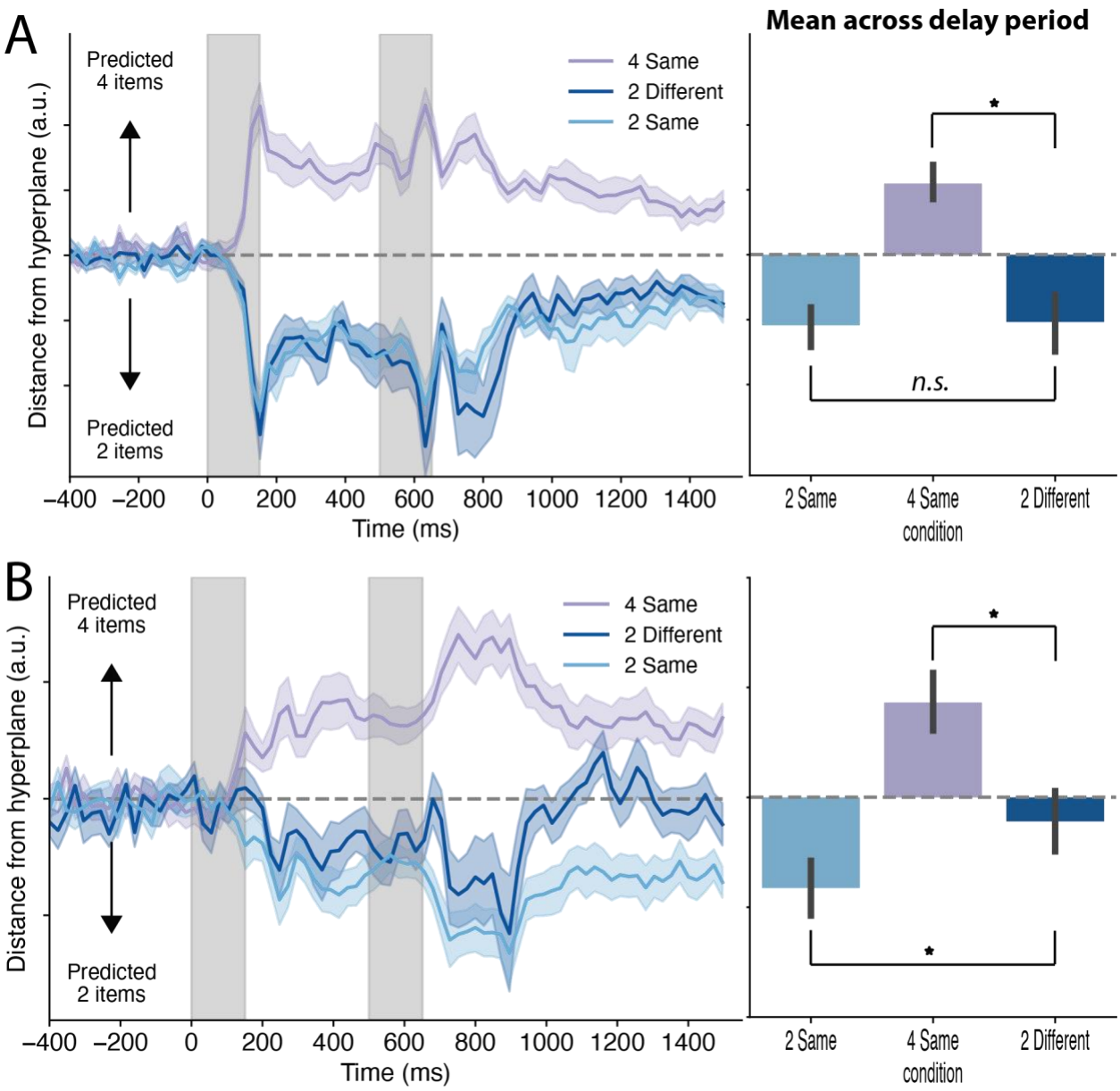

Figure 5. Distance from the classification hyperplane for Set Size 2 Same, Set Size 4 Same, and Set Size 2 Different trials across time for Experiment 1a (A) and 1b (B). Classification is trained on Set Size 2 and 4 Same trials and tested on all three conditions. Dashed gray line indicates hyperplane. Left, distances across time. Right, mean distances during the delay period, with

significance markers comparing the Set Size 2 Different distances to those of the training conditions.

Discussion of Experiment 1

In experiment 1, we found that alpha power was sensitive to the number of attended locations, but not the number of items stored in working memory. In contrast, mvLoad decoded the total number of items stored, regardless of the number of attended locations. In experiment 2, we extended this pattern of results with a refined modeling approach to quantify both the locus and *precision* of covert spatial attention. This allowed a more sensitive test of the influence of spatial attention on our multivariate measure of WM load. We employed a spatial change detection task using 'dot clouds' as stimuli. Previous work has shown that people can approximately enumerate overlapping sets of dots in parallel, with a capacity of around 3 sets (Halberda et al., 2006). Therefore, we anticipated that participants would be able to identify overlapping dot clouds as distinct objects to be maintained in working memory. These stimuli also removed confounds found in previous work using mvLoad. The total number of dots could be controlled to equate stimulus energy without requiring irrelevant placeholders. In turn, eliminating placeholders avoids a potential confound between the number of relevant targets and the number of irrelevant objects to be ignored. Finally, the dot clouds afforded strong variations in the spatial extent and position of each item, enabling a clear test of whether substantial changes in covert spatial orienting had any influence on multivariate measures of WM load.

Experiment 2

Materials & Methods

Participants

33 participants (25 female, 8 male; mean age = 25.6 years, SD = 3.7) were recruited for this study and were compensated $20 per hour for their participation. Four participants ended the session early due to scheduling conflicts, leaving 29 who completed a full data acquisition session. Of these, six were excluded for excessive eye movements or artifacts (see preprocessing and artifact rejection below), and one was excluded for chance performance in set size 2, leaving a total of 23 participants whose data was included for analysis (17 female, 6 male; mean age = 25.7, SD = 3.1). Our target sample was 20 participants, as we assumed the difference in precision of spatial attention when attending single narrow clouds compared to attended single broad clouds would be the smallest effect size. This comparison is a conceptual replication of Feldman-Wüstefeld & Awh, 2020, whose experiment 1 data included 22 participants.. We overshot our target as we completed pre-processing and rejection in batches.

Participants were between 18 and 35 years old, reported normal or corrected-to-normal visual acuity, and provided informed consent according to procedures approved by The University of Chicago Institutional Review Board. Participants were recruited via online advertisements and fliers posted on the university campus.

Apparatus

Participants were tested in a dimly lit, electrically shielded chamber. For the first 8 data sessions, stimuli were generated using PsychToolbox (Brainard, 1997, Pelli, 1997). For the remaining 25 data sessions, stimuli were generated using PsychoPy (Peirce et al., 2019). Participants viewed the stimuli on a gamma-corrected 24-in. LCD monitor (refresh rate = 120

Hz, resolution = 1,080 × 1,920 pixels) with their chins on a padded chin rest at a viewing distance of approximately 80 cm.

Stimuli

Displays consisted of a neutral gray background (RGB values = [127.5, 127.5, 127.5]) containing a central fixation point, and one or two sets of colored dots, referred to as dot clouds. In the initial acquisition sessions, the fixation point was a black (RGB values = [0,0,0]) cross spanning .3 visual degrees. In the remaining sessions, a modified version of a validated fixation point (Thaler et al., 2013) was used containing a black circle spanning .5 visual degrees across the diameter with a fixation cross matching the background gray overlaid on top, and a central black circle spanning .15 visual degrees.

To generate the dot clouds, locations were drawn from an imaginary circular grid, consisting of 40 radial columns and 10 circular rows. The innermost row began 1.75 visual degrees from the center of the screen, and the outermost row ended 5.25 visual degrees from the center of the screen. The 40 columns were divided into 8 location bins, each containing 5 columns and spanning 45 degrees. Dot clouds spanned either 1 or 3 bins (i.e., either 45 or 135 degrees). In the case of a dot cloud spanning 3 bins, its location is defined as the central bin. For a given dot cloud, individual dot locations are chosen by randomly sampling cells from the circular grid in the location bin(s), with the following constraints. To ensure that each cloud spans the bin completely, one dot each is assigned to a row in the outermost columns of its assigned bin(s), and the remaining dots are randomly distributed across the remaining cells. For the first cloud to be drawn (regardless of set size), a cell is reserved from each column that could potentially be an outermost column for the second cloud to be drawn. For the second cloud in set

size 2 trials, those reserved cells are used if they are in fact on an outermost column. Otherwise, the above process is repeated for the second cloud, excluding already selected cells, and without reserving any additional cells. The two cloud locations and sizes were independently drawn on each set size 2 trial.

Dots in the innermost row were .25 visual degrees, and grew by .01 visual degrees for each row further from fixation. Within their cells, dots were randomly slightly jittered with the constraint that they could not cross into an adjacent cell.

The number of dots for a given cloud was drawn from a uniform distribution on each trial. For set size 1 trials, the number ranged from 12 to 48 dots inclusive; in set size 2 trials, the number ranged from 12 to 24 inclusive for each cloud (a total of 24 to 48 dots). Dot clouds were either blue (RGB values = [0, 0, 255]) or a luminance-matched green ([0, 52, 0]). The color was randomly chosen on set size 1 trials, and the color of the probed cloud was randomly chosen on set size 2 trials, with the unprobed cloud being the other color.

While encoding displays could consist of 1 or 2 dot clouds, probe displays only consisted of a single cloud. On no-change trials, the probed dot cloud was redrawn identically to the encoding display. On change trials, the dot cloud could change in 2 ways. First, the cloud could shift by 1 bin, either clockwise or counterclockwise, while maintaining its spatial configuration. Alternatively, the cloud could change in size, shifting from 1 bin width to 3 bin width, or vice versa. In this latter case, the cloud was redrawn using the procedure described above, with the same number of dots as during the encoding display.

Task Procedure

Participants completed a change detection task, based on the spatial configuration of dot clouds (Figure 6). Trials consisted of an 250ms encoding display, in which participants would see 1 or 2 colored dot clouds placed among an imaginary ring encircling a fixation point (see above). Dot clouds were centered on 1 of 8 45 degree bins dividing the ring, and spanned either 1 bin (45 degrees) or 3 bins (135 degrees). In set size 2 conditions, the dot clouds' locations were independent, allowing for no, partial, or complete spatial overlap, depending on the location and width of the 2 clouds. During overlap conditions, the individual dots of the clouds were interleaved. Clouds would either be luminance-matched blue or green.

A delay period lasting 1000ms would follow the encoding display, after which participants would be probed on the spatial configuration of a cloud. On 50% of trials there was no change, and the probed cloud would be redrawn identically to its configuration in the encoding display. On the remaining 50% of trials, the cloud could change in one of two ways. On 50% of change trials, clouds could shift by 1 bin, either clockwise or counterclockwise. On the remaining 50% of change trials, clouds could change in width, either broadening from 1 bin to 3 bins, or narrowing from 3 bins to 1 bin. Participants used the left and right arrow keys to respond, with the mapping to change and no-change being randomly assigned for each participant.

Participants began with practice blocks consisting of 10 trials each in which they were given immediate feedback after each response, and they completed practice blocks until they comprehended the task. During the test phase, participants completed 16 blocks of 96 trials, for a total of 1,536 trials. After the first 8 acquisition sessions, participants received real-time feedback for eye movements made during the encoding and delay periods of the trial (see below). In cases where an eye movement was detected, the trial was aborted and a version of the trial (containing

the same high-level information, but with the exact dot locations randomly redrawn) was appended to a set of makeup blocks at the end of the session. Acquisition sessions lasted approximately 4 hours, including preparation.

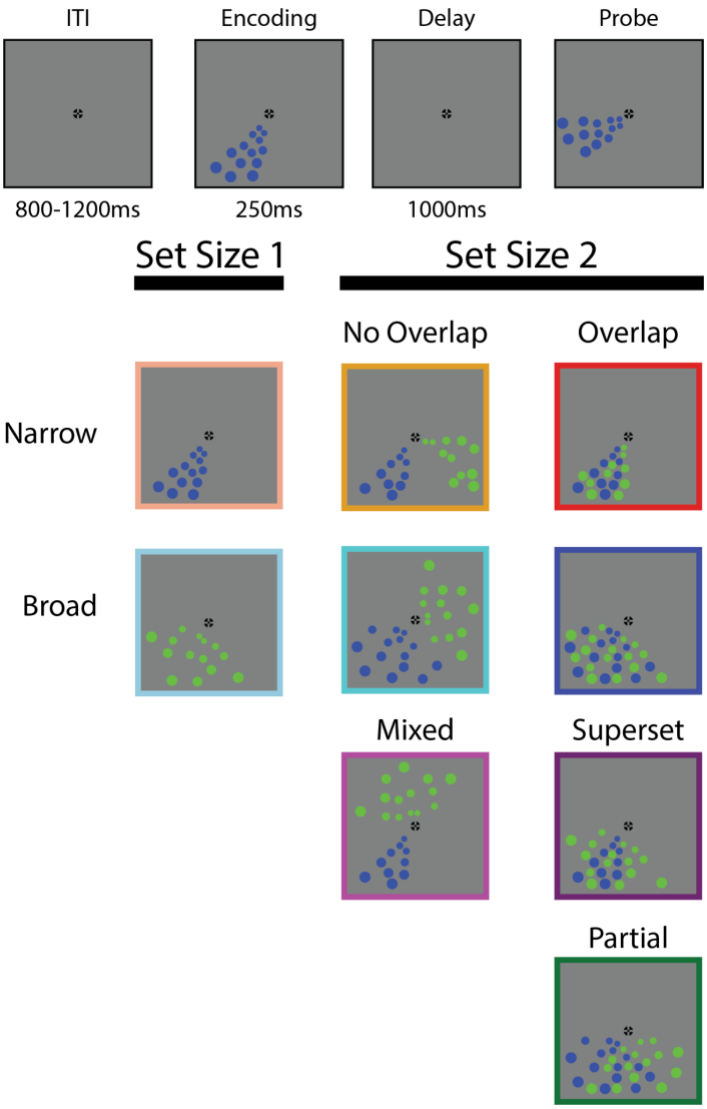

Figure 6. Top: illustrative schematic trial from the dot cloud change detection task. In this case, the cloud changed in position by shifting one bin clockwise. The alternative change cases are shifting one bin counterclockwise, and changing in width (from spanning 1 bin to 3 bins, or vice versa). Bottom: exemplars for each of the 9 conditions across cloud size, set size, and forms of

overlap in set size 2 conditions. The colored borders correspond to those used in the IEM and hyperplane results below. For illustrative purposes, the dots have been enlarged and the contrast between the colors has been amplified, relative to their actual display for the experiment.

EEG Acquisition

We recorded EEG activity from 30 active Ag/AgCl electrodes mounted in an elastic cap (Brain Products actiCHamp, Munich, Germany). We recorded from international 10-20 sites Fp1, Fp2, F7, F3, Fz, F4, F8, FT9, FC5, FC1, FC2, FC6, FT10, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1, Oz, and O2. Two additional electrodes were affixed with stickers to the left and right mastoids, and a ground electrode was placed in the elastic cap at position Fpz. All sites were recorded with a right-mastoid reference and were rereferenced off-line to the algebraic average of the left and right mastoids. We recorded electrooculogram (EOG) data using passive electrodes, with a ground electrode placed on the left cheek. Horizontal EOG data were recorded from a bipolar pair of electrodes placed ~1 cm from the external canthus of each eye. Vertical EOG data were recorded from a bipolar pair of electrodes placed above and below the right eye. Data were filtered on-line (low cut-off = 0.01 Hz, high cut-off = 80 Hz, slope from low to high cut-off = 12 dB/octave) and were digitized at 500 Hz using BrainVision Recorder (Brain Products, Munich, Germany) running on a PC. Impedance values were brought below 10 kΩ at the beginning of the session.

Eye Tracking

We monitored gaze position using a desk-mounted EyeLink 1000 Plus infrared eye-tracking camera (SR Research, Ontario, Canada). Gaze position was sampled at 1000 Hz.

According to the manufacturer, this system provides spatial resolution of .01° of visual angle and average accuracy of 0.25 to 0.50° of visual angle. We calibrated the eye tracker before each test block and between trials during the blocks if necessary. For participants run using PsychoPy, we ran a drift correction procedure every 6 trials[1]. We drift-corrected the eye-tracking data for each trial by subtracting the mean gaze position measured during a 200-ms window from -250 ms to -50 ms before the encoding display.

Artifact Rejection

Our preprocessing pipeline combines functions from EEGLab (Delorme & Makeig, 2004), ERPLab (Lopez-Calderon & Luck, 2014), and custom MATLAB scripts. We segmented the EEG data into epochs time-locked to the onset of the memory array (250 ms before until 1,650 ms after stimulus onset). We baseline-corrected the EEG data by subtracting mean voltage during a 200-ms window from -250 ms to -50 ms before the encoding display, and examined the epoch from the beginning of the baseline period (-250 ms) to the onset of the test display (1250 ms) for artifacts. Eye movements, blinks, blocking, drift, and muscle artifacts were first detected by applying automatic criteria. After automatic detection, we visually inspected the segmented EEG data for artifacts (amplifier saturation, excessive muscle noise, and skin potentials) and the eye-tracking data for ocular artifacts (blinks, eye movements, and deviations in eye position from fixation), and we discarded any epochs contaminated by artifacts. Participants were excluded if they lacked 15 trials in any of the held-out test conditions (see below), resulting in 4 exclusions. On average, 9.5% of trials were rejected in the remaining 24 participants.

[1] For the first two participants using PsychoPy, this drift correction procedure was every 10 trials.

Eye artifacts

Using eye-tracking data, we rejected trials that contained ocular artifacts such as blinks, saccades, and deviations from fixation. Blinks were rejected as any time point without a value. Saccades were identified as a .5 visual-degree shift in location between the first half and the last half of a 80ms window, which was slid across the epoch in steps of 10 ms. Movements from fixation were identified as a 1 visual-degree shift from fixation at any time point. For participants with both eyes being tracked, an artifact had to be identified in both eyes. The first 8 participants had only 1 eye tracked, in which case an artifact just needed to be identified for that eye.
For 1 participant, eye-tracking data were not available, so EOG was used. Saccades and blinks were identified via a 20 μV difference between the first half and the second half of a 150 ms window, which was slid across the epoch in steps of 10 ms. Large movements were identified via a 50 μV difference from baseline at any time point.

EEG artifacts

We checked for drift (e.g., skin potentials) by fitting a line to each channel. Trials were excluded if the line had a slope greater than a certain threshold (slope = 75, minimal $r^2 = .3$). To check for muscle artifacts, we excluded trials with peak-to-peak activity greater than 100 μV in a 250-ms window that was moved across the epoch in 20-ms steps. We also excluded trials that differed by 60 μV between the first half and the last half of a 250-ms window, also by comparing peak-to-peak activity using a 250-ms window that was moved across the epoch in 20-ms steps. Lastly, we excluded trials with any value beyond a threshold of 100 μV.

Analyses

To decode working memory load and spatial attention, we relied on two machine learning based approaches: mvLoad (Adam et al., 2020, Thyer et al., 2022) and inverted encoding models (IEMs). Both correct and incorrect trials were used during training and testing (Feldmann-Wüstefeld & Awh, 2020, Thyer et al., 2022). For all analyses, set size 1 trials containing less than 24 dots were excluded. This ensured that the set size 1 and set size 2 trials examined in these analyses had the same range of dots (24-48) and central tendency (mean and median = 36), making the number of dots nondiagnostic of set size. Analyses were run using Python's Scikit-Learn package (Pedregosa et al. 2011), along with custom Python scripts.

Behavior

As this was an unspeeded task, we focused on accuracy measures. To support comparison with the neural results below, we excluded trials containing artifacts and set size 1 trials containing fewer than 24 dots. We began by testing whether the mean choice accuracy was above chance in both set size conditions using a two-tailed t-test. All between-condition comparisons were completed as two-tailed repeated measures t-tests. As in Experiment 1, we computed Bayes factors (with a standard Cauchy scale of .707) for all t-tests to provide complementary evidence against the null (BF_{10}), the reciprocal of which reflects evidence in favor of the null.

Inverted Encoding Models (IEMs)

To examine how spatial attention was impacted across conditions, we applied inverted encoding models. Previous work has fit IEMs to the topography of alpha power across the scalp to reconstruct a location-selective channel response function (CRF) reflecting the allocation of

attention across space (Foster et al., 2017). This technique has been used to decode the location spatially attended during the delay period of WM tasks, even when participants must internally transform the location in response to auditory cues (Günseli et al., 2022). In addition, researchers have found that the shape of the tuning function broadens with increasing number of relevant locations (Sutterer et al., 2019), and with increasing size of the relevant location (Feldmann-Wüstefeld & Awh, 2020).

We began by replicating these latter two findings. Both analyses relied on comparing the precision of the reconstructed channel response functions (CRFs) between two conditions, so the overall architecture of the two analyses is the same. For a given pair of conditions (e.g., set size 1 vs set size 2), we performed the following model fitting and testing procedure. First, we isolated activity within the alpha band (8-12Hz) for each trial by applying an FIR filter implemented by MNE (Gramfort et al., 2013). We then extracted the complex analytic signal of the isolated activity for each via a Hilbert transform and squared its complex magnitude to compute alpha power. Before this process, we set timepoints from the test period to 0 to avoid contaminating the end of the delay period. Within each trial, we grouped timepoints into non-overlapping 25-ms windows, and averaged alpha power within each time window to increase SNR. We completed the following analysis at each time window independently, using all electrodes.

We grouped trials within each condition by the location bin of the to-be-probed cloud. For clouds spanning 3 bins, this was the middle bin. Across the 16 groupings (8 locations x 2 conditions), we identified the grouping with the fewest trials and randomly trimmed the other groupings to match the minimum. This equated the number of trials across locations and conditions. From this, we randomly sorted the data for each condition into three folds, each containing one-third of the trials at each location for each condition, and averaged the data within

each fold. This produced a matrix of shape (number of locations) x (number of electrodes) x (number of time windows) for each of the 3 folds, for each condition.

Following the standard IEM approach, we assumed that alpha power at each electrode is the result of weighted contributions from various subpopulations of cells (hereafter referred to as channels). Each channel has a preferred location with a graded response profile, such that its activity is highest at its preferred location and decreases with increasing distance from its preferred location. We modeled each channel's response function as a half cosine raised to the 25th power, with one channel for each location bin. From this, we computed a response for each channel to each location. This produced a matrix of shape (number of locations) x (number of channels), containing the response of each channel for each of the locations.

In a k-fold procedure, we fit a linear regression model to 2 folds for each condition (4 folds total) to predict alpha power at each electrode from channel responses for each location, at each time point. At a given time point, this took the following mathematic form:

$$P = WC$$

Where P is the matrix of alpha power across the 4 folds of shape (number of electrodes) x (number of locations * 4), C is a matrix describing the channel responses of shape (number of channels) x (number of locations * 4), and W is a weight matrix of shape (number of electrodes) x (number of channels), which is found via the linear regression fitting procedure. W describes the individual contributions of each channel to alpha power at each electrode. Once fit, we inverted the weights of the model (W^{-1}) to produce a matrix that produces channel responses from the topography of alpha power. We applied W^{-1} to the held-out fold of each condition

separately to reconstruct channel response functions (CRFs) at each location in the held-out folds. Within a condition, we rotated the CRFs for each location to be centered on 0°, and averaged them together. We repeated this process for each of the 3 folds, and averaged the resulting CRFs together. Because some trials were randomly trimmed to balance the groupings, we repeated this whole process 100 times, with random trimming and randomly grouping into folds, and averaged the CRFs across the 100 permutations together.

We then computed the slope of each condition's CRF. To do so, We folded the CRFs in half and averaged pairs of values equidistant from the center (e.g., the estimated responses for channels preferring locations +45° and -45° from 0 were averaged together). We fit a line to this folded CRF and recorded the slope. To increase power, slope estimates were averaged across the delay period, excluding the final 100ms to avoid bleed over from the censored test timepoints. Slopes were compared with a repeated-measures t-test, along with a Bayes Factor to provide complementary evidence against or for the null.

In addition to replicating the previous two findings, we tested the slopes of 2 conditions which varied in set size, but not in the total attended area. Specifically, we repeated the above process to compare the CRF slope between set size one trials containing a broad cloud, and set size 2 trials containing one broad cloud (SS1Broad), with one narrow cloud superset within it (SS2Superset), both of which contained clouds covering 3 bins[2].

Multivariate load decoding (mvLoad)

[2] SS2 Broad also contains clouds spanning 3 bins, but there is much more data for SS2 Superset compared to SS2 Broad, allowing us to produce much more reliable slope estimates.

Set size was decoded within-participant. First, we replicated previous work decoding set size while ignoring conditional differences like cloud size and overlap. To increase the signal-to-noise ratio, we randomly assigned trials of the same set size into groups of 20 and averaged within each group. These grouped-and-averaged timeseries were the foundation of our testing and training procedure. As above, we divided the epoch into non-overlapping 25ms windows and found the average voltage for each electrode in each window.

We tested classification using a cross-validation procedure at each time point. At a given timepoint, we selected 80% of the grouped timeseries as training data, and reserved the remaining 20% for testing, stratifying by set size. We then trimmed the training data to equate the number of grouped timeseries for each set size, and then z-standardized the training set. We fit a logistic regression model to the training data to predict set size. We then rescaled the test data using the mean and standard deviation of the training set before applying the fitted model, and recording the accuracy of its predictions. Lastly, we generated an empirical null accuracy by testing the model on the same test data, but with randomly shuffled labels. This procedure was repeated 1,000 times (including the random grouping and averaging to improve SNR and the trimming to balance the training data) for each time point, for each participant, and the average accuracy and shuffled accuracy were recorded.

For the main decoding approach, we tested classification accuracy for each 25-ms time window using a one-tailed t-test, comparing the test accuracy against the shuffled test accuracy. To correct for multiple comparisons, we applied the Benjamini-Hochberg procedure to control the false-discovery rate (FDR) at .05. We also computed the Bayes Factor at each time window. However, as the BF_{10}s are aligned with the statistical results, we don't present them for simplicity.

Predicting load regardless of spatial conditions

To assess whether load can be decoded regardless of the number and shape of spatial envelopes, we fit a second logistic regression model at each 25-ms window. For this analysis, trials were labeled both by set size, and by spatial conditions, specifically the size of the dot clouds and overlap. Set size 1 trials were labeled as either narrow or broad. Set size 2 trials were labeled by cloud size (both narrow, both broad, or mixed), and whether there was complete overlap (both clouds the same size and in the same location), superset overlap (one cloud is broad and subsuming a second, narrow cloud), partial overlap (both clouds are broad and overlap in 1 or 2 bins), or no overlap. The no-overlap trials were distinguished by the size of the clouds, which could both be narrow, both be broad, or could contain one narrow cloud and one broad cloud.

We randomly selected groups of 15 trials to average together within each condition. For training, we selected set size 1 trials with broad clouds (SS1Broad) and set size 2 trials in which 1 cloud subsumed another (SS2Superset). Both of these trial types spanned the same spatial area (3 bins) and contained the same number of dots on average, so differences in these conditions should be driven by the number of perceived dot clouds held in WM. Note that this comparison would also work for SS1Broad trials and set size 2 trials with both clouds being broad and perfectly overlapping (SS2Broad), but there were much fewer SS2Broad trials than SS2Superset trials, making model fits much less stable.

To train the model, we randomly reserved 1 group from SS1Broad and SS2Superset each for test, and used the remaining groupings for each of the 2 conditions for training. If one condition contained more groups, then we would trim a random subset of groupings to balance

the two conditions before z-standardizing the final training set and fitting the model. We applied the fitted model to the 2 held-out groupings along with all groupings for the remaining conditions (e.g., SS1Narrow, SS2NoOverlap, etc.), standardized using the mean and standard deviation of the training set. For each test grouping, we recorded the distance from the model's classification hyperplane set by the training data. This process was repeated 1,000 times (including the random grouping and averaging within each condition), and the final distances were recorded for each participant in each time window.

To increase power, results were averaged across the delay period, using the same time window as above (250 ms to 1150 ms after stimulus onset). The distances of the training conditions from the hyperplane were compared against 0 with a two-tailed t-test. Each held out condition was compared against each training condition using a two-tailed repeated-measures t-test. We also computed bayes factors describing evidence against the null (BF_{10}) for each of these tests.

Representational Similarity Analysis

While decoding methods can be applied to distinguish between pairs of conditions, it can be difficult to determine what signals the model is seizing upon. For example (as will be discussed in the results section), it is possible that SS2Superset contains an additional signal reflecting a cognitive mechanism (or set of mechanisms) responding to overlapping clouds (i.e., a separation process, or interference processing). A model trained to separate SS1Broad from SS2Superset could make use of both (or either) the set size signal and this overlap-related signal to separate the conditions. To address this issue, we applied representational similarity analysis (RSA) to simultaneously estimate the influence of multiple signals on EEG voltage patterns.

RSA provides a means for exploring the similarity structure across conditions of an experiment, and the factors which contribute to said structure (Kriegskorte et al., 2008). The base assumption is that conditions that are more similar to each other should also produce more similar neural signatures, measured via some distance measure. To return to our example above, we can make distinct predictions on the structure of distances between conditions based on our two hypothetical signals. If there is a pure set size signal, we should expect all pairs of conditions which differ in set size to be more neurally distinct from each other, and pairs of conditions with the same set size to be more neurally similar to each other. If there is some overlap-related process, then pairs of conditions which either both lack or both contain overlap should be neurally similar to each other, and pairs of trials where one contains overlap and the other doesn't should be more neurally distinct. Critically, the influence of different factors is not mutually exclusive, and RSA provides a way to simultaneously estimate the contributions of multiple factors to the similarity structure of the neural signals. Thus, we tested for the presence of these signals using the following RSA procedure.

To increase SNR, we binned trials in non-overlapping 50-ms windows, and averaged the values within each window. The distance between each pair of conditions was measured using a linear discriminant contrast (LDC) within each participant, also known as cross-validated Mahalanobis distance. We chose LDC rather than another measure for two reasons. First, LDC has been shown to be highly reliable (Walther et al., 2016). Second, while previous work applying RSA has used correlation as a distance measure (Kiat et al., 2022), correlation is "scale-invariant". In other words, correlation would not distinguish between a vector representing some process (v), and a vector representing "twice" that process (2*v). However, it is very possible that changes in EEG activity between set sizes are reflected by scaling of a given pattern, rather

than a shift in topography. For example, the Contralateral Delay Activity (CDA) becomes increasingly negative with set size, at least until it plateaus around 3 or 4 items (Vogel & Machizawa, 2004). In addition, previous work using mvLoad has identified increasing negativity in posterior electrodes and increasing positivity in frontal electrodes with increasing set size (Thyer et al., 2022). Because our experiment is sub-capacity, it seems reasonable to assume that changes in set size could be reflected in scaling. LDC is a measure of the Euclidean distance between two conditions, and therefore can still detect changes in scale.

To compute the representational dissimilarity matrix (RDM) for a given participant and timepoint, we first randomly split the data into training and test sets. For each of the 9 conditions, half of the trials were assigned as training data, and half as test. For each condition, we found the mean train-set value and the mean test-set value at each electrode. To adjust for the covariance between electrodes, we computed the covariance within the training set by first demeaning each trial by its condition's mean, then computing the covariance across trials. This covariance estimate was regularized using the Ledoit-Wolf procedure (Walther et al., 2016, Ledoit and Wolf, 2004). The distance between each pair of conditions (i and j) was computed as:

$$d^2_{LDC}(i,j) = (m_i - m_j)_{Train} * \Sigma^{-1}_{train} * (m_i - m_j)_{Test}$$

Where mi and mj are the vectors of mean values for conditions i and j in the train or test sets, and Σ_{train}^{-1} is the inverse of the regularized covariance matrix (Walther et al., 2016). To produce stable distance estimates, we repeated this procedure with 10,000 train-test splits and averaged across the resulting RDMs for each participant. We computed an RDM for each participant at each time window.

We next compared each participant's empirical RDM against a set of 4 theoretical RDMs, enabling a quantification of the amount of variance in the EEG signal explained by each (putative) cognitive process. First, our *set size* RDM predicted that trials with the same set size should look more similar to each other, and distinct from trials with the other set size. Our *overlap-related process* RDM predicted that trials containing overlap should look similar to each other, trials without overlap should look similar to each other, and pairs of trials crossing these sets should look different. We also included two additional RDMs related to spatial attention. First, we included an RDM of the absolute differences in the *total amount of attended area* between each pair of conditions. This RDM reflects the assumptions that trials with the same attended area should look similar to each other, trials with different attended areas should look different from each other, and that difference should grow with the difference in size of attended areas (e.g. trials covering 6 bins and trials covering 1 bin should look more different from one another than trials covering 4 bins and trials covering 3 bins). Note that attended area also scales with the average density of dots, as the number of dots was equated across set sizes, so this RDM could also be considered to capture density. Lastly, we added an RDM for the *number of non-contiguous spatial locations* or spatial envelopes attended. This assumed that both set size 1 and set size 2 conditions with overlap contained only 1 relevant spatial location. For set size 2 conditions without overlap, we computed the expected number of non-contiguous locations, based on the sizes of the two clouds. For example, with two narrow clouds, they would be contiguous (number of locations = 1) on 2/7 of trials and noncontiguous (number of locations = 2) on 5/7 of trials, so the expected value is 12/7. This RDM was computed as the absolute difference in the expected number of noncontiguous attended locations between each pair of conditions. These 4 theoretical RDMs were minimally related, with the largest positive

correlation (r=.23) between the attended area RDM and the number of attended locations RDM, and the largest negative correlation (r=-.17) between the attended area RDM and the overlap-related process RDM.

At each time point, we used a rank regression procedure to account for the independent contribution of each theoretical factor (Iman & Conover, 1979, Kiat et al., 2022). For each condition, we computed the semipartial rank correlation by comparing the R^2 of a full model predicting the rank of empirical distances from the ranked distances of all theoretical factors and comparing this against the R^2 of a model excluding the condition. We then multiplied the square root of the difference in R^2s by the sign of the condition's coefficient in the full model to produce the condition's semipartial correlation. As in (Kiat et al., 2022), we choose rank correlation as we did not assume a linear relationship between any of our theoretical factors and the observed condition distances. We tested these correlations at the group level using a Wilcoxon sign-rank test against zero. We tested each time bin with FDR correction using the Benjamini-Hochberg procedure. We also tested each factor's semipartial correlation after averaging across the delay period for maximal power. We also computed Bayes factors for the Wilcoxon tests using the procedure outlined in van Doorn et al., (2020), and the publicly available R code. For the tests at each time window, we sampled 5,000 points from the posterior for each factor. As these were highly aligned with the statistical tests, we omit them for simplicity. For the tests using data averaged across the delay period, we sampled 10,000 points from the posterior for each factor.

Controls for eye movements

We examined whether our IEM analysis was detecting changes in eye position rather than the deployment of covert attention to the dot clouds. This analysis relied on the 22

participants with eye tracking data. For each participant, eye gaze was drift corrected using baseline data from -250ms to 50ms before stim onset. The average gaze location during the trial was found, and converted to a location bin. We then repeated the set size 1 vs set size 2 IEM comparison, using the eye-gaze bins as the target locations, rather than the cloud locations. Because eye-gaze bins are not controlled by experimenters, it's possible that individuals expressed a large amount of bias, favoring some locations over others. This would result in some bins rarely being set as the target, and overall a smaller amount of data being used in the model after trimming. This lack of data could potential explain any unreliable CRF reconstructions. To account for this, we also repeated the set size 1 vs set size 2 IEM comparison while setting the number of trials per cell equal to that of the eye-gaze model.

We also examined whether load could be decoded across the delay period using eye gaze information from participants. For each participant, we fit a logistic regression model to classify set size using all eye data available for that participant. The process was identical to the set size decoding process above. We set a delay-period decoding threshold of 60% accuracy for whether a participant might have informative eye movements, which identified 4 participants (mean delay accuracies: 71.7%, 65.8%, 62.0%, and 62.2%). We then repeated all of the above analyses excluding those 4 participants to confirm that no patterns of results changed.

Results

Behavior

Participants performed the task well, with an average accuracy of 91.9% (SD=6.42%). Participants were more accurate for set size 1 trials (mean accuracy = 96.7%, SD=5.34%) than

for set size 2 trials (mean accuracy = 88.7%, SD = .08%; difference = 8.1%, t(22)=6.67, p=1e-6, BF_{10}=17280, d=1.171).

In addition, within set size 2 trials, participants were worse for trials containing overlap (mean accuracy = 86.1%, SD = 8.7%) than trials without overlap (mean accuracy = 90.2%, SD = 8.1%, t(22) = 5.29, p=2.6e-5, BF_{10}=911.6, d=.482). It is worth noting that these trial types differ in a few ways. In addition to the presence or absence of overlap, they also differ in the total attended area, as clouds without overlap span more bins on average than those with overlap, and in the number of non-contiguous spatial locations that should be attended to. We will return to these differences in the sections on multivariate load decoding and representational similarity analysis. Table 1 provides a description of the mean accuracies and standard deviations across each condition.

Set Size 1		Set Size 2						
Narrow	Broad	Narrow Overlap	Broad Overlap	Superset Overlap	Partial overlap	No Overlap, Mixed	No Overlap, Narrow	No Overlap, Broad
97.0 (5.1)	**96.5** (5.7)	**91.9** (11.3)	**90.0** (12.4)	**86.2** (9.5)	**83.6** (9.1)	**89.4** (8.6)	**92.0** (7.8)	**88.5** (8.6)

Table 1. Mean accuracies and standard deviations for each condition.

Inverted Encoding Models

We began by replicating previous work, which found that spatial attention, as measured by CRF slope, is deployed less precisely as the number of items stored increases (Sutterer et al., 2019). We fit encoding models to the total alpha power set size 1 (SS1) and set size 2 (SS2) trials with an equal number of trials per location bin and set size. For set size 2 trials, the to-be-probed dot cloud was arbitrarily selected as the relevant location. We then inverted the models and applied them to held out trials for each condition separately, and reconstructed the channel response functions (CRFs). We folded each CRF in half and fit a line, recording the slope as a measure of precision of spatial attention. To increase power, we averaged slopes across the delay period within each participant, excluding the last 100ms to avoid bleed over of information from the test period. We found that CRFs for SS1 trials had a significantly higher slope than SS2 trials ($t(22)=5.66$, $p=1.1e\text{-}5$, $BF_{10}=2033.46$, $d=.872$; Figure 7, top).

We next assessed whether our task design was sensitive enough to capture previously identified changes in precision of spatial attention in response to the size of the relevant spatial area (Feldmann-Wüstefeld & Awh, 2020). We repeated the above analysis, but compared trials containing set size 1 narrow clouds (SS1Narrow) against trials set size 1 broad clouds (SS1Broad). In line with past findings, CRFs during the delay period were more precise with 1 narrow cloud than 1 broad cloud ($t(22)=3.32$, $p=.003$, $BF_{10}=13.20$, $d=.375$; Figure 7, middle).

One possibility is that this difference in the precision of CRFs is driven by cross-hemifield effects. Specifically, the broad clouds span the vertical midline when centered on location bins next to the midline (four of the eight possible locations) and may be more likely to recruit cross-hemisphere processing, while the narrow clouds never do. To test this, we split the CRFs for both conditions into locations for which broad clouds would span the vertical midline and locations for which they would not, before aligning, averaging, and computing the slope

(Figure S1). An ANOVA found a main effect of condition ($F(1,22) = 10.34$, $p=.004$, $\eta_g^2=.033$), but no main effect of location nor an interaction (both $Fs < 3.1$, $p > .05$). A Bayesian rmANOVA also supported a model with only a main effect for condition ($BF_M=2.54$, next highest =2.24; $BF_{10}=12.26$). Thus, broad clouds produce broader CRFs than narrow clouds at every location.

While previous analyses have suggested that spatial attention is deployed with less precision with increasing set size, the work has typically conflated the number of relevant objects with the number of relevant locations. In the current study, however, we had the opportunity to compare a pair of conditions that varied in set size while the relevant spatial envelope was held constant. We compared SS1Broad trials against set size 2 trials containing a "superset" overlap pattern, i.e., 1 broad cloud with 1 narrow cloud within 1 of the 3 bins occupied by the broad cloud (SS2Superset). We followed the above analyses, with the center of the broad cloud being the target location for the set size 2 superset condition. These conditions are matched on the total relevant spatial area, and differ only in the number of perceivable objects. Therefore, we predicted that IEMs should show no difference between these two conditions. We found that there was a trend towards SS2Superset slopes being steeper than SS1Broad slopes, but it was not significant ($t(22)=2.04$, $p=.053$, $BF_{10}=1.26$, $d=.211$; Figure 7, bottom). Note that the size of this marginal effect was very small, equaling about half that of comparing set size 1 narrow and broad clouds, and less than a quarter of the size of comparing set size 1 and set size 2 across all conditions.

Though the slopes between SS1Broad and SS2Superset were not significantly different when averaged across the whole delay period, visual inspection of the slopes suggests that the conditions started to diverge towards the end of the delay period (Figure 8). To test this, we

divided the delay period into the first and last half and ran a repeated measures ANOVA (rmANOVA) with 2 factors: condition (SS1Broad and SS2Superset) and delay half (first and last). While there was a main effect of delay half ($F(1,22) = 4.82$, $p=.039$, $\eta_g^2=.025$), the effect of condition was only trending towards significance ($F(1,22) = 4.28$, $p=.051$, $\eta_g^2=.010$), as was the interaction between delay half and condition ($F(1,22) = 3.13$, $p=.091$, $\eta_g^2=.0033$). A Bayesian rmANOVA provided different results, supporting a model with main effects for delay half *and* condition, though the evidence was weak ($BF_M=1.65$, next highest=1.41; $BF_{10}=2.92$). Together, it appears that the precision of spatial attention is minimally impacted by the number of dot clouds within a fixed spatial area. Thus, although multiple past studies have documented reduced spatial selectivity as the number of items encoded into working memory increases (e.g., Sutterer et al., 2019; Sprague et al., 2016), in line with past findings of a tight relationship between working memory and attention (Awh and Jonides, 2001; Awh, Vogel and Oh, 2006; Chun, 2011), our findings suggest that this empirical pattern may have been driven by changes in the spatial extent of relevant positions rather than by the effects of storing additional items.

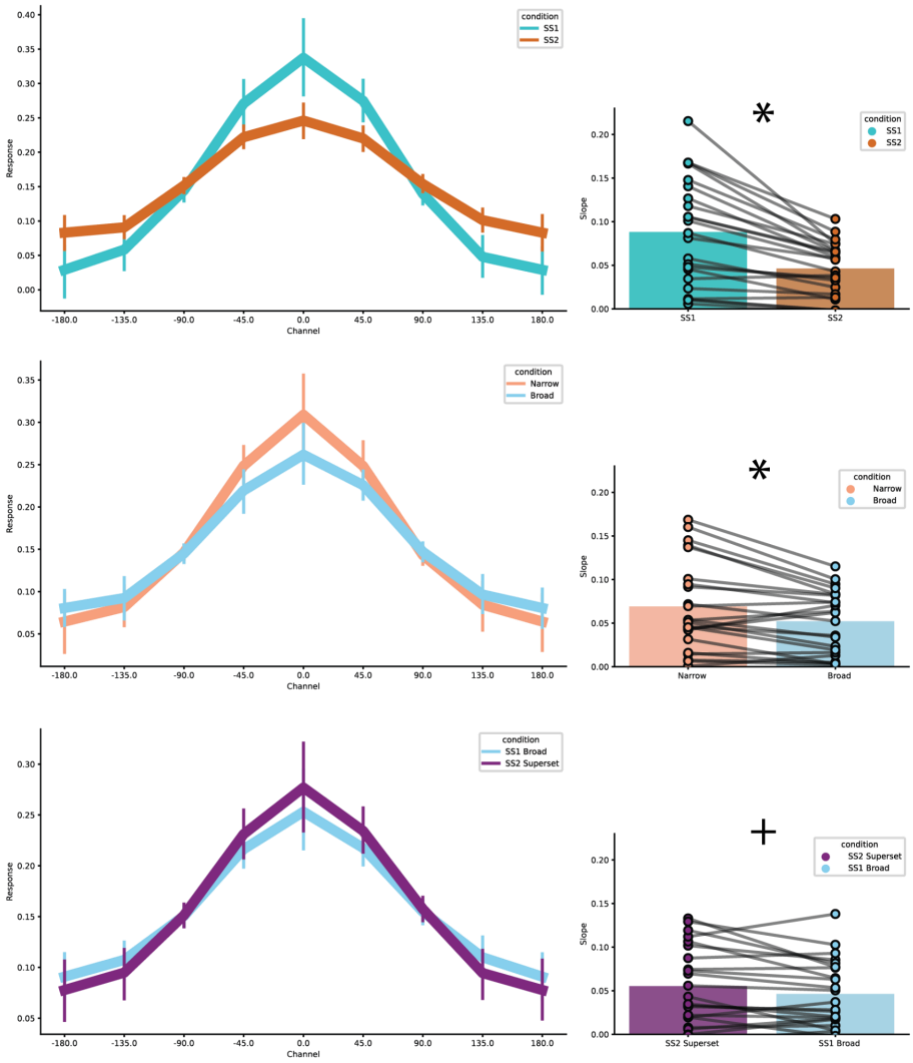

Figure 7. IEM results, averaged across the delay period (250ms-1150ms). Left, reconstructed CRFs. Right, average slope, with individual participants superimposed. Top, set size 1 and set size 2. Middle, narrow and broad clouds within set size 1. Bottom, set size 1 broad clouds and set size 2 superset clouds.

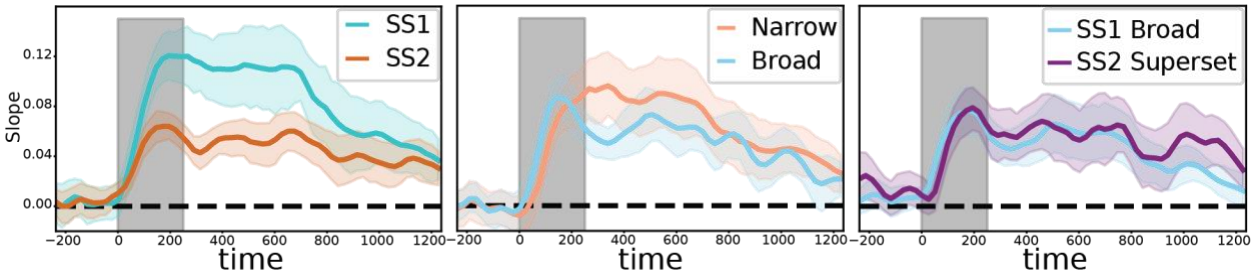

Figure 8. CRF slopes across time. Left, set size 1 vs set size 2. Middle, set size 1 narrow vs set size 1 broad. Right, set size 1 broad vs set size 2 superset. Shaded intervals reflect 95% confidence intervals of the slopes. Note that comparisons between conditions are within-participant comparisons, so the confidence intervals presented here do not indicate whether conditions are significantly different from one another.

Multivariate load decoding

As in previous work, we found reliable decoding of set size across the delay period (Figure 9; mean delay accuracy=.68, SD=.052). This decoding accuracy was significant despite matching the number of dots and luminance of the two colors, minimizing the impact of stimulus energy. However, other confounds remained. For example, set size two trials covered a larger spatial area on average, and the classifier could make use of this difference to separate the two conditions.

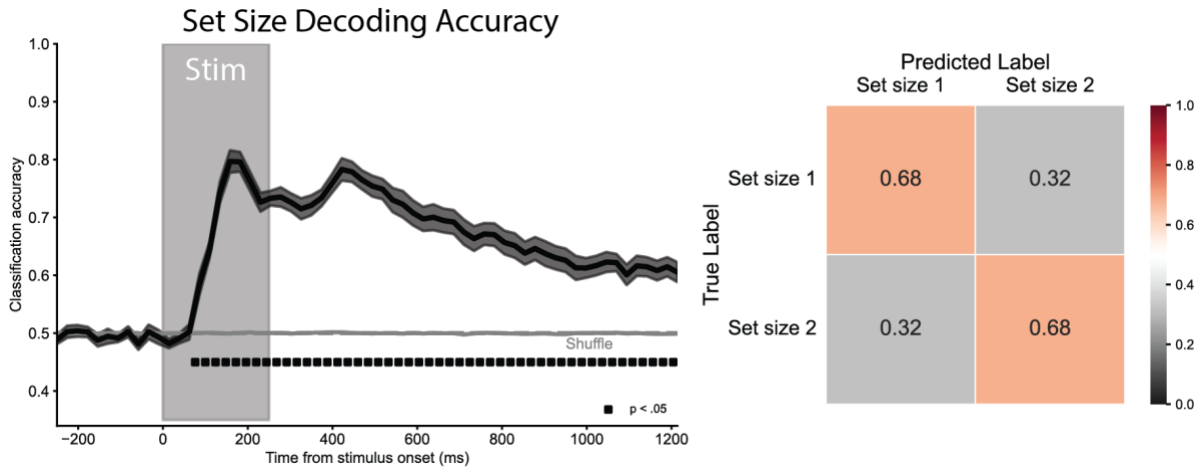

Figure 9. Set Size decoding accuracy across the trial. Left, the decoding accuracy across time, with significance values for each time window. Right, the confusion matrix for the model, averaged across the delay period.

To control for this, we tested whether a classifier trained on conditions with matched spatial areas could still decode set size, and compared its predictions to held out conditions varying in spatial area. First, we fit a model to predict load from SS1Broad and SS2Superset conditions, both of which spanned 3 location bins. We then found the predictions of the model for held out trials of these conditions by computing their distance from the model's decision boundary, and average these values across the delay period (Figure 10). We found that held out trials for both conditions were significantly far from the decision boundary, and in opposite directions of the boundary (SS1Broad: $t(22)=-8.55$, $p=1.9e-8$, $BF_{10}=69260.0$, $d=1.78$; SS2Superset: $t(22)=8.10$, $p=4.79e-8$, $BF_{10}=29730.0$, $d=1.69$). We next tested the predictions of the model to held out conditions by finding their distance to the hyperplane, averaged across the delay period. We tested the following conditions: SS1Narrow, set size 2 with two overlapping narrow clouds (SS2NarrowOverlap), set size 2 with two overlapping broad clouds

(SS2BroadOverlap), set size 2 with two broad, partially overlapping clouds (SS2PartialOverlap), and set size 2 with no overlap, broken into 3 size conditions: both broad (SS2NoOverlapBroad), both narrow (SS2NoOverlapNarrow), or one broad and one narrow (SS2NoOverlapMixed). We compared the distance estimates for each of these held out conditions against the test distances of the two training conditions.

We found that the mean distance across the delay period for SS1Narrow trials was significantly different from that of SS2Superset ($t(22)=5.67$, $p = 1.1e\text{-}5$, $BF_{10}=2072.12$, $d=2.03$), but not significantly different from that of SS1Broad ($t(22)=.38$, $p=.71$, $BF_{10}=.233$, $d=.085$). In contrast, the distances for SSS2BroadOverlap, SS2NarrowOverlap, and SS2PartialOverlap showed the opposite pattern: each was significantly different from set size 1 broad (all $ts > 5$, all $ps < .0001$, $BF_{10}s > 1000$), and not significantly different from set size 2 superset (all $ts < 1.6$, all $ps > .1$, $BF_{10}s < 1$).

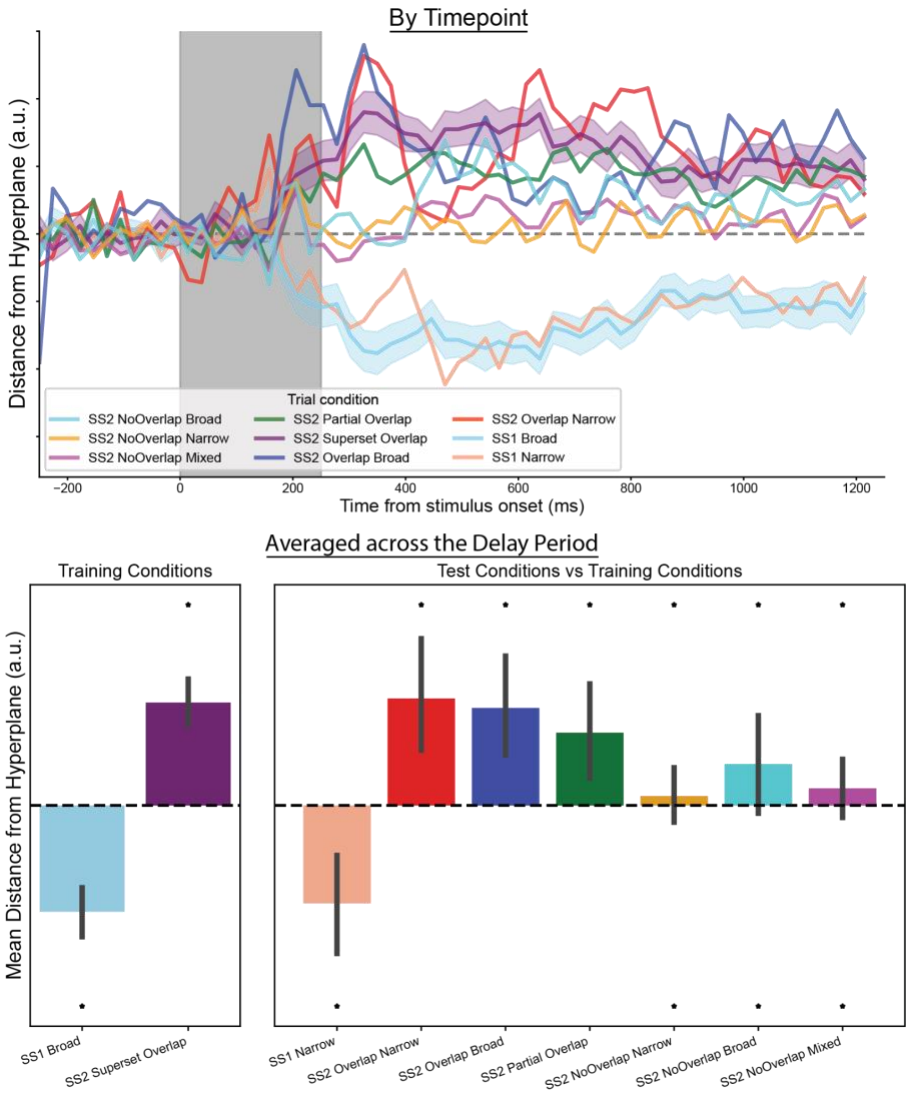

Figure 10. Distance from the Hyperplane across conditions. Top, timeseries of distances to test trials across conditions. 95% CI is included for the two conditions which were used as the training set. Bottom, the average distances from the hyperplane across the delay period. On the left side, asterisks indicate significant difference from the hyperplane (0). On the right side, asterisks indicate significance difference from the training conditions on the opposite side of the hyperplane. Asterisks above indicate difference between a given condition and SS1 Broad (i.e., it looks more set size 2 like). Asterisks below indicate significant difference between a given condition and SS2 Superset (i.e., it looks more set size 1 like).

This set of results suggest that the trained model can decode set size regardless of the size of the attended area. However, set size 2 conditions without overlap were not as cleanly decoded. All 3 conditions without overlap were numerically on the set size 2 side of the hyperplane, but all 3 conditions were significantly different from *both* training conditions (all ts > 2.5, and ps < .02). The pattern of distances was qualitatively identical between the first half (250-700 ms) and the second half (700-1150) of the delay period. Below, we offer a straightforward explanation for why these conditions were less confidently classified as set size 2, with an eye towards EEG signals that track spatial overlap.

Overall, we found that the classifier could separate set size regardless of attended area throughout the delay period. However, the distances of some held out set size 2 conditions were closer to the decision boundary than we predicted, and significantly different from both training conditions. There are a few possible reasons for a condition's predictions being near the hyperplane. First a condition may have too few trials, making it difficult to produce a reliable distance estimate. However, this is unlikely. Conditions with the fewest number of trials (SS2NarrowOveral, and SS2BroadOverlap, each making up 24 trials or 1.5625% of all trials, before artifact rejection) were as far from the hyperplane as the training conditions, which contained many more trials, while the two conditions closest to the hyperplane accounted for 168 trials and 240 trials, or 10.9375% and 15.625%.

One might argue that these results could be driven by differences in effort across trials. Under this explanation, participants used less effort for SS1Broad compared to SS2Superset trials and the remaining conditions are sorted based on the amount of effort participants

expended. While we cannot measure effort directly, we can look at task performance as a measure of the difficulty of each condition, which may modulate effort (and if it does not, then effort is not a confound). To assess whether difficulty (and therefore effort) impacted these results, we ran the following post-hoc analysis. The assumption is that, if the hyperplane results are driven by effort, then conditions with similar hyperplane distances should also have similar accuracies. We compared participants' accuracy between each training condition and those conditions whose distances from the hyperplane were not significantly different from that training condition. SS1Broad and SS1Narrow did not significantly differ in distance from the hyperplane (see above), nor in accuracy ($t(22)=1.36$, $p=.19$, $BF_{10}=.493$, $d=.0996$), potentially in line with the possibility that effort and difficulty are driving the hyperplane results, but this hypothesis does not bear out when considering comparisons against SS2Superset. In the hyperplane distances, SS2Superset was not significantly different from SS2OverlapNarrow ($p=.87$, $BF_{10}=.219$), SS2OverlapBroad ($p=.84$, $BF_{10}=.305$), nor SS2PartialOverlap ($p=.14$, $BF_{10}=.570$). However, accuracy between SS2Superset and these conditions was either trending towards significantly different (against SS2PartialOverlap: $t(22)=1.89$, $p=.072$, $BF_{10}=.99$, $d=.275$), or was significantly different (against SS2NarrowOverlap: $t(22)=-2.52$, $p=.02$, $BF_{10}=2.84$, $d=.542$; against SS2BroadOverlap: $t(22)=-2.3$, $p=.03$, $BF_{10}=1.93$, $d=.351$). Thus, it does not appear that the hyperplane results can be explained by variations in difficulty or effort.

Ruling out imprecision due to low trial count and the combination of effort and difficulty, two possibilities remain to explain why the set size 2 conditions without overlap were less confidently classified as set size 2. The first is that the conditions lie in a space that is largely orthogonal to the hyperplane drawn by the classifier. The second is that the conditions do lie along the long axis separating the two training conditions, and are falling somewhere in the

middle. Both cases would imply that, even if there is an axis that maximally separates set size, our classifier has not purely identified it.

We hypothesized that an additional signal may be present in the training set which the classifier is relying on. Specifically, there may be an additional process that is engaged when the two clouds overlap in space. This signal would be present in the SS2Superset training condition, and could contribute to separation from the SS1Broad training condition. The pattern of results could then be explained by both set size and this additional overlap-related process. Conditions containing both two objects and an overlap-related process, like SS2NarrowOverlap, SS2BroadOverlap, and SS2PartialOverlap, would be most like the training set size 2 condition, and the condition containing only 1 object and no overlap-related process, like set size 1 narrow, would be most like the training set size 1 condition. Conditions which contained 2 objects but no overlap-related process would fall in the middle. Although our mvLoad analysis could not directly assess these possibilities, the following section reports how we used representational similarity analysis (RSA) to identify the independent contributions of these signals to our decoding results.

Representational similarity analysis

We investigated whether both a set size signal and an overlap-related process signal could be contributing to these results using RSA. In addition to set size and overlap, we also considered the impact of the total attended area, as well as the expected number of non-contiguous attended locations, as potential confounds to be controlled for. The aim of this approach was to compare empirical representational dissimilarity matrices (RDMs), which capture the similarity structure across conditions, against theoretical RDMs reflecting how these

factors of interest would affect the similarity structure (Figure 11). Thus, this RSA analysis revealed the degree to which each theoretical process was a reliable influence on the overall similarity structure.

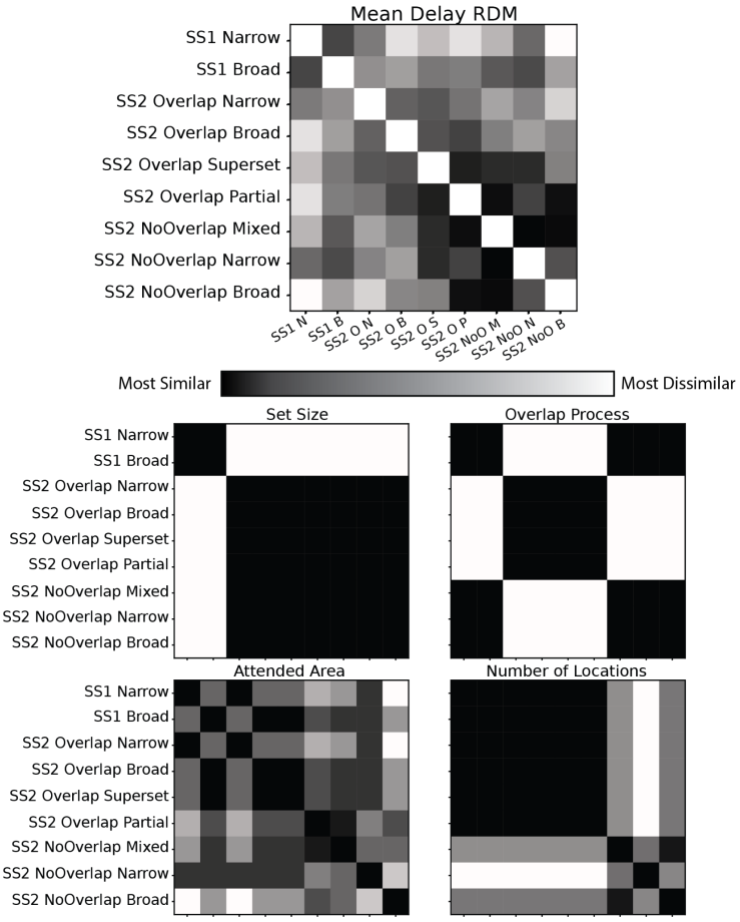

Figure 11. Theoretical and Empirical RDMs. The top RDM shows the average ranked distance between each pair of conditions during the delay period, averaged across participants. The bottom two rows reflect the theoretical RDMs.

In each non-overlapping, 50-ms window, we computed the linear discriminant contrast (LDC) (Walthers et al., 2016) for each pair of conditions for each participant using the average

activity of the window. This produced an empirical RDM for each participant in each time window. We then compared this empirical RDM against a set of theoretical RDMs by computing the semipartial correlation between the empirical RDM and each factor.

To identify each factor's contribution to the structure of distances between conditions, we computed the semipartial rank correlation for each RDM (Kiat et al., 2022). Using a rank correlation avoids assuming a linear relationship between changes in factor values and changes in the empirical distances. Using a semipartial correlation assures that we are only capturing each factor's independent contribution to the empirical RDM. Thus, the presence of a significant semipartial correlation can be taken as strong evidence for the presence of the factor. We repeated this procedure for each participant at each timepoint. We tested these correlations at the group level using a Wilcoxon sign-rank test against zero.

The results are presented in Figure 12. We found that, averaged across the delay period, an overlap-related process independently contributed to differences between conditions (mean semipartial r=.133, p=.0003, BF_{10}=205.92), as did differences in the total attended area (mean semipartial r=.212, p=4.8e-7, BF_{10}=5,173.48). There was no significant contribution from the number of non-contiguous attended locations, with moderate evidence for the null (mean semipartial r=-.007, p=.64, BF_{10}=0.246). Critically, set size also produced an independent contribution to the difference between conditions across the delay period (mean semipartial r=.22, p=4.5e-6, BF_{10}=1,445.67). This pattern of results was qualitatively identical when subsetting to just the first or second half of the delay period[3]. Thus, RSA confirmed the presence

[3] In the second half of the delay period, there was a significant inverse contribution of the number of locations (mean semipartial r=-.05, p=.048, BF_{10}=2.48). However, this effect was no longer significant when excluding participants with informative eye movements (mean semipartial r=-.038, p=.145, BF_{10}=.92), unlike all other effects, which remained significant (and which were significant during both halves of the delay period).

of a set size factor that tracked the number of items stored in memory, regardless of variations in the spatial extent of the items or the overlap between those items.

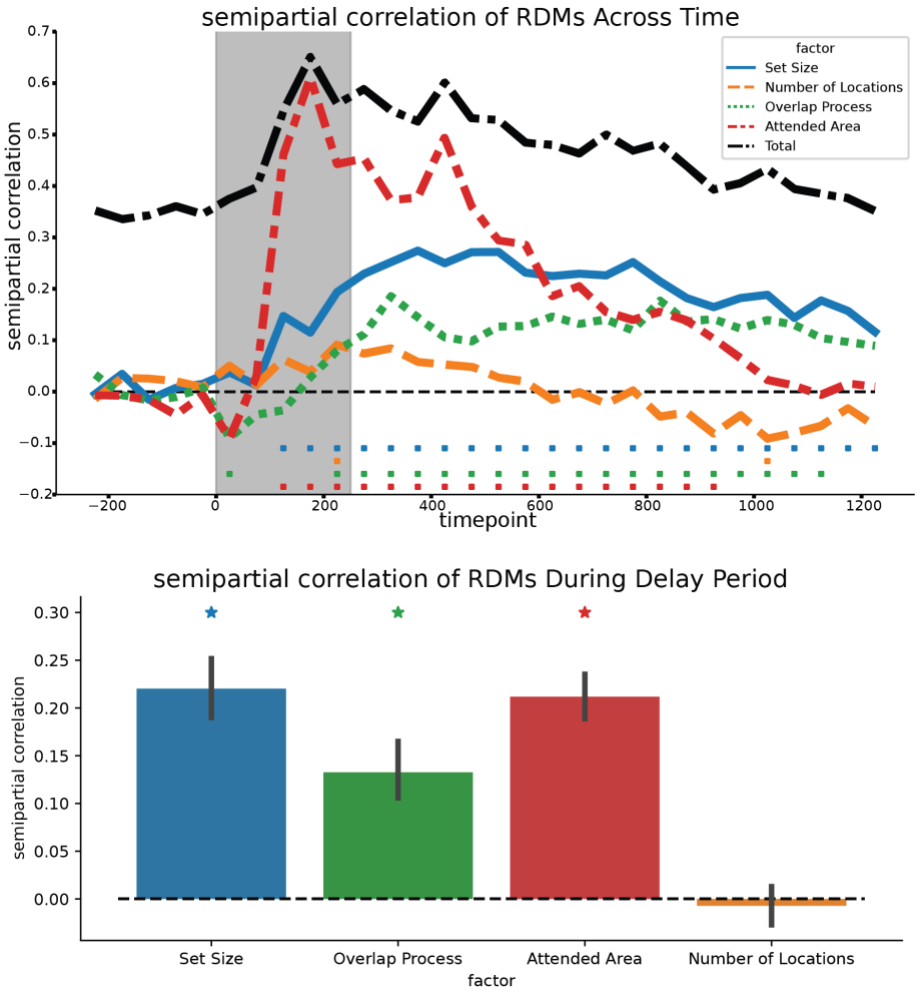

Figure 12. Semipartial rank correlations between theoretical RDMs and the empirical RDM across the delay period. Top, the semipartial rank correlations across time. Color-matched squares indicate correlations significantly greater than 0, Benjamini-Hochberg corrected for comparisons at each timepoint. Bottom, the average correlation during the delay period. Stars indicate statistical significance.

Controls for eye movements

We examined whether systematic eye movements could be the driver of these EEG results using two sets of analyses.

First, we asked whether our IEM analysis was influenced by eye position rather than by the locus of covert spatial attention. To this end, we re-coded the location for each trial using the average gaze position of the trial. We then compared set size 1 vs set size 2 trials using these re-coded locations. We also reran the original IEM comparison using total alpha power, matching the number of trials per location with that of the eye gaze locations. The results are plotted in Figure 13. We found extremely small slopes for eye locations, with no distinction based on set size, and moderate evidence for the null ($t(22)=.025$, $p=.98$, $BF_{10}=0.219$, $d=.005$). In contrast, using a matched number of trials per location, we continued to find clear differences in slope ($t(22)=4.90$, $p=6.8e\text{-}5$, $BF_{10}=387.7$, $d=1.02$). Given this set size contrast produced the largest difference in slopes between conditions, it is unlikely that eye movements are contributing to the differences in conditions that we find here.

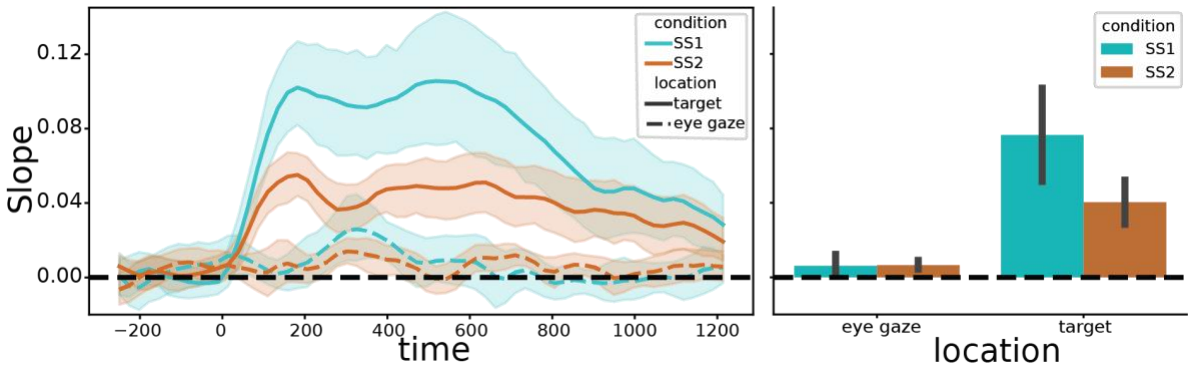

Figure 13. CTF slopes across time, and averaged across the delay period. Left, dashed lines reflect IEMs based on eye gaze, while solid lines reflect IEMs based on target conditions, with a matched number of trials.

Next, we attempted to decode set size (regardless of cloud size and overlap) using just information from eye movements, combining all eye tracking and EOG data available for each participant. Doing so revealed slight but significant decoding (mean decoding accuracy = 53.3%, SD = 1.6%), beginning around 600ms after stimulus onset (Figure 14). This contrasts with the much higher decoding accuracy that is capable when using EEG (mean decoding accuracy = 67.8%, SD = 5.5%), and which begins around 100ms after stimulus onset. We next identified 4 participants with informative eye movements by setting a threshold of 60% decoding accuracy across the delay period. We replicated all analyses excluding those 4 participants. All significant analyses remained significant or trending towards significance, and in no cases did the pattern of results meaningfully change (but see footnote 3).

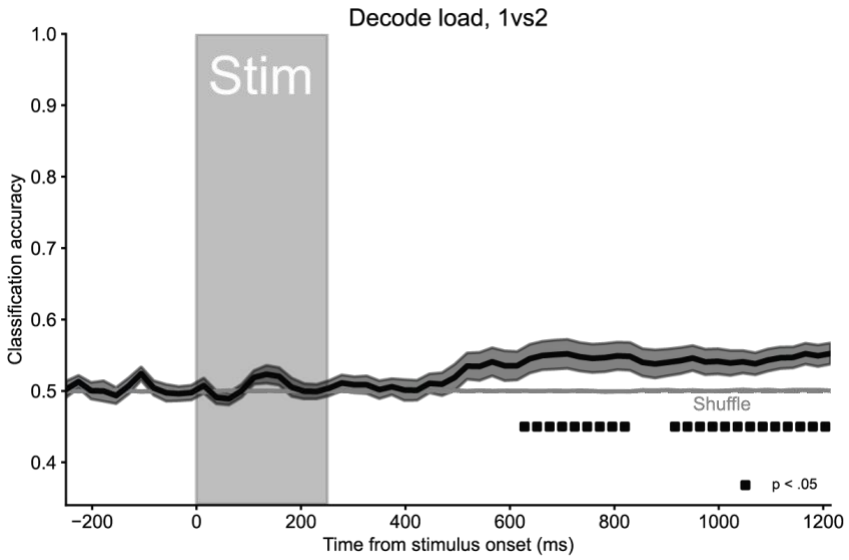

Figure 14. Decoding of Set Size based on eye movement data available for each participant (EOG and gaze data).

Overall, these results suggest that eye movements made little if any contribution to the signals of interest here.

Discussion of Experiment 2

In experiment 2, we examined whether neural activity that precisely tracks the locus and breadth of covert spatial attention was tied to neural activity that indexes the number of individuated objects stored in visual WM. Using dot cloud stimuli to control for variations in spatial extent, we found that the precision of spatially selective alpha activity was unaffected by the number of objects stored in working memory. At the same time, when the spatial extent across different loads was perfectly controlled, a multivariate model still robustly decoded the number of items encoded into visual WM. Moreover, representational similarity analyses confirmed the conclusion that there was a pure load signal – sustained throughout the delay period – that indexed the number of items in working memory, even while simultaneously documenting the unique variance that was explained by other factors.

General Discussion

Past work has shown a close intertwining between spatial attention and working memory (e.g., Awh et al., 2006; Awh & Jonides, 2001; Chun, 2011). Observers direct spatial attention toward the position of items held in working memory, even when position is irrelevant to the task (Foster et al., 2017), and working memory performance declines when covert orienting toward those positions is hindered (Awh et al., 1998; Williams et al., 2013). Nevertheless, here we found strong evidence that these signals tap into dissociable aspects of voluntary attentional control. In Experiment 1, spatial attention, as measured by posterior alpha power, was sensitive

to the number of relevant locations regardless of the number of items occupying those positions, whereas our load signal was sensitive to the number of relevant objects, regardless of the number of locations. In Experiment 2, the precision of spatial attention, as measured by inverted encoding models applied to the topography of alpha power, was sensitive to the size of the attended region but not the number of relevant objects. Simultaneously, we observed a load signal that was sensitive to the number of relevant objects, regardless of the breadth of the attended region. Thus, although both processes are subject to voluntary control, distinct experimental factors had a selective influence on each aspect of attention. This strongly implies a functional dissociation between covert spatial attention and working memory storage. These results are aligned with past work (Bae & Luck, 2018; Diaz et al., 2021; Fukuda et al., 2015; Günseli et al., 2019; Hakim et al., 2019, 2021; Thyer et al., 2022). Our work shares similar logic to that of Bae and Luck (2018), which separately manipulated and decoded the location and orientation of remembered teardrop stimuli using low-frequency event-related potential signals (6 Hz) but could only decode location, and not orientation, with alpha power. However, rather than dissociating spatial position from specific feature information (teardrop orientation) maintained in working memory, we are dissociating spatial position from the number of items held in working memory. This rules out spatial attention as a compelling alternative explanation for past observations of working memory load signals that generalized across variations in the type of visual features maintained (Thyer et al., 2022).

Importantly, we are not claiming that alpha power only reflects spatial attentional signals and that raw voltage only reflects signals of working memory load. Although it is true that precise spatial information is predominantly reflected by variations in alpha power (as seen in Experiment 1b; Bae & Luck, 2018; Foster et al., 2017), it is likely that activity in the alpha band

indexes other cognitive processes as well. For example, unpublished work in our lab found that robust mvLoad models can be constructed on the basis of the topography of alpha activity. Moreover, RSA in the present work revealed the presence of spatial information in raw voltage. Thus, rather than arguing for a distinction between alpha power and raw voltage, we are focused on the distinction between spatial attention and working memory gating. In that context, alpha power and raw voltage provided the most robust analytic approaches for tracking spatial attention and working memory encoding, respectively.

Although the methods we employed were chosen because they provide a precise measurement of spatial attention and working memory load, they are still limited by the information in the signals to which they are applied. Thus, it is possible that some of our conclu-sions could change if we were to use alternative meth-ods. For instance, the amplitude of the P1 event-related potential indexes whether spatial attention is allocated toward a position (Mangun & Hillyard, 1991). Future work could examine whether dissociations between the breadth of covert spatial attention and the number of items encoded into working memory will generalize with alternative measures of these processes.

Although these findings help to refine our taxonomy of voluntary attentional control, they do not affect the importance of interactions between spatial attention and working memory storage. As noted above, spatial attention is consistently oriented toward the position of items in working memory, suggesting that spatial selection could be a prerequisite for working memory encoding, or a component of rehearsal during the delay period (Awh & Jonides, 2001; Williams et al., 2013). Thus, these attentional processes may interact with one another in a mutually reinforcing manner to prioritize the processing of relevant information across stages of processing. Basic questions remain regarding the specific role (or roles) of spatial attention in

visual working memory. Here we consider three possible answers, which are not mutually exclusive. First, space may act as a key organizing dimension in the building and maintenance of object representations, (i.e., for feature binding; Abrahamse et al., 2014; Pertzov & Husain, 2014; Schneegans & Bays, 2017; Treisman & Zhang, 2006). If so, it may be that the allocation of spatial attention is a prerequisite to working memory encoding. Second, the allocation of spatial attention may improve the signal-to-noise ratio of early sensory processing and representations (Martínez et al., 2006), thereby shaping the fidelity of representations available for working memory storage. Finally, spatial attention may be a specific instance of feature-based attention. Previous work suggests that humans can prioritize the maintenance of the relevant feature during the delay period of a working memory task (Serences et al., 2009; Woodman & Vogel, 2008). The task employed in Experiment 2 required participants to remember the spatial location and envelope of each cloud; the reconstructed CRFs could reflect attending to spatial information as a feature of the clouds. As stated above, these possibilities are not mutually exclusive. Attentional modulation of early sensory processing may be a form of feature-based attention (e.g., Müller et al., 2006). Thus, future work can examine whether similar dissociations might be observed between nonspatial forms of feature-based attention (e.g., attention to color or motion) and the gating of working memory representations.

Finally, our proposed dissociation raises interesting questions for the broader literature on attentional control. For example, much past work has examined the neural mechanisms of spatial attention, often focusing on parietal regions (e.g., Bisley & Goldberg, 2003; Fiebelkorn et al., 2019; Karnath & Rorden, 2012). However, this literature often fails to differentiate between the selection of spatial locations or the encoding into working memory of the objects in those locations. Likewise, an emerging body of work has focused on how experience with distractors

of a given color or location can yield increased resistance to interference from those irrelevant stimuli (e.g., Gaspelin & Luck, 2018; Wang et al., 2019). But reduced processing of distractors could be explained either by reduced capture of spatial attention or by a reduced probability of encoding those distractors into working memory. Thus, the dissociation we are suggesting between spatial attention and working memory gating may enable a refined understanding of which aspects of voluntary control are critical in different contexts.

Chapter 2: Content-independent storage mechanisms underlie working memory

Chapter 2a: Cortically Disparate Visual Features Evoke Content-Independent Load Signals during Storage in Working Memory

Abstract

It is well established that holding information in working memory (WM) elicits sustained stimulus-specific patterns of neural activity. Nevertheless, here we provide evidence for a distinct class of neural activity that tracks the number of individuated items in working memory, independent of the type of visual features stored. We present two EEG studies of young adults of both sexes that provide robust evidence for a signal tracking the number of individuated representations in working memory, regardless of the specific feature values stored. In Study 1, subjects maintained either colors or orientations across separate blocks in a single session. We found near-perfect generalization of the load signal between these two conditions, despite being able to simultaneously decode which feature had been voluntarily stored. In Study 2, participants attended to two features with very distinct cortical representations: color and motion coherence. We again found evidence for a neural load signal that robustly generalized across these distinct visual features, even though cortically disparate regions process color and motion coherence. Moreover, representational similarity analysis provided converging evidence for a content-independent load signal, while simultaneously showing that unique variance in EEG activity tracked the specific features that were stored. We posit that this load signal reflects a content-

independent “pointer” operation that binds objects to the current context while parallel but distinct neural signals represent the features that are stored for each item in memory.

Introduction

It is well established that WM storage elicits sustained patterns of neural activity that track the contents of the stored items, even in the absence of the remembered stimulus (Fuster and Alexander, 1971; Fuster and Jervey, 1981; Funahashi et al., 1989; Goldman-Rakic, 1995; Harrison and Tong, 2009; Serences et al., 2009; D’Esposito and Postle, 2015; Rademaker et al., 2019). These stimulus-specific neural signals are contingent on voluntary storage goals, and they track behavioral estimates of the precision of the stored memories (Emrich et al., 2013; Ester et al., 2013; Wimmer et al., 2014). Thus, there has been a clear motivation for the strong focus on the content-specific neural signals that are sustained during WM storage. Nevertheless, the present work will highlight a distinct class of storage-related neural activity that is functionally separable from the representation of content, per se.

Specifically, we are referring to neural signals that track the number of items stored in working memory, without respect to the specific features of those items (Todd and Marois, 2004; Vogel and Machizawa, 2004; Xu and Chun, 2006; Adam et al., 2020; Thyer et al., 2022). For example, Vogel and Machizawa (2004) discovered a contralateral EEG waveform that tracks the number of items stored on the attended side and is highly predictive of individual WM capacity (Luria et al., 2016). Critically, the amplitude of the CDA is similar for single- and multi-featured objects (Woodman and Vogel, 2008), suggesting that it tracks the number of items stored rather than the total amount of information associated with those items. More recently, Adam et al. (2020) used a machine learning approach that uses the full scalp topography of EEG voltage to

decode the number of individuated items stored in WM. This approach is sensitive enough to provide above-chance decoding with single trials of data and reveals a signature of WM storage that generalizes across novel observers. Importantly, Thyer et al. (2022) showed that this multivariate load signature (mvLoad) generalizes across stimuli that vary in both the type (color vs orientation) and the number (single vs dual-feature objects) of features within the stored items. These findings provide evidence for an EEG signature of WM load that tracks the number of individuated representations stored rather than the type or amount of information associated with each item.

What is the computational role of content-independent load signals? We hypothesize that they may reflect an indexing operation that enables the online tracking of items through time and space. For example, Kahneman et al. (1992) proposed the object file as a temporary episodic representation that enables the continuous tracking of items through time and space, despite possible changes in the visual appearance and position of those items. Kahneman et al. (1992) argued that this was essential for binding the representation of an attended item to the context of an unfolding event. Likewise, Pylyshyn (1989) described "fingers of instantiation" (FINSTs) that enable the dynamic tracking of items despite changes in their appearance or location. With both object files and FINSTs, the core insight is that observers require a flexible indexing system that can support the tracking and storage of objects, even though the appearance and position of a relevant item can change dramatically during an unfolding event. Thus, object files and FINSTs were proposed as a mechanism to ensure the continuity of an item's representation despite these challenges. Here we refer to object files and FINSTs as spatiotemporal "pointers" to highlight the process of binding an item to a specific set of spatiotemporal coordinates, and we argue that this may be a fundamental requirement for storage in visual working memory. This perspective

aligns with prominent models of WM that propose separate neural processes for the maintenance of features and binding of items to a specific context (Xu and Chun, 2006; Swan and Wyble, 2014; Oberauer and Lin, 2017; Balaban et al., 2019; Bouchacourt and Buschman, 2019; Hedayati et al., 2022). For example, Xu and Chun (2006) argued for two separate neural pathways for object identification and individuation. Similarly, Swan and Wyble (2014) proposed a "neural binding pool" that binds the features of an object together so that it can be represented as an individuated token within a specific scene. Critically, both theories argue for a clear separation between the processes that represent the specific details of an item and the processes that bind that representation to the surrounding context. Thus, clear evidence for content-independent load signals that track the number of stored items without tracking their featural content would support the hypothesis that the binding of items to context reflects a distinct aspect of working memory from the maintenance of each item's features.

As noted above, Thyer et al. (2022) used multivariate decoding approaches to show that EEG signatures of WM load generalized across variations in both the type and number of visual features stored. Although this suggests a clear dissociation between WM load and featural content, there are two key limitations of the Thyer et al. demonstration that merit discussion and follow-up. First, the generalization shown by Thyer et al. (2022) was imperfect, such that decoding accuracy dropped when a model trained on one condition (e.g., color) was tested on another condition (e.g., orientation). This decline in decoding could indicate that there were feature-specific aspects of the EEG load signature. However, each condition in the Thyer et al. study was collected during a different experimental session, opening the possibility that small differences in electrode placement and noise levels reduced generalization. Thus, further work is needed to determine the degree of overlap between load signatures for distinct visual features.

Second, the choice of color and orientation as the stored features might have led to the overestimation of content independence, simply because the neural populations representing color and orientation are so interdigitated in the visual cortices. Given the coarse spatial resolution of scalp-recorded EEG activity, it is possible that distinct populations of color and orientation cells could still produce similar enough EEG signals to mimic a general load signal (Sandhaeger and Siegel, 2023). In the present work, we address both of these concerns by collecting all conditions within a single experimental session (Experiment 1) and by examining generalization across color and motion coherence (Experiment 2), features that are known to have highly separable cortical populations (Zeki, 1978; Felleman and Van Essen, 1991; Vaina, 1994). Finally, we applied representational similarity analysis (RSA) to obtain converging evidence for a common load signature across disparate stimulus types, while showing that distinct variance in ongoing EEG activity was explained by the specific features that were maintained.

To anticipate the results, Experiment 1 revealed effectively perfect generalization between the load signatures generated during color and orientation blocks, in line with the content-independent pointer hypothesis. Moreover, in the same dataset we could robustly decode which feature (color or orientation) was being attended and stored, corroborating our assumption that observers were selectively storing distinct aspects of the stimuli in the color and orientation conditions. In Experiment 2, we again observed strong generalization between the load signatures for two cortically disparate features, color, and motion coherence, as well as robust decoding of the attended feature. Finally, RSA provided complementary evidence of a load signal that was independent of the attended feature, even while distinct variance in EEG activity tracked the specific features that were stored in working memory. Together, these results provide strong

evidence for content-independent pointers as a key component of storage in visual working memory.

Materials & Methods

Subjects

Experiments included 29 volunteers (Experiment 1, n = 13; Experiment 2, n = 16) participating for monetary compensation ($20 per hour). Subjects were between the ages of 18 and 35 years old, reported normal or corrected-to-normal visual acuity, and provided informed consent according to procedures approved by The University of Chicago Institutional Review Board. Subjects were recruited via online advertisements and fliers posted on the university campus.

Experiment 1. Our target sample in Experiment 1 was 12 subjects. Seventeen volunteers participated in Experiment 1 (8 females; mean age = 24.9 years, SD = 3.8). Four subjects were excluded from the final sample for the following reasons: the session was ended early due to eye movements (n = 2); the subject's data was corrupted or otherwise unusable (n = 2). The final sample size was 13 (6 female; mean age = 25.0 years; SD = 4.1).

Experiment 2. Our target sample in Experiment 1 was 16 subjects. Nineteen volunteers participated in Experiment 2 (9 females; mean age = 26.6 years; SD = 4.0). Three subjects were excluded from the final sample for the following reasons: The session was ended early due to eye movements (n = 2); the subject did not have enough data after artifact rejection (n = 1). The final sample size was 16 (9 female; mean age = 26.7 years; SD = 2.2).

Apparatus

We tested the subjects in a dimly lit, electrically shielded chamber. Stimuli were generated using PsychoPy (Peirce et al., 2019). Subjects viewed the stimuli on a gamma-corrected 24 in LCD monitor (refresh rate = 120 Hz; resolution = 1,080 × 1,920 pixels) with their chins on a padded chin rest at a viewing distance of 75 cm.

Task procedures

Luminance-balanced displays. For each experiment, the luminance of the target color set was measured using a LS-150 Luminance Meter. Next, the luminance meter was used to identify an RGB value that produced a gray matching the average luminance of the target set in each experiment. Placeholders were colored with this gray and set to cover the same area as the targets, controlling the total area and luminance across set sizes.

Experiment 1. Experiment 1 used a whole-field change detection task (Fig. 15*a*). On each trial, a memory array appeared containing four total elements. There were one or three targets to be remembered, and the remaining items were gray placeholders. Stimuli were presented against a mid-gray background (~61 cd/m^2). Items were positioned with a maximum of one item per quadrant, and all items were placed at least 4° apart from fixation and from one another.

Memory targets were colored circles (radius, 1.3°) with oriented bars cut out of the middle (height, 2.6°; width, 0.5°). The possible orientations were 0°, 45°, 90°, and 135°, and they were sampled without replacement for each trial. The possible colors were randomly sampled without replacement from a set of four colors (RGB values: red, 255, 0, 0; green, 0, 255, 0; blue, 0, 0, 255; yellow, 255, 255, 0). The placeholders were filled circles (radius, 1.13° to match the same total area) in a shade of gray [red, green, blue (RGB) value = 149, 150,149]

that matched the average luminance of all possible colors in the color set (see above, Luminance-balanced displays).

On each trial, subjects viewed a memory array (250 ms), remembered items across a delay (1,000 ms), were probed on one item, and reported whether the probed item was the same as or different from the remembered item (unspeeded). Alternating each block, subjects were instructed to attend to either the color or the orientation of the memory items. Only the attended feature dimension could change. On a change trial, the relevant feature could change into any other value from the set (i.e., any other color or orientation), regardless of whether that feature value was present in another item in the display.

Experiment 2. Experiment 2 used a whole-field change detection task (Fig. 15*b*). On each trial, a memory array appeared containing three total items. There were one or two targets to be remembered, and the remaining items were gray placeholders. Stimuli were presented against a mid-gray background (~61 cd/m^2). Items were positioned with a maximum of one item per quadrant, with the center of each item randomly selected between 1.5 and 3° away from fixation on both the horizontal and vertical axes.

Memory targets were colored random dot kinematograms (RDKs; radius, 1.5°). The RDKs were presented as 100 individual dots moving through a circular area. Dots were 7.5 pixels in size, lasted four frames, and moved at a speed of 0.06°/s through the circular area. The possible colors were randomly sampled without replacement from a set of seven colors (RGB values: red, 255, 0, 0; green, 0, 255, 0; blue, 0, 0, 255; yellow, 255, 255, 0; purple, 255, 0, 255; teal, 0, 255, 255; orange, 255, 128, 0). The moving dots within each item were either moving coherently (all in one direction) or incoherently (moving in random directions). For coherently

moving dots, the direction of movement was randomly sampled from a uniform distribution between 0 and 359° (inclusive) with 1° steps. To discourage subjects from encoding coherence as specific orientations, the orientation of the probed cloud was again randomly sampled at test so that it would be independent of the original orientation. The place-holders were also RDKs with the same dimensions, shown in a shade of gray (RGB value = 166, 166, 166) that matched the average luminance of all possible colors in the color set (see above, Luminance-balanced displays).

After data collection, a coding error was discovered in how coherence was assigned, as the options were sampled from a repeated set, such that there were always one or two coherent dot patches in the display. In set size 1, the probability of the target RDK being coherent was 50%, but, due to this error, when the target was coherent, at most 1 distractor was coherent, and when the target was incoherent, at least 1 distractor was coherent. Probabilities in the set size two condition mirrored that of set size one. For example, when both targets were coherent, there was no possibility that the distractor would be coherent and vice versa. To address this, we examined whether we could decode coherence. As described in the Results section, the amount of coherence is only decodable for targets and only when subjects are attending to coherence. Given that coherence was not decodable during the color blocks, stimulus-driven effects of coherence cannot explain the common load signature between the color and motion coherence conditions.

On each trial, subjects viewed a memory array (500 ms), remembered items across a delay (1,000 ms), were probed on one item, and reported whether the probed item was the same as or different from the remembered item (unspeeded). In alternating blocks, subjects were instructed to attend to either the color or the motion coherence of the memory items. Only the

attended feature dimension could change. As in Experiment 1, on a change trial, the relevant feature could change into any other value from the set (i.e., any other color, or the alternate coherence level), regardless of whether that feature value was present in another item in the display.

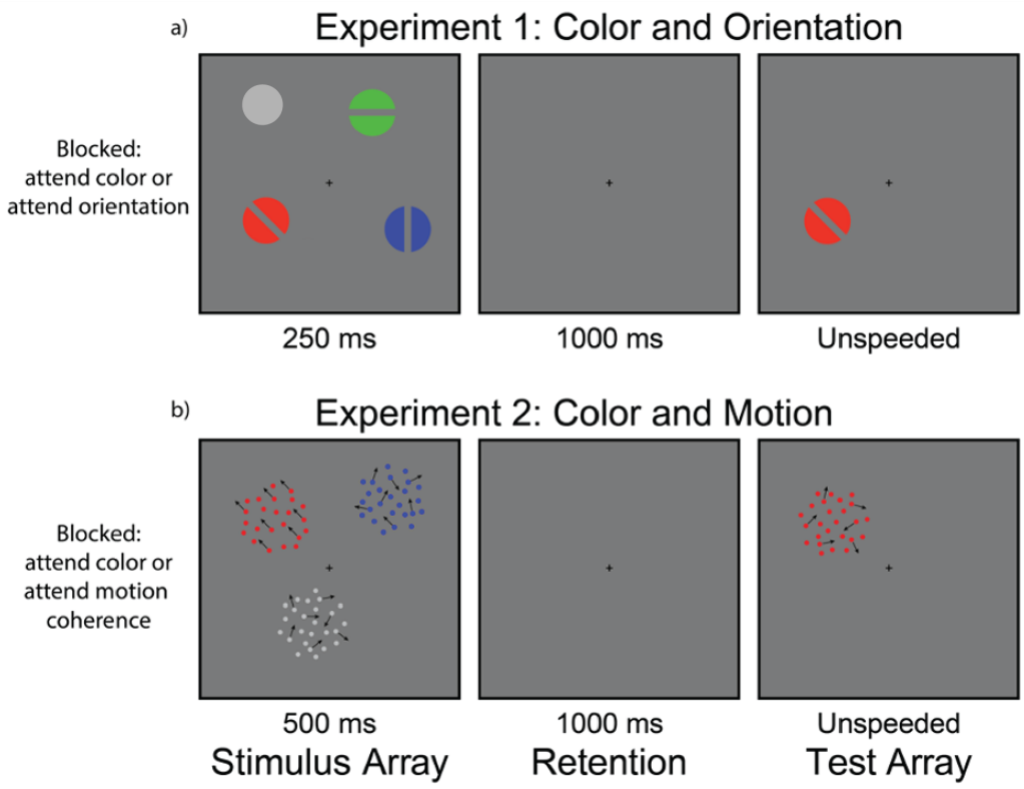

Figure 15. Task schematics for example set size 3 and set size 2 trials in the whole-field change detection task used in both experiments. In Experiment 1 (a), subjects were cued at the beginning of each block to either remember the color or orientation of the memory items while ignoring the gray distractor items. In Experiment 2 (b), subjects were cued at the beginning of each block to either remember the color or the motion coherence of the memory items while ignoring the gray distractor items. During change trials, the attended feature changes 50% of the time. The Experiment 1 example depicts a no-change trials, while the Experiment 2 trials depicts a change trial (change in motion coherence).

EEG acquisition and preprocessing

EEG acquisition. We recorded EEG activity from 30 active Ag/AgCl electrodes mounted in an elastic cap (Brain Products, actiCHamp). We recorded from international 10–20 sites Fp1, Fp2, F7, F3, Fz, F4, F8, FT9, FC5, FC1, FC2, FC6, FT10, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1, Oz, and O2. Two additional electrodes were affixed with stickers to the left and right mastoids, and a ground electrode was placed in the elastic cap at position Fpz. All sites were recorded with a right-mastoid reference and were rereferenced off-line to the algebraic average of the left and right mastoids. Data were filtered on-line (low cutoff, 0.01 Hz; high cutoff, 80 Hz; slope from low to high cutoff, 12 dB/octave) and were digitized at 500 Hz using BrainVision Recorder (Brain Products) running on a PC. Impedance values were brought below 10 kΩ at the beginning of the session.

Eye tracking. We monitored gaze position using a desk-mounted EyeLink 1000 Plus infrared eye tracking camera (SR Research). Gaze position was sampled at 1,000 Hz. According to the manufacturer, this system provides spatial resolution of 0.01° of visual angle and average accuracy of 0.25–0.50° of visual angle. We calibrated the eye tracker every one to two blocks of the task and between trials during the blocks if necessary. We drift-corrected every five trials. Additionally, we drift-corrected the eye tracking data for each trial by subtracting the mean gaze position measured during the 200 ms window immediately preceding the memory array.

Artifact rejection. We segmented the EEG data into epochs time locked to the onset of the memory array (200 ms before until 1,000 ms after stimulus onset). We baseline-corrected

the EEG data by subtracting mean voltage during the 200 ms window immediately prior to stimulus onset. Eye movements, blinks, blocking, drift, and muscle artifacts were detected by applying automatic criteria, and we discarded any epochs contaminated by artifacts. All subjects included in the final sample had at least 140 trials of each condition.

Eye movements and blinks. We employed real-time eye movement detection. If subjects moved their eyes >1.25° from fixation, the trial was interrupted and their eye position was shown to them for feedback purposes. Interrupted trials were made up at the end of each block. During preprocessing, we rejected trials that contained eye movements beyond 1° of visual angle using the pop_artextval function in ERPLAB (Lopez-Calderon and Luck, 2014).

Drift and muscle artifacts. We checked for drift (e.g., skin potentials) with the pop_rejtrend function in ERPLAB.). We checked for muscle artifacts with the pop_artextval function in ERPLAB. We excluded trials where EEG activity was >80 μV or less than −80 μV. We excluded trials with peak-to-peak activity >100 μV within a 200 ms window with 100 ms steps. We also excluded trials with any value beyond a threshold of 80 μV.

Multivariate decoding and generalization procedure

All decoding analyses followed the framework of the multivariate load analysis (mvLoad), a within-subject decoding approach using the Scikit-Learn Logistic Regression model (Pedregosa et al., 2011), which has been previously described (Adam et al., 2020; Thyer et al., 2022). First, we divided each trial into nonoverlapping 25 ms windows and calculated the average voltage for each electrode in the window. Next, we randomly binned trials of the same

condition into sets of 15 trials and averaged within them to increase signal-to-noise. Next, the binned trials were sorted into training and testing sets, stratified by trial condition, using the train_test_split function from Scikit-Learn (Pedregosa et al., 2011). This cross-validation procedure splits the binned data into 80% training and 20% testing sets while balancing the percentage of samples for each condition. Training data were *z*-score normalized by their mean and standard deviation at each time point using the StandardScaler Scikit-Learn function, and test data were *z*-score standardized using the mean and standard deviation of the training set, before training and testing the model at each time point. Our measure of decodability on the test sets is described below. A measure using the test set with shuffled labels was also recorded to produce an empirical null distribution. This process was repeated 1,000 times, including the initial random binning, and the results for each subject at each time point were averaged across repetitions.

Decodability measure: hyperplane contrast. In past work (Thyer et al., 2022), we tested the generality of mvLoad EEG patterns by training the model on one condition and measuring decoding accuracy in a different condition. High decoding accuracy was taken to indicate reliable generalization of the load patterns across the two conditions. Although this is a very common approach for examining the generalizability of a multivariate pattern, it is possible for generalization to be imperfect even if there is a perfectly overlapping load signal. For instance, if there are differences in EEG activity due to feature-specific activity, and those differences are not fully orthogonal to the axis that encodes WM load, then generalization could be undermined despite a shared representation of load. To provide an analogy, there is an additive shift in the CDA (i.e., an increase that is constant across all set sizes) when people hold orientations in WM

rather than colors (Woodman and Vogel, 2008). If a model's decision boundary was fit to separate set size 1 from set size 2 based on the amplitude of the CDA with colors, it would be biased to call all orientation trials set size 2, even though the same CDA difference between set sizes is present. A better alternative is to estimate the vector differentiating pairs of conditions (e.g., the CDA set size difference) and ask how similar it is to the vector differentiating a new pair of conditions.

Given those considerations, we attempted to isolate the multivariate axis that was specific to WM load by estimating a "hyperplane contrast." We recorded each test trial bin's signed distance from the fit model's decision boundary, also known as a hyperplane, and distances for trials of the same condition were averaged together. Decodability was defined as the contrast (difference between) the two conditions' distances from the hyperplane. This is a cross-validated estimate of the distance between two conditions along the discriminating axis which most separates them. When this procedure is performed using a linear discriminant analysis (LDA) as the classifier, it is called the linear discriminant contrast (LDC) and is equivalent to cross-validated Mahalanobis distance (Walther et al., 2016). The only difference between LDC and this procedure is that we use logistic regression to identify the hyperplane, rather than LDA. To test generalization, we record the signed distance of test data from a pair of conditions from a held-out context (i.e., the other attended feature), and the contrast between the two conditions is recorded. The random binning and averaging of the held-out context data happens at the same time as it does for the trained context data.

This hyperplane contrast is beneficial for a few reasons. First, it is less coarse than accuracy, which requires recoding the test scores into binary categories. This binarization treats trials adjacent to the hyperplane the same as trials extremely far from the hyperplane. Second, it

is not impacted by the exact placement of the hyperplane along the discriminant axis, which is beneficial when estimating the generalization on held-out conditions. If there is a signal that indicates a change in context that is not parallel to the hyperplane (orthogonal to the discriminant axis, e.g., the additive shift for orientations described above), this would bias classification, even if the true signal was present in the new context and identical to the original signal that the model was trained on. Third, it is unbounded, in contrast to accuracy and area-under-the-curve, so it should not be warped by ceiling effects and should grow proportionally with the signal separating the conditions. This makes the interpretation of the contrast distance straightforward: if the contrast distance between the training context and the held-out context are the same and vice versa when the contexts are flipped, then the discriminating axis and the magnitude of the signal separating the conditions in each context are the same. That is, it is the same signal.

Attended feature decoding. In both experiments, we assessed whether we could decode which feature was attended. When setting the training data, the number of trials for each set size was equated to prevent the model from leveraging a potential content-independent WM load signal.

Load decoding and generalization. In both experiments, we assessed the decodability of load within each feature and whether the load signal generalized to the other feature (e.g., whether a model trained on color data would accurately decode orientation data, and vice versa). Trials for the held-out feature were also randomly binned, and 20% were randomly selected for testing on each permutation following the same procedure as the trained feature (see above).

Coherence level decoding. We assessed whether we could decode the coherence level in Experiment 2. First, we asked whether we could decode the number of coherent items in the display while controlling for the number of attended coherent items. To maximize power, we combined three pairs of conditions, which differed only in the number of coherent distractors: set size 1 trials with no coherent targets and either 1 or 2 coherent distractors, set size 1 trials with 1 coherent target and either 0 or 1 coherent distractors, and set size 2 trials with 1 coherent target and either 0 or 1 coherent distractors. Within each pair of conditions, the number of trials was equated by randomly downsampling the condition with more trials on each permutation. Once the number of trials was equated for each pair, all conditions with fewer coherent distractors were randomly binned together, as were all conditions with more coherent distractors. As the proportion of trials contributed by each pair, after downsampling, was broken roughly in ¼, ¼, and ½, we set the number of trials per bin to be 16. After binning, decoding proceeded the same as the above analyses.

Next, we asked if we could decode the number of attended coherent targets. The above analysis showed no evidence for a signal reflecting the number of coherent distractors (see Results). Therefore, to maximize power, we focus just on the number of coherent targets. We did this by using set size 1 trials and contrasting trials with a coherent target against trials with an incoherent target. Both analyses were run separately for the attend coherence and the attend-color blocks.

Significance testing. In all decoding analyses, we tested whether decodability was significantly above chance at each time point using a paired-samples, one-tailed *t* test against the empirical null. This empirical null was defined by testing the trained models on randomly shuffled trial labels (see above). To test for imperfect generalization, we assessed whether the

difference in the hyperplane contrast between the trained feature and the held-out feature was larger than the difference between the two empirical nulls, also with a paired-samples, one-tailed *t* test. Because we tested for significance at each time point, we used the Benjamini–Hochberg procedure to control the false discovery rate (FDR) at 0.05.

Eye movement controls

Eye-based decoding and correlations with EEG results. To examine whether decoding was driven by neural activity related to eye movements, we repeated all decoding analyses using the eye movement data available for each subject to see if any signals were decodable. These analyses were repeated simultaneously with the main decoding analyses, so that the same trials were randomly binned together, and the same train–test split occurred. This allowed us to also ask whether the individual test sets produced similar predictions between the eye and EEG data. For each participant, we correlated the predictions across all 1,000 repetitions to assess the amount of mutual information between them. These correlations were transformed to Fisher's *z* scores and tested at the group level against 0 at each time point using a one-tailed *t*-test. *p*-values were FDR corrected, as above. In Experiment 2, one subject's eye data could not be aligned, so they were excluded from these analyses. All of the above analyses were rerun without the subject to ensure there were no qualitative differences. The only change is that the target coherence level in attend-motion blocks is no longer significantly decodable at any time point after FDR correction, though there are several time points which are trending toward significance ($p < 0.1$).

Associating EEG-based decodability and eye-based decodability across participants. In the above analyses, we tested whether EEG-based decodability aligned with eye-based decodability across trials, within each subject. As an additional test of whether EEG-based

decodability was driven by eye-based decodability, we tested whether EEG-based decodability aligned with eye-based decodability across subjects. As the major motivation of this work is whether WM load signals generalize across attended features, we focused on the strength of the across-feature contrast (training on one feature and finding the size of the contrast in the other), averaging across the two models per experiment. To increase SNR we average time across the last 500 ms of the delay period in each experiment, as these appeared to be times when both EEG-based and eye-based contrasts were stable, and combined the data across experiments. We fit a linear regression model predicting an individual's average EEG-based cross-feature contrast from their eye-based cross-feature contrast, with a separate intercept for each experiment.

Representational similarity analysis

While decoding gives us an indication of which signals are present, and their similarity across contexts, decoding results can be difficult to interpret, as models will leverage all available signals to discriminate between conditions. For example, the ability to decode which feature is attended could result from the model leveraging task-set like signals, feature-specific signals, or a mixture of both signals. RSA provides an alternative means of assessing which signals are present in the data. Here, we leverage RSA to assess whether feature-specific load signals are present in the data and whether a content-independent load signal remains while controlling for other confounds.

We only ran this analysis in Experiment 2, as RSA benefits from condition-rich designs (Kriegeskorte et al., 2008). Experiment 1 can only be divided into four total conditions (2 set sizes × 2 attended features). In contrast, the decoding results below provide initial evidence that

the coherence level of the memory targets can be decoded when participants are tasked with holding the coherence level in WM, suggesting Experiment 2 can be meaningfully divided into 10 conditions once the coherence level of memory targets in the displays are included. The division into specific coherence levels is beneficial because it allows examination of the content-independent load signal while simultaneously providing a complementary means of assessing the apparent coherence level signal that was identified via decoding. Such convergence would confirm that scalp EEG activity tracks the specific coherence values that are held in WM. This could be a useful signal for future studies focused on feature-specific activity and its time course.

At a high level, we computed the empirical representational dissimilarity matrices (RDMs) between every pair of conditions at each time point for each subject, using the same time windows as the decoding analyses. We then computed the semipartial correlation between these empirical RDMs and predicted dissimilarities based on a set of theoretically relevant factors: feature-specific load signals, an attended feature signal, an attended coherence level signal, and a content-independent load signal. We assessed whether these semipartial correlations were significantly greater than 0 at the group level. Because some regressors were based on pupil sizes, only the 15 subjects with available eye data were analyzed.

Computation of empirical dissimilarities. To estimate the empirical RDMs, we computed the cross-validated Mahalanobis distance, also known as the LDC, between every pair of conditions. This is known to be a reliable measure for constructing RDMs (Walther et al., 2016). In addition, as mentioned above, it is a variant of the hyperplane contrast used in our decoding analyses that effectively replaces the logistic regression classifier with a linear discriminant

classifier. The LDC between each pair of conditions was simultaneously computed with the following formula:

$$d^2_{LDC}(i,j) = (m_i - m_j)_{Train} * \Sigma^{-1}_{train} * (m_i - m_j)_{Test}$$

where m_i and m_j are the mean vectors (i.e., average voltage at each electrode) for conditions *i* and *j*, computed separately for the train and test set, and $\Sigma_{\text{Train}}^{-1}$ is the inverse of the variance–covariance matrix between electrodes in the training set. The training and test sets were defined by randomly splitting the data in half, stratified by the 10 conditions. We computed the variance–covariance between electrodes within the training set by first demeaning each trial by its condition's training mean and then computing the variance–covariance matrix across the training trials. This matrix was then regularized using the Ledoit–Wolf procedure, which identifies the optimal shrinkage coefficient in a data-driven manner (Ledoit and Wolf, 2004; Walther et al., 2016). We repeated this procedure with 10,000 train–test splits for each subject at each time window and averaged across these permutations.

Theoretical dissimilarities. We compared the empirical RDMs to a set of RDMs constructed to reflect predicted differences based on theoretical factors. Four of the RDMs were constructed based on theory, and four were based on empirical results. Our content-independent load RDM predicted that all set size 1 trials would look identical, all set size 2 trials would look identical, and all set size 1 trials would look equally distinct from all set size 2 trials. Our attended-feature RDM predicted that all attend-color conditions would be equally similar to one another and equally dissimilar to all attend-motion conditions (which would all be equally similar to one another). We also included two feature-specific load signals, one for the number

of maintained colors and one for the number of maintained motion patches. Each assumed that the amount of feature-specific information for the attended feature increased stepwise with set size and was 0 (or at least less, as the distances are converted to ranks, see below) when the other feature was attended. Note, because the distances are rank transformed (see below), only the relative size of the steps matter. While we present results assuming an equal step from load 0 to load 1 and from load 1 to load 2 for the feature-specific load signals, we confirmed that our results are qualitatively identical if one assumes compressive steps (the load-1–load-2 step is smaller than the load-0–load-1 step) or expansive steps (the load-1–load-2 step is larger than the load-0–load-1 step).

We also included four additional regressors, based on empirical results. First, we included an RDM for the attended coherence level, as we found slight evidence for an attended coherence level signal (see Results). We did not have strong predictions for how the attended coherence level signal would change at set size two. It might reflect the number of coherent targets, in which case the signal might be confounded with set size and mimic a motion-specific load signal, as the number of coherent targets will increase with the number of targets. Alternatively, the signal might reflect the proportion of coherent targets, in which case it would be deconfounded from set size. Thus, we empirically estimated the RDM. Using the set size 1 model to decode the target coherence level in attend-motion blocks, we found the distance from the hyperplane for the two training conditions and the eight remaining conditions, averaged across subjects. These distances were converted into pairs of distances between all 10 conditions. This was done at each time window separately, as the coherence level signal may change across time. To avoid data leakage, each subject's data was excluded before averaging

and converting the hyperplane distances into pairwise distances, resulting in a unique coherence level RDM for each subject, at each time window, that did not contain their data.

Second, we included a set of three regressors to examine whether some effort signal contributed to the apparent content-independent load signal. As a proxy for effort, we included condition-level accuracy, pupil size, and the interaction of accuracy and pupil sizes to generate RDMs. We chose accuracy as one potential proxy for effort under the assumption that it may reflect the difficulty of conditions. Some subjects may expend more effort in more difficult conditions, in which case the RDMs for accuracy and effort would be the same, or they may not expend different levels of effort, in which case effort could not explain the apparent content-independent load signal. Relatedly, we chose pupil size as an alternative proxy to effort, based on previous work linking pupil size both to WM load (Koevoet et al., 2023) and to effort (van der Wel and van Steenbergen, 2018). Pupil size RDMs were built in each time window via the following procedure. First, the pupil diameter was baselined using the same window as the EEG data, i.e., the 200 ms preceding stimulus onset, then binned in 25 ms time windows. To facilitate intersubject averaging, the pupil data was *z*-score standardized across all trials and time points (Naber et al., 2013), and then averaged within each condition across time. Finally, the data were averaged across subjects within each condition, and the RDMs were constructed for each time window. Lastly, we included an RDM reflecting the interaction of pupil size and accuracy. To do this, we *z*-score standardized both the group-averaged accuracy across conditions and the group-averaged pupil size across conditions for each time window and multiplied the two together before constructing the final RDM at each time window.

We examined the unique information for each regressor relative to other regressors by computing the variance inflation factor (VIF) for each subject at each time point (as the

coherence level, pupil, and accuracy* pupil RDMs changed across time). The VIF of a given regressor (*i*) reflects the proportion of a given regressor's variance that is explained by other regressors and is computed as follows:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R^2 is the R^2 of a linear regression model predicting the values of regressor *i* from the remaining regressors. Common VIF thresholds for concern or exclusion are 5 and 10, which mean that 80 or 90% of a given regressor's variance are explained by the other regressors in the model.

Identifying theoretical factors that explain reliable neural variance. To compare the overall empirical RDM to our factor RDMs, we used a rank regression procedure (Iman and Conover, 1979; Kiat et al., 2022). For each theoretical factor, we computed the semipartial rank correlation by subtracting the R^2 of a submodel excluding that factor from the R^2 of the full model and multiplied the square root of this difference by the sign of the factor's coefficient in the full model. As in Kiat et al. (2022), we chose rank correlation as we did not assume a linear relationship between any of our theoretical factors and the observed condition distances. We tested these correlations at the group level using a one-tailed Wilcoxon sign-rank test against zero. We tested each time window, with FDR correction using the Benjamini–Hochberg procedure.

Estimating spatial attention signals. One possible explanation for a generalizable, content-independent load signal is that it reflects the allocation of spatial attention which differs across set sizes but is consistent across attended feature conditions. To examine this possibility, we assessed whether we could decode how spatial attention was distributed across trials and whether it differed

across set sizes, in both experiments. For set size 1 trials, we attempted to decode which quadrant the colored target was in. For set size 3 (Experiment 1) or set size 2 (Experiment 2) trials, we attempted to decode which quadrant the irrelevant placeholder was in. The broader question we examined was whether these spatially specific patterns could provide an alternative explanation for why load decoding generalized across distinct features.

Given there are four possible locations for each set size and each attended feature (16 total conditions per experiment), we used an analytical pipeline similar to the RSA above. Namely, at each time point, we computed the cross-validated Mahalanobis distance between all conditions simultaneously across 1,000 train–test splits, following the same procedure as above. Across splits, the different locations in the training set were balanced via down-sampling to equally contribute to the estimate of the covariance matrix. We next took the signed square root of these distance measures, as a cross-validated estimate of the *d*', or SNR between conditions. At each time point and for each participant, we averaged the *d*'s between locations together, separately for each set size of each attended feature.

For each set size of each attended feature, we tested whether spatial decodability between locations (average *d*') was greater than 0 using a one-tailed *t*-test at each time point, with FDR correction using the Benjamini–Hochberg procedure as above. At each time point, we also ran a two-way repeated-measures analysis of variance (rmANOVA) with a factor for attended feature (color vs orientation/motion coherence) and set size (1 vs 3/2). *p*-values were FDR corrected using the Benjamini–Hochberg procedure.

Results

Behavior is impacted by WM load and remembered feature

Across both experiments and conditions, subjects performed the task with above-chance accuracy (Fig. 16). In Experiment 1, an rmANOVA revealed a significant main effect of set size, indicating that accuracy declined as set size increased, $F_{(1,11)} = 20.59$, $p < 0.001$. There was also a significant main effect of attended feature, $F_{(1,11)} = 32.74$, $p < 0.001$, and a significant interaction of attended feature and set size, $F_{(1,11)} = 16.72$, $p < 0.001$. Accuracy was high for color ($M = 0.95$; SD = 0.069) than orientation ($M = 0.89$; SD = 0.099), and this effect was larger in the set size 2 condition than in the set size 1 condition ($t_{(11)} = 4.09$; $p = 0.0018$).

In Experiment 2, an rmANOVA revealed a significant main effect of set size, indicating that accuracy declined as set size increased, $F_{(1,12)} = 31.47$, $p < 0.001$. There was also a significant main effect of attended feature, $F_{(1,12)} = 26.02$, $p < 0.001$, and a sign-ificant interaction of attended feature and set size, $F_{(1,12)} = 27.76$, $p < 0.001$. Accuracy was higher for color ($M = 0.98$; SD = 0.027) than for motion coherence ($M = 0.848$, SD = 0.097), and this effect was larger in the set size 2 condition than in the set size 1 condition ($t_{(15)} = 4.68$; $p = 0.0003$).

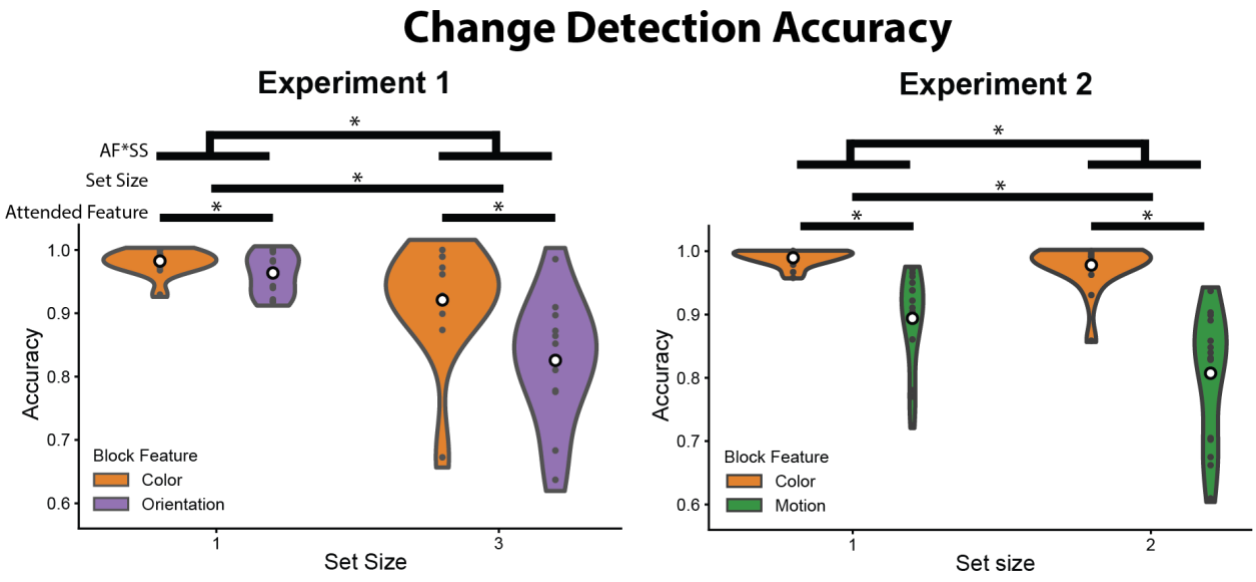

Figure 16. Change detection accuracy for each set size in both experiments. Black dots indicate individual data, white dots indicate means, and shaded regions indicate the density of the data.

The attended feature can be reliably decoded across time

EEG-based decodability

In each experiment, we first assessed whether subjects were attending to the relevant feature by computing the decodability of the attended feature (Figs. 17, 18). In both experiments, decodability, as measured by the hyperplane contrast, was sustained throughout the entire delay period (Fig. 18). In Experiment 1, the mean hyperplane contrast during the delay period was 0.730 (arbitrary units, SD = 0.756). In Experiment 2, the mean hyperplane contrast during the delay period was 3.696 (arbitrary units, SD = 1.564). We note that this is a very large signal—if the predictions were binarized by category and scored as accuracy instead of the contrast, it would peak ~90% and end ~70%. In line with this, remarkably different ERPs were evoked by physically identical stimuli when the relevant feature switched between color and motion coherence (Fig. 17).

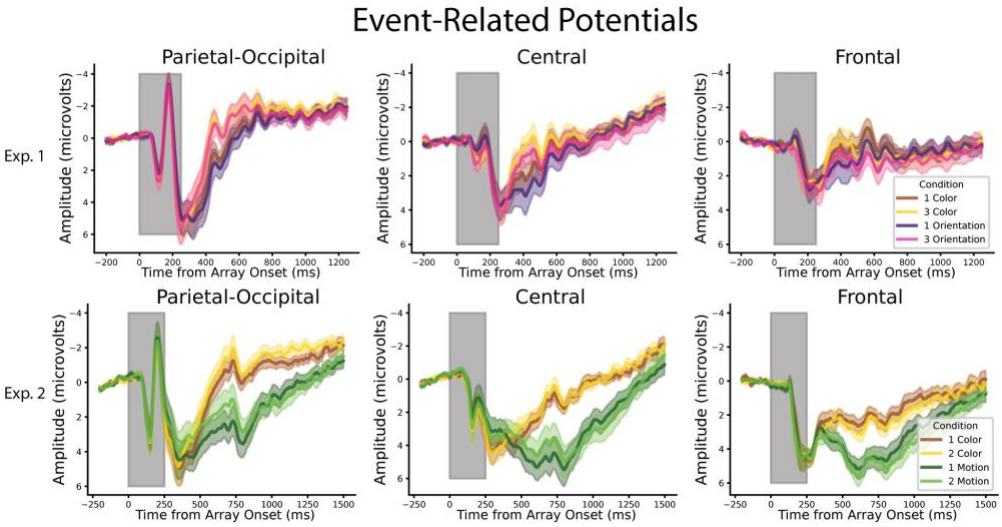

Figure 17. Event-related potentials (ERPs) for each condition in both experiments at parietal/occipital, central, and frontal electrodes. The gray region indicates stimulus presentation.

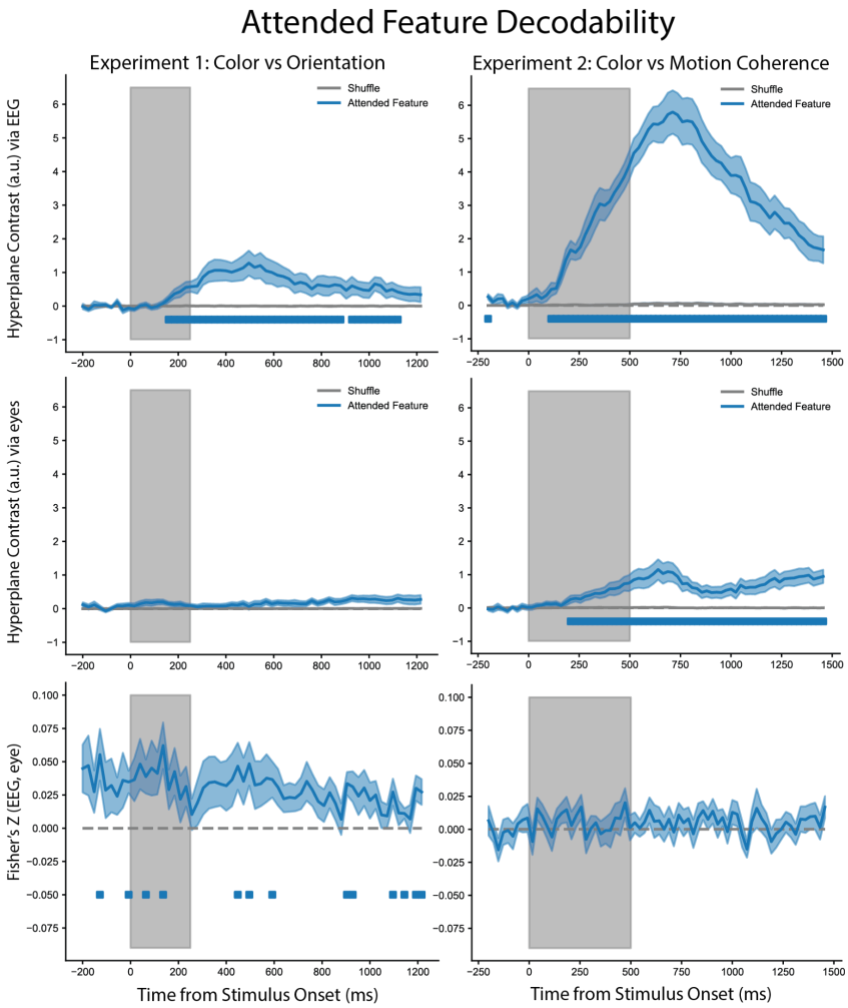

Figure 18. Decodability of attended feature in (a) Experiment 1 and (b) Experiment 2. Top row, The blue line represents decodability for which feature the subject is attending to (and subsequently remembering) in a given trial. The gray line indicates empirical chance hyperplane contrast. The blue squares represent times when the hyperplane contrast is significantly greater than empirical chance (corrected $p < 0.05$, FDR = 0.05 with Benjamini–Hochberg procedure). The gray vertical rectangle indicates the time period during which the memory array was displayed. The shuffle condition reveals empirical chance accuracy, obtained by training the model on nonpermuted data then testing on data with permuted trial labels. Middle row, Same as top row, but the model is trained and tested on eye tracking data.

Bottom row, The correlation between the EEG-trained model and the eye-trained model of hyperplane contrasts across permutations, transformed to Fisher's Z. The blue squares represent times when the correlation is significantly above 0 (corrected $p < 0.05$, FDR = 0.05 with Benjamini–Hochberg procedure).

Eye-based decodability

We also assessed whether the decodability of the attended feature was driven by eye movements. In Experiment 1, attended feature could not be decoded from eye movements at any time point. Despite this, there was a correlation of the contrasts across permutations between the two models, which was significant at times. However, this correlation was very small (mean *Z* across the delay = 0.0263, *r* also = 0.0263). In Experiment 2, there was sustained decodability of attended feature from eye movements throughout the delay. However, the time courses of decodability were different, with eye movement having a secondary increase in decodability late in the delay period that was missing from the EEG results. In addition, there was no significant correlation of the contrasts across permutations between the two models. Together, this suggests that eye movements had little impact if any on the decodability of the attended feature.

The amount of motion coherence can be decoded, but only when attended

In addition to attended feature decodability, we examined whether we could decode the level of coherence (i.e., the number or proportion of coherent stimuli) in the display. We focus on

decoding coherence in the attend-motion blocks, though we repeated the same analyses in the attend-color blocks.

First, we asked whether we could decode the total amount of coherence on the screen, regardless of how many coherent items were stored, by identifying pairs of conditions that differed only in the number of coherent distractors and pooling them together. We found no decodability for the total amount of coherence, either via EEG signals or eye movements (Fig. 19, left column).

We next asked whether we could decode the number of coherent stimuli that were stored in working memory by contrasting set size 1 trials with and without a coherent target (Fig. 19, right column). We were able to decode the attended coherence level via both EEG and eye movements, though their contrasts were not correlated across permutations. Notably, there were only two significant time points overall after FDR correction, and there were no longer any significant time points after excluding the participant for which we did not have eye tracking, indicating that the signal-to-noise ratio is low.

We repeated these analyses using attend-color trials and found no decodability of the amount of coherence for either signal via EEG or eye movements and no correlations in the contrasts across permutations. This suggests that subjects are only maintaining information about the number of coherent clouds when the information is relevant to the task at hand, in line with previous findings that observers have voluntary control over which features are stored from an object (Woodman and Vogel, 2008; Serences et al., 2009). Future work may confirm the presence of this signal with a larger sample and leverage it to study feature-specific activity and its time course in WM.

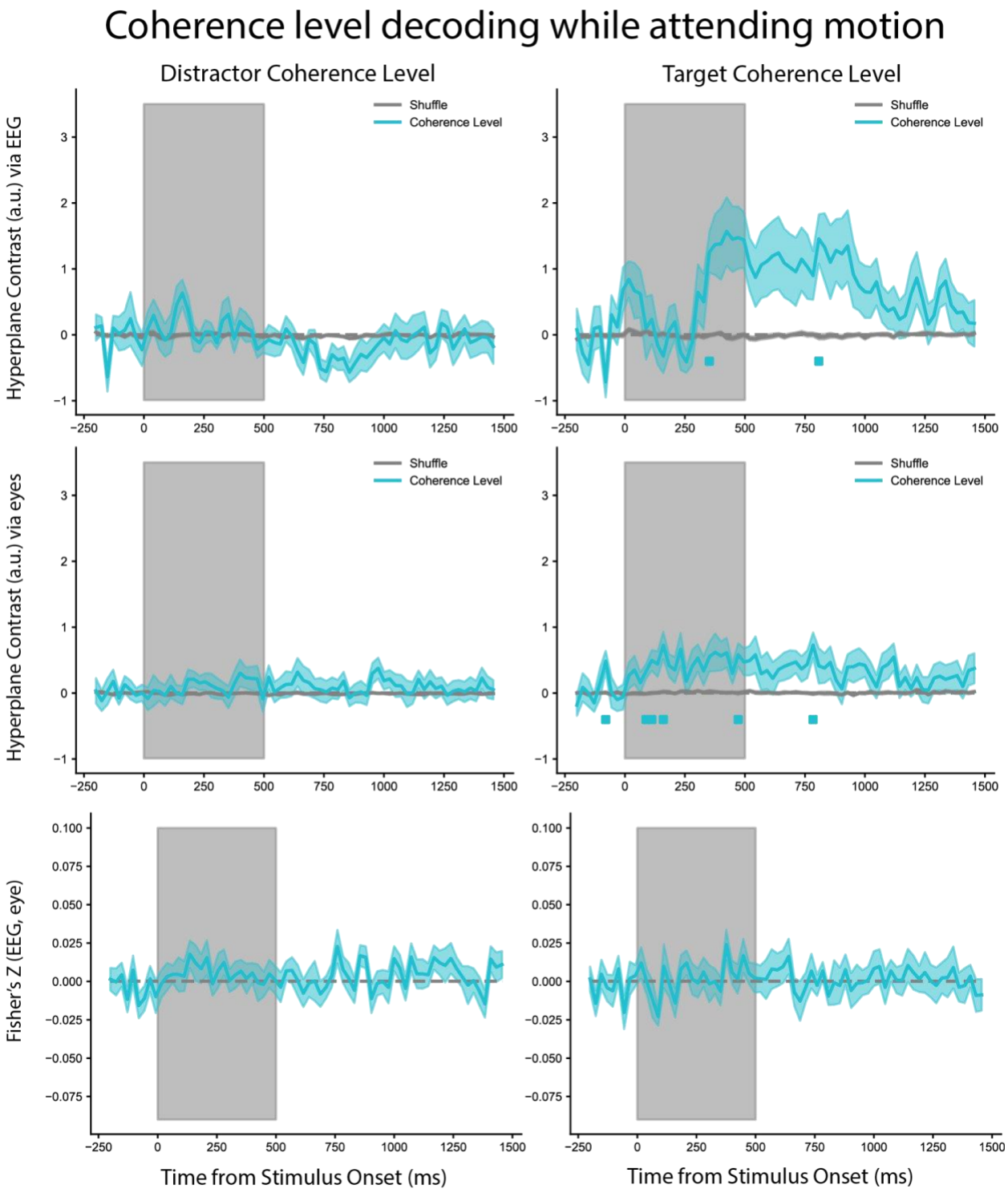

Figure 19. Decodability of motion coherence, within attend-motion blocks. The left column presents results from contrasting pairs of conditions that differ only in the coherence level of distractors; the right column presents results from contrasting the target coherence level within

set size 1 trials. Top row, Colored lines represent decodability of coherence level via EEG. The gray lines indicate empirical chance hyperplane contrast. The colored squares represent times when decodability is significantly above empirical chance (corrected $p < 0.05$, FDR = 0.05 with Benjamini–Hochberg procedure). The gray vertical rectangle indicates the time period during which the memory array was displayed. Middle row, Same as top row, but the model is trained and tested on eye tracking data. Bottom row, The correlation between the EEG-trained model and the eye-trained model of hyperplane contrasts across permutations, transformed to Fisher's Z.

WM load can be decoded and robustly generalizes across features

EEG-based decodability

We next assessed whether we could decode WM load. We assessed the decodability and generalizability simultaneously; we trained a model to decode load within each attended feature and assessed the decodability of load both for that feature via held-out test data (within-feature) and for the other attended feature (across-feature).

Figure 20 summarizes our results in Experiment 1. We were able to decode load for both color and orientation, with sustained decodability within-feature across the delay period. In addition, we saw effectively perfect generalization. Across-feature decodability was also sustained throughout the delay period, and there was no time point in which the across-feature contrast was significantly smaller than the within-feature contrast in either case.

Figure 21 summarizes our results in Experiment 2. We again were able to decode load for both color and motion coherence, with sustained decodability within-feature across the delay period. When training on color data and testing on coherence, we again saw effectively perfect generalization, as there was no time point in which the color-based contrast was significantly

bigger than the coherence-based contrast. When training on coherence, we again saw generalization to color data, but it was imperfect. The color-based contrast was significantly smaller than the coherence-based contrast, especially during the delay period. This suggests that the model was able to pick up on motion-specific load signals during training, hurting the generalization to color. This may also affect our ability to decode the attended feature, as the model may leverage feature-specific load signals to differentiate between the attended feature conditions.

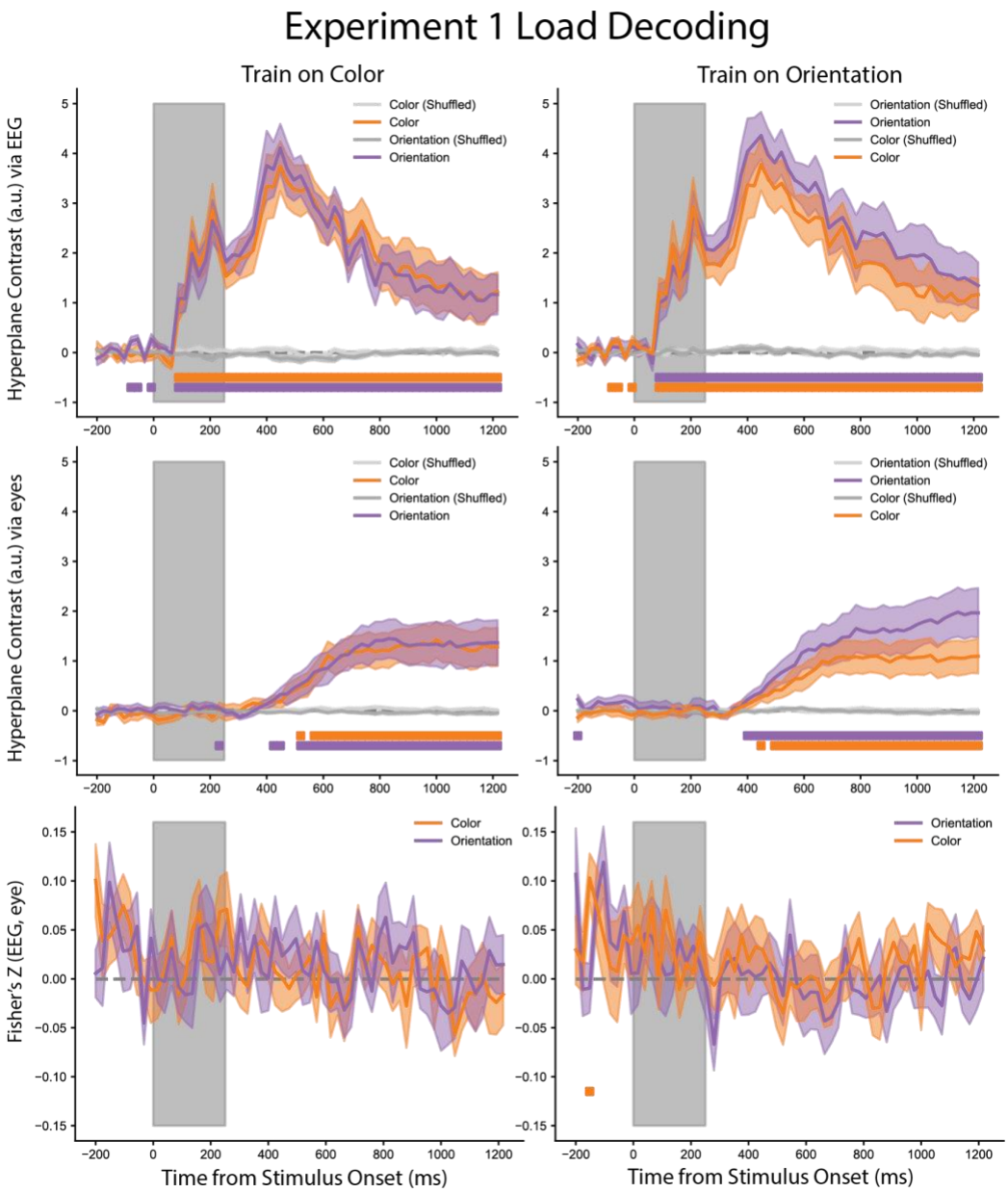

Figure 20. Decodability of WM load in Experiment 1, within and across attended features. The left column presents results from training on color data; right column presents results from training on orientation data. Top row, Colored lines represent decodability of WM load. The gray lines indicate empirical chance hyperplane contrast. The colored squares represent times when decodability is significantly above empirical chance (corrected $p < 0.05$, FDR = 0.05 with Benjamini–Hochberg procedure). The gray vertical rectangle indicates the time period during

which the memory array was displayed. Middle row, Same as top row, but the model is trained and tested on eye tracking data. Bottom row, The correlation between the EEG-trained model and the eye-trained model of hyperplane contrasts across permutations, transformed to Fisher's Z. The colored squares represent times when the correlation is significantly above 0 (corrected $p < 0.05$, FDR = 0.05 with Benjamini–Hochberg procedure).

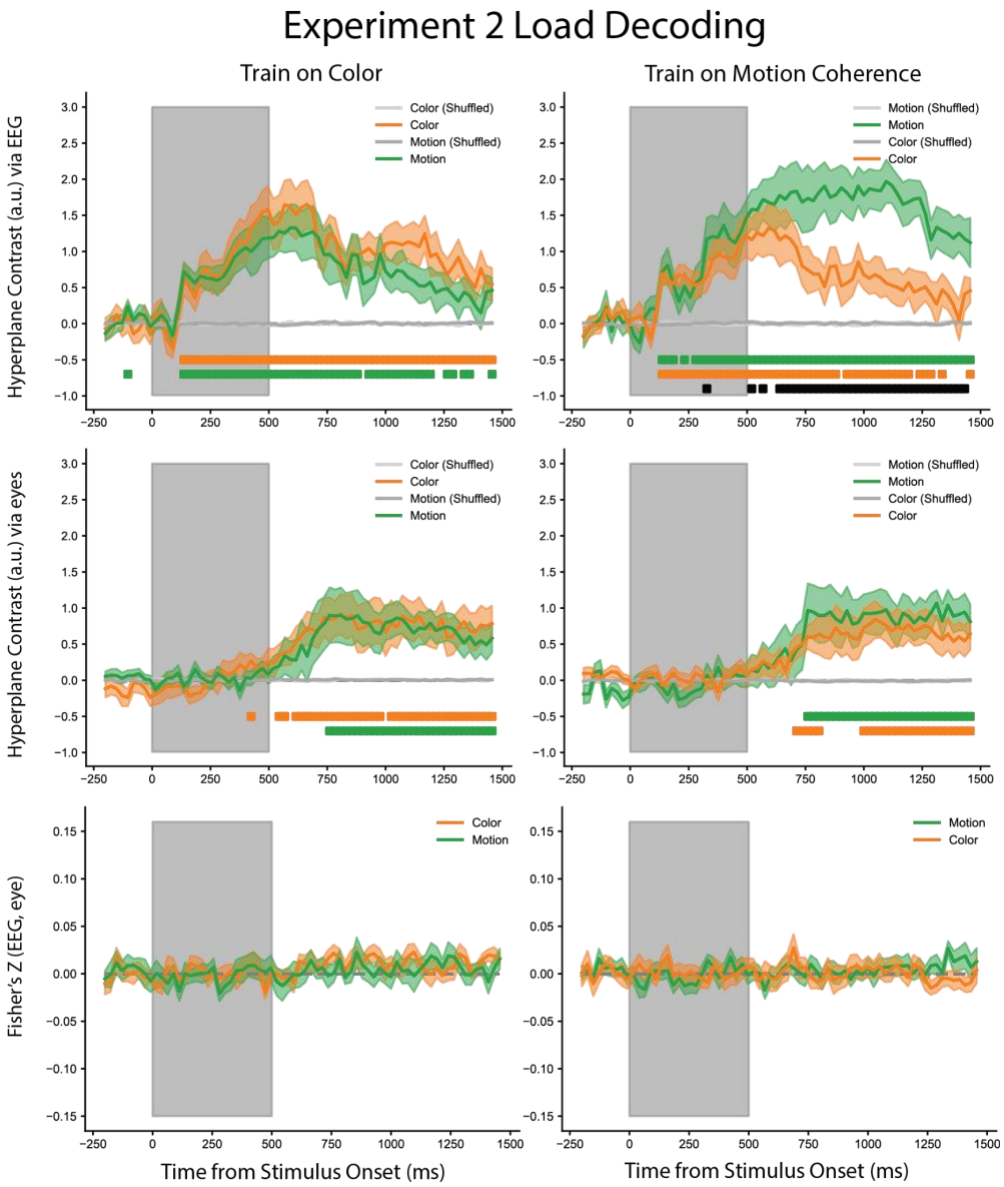

Figure 21. Decodability of WM load in Experiment 2, within and across attended features. The left column presents results from training on color data; right column presents results from

training on orientation data. Top row, Colored lines represent decodability of WM load. The gray lines indicate empirical chance hyperplane contrast. The colored squares represent times when decodability is significantly above empirical chance (corrected $p < 0.05$, FDR = 0.05 with Benjamini–Hochberg procedure). Black squares indicate time periods when the within-feature decodability was greater than across-feature decodability. The gray vertical rectangle indicates the time period during which the memory array was displayed. Middle row, Same as top row, but the model is trained and tested on eye tracking data. Bottom row, The correlation between the EEG-trained model and the eye-trained model of hyperplane contrasts across permutations, transformed to Fisher's Z.

Eye-based decodability

We also assessed whether eye movements drove load decoding in the two experiments (Figs. 20, 21, bottom 2 rows). In both experiments, load was decodable across the delay period within features, with effectively perfect generalization across features. However, the time courses of decodability and generalization differed from that of the EEG-based results, beginning in the delay period and plateauing. In contrast, EEG-based decodability began during the stimulus presentation and decreased over the delay period. In addition, the contrasts between EEG-based and eye-based models trained and tested on the same data across permutations were not correlated. The mean Fisher's Z for the eight models (2 experiments × 2 training conditions × 2 testing conditions) across the trial epochs ranged between −0.004 and 0.013, with only a single significant time point (Experiment 1, training on orientation and testing on color, during the baseline period, after FDR correction via Benjamini–Hochberg procedure).

Eye-based decodability does not drive EEG-based results across individuals

The preceding analysis examined whether eye-based decoding results corresponded to EEG-based results across permutations within individuals, finding no evidence of correlations between the two. However, the permutation outputs that were used for the contrasts may be too noisy to reliably identify correlations.

Therefore, we also examined whether eye-based results drove EEG-based results across individuals. Specifically, we fit a linear regression to predict the average across-feature EEG-based contrast from the average across-feature eye-based contrast. To increase SNR, we average time across the last 500 ms of each experiment's delay period and pooled across experiments (see Materials and Methods). The model included a unique intercept per experiment, but a shared slope association the EEG-based results to the eye-based results.

The model was significant (Fig. 22; $F_{(25,2)} = 3.616$; $p = 0.042$), with a significant regressor for eye-based decoding ($t = 2.35$; $p = 0.027$). However, both intercepts were also significant, meaning that EEG-based decoding was above chance even after regressing out the impact of eye-based decoding across individuals (Exp1: $t = 2.067$, $p = 0.049$; Exp2: $t = 2.082$, $p = 0.048$). Furthermore, visual inspection of Figure 22 suggests that the significant relationship between EEG-based decoding and eye-based decoding was driven by 2 participants in Experiment 1 (Fig. 22, top right points). Removing these subjects from the model removed the association, but did not impact the significance of the intercepts ($F_{(23,2)} = 0.114$, $p = 0.89$; $t_{eye} = 0.268$, $p = 0.791$; $t_{Exp1} = 2.671$, $p = 0.014$, $t_{Exp2} = 2.548$, $p = 0.018$). The results were preserved if we additionally removed the remaining subject that was furthest from the remaining data points (Fig. 22, top left point; also from Experiment 1) along with the other two subjects. Lastly, we confirmed that the EEG-based load decoding and generalization results of Experiment 1 are qualitatively identical

after excluding these participants. Taken together with the differences in time courses between the EEG-based and eye-based decodability, these results suggest that, while eye movements also vary systematically across WM loads in a way that is independent of the maintained features, eye movements cannot explain the generalization of WM load signals that we observe in EEG.

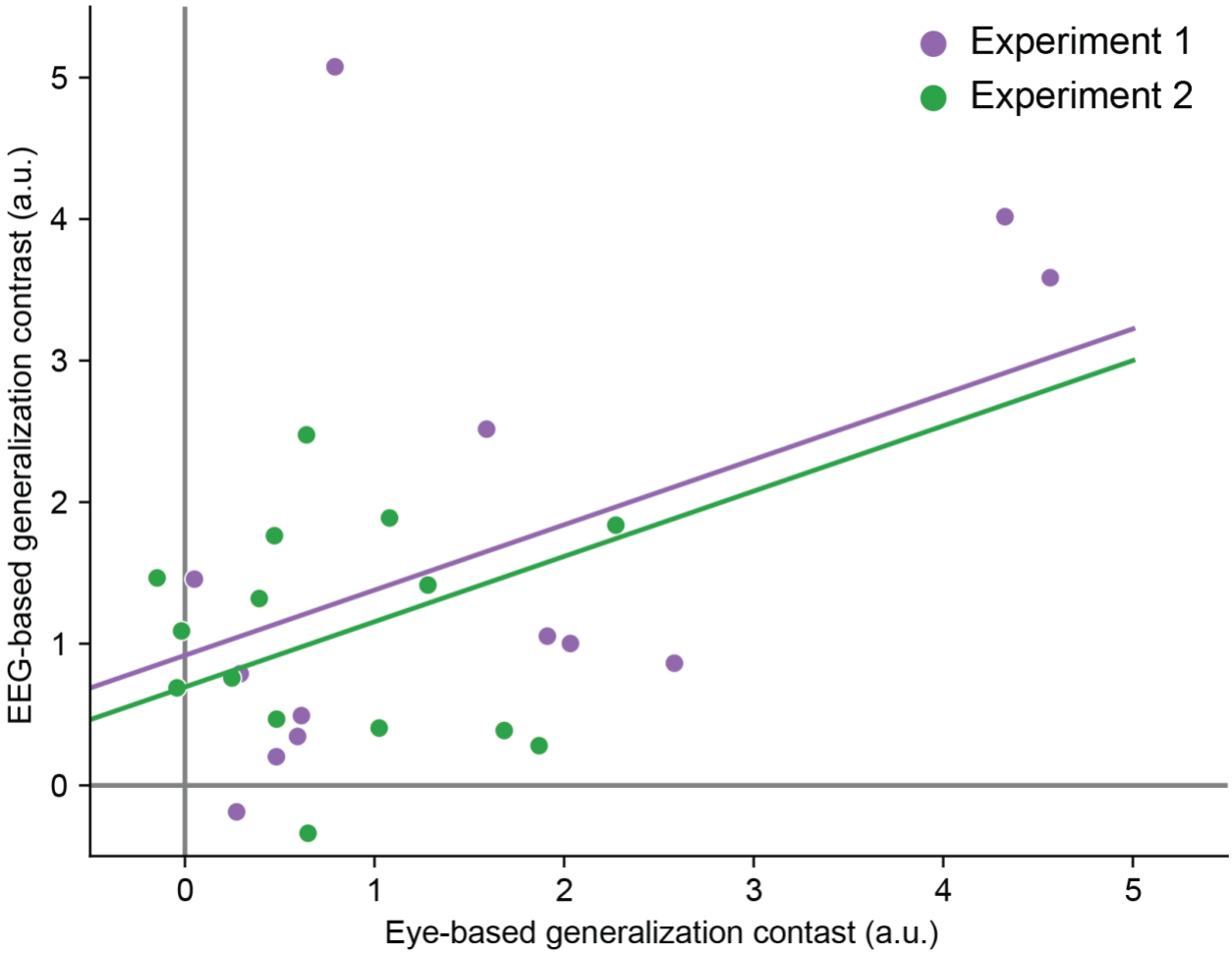

Figure 22. Comparison of EEG-based and Eye-based decodability across individuals. Dots reflect the average across-feature decodability during the last 500 ms for a given individual. Purple dots reflect participants in Experiment 1; green reflect Experiment 2.

Representational similarity analysis identifies feature-specific and content-independent signals

In addition to the decoding analyses described above, we also examined the similarity structure of the Experiment 2 conditions across time, using RSA. RSA is complementary to our decoding analyses in two important ways. First, decoding models will leverage all available signals for differentiating a pair of conditions, making interpretation difficult if multiple features are changing. Second, RSA forces researchers to make relatively precise predictions about the nature of signals of interest, and RSA enables concurrent measurement of the unique variance in neural activity that is associated with each predicted signal. Thus, we used RSA to confirm the presence of a content-independent load signal, while simultaneously examining the presence of other related signals. Within each time window, we computed the semipartial rank correlation between our empirical RDMs and a set of RDMs based on theoretical factors (see Materials and Methods). This allows us to examine the variance that is uniquely explained by a given factor, while controlling for the other, potentially confounding factors. For example, it is unclear to what extent attended-feature decodability is driven by task-set-like signals or the ability of the models to leverage feature-specific load signals. RSA provided converging evidence for feature-specific and content-independent signals.

We found evidence for the presence of multiple signals in the EEG data (Fig. 23). First, we again found evidence for an attended feature signal, which onset shortly after stimulus onset (window centered on 52.5 ms) and lasted through a large portion of the delay period (final time window centered on 1,012.5 ms). Second, while we did not find evidence for a color-specific load signal, we did find evidence for a motion-specific load signal, which achieved significance later in the delay period (first window centered on 892.5 ms). Notably, the time course of the motion-specific load signal appeared to match that of the drop in generalization in decodability when training on attend-motion data and testing on the attend-color data. This suggests that the imperfect

generalization we observed when training on motion data and testing on color data was driven by the existence of motion-specific load signals. As we did not have a large enough sample for an individual differences analysis, we instead ran a within-subject analysis to investigate this, post hoc. For each subject, we correlated the time course of the motion-specific load signal with that of the drop in generalization when training on attend-motion data and testing on attend-color data ($contrast_{motion} - contrast_{color}$), using the 70 time windows, transformed the correlations to Fisher's *Z*, and assessed the correlations at the group level with a two-tailed *t*-test. There was a significant correlation between the two time courses at the group level (mean $Z = 0.411$; mean $r = 0.390$; $t = 7.08$; $p = 5.48 \times 10^{-6}$; note that this comparison and the whole RSA were run using the 15 subjects with available eye tracking and pupil size data), supporting the idea that the drop in generalization in decodability was driven by the leveraging of motion-specific load signals.

In this analysis, the coherence signal was not significant at any time point. We compared it to the other regressors by examining its VIF, a measure of the proportion of variance in the coherence RDM that was explained by the other RDMs. When a VIF is high, it can be difficult to recover accurate estimates of a regressor's contribution to a model. With this said, the VIF for coherence was low, peaking ~750 ms at 3 (roughly two-thirds variance explained), but averaging at 1.41 (roughly 30% variance explained), meaning that it can only be partially explained by other regressors. It may be that previous decoding results simply reflected the same variance captured by these other regressors, or it may be that this partial overlap with other regressors, combined with the exclusion of the subject without available eye tracking data, reduced our sensitivity to an already subtle signal in this analysis. In addition, it may be that the signal reflects each individual's ability to extract coherence and therefore is hindered by using predictions based on group averages.

We also examined accuracy, pupil size, and their interaction as proxies for the varying difficulty or effort across conditions. To increase SNR, given the relatively low trial count per condition, we averaged measures across participants before constructing the predicted RDMs. The accuracy-based predicted differences were significant at 1 time point (window centered on 604.5 ms), but the pupil size based predicted differences were never significant, nor were the predicted differences based on the interaction of accuracy and pupil size. We also examined predictions based on individual-specific differences in accuracy, pupil, size, and their interaction; these produced qualitatively identical results.

Finally, we saw evidence for a sustained content-independent load signal (first significant window centered on 124.5 ms, last centered on 1,372.5 ms). This result was qualitatively unchanged if we excluded the color-specific load factor, which appeared to have the worst fit to the empirical RDM, or if we excluded any other subset of regressors. We also examined an alternative coding of accuracy, using the level of the behavioral analysis (2 set sizes × 2 attended features). This regressor caused the motion-specific load signal to no longer be significant at any time point. Examination of the VIF revealed that this accuracy regressor was highly similar to the motion-specific load regressor, producing extremely high VIFs for both accuracy (mean VIF = 246.44) and motion-specific load signals (mean VIF = 270.71), with over 99.5% of each factor's variance being explained. In this model, the attended feature and content-independent load RDMs were again qualitatively unchanged. Overall, RSA provided clear converging evidence for the presence of a content-independent load signal that is independent of the attended feature of the stored items.

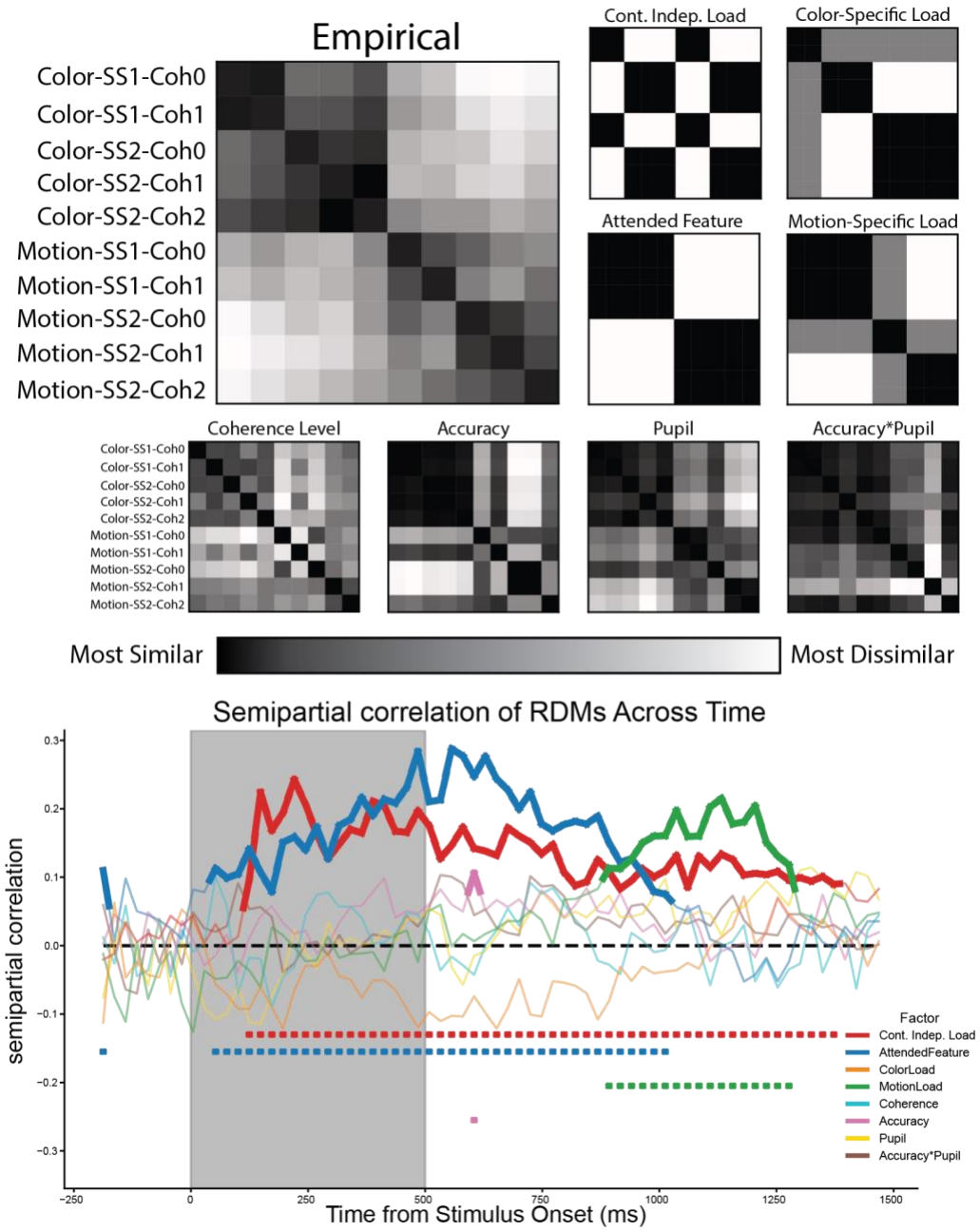

Figure 23. RSA results. Top, The Empirical RDM, averaged across the delay period and across subjects, along with representations of the eight regressors. All RDMs are organized by (1) attended feature (color, motion), (2) set size (1, 2), and number of coherent targets (0,1 for set size 1; 0, 1, 2 for set size 2). The content-independent load, attended feature, and feature-specific load RDMs are all based on theoretical predictions, while the coherence-level RDM is based on

*a hyperplane analysis (see Materials and Methods), the accuracy RDM is based on the condition-level behavioral performance, and the pupil and accuracy*pupil RDMs are based on the average normalized pupil sizes across conditions. Note that the bottom four RDMs visualized here reflect the average of all subjects, averaged across the delay period, but the coherence, pupil, and accuracy*pupil RDMs were computed separately in each time window, and the coherence RDM for each subject exclude their own data as it was also based on EEG data. Bottom, The semipartial correlations between the theoretical factors and the empirical RDMs estimated at each time window, averaged across subjects. Colored lines reflect the semipartial correlations of a given theoretical factor. The colored squares, and thicker segments of the corresponding line, indicate that the correlation for that factor is significantly greater than 0 (one-tailed Wilcoxon sign-rank test) after FDR correction via the Benjamini–Hochberg procedure.*

Spatial attention signals are transient in raw voltage, unlike WM load signals

One possible explanation for an apparent content-independent load signal is that is driven by changes in spatial attention across set sizes. To examine whether this is the case in the current dataset, we tested our ability to decode different configurations of targets across set sizes. We took advantage of the fact that each stimulus was presented in a distinct quadrant, meaning that each target or placeholder can be coded by the quadrant it was presented in. For each experiment and each attended feature, we simultaneously computed the cross-validated *d*' (a similar contrast measure to the hyperplane contrast used above; see Materials and Methods) between every pair of target locations for set size 1, and every pair of distractor locations for the higher set size. We

compared each decodability trace against 0 using one-tailed *t* tests and ran two-way rmANOVAs comparing decodability across set sizes and attended features.

In both experiments, we could decode the location of both the set size 1 target and the higher set size placeholder (Fig. 24). In addition, a two-way rmANOVA revealed set size effects when examining differences in spatial decodability in both experiments. However, these effects were transient, ending soon after stimulus onset (final time bins centered on 460.5 ms in Experiment 1 and 724.5 ms in Experiment 2). This suggests that the generalizable WM load signals during the delay period are not driven by differences in the strength of spatial attention signals.

One may argue that even if the overall strength of spatial signals is equated across set sizes, there may still be a difference in format of the spatial signal between set sizes (e.g., an enhancement signal for set size 1 targets, and a suppression signal for higher set size place holders), which may drive load decoding. However, the overall spatial decodability time course is also too transient for this account. In both experiments, this spatial decodability peaked during or soon after the initial stimulus presentation, before dropping (Fig. 24). In Experiment 1, the orientation-based spatial decodability was sustained, but color-based spatial decodability was sparse after ~650 ms (time bin centered on 628.5 ms), with a few more significant moments between ~800 and 1,050 ms (final time bin centered on 1,036.5 ms). In Experiment 2, motion-based spatial decodability of set size 1 targets lasted until ~1,100 ms (final time bin centered on 1,108.5 ms), and color-based spatial decodability was sparse starting ~900 ms (window centered on 892.5 ms), with a few more significant time points between ~1,000 and 1,150 ms (final time bin centered on 1,156.5 ms). Thus, the time windows in which spatial decoding succeeded could not explain the sustained time course over which content-independent load activity was observed.

These time courses also align with recent work which attempted to actively separate spatial attention signals from WM load signals through the use of "dot cloud" stimuli, which could change in overall area and overlap with one another, enabling a separation between the number of items stored and the breadth of the attended region (Jones, Diaz et al., 2024). This work found that indeed the two signals are dissociable and that signals which reflected the breadth of spatial attention fade rapidly across the delay period while WM load signals are sustained. We note that, both this previous work and the current work find this drop in spatial signals when examining raw EEG voltage, rather than other signal bands such as alpha power, as we are curious about whether spatial attention signals can drive WM load signals, which are also estimated using raw EEG voltage. Taken together, these results suggest that spatial attention signals cannot explain the sustained and generalizable WM load decoding that we find throughout the delay period.

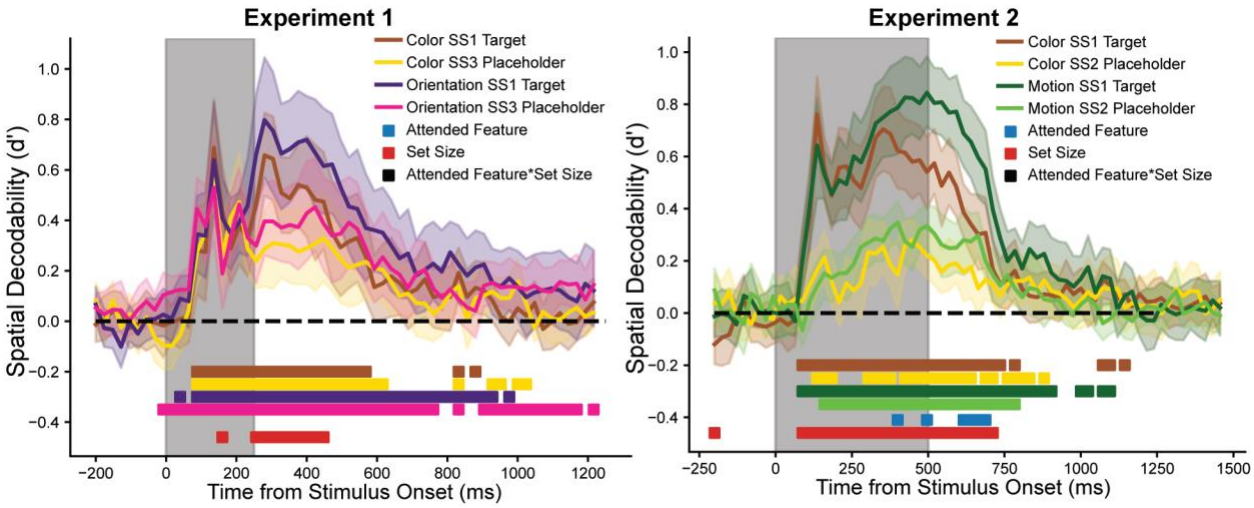

Figure 24. Spatial decodability across attended features, set sizes. Time courses reflect the average decodability (cross-validated d'; see Materials and Methods) between pairs of locations. Darker traces reflect decoding of set size 1 target locations, and lighter traces reflect decoding of the placeholder location at higher set sizes. Corresponding colored squares reflect above-change decodability, after FDR correction via the Benjamini–Hochberg procedure. Additional

colored squares reflect the significant of ANOVA results, also after correction. The blue squares reflect a main effect of attended feature. Red squares reflect of main effect of set size. The black squares would reflect an interaction between attended feature and set size; this interaction was never significant.

Discussion

Our findings provide evidence for a content-independent WM load signal that tracks the number of items stored regardless of the features that define those items. At the same time, distinct aspects of the EEG signal tracked the attended features, in line with many past studies that have reported stimulus-specific neural activity during WM storage. Thus, we are not challenging the importance of stimulus-specific delay activity. Instead, we are highlighting the importance of a distinct class of storage-related neural activity that is not tied to feature-specific activity.

Earlier work from Thyer et al. (2022) provided evidence for a content-independent load signal by showing strong generalization between WM load models for color and orientation. Here, we provide three critical extensions of the prior work. First, we replicated the generalization between color and orientation load signals using a within-session manipulation of the attended feature, showing near-perfect generalization of the load signature between color and orientation. Second, we show that WM load signatures generalize across color and motion coherence stimuli, features that are known to be processed in cortically disparate regions. Thus, content-independent load patterns could not be explained by the hypothesis that to the neural populations processing the two features were too finely interwoven to be distinguished by EEG. Third, we provide converging evidence for a content-independent load signal from a different analytic approach,

RSA, which showed that the unique variance in the EEG signal was explained by a content-independent load signal while simultaneously identifying feature-specific signals.

We considered the possibility this content-independent load signal actually reflects the load for a single consistent feature that is encoded across conditions. For example, in Experiment 2, color was the target-defining feature. If participants must attend to color to select targets for both feature conditions, this signal may reflect the number of colors in memory. However, there are both theoretical and empirical arguments against this possibility. First, if participants must encode colors into working memory in order to assess target status, they would also need to encode the colors of distractors into working memory, leading to equivalent color load between the set sizes. Relatedly, there is growing empirical evidence that people can attend to target-defining features, requiring a relatively deep amount of processing, without appearing to maintain the featural information in WM, leading to a phenomenon known as attribute amnesia (Chen and Wyble, 2016). Thus, even if participants are only attending to the color of targets and not the distractors to make a target judgement, the literature suggests that they do not encode it into (or else rapidly remove it from) WM if it is not relevant.

Second, we found consistent evidence in this work that participants are selectively attending to the relevant feature. In both experiments, we found that we could decode the attended feature. Further, in Experiment 2, while RSA showed evidence for motion-specific load signals, and decoding results even provided initial evidence that we can decode the amount of coherence present, but only when it was the relevant feature (though future work with a larger sample size is needed to confirm its presence). The latter finding falls in line with past studies showing that observers can selectively store the relevant features from multifeature objects (Woodman and Vogel, 2008; Serences et al., 2009). For instance, Serences et al. (2009) presented observers with

oriented gratings of different colors and found that patterns of activity in the primary visual cortex faithfully decoded the attended but not the unattended features of the gratings. Taken together, this pattern of results strongly argues against the possibility that participants are simply maintaining one or both features equally regardless of the experimental condition and supports our interpretation of the signal as truly content independent.

What underlying process or processes does this signal reflect? Although we offer a working hypothesis that focuses on the binding of items to context, we acknowledge the need to consider alternative explanations based on nonmnemonic processes that might be confounded with WM load. Here, we examined two prominent possibilities, cognitive effort and spatial attention (shown in past work to support rehearsal in visual WM; Awh and Jonides, 2001; Williams et al., 2013). We attempted to capture putative changes in effort across conditions by including accuracy, pupil size, and their interaction as proxies for effort in the RSA. We continued to find evidence for a sustained content-independent load signal, arguing against effort as an explanation. To examine whether content-independent decoding of load could be explained by a common spatial attention signal, we examined whether EEG activity tracked the locus of spatial attention and whether this spatially specific signal distinguished between set sizes. Although we did observe differences in spatial decodability across set sizes, these effects were transient and did not align with the time course of the WM load signal. This is consistent with recent work (Jones, Diaz et al., 2024), in which we have found that spatial attentional signals were more transient than load signals and explained distinct variance in EEG activity. Thus, differences in effort or covert spatial orienting do not explain the sustained content-independent load signals observed here. Given these results, we next offer a working hypothesis that these content-independent signals may reflect a content-general process for binding item representations to the surrounding context.

Prominent theoretical accounts of WM storage propose separable neural processes for the maintenance of item representations, on the one hand, and the individuation and binding of the stored representations to the current context, on the other hand (Xu and Chun, 2006; Swan and Wyble, 2014; Balaban et al., 2019; Bouchacourt and Buschman, 2019; Oberauer, 2019). These proposals align with past theories of dynamic visual cognition (Kahneman et al., 1992; Pylyshyn, 2009), which require a spatiotemporal indexing process that enables the continuous monitoring of items through time and space (Hakim et al., 2019; Thyer et al., 2022). Thus, our hypothesis is that content-independent load signals may reflect the number of spatiotemporal pointers that are deployed to bind items to context, enabling their maintenance and retrieval from WM.

This distinction between content-independent pointers and feature-specific neural activity may provide a productive perspective for a number of findings in the working memory literature. For example, Fukuda et al. (2010) found that the number of items that each individual could store in working memory was correlated with fluid intelligence, while no similar link was observed between intelligence and the precision of the stored memories. It is possible that the number of items that can be stored is limited by the number of pointers that can be simultaneously deployed, while mnemonic precision is determined by parallel but separate neural processes that maintain precise feature representations for each stored item. Another key finding in the visual WM literature is that storage is limited by the number of objects instead of the total number of feature values to be stored (Luck and Vogel, 1997). Although there is some behavioral cost of adding new features to objects that are stored in working memory (Olson and Jiang, 2002), there are robust "object-based benefits," such that substantially more features can be stored when they are integrated within a smaller number of objects. Robust evidence for object-based limits in visual WM storage (Ngiam et al., 2023) argues against a storage ceiling that is determined by the total

amount of information stored. However, if there is a limit to the number of content-independent pointers that can be concurrently deployed, this could explain why visual WM performance is better predicted by the total number of objects to be stored than by the total amount of feature information contained within those objects. Finally, a dichotomy between the maintenance of stimulus-specific details and content-independent indexing operations may explain why distinct regions of the cortex have been shown to track the number and complexity of objects in visual WM (Xu and Chun, 2006). For example, the motion-specific load signals found in this work may reflect the maintenance of those coherence representations in downstream visual regions (Zeki, 1978; Felleman and Van Essen, 1991; Vaina, 1994), which are pointed at (producing the content-independent load signal) in order to be actively maintained and accessible in working memory.

It is important to note that different theories of WM storage may describe different content-independent processes that scale with WM load. One possibility is that WM storage relies on the oscillation of the focus of attention between WM representations which would otherwise rapidly decay (Mongillo et al., 2008; Landau and Fries, 2012; Fiebelkorn et al., 2013; Chota et al., 2022). Under this scenario, attention may change in sampling frequency as more items are oscillated between with increasing WM load (Holcombe and Chen, 2013). The signal we observe may reflect this change in sampling frequency, and this theory may be extended to incorporate the object-based benefits described above. Thus, while there is a broader theoretical motivation for an item-based binding operation that scales with WM load, more work is needed to test this computational account of content-independent load activity. For example, future work can examine whether these content-independent load signals are a necessary precursor for accessing item representations based on contextual retrieval cues, as predicted by models that distinguish between recollective and familiarity-based modes of memory retrieval (Yonelinas, 2023).

In summary, we present strong evidence for a content-independent load signature that generalizes across color and orientation as well as color and motion coherence, two cortically disparate visual features. This conclusion was supported by both generalization of classifier models across distinct features, as well as by RSA analyses that identified a pure WM load signal. We argue that this signal is distinct from spatial attention and effort and that theories of WM must include content-independent processes. These findings align with a broad class of models that argue for a distinction between the maintenance of stimulus-specific details and the binding of stored items to the current context. We propose that this signal may reflect the allocation of spatiotemporal pointers for binding items to context, providing a unifying perspective on previously identified neural and behavioral effects.

Chapter 2b: Neural evidence for modality-independent storage in working memory

Abstract

Working memory (WM) is a core component of intellectual ability. Traditional behavioral accounts have argued that there remain distinct memory systems based on the type and sensory modality of information being stored. However, more recent work has provided evidence for a class of neural activity that indexes the number of visual items stored in a content-independent fashion. Here, across 2 electroencephalogram (EEG) experiments, we demonstrate an item-based signature of WM storage that generalizes across visual and auditory sensory modalities. Using multivariate techniques, we observed parallel but separate neural patterns that independently track stimulus modality, the number of items stored in WM (regardless of modality), and the number of spatially attended positions. We propose that these load signals reflect a modality-independent process for binding item representations to context, reinforcing accounts arguing for a distinction between the maintenance of the content of one's thoughts and the manipulation and gating of those thoughts.

Introduction

Neural studies of working memory (WM) have made major progress by examining stimulus-specific activity that tracks the content of stored items (Fuster et al., 1971; Fuster & Jarvey, 1981; Golman-Rakic, 1996; Harrison & Tong, 2009; Serences et al., 2009). These neural signals track the voluntarily stored aspects of relevant items (Serences et al., 2009, Yu & Shim, 2017), predict mnemonic fidelity (Ester et al., 2013; Vo et al., 2022; Zhao et al., 2020), and

provide insight into the dynamics of selection and access to stored content (Panichello & Buschman, 2021; Libby & Buschman, 2021; Lewis-Peacock et al., 2012). Nevertheless, recent work has highlighted evidence for a distinct class of neural activity that indexes the *number* of items encoded into WM, independent of the specific content associated with those items (Vogel & Machizawa, 2004; Xu & Chun, 2006; Rajsic, Burton & Woodman, 2019; Thyer et al., 2022; Jones, Diaz et al., 2024; Jones, Thyer et al., 2024). For instance, multivariate analyses of electroencephalogram (EEG) activity reveal a common signature of the number of items stored in WM, despite variations in both the *type* (e.g., color, orientation, and motion) and the *number* of feature values stored for each item (Thyer et al., 2022; Jones, Thyer et al., 2024). Moreover, when perceptual grouping encourages multiple elements to be perceived as a unit, this neural load signal tracks the number of *perceived items*, not the number of attended elements or positions (Thyer et al., 2022; Diaz et al., 2021). Thus, these findings reveal an *item-based* neural signature of WM storage that generalizes across highly distinct visual features.

Our working hypothesis is that these content-independent signatures of WM load are generated by a spatiotemporal "pointer" operation that binds item representations to the surrounding event context (Awh & Vogel, 2025). This kind of contextual binding has been highlighted both in major models of WM (Baddeley, 2000; Swan & Wyble, 2014; Hedayati, Donnell & Wyble, 2022; Oberauer & Lin, 2024; Yonelinas, 2024) as well as in longstanding theories of dynamic visual cognition (Pylyshyn, 2009; Kahneman, Treisman & Gibbs, 1992). Moreover, the latter theories embrace a clear separation between the indexing operation supported by pointers, on the one hand, and parallel operations that maintain the featural content of the selected items, on the other. Here, there is a useful analogy between pointers and demonstrative words such as "this" and "that": demonstratives hold no meaning by themselves,

but their function is to refer to other meaningful words. Likewise, pointers may support the contextual binding of items, while parallel operations maintain the contents of the indexed representations. Thus, pointers provide an attractive explanation for the presence of content-independent signatures of WM storage.

Extant demonstrations of content-independent load signals, however, have been restricted to the visual modality, leaving open the possibility that distinct sensory modalities recruit distinct pointer systems. Indeed, prominent WM models have argued that there are independent resource pools for the storage of visual and auditory/phonological information (Baddeley, 1986; Saults & Cowan, 2007; Cowan, Saults & Blume, 2014) and multiple studies have found minimal interference between auditory and visual loads competing for WM storage (Fougnie et al., 2015; Cocchini et al., 2002). Moreover, past EEG studies have highlighted distinct electrode sites where activity tracks auditory and visual WM loads, with anterior electrode sites tracking auditory loads (Guimond et al., 2011) while posterior parietal electrodes track the number of visual items stored (Vogel & Machizawa, 2004; Luria et al., 2016). Thus, both behavioral and neural studies have pointed toward dissociable WM systems for visual and auditory information. In the present work, we replicate the observation that ongoing EEG activity is shaped by the sensory modality that is stored, but we also present clear evidence for a modality-general component of WM storage.

Pilot work revealed that we could decode the number of discrete sounds being held in WM by applying multivariate classifiers to ongoing EEG activity, similar to recent work with visual stimuli (Thyer et al., 2022; Jones, Diaz et al., 2024; Jones, Thyer et al., 2024; Adam, Vogel & Awh, 2016). Here, we manipulated the number and sensory modality of items stored in WM, using audiovisual displays that controlled sensory energy across experimental conditions.

Experiment 1 revealed robust generalization of EEG load signatures across auditory and visual features, such that training on one sensory modality enabled precise decoding of load in the other sensory modality. Moreover, the load signature for single-feature objects (colors or sounds alone) generalized to dual-feature audiovisual objects, showing that this analysis provides an *item-based* rather than a feature-based measure of WM storage. Representational similarity analysis provided clear evidence for a sustained neural signal that tracked the number of items stored in WM, regardless of the sensory modality or informational complexity of those memories. Critically, this signal explained distinct variance in ongoing EEG activity from another robust signal that tracked the stored sensory modality. Experiment 2 replicated these key findings while independently manipulating the number of attended positions and the number of items stored in those positions. We saw clear evidence for a neural signal tracking the number of attended positions, but this explained distinct variance in EEG activity from the modality-general signal that tracked the number of stored items. Thus, our findings reinforce recent work arguing for a functional dissociation between the voluntary control of spatial attention and the encoding of items into WM (Jones, Diaz et al., 2024; Hakim et al., 2019), and they provide direct neural evidence for a modality-general component of this online memory system.

Results

28 healthy human participants performed a WM task in which both the number of items to be stored and the sensory modality of the stored information were manipulated (Figure 25). Our task employed sequentially presented audiovisual stimuli that consisted of a colored diamond and a real-world sound that onset simultaneously in one of three positions (left, middle, or right). To match sensory energy across set sizes, all sample displays included two audiovisual

presentations. For the set size 2 trials, all audiovisual stimuli were comprised of a saturated color and a real-world sound. For set size 1 trials, 1 of those audiovisual stimuli was randomly selected and replaced with energy-matched "placeholders." Participants were instructed to store the targets and ignore the placeholders in preparation for a change detection probe. Critically, in different blocks, participants were instructed to store either the color, the sound, or the conjunction of color and sound for each audiovisual target. Following a 1,000 ms delay period, a single probe stimulus was presented, and subjects reported whether it was the same or different from the stimulus presented at that location. Test probes were in the modality that matched the storage instructions. When both visual and auditory features were stored, visual and auditory probes were randomly intermixed and equally likely.

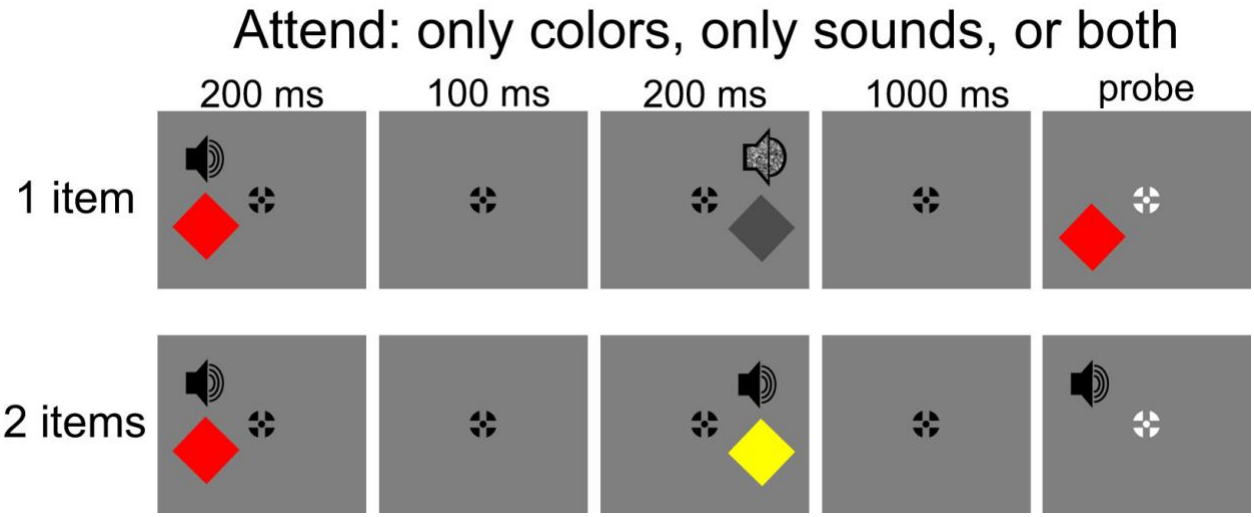

Figure 25. Experiment 1 task. Participants attended to colors, sounds, or both features. Stimuli were color-sound pairs. In set size 1 trials, one pair was replaced with white noise and a gray diamond.

We had behavioral data for 22 out of 24 subjects (behavioral data for subjects 1 and 2 were missing due to a programming error). Mean accuracy across subjects was high for all conditions (range 92.3% for auditory set size 2%–97.7% for visual set size 1). A repeated-measures ANOVA revealed a significant effect of attended modality ($F_{(2,42)} = 8.95$, $p = 5.8 \times 10^{-4}$), WM load ($F_{(1,21)} = 18.43$, $p = 3.2 \times 10^{-4}$), and a non-significant interaction term ($F_{(2,42)} = 2.52$, $p = 0.092$).

Neural signatures of WM load generalize across sensory modalities

Consistent with past work, multivariate analysis of EEG activity enabled robust and sustained decoding of the number of items stored in WM for all 3 conditions (visual, auditory, and conjunction). To test whether load decoding generalized across these conditions, we measured classifier performance when training and testing across different modalities (Figures 26A and 26B). Thus, we trained classifiers to discriminate between set sizes 1 and 2 within each modality, and we tested those classifiers on independent data in which either the same or a different modality was stored. Instead of using raw classification accuracy to measure decodability, we employed a metric called "hyperplane contrast" (Jones, Thyer et al., 2024). Hyperplane contrast is conceptually similar to decoding accuracy, with higher numbers indicating better decoding. Hyperplane contrast has the advantage of being more robust to non-orthogonal but potentially confounding neural signals. Moreover, this signal grows proportionally to the signal-to-noise ratio, improving interpretability (see Materials & Methods for additional rationale).

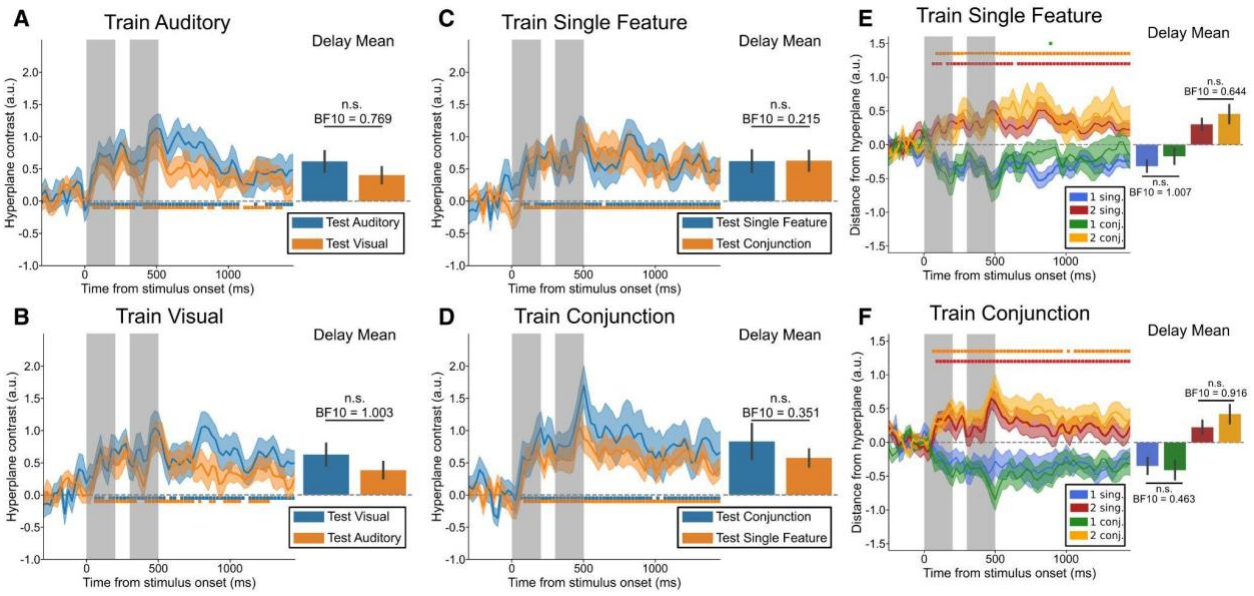

Figure 26. Decoding results for experiment 1 (n = 24). Stimuli were presented from 0 to 200 ms and 300 to 500 ms (shown in gray). Bar graphs are the average value (hyperplane contrast for A–D, hyperplane score for E and F) over the delay period (500–1,500 ms). Lines denote mean value, and the shaded region = SEM. Squares indicate hyperplane contrast values (A)–(D) greater than zero (FDR-corrected one-tailed t test). BFs for each delay period comparison (A and B: H_A = intramodal > crossmodal; C and D: H_A = single feature ≠ conjunction; E and F: conjunction > single feature) are also given. (A and B) Hyperplane contrast within and across sensory modalities. Blue: same modality as training; orange: opposite modality. (C and D) Hyperplane contrast for single-feature vs. conjunction (dual-feature) trials. Blue: same number of features as training, orange: different number of features. (E and F) Hyperplane distances for single-feature vs. conjunction trials. More negative values are higher confidence set size 1, and more positive values are higher confidence set size 2. Yellow squares indicate 2 conjunctions > 1 conjunction. Red indicates 2 singles > 1 single. Green indicates 1 conjunction > 1 single (only present at one time point). See also Figure A5 for an analysis of the effects of trial bin size on generalization.

Overall, we observed robust decoding of WM load for nearly all time points within modalities (51/58 auditory, 52/58 visual) and for most, but not all, time points across modalities (45/58 auditory to visual, 40/58 visual to auditory). When averaging over the delay period, hyperplane contrast (a measure of classifier performance) over the delay period was significantly greater than zero both when training and testing within a modality (auditory, $t(23) = 3.91$, $p = 3.48 \times 10^{-4}$, Bayes factor [BF] = 96.4; visual, $t(23) = 3.74$, $p = 5.33 \times 10^{-4}$, BF = 66.4) and when training and testing across modalities (auditory to visual, $t(23) = 3.18$, $p = 2.10 \times 10^{-3}$, BF = 20.2; visual to auditory, $t(23) = 3.01$, $p = 3.16 \times 10^{-3}$, BF = 14.288). Hyperplane contrast when testing across modalities (crossmodal) was not significantly lower than within modalities (intramodal) at any time point (one-tailed t test, false discovery rate [FDR] corrected). In the aggregate, BFs for the difference between intramodal and crossmodal contrast were 0.77 (train auditory, $t(23) = 1.143$, $p = 0.132$) and 1.0 (train visual, $t(23) = 1.388$, $p = 0.089$), indicating ambiguous evidence for either the null or alternative hypotheses. While the ambiguous Bayes factor does not permit us to rule out some domain-specificity (in fact, we show later that attended modality substantially impacts the load signal), the presence of robust cross-decoding across modalities provides strong evidence for a common neural signature of load across visual and auditory modalities.

Neural signatures of WM load are item-based, not feature-based

To test whether these load classifiers were operating at the level of items or features, we examined whether load classifiers based on single-feature stimuli generalized to conjunction conditions in which both visual and auditory features were retained. Cross-decoding between

single-feature and conjunction conditions was robust and nearly identical (Figures 26C and 26D), consistent with prior findings with conjunctions of color and orientation (Thyer et al., 2022). Hyperplane contrasts for conjunction and single-feature trials were not significantly different at any point (two-tailed t test, FDR corrected; delay average, train single-feature: $t(23) = -0.029$, $p = 0.976$, BF = 0.215; train conjunction: $t(23) = 1.049$, $p = 0.305$, BF = 0.351). While successful cross-decoding provides evidence for a common load signature, we also used a more specific analysis to distinguish whether load decoding was item-based or feature based. We trained a model to distinguish between set sizes 1 and 2 using only single-feature stimuli and then tested this model with a single conjunction stimulus. If load decoding is based on the number of feature values stored, then models trained on single-feature stimuli should be biased toward higher load classification when the number of features stored per item is doubled. This prediction was not supported. Even though twice as many features were stored in the conjunction condition, a single conjunction stimulus did not show an upward shift from the single-feature set size 1 condition in a hyperplane analysis (Figure 26E), with the exception of a single time point. No significant increase in load was observed in the aggregate (set size 1: $t(23) = 1.391$, $p = 0.089$, BF = 1.007; set size 2: $t(23) = 0.949$, $p = 0.176$, BF = 0.644). The same pattern when training the model on conjunction stimuli (Figure 26F, set size 1: $t(23) = -0.407$, $p = 0.656$, BF = 0.463; set size 2: $t(23) = 1.309$, $p = 0.102$, BF = 0.916). Thus, this modality-general load signal tracks the number of *items* stored in WM, not the number of feature values.

Generalization across modalities is not due to storage of irrelevant features

An alternative explanation for the generalization of load models across the auditory and visual conditions is that observers tended to store both auditory and visual features, regardless of

instructions. In this case, perfect cross training would be expected even if independent neural signals tracked each modality. Two pieces of evidence contradict this hypothesis. First, attended modality (auditory/visual) was robustly discriminated by our classifier (Figure 27A) to an even greater degree than load, and this feature-specific activity was sustained across the delay period. This provides direct neural evidence that observers selectively stored the relevant features. Second, we reasoned that if cross training was driven by storage of the irrelevant feature in WM, then individuals with higher decodability of the attended feature should show worse generalization between auditory and visual load models. We evaluated this in two ways. First, we correlated attended feature discriminability with crossmodal decoding performance (Figures 27B and 27D) to test the prediction that subjects worse at filtering would show better generalization across modalities. Second, we replaced crossmodal discriminability with a "penalty" metric (Figures 27C and 27E), equal to the difference between the mean intramodal and crossmodal contrast. This hypothesis predicts that poor filterers (low attended feature discriminability) should have a smaller penalty because they were in more similar attentional states during the auditory and visual conditions. However, three of these correlations were not significant (Figure 27B $p = 0.912$, BF = 0.255; Figure 27C $p = 0.562$, BF = 0.297; Figure 27D $p = 0.954$, BF = 0.254), while one was statistically significant in the opposite direction of the prediction (Figure 27E, $p = 0.016$, BF = 3.93). However, after excluding outlier values, this association was no longer significant ($p = 0.595$, BF = 0.302). Thus, storage of irrelevant features cannot explain our results.

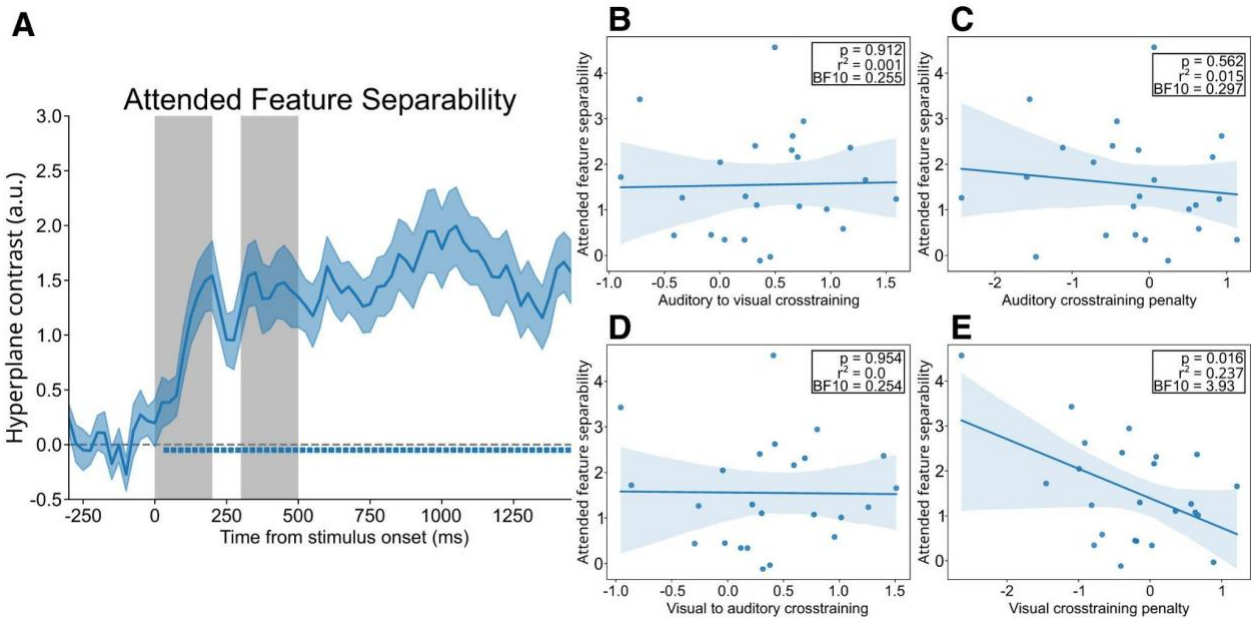

Figure 27. Leakage does not explain generalization (n = 24). (A) Hyperplane contrast between attend auditory and attend visual trials (set sizes combined). Significant time points (FDR corrected $p < 0.05$) are denoted by blue squares. (B and D) Correlation between attended feature discriminability and cross-decoding (B: auditory to visual, D: visual to auditory). (C and E) Correlation between attended feature discriminability and crosstraining penalty equivalent to within modality—cross modality hyperplane contrast, (C) auditory—auditory to visual, (E) visual—visual to auditory. All values are averaged across the delay period. Correlations were evaluated by linear regression (Wald test) across subjects (line = line of best fit, shaded region = 95% CI).

Using Representational similarity analysis for a concurrent analysis of modality-general pointers and feature-specific activity

One limitation of our decoding approach so far is that each analysis targeted one construct of interest (i.e., modality-general load vs. feature-specific activity) while ignoring the presence of

the others. Moreover, even good classifier generalization does not conclusively establish overlap in neural codes (Sandhaeger & Siegel, 2023). Thus, to obtain converging evidence regarding these distinct classes of neural activity, we used representational similarity analysis (RSA) to simultaneously model the effects of load and stimulus modality. The logic of RSA is to examine whether specific theoretical models predict the pairwise similarity across all conditions of the experiment. We tested four separate regressors that might predict this similarity structure (Figure 28C): (1) pointer load model—predicts similarity based on the number of *items* stored, regardless of the selected sensory modality. (2) Feature load model—predicts similarity based on the total number of *feature* values stored (regardless of modality). (3) Graded feature model—predicts similarity based on the degree of overlap in attended features, such that the conjunction condition falls in between the single-feature conditions. (4) Task-set model—predicts similarity based on which task set the observer has adopted (color, sound, or conjunction), with each task set being equally distinct from the others. Thus, regressors 1 and 2 contrasted item-based vs. feature-based models of WM load activity, while regressors 3 and 4 examined whether feature-selective signals tracked the specific constellation of features stored or the discrete task sets that corresponded to each experimental condition.

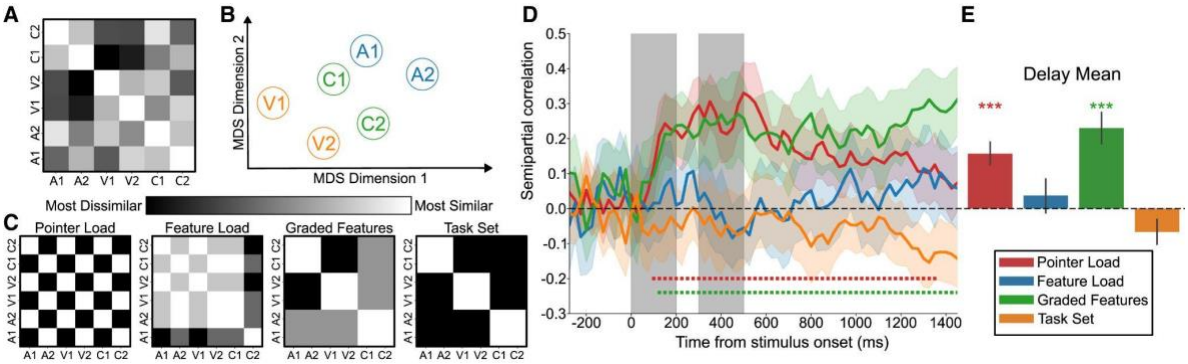

Figure 28. RSA results for experiment 1 (n = 24). Condition labels are denoted by a letter to indicate modality condition (A: Auditory, V: Visual, C: Conjunction), as well as a number to

*denote set size (1 or 2). (A) Empirical representational dissimilarity matrix (RDM), averaged across subjects over delay period (t > 500 ms). (B) MDS projection of RDM, color coded by modality, averaged over the delay period. (C) Graphical description of tested theoretical models. Darker = more dissimilar, lighter = more similar, arbitrary scale. (D) Semipartial correlations of each factor to the empirical RDM over time. Stimulus period denoted by the gray-shaded region, significant time points (Wilcoxon signed-rank test, FDR corrected) denoted by colored boxes under the graph.(E) Semipartial correlations averaged across the delay period (t > 500 ms). ***p < 0.001, **p < 0.01, *p < 0.05 (Wilcoxon signed-rank test, uncorrected). See also Figure A3 for a replication with a pupil size regressor.*

Item-based pointers and feature-specific activity explain unique variance in EEG activity

By calculating the semipartial correlations between each regressor and ongoing EEG activity, we measured the unique variance explained by each regressor at each time point (Figures 28D and 28E). Two regressors reliably explained variance in neural activity. First, the pointer load model explained robust variance throughout the entire delay period (Wilcoxon signed-rank test, $W = 273$, $p = 7.48 \times 10^{-5}$), while the feature load model explained no unique variance ($W = 183$, $p = 0.180$). This reinforces the prior conclusion that these neural load signals are item-based, not feature based. Second, the graded feature model explained robust variance throughout the delay period ($W = 278$, $p = 3.19 \times 10^{-5}$), while the task-set model failed to explain unique variance ($W = 92$, $p = 0.952$). Thus, feature-selective activity in this study was better explained by activity within modality-specific sensory regions than by activity that tracked the three discrete task sets required by our instructions (Kikumoto & Mayr, 2020). Critically, the pointer load and graded feature models explained *distinct variance* in EEG activity. Thus, RSA

reinforced our initial evidence for modality-general pointers while simultaneously controlling for feature-specific neural activity. Finally, multidimensional scaling (MDS) allowed us to visualize the similarity relationships between all experimental conditions within a 2D space (Figure 28B). Here, two axes of separation are apparent. First, a "sensory modality" axis divides the conditions from left to right. Visual and auditory conditions are on the left and right, respectively, while the conjunction condition falls in between them, in line with the graded feature model. Second, another axis separates load 1 conditions (top) from load 2 conditions (bottom), regardless of modality. Thus, this separability aligns with our hypothesis and the semipartial correlations described above.

Experiment 2

Although the findings so far reveal a clear dissociation between load-sensitive neural signals and feature-specific neural activity, the number of items stored in WM was confounded with the number of relevant positions in the display. Could changes in the breadth of spatial attention explain the common load signature between visual and auditory modalities? Indeed, recent work has shown that RSA is sensitive to EEG signatures of spatial attention,[18] enabling a direct test of this question. Thus, we independently manipulated WM load, attended modality, and the number of spatially attended positions (Figure 29). Subjects attended to either auditory or visual features and stored one or three targets that were sequentially presented in either a single location or across three unique locations.

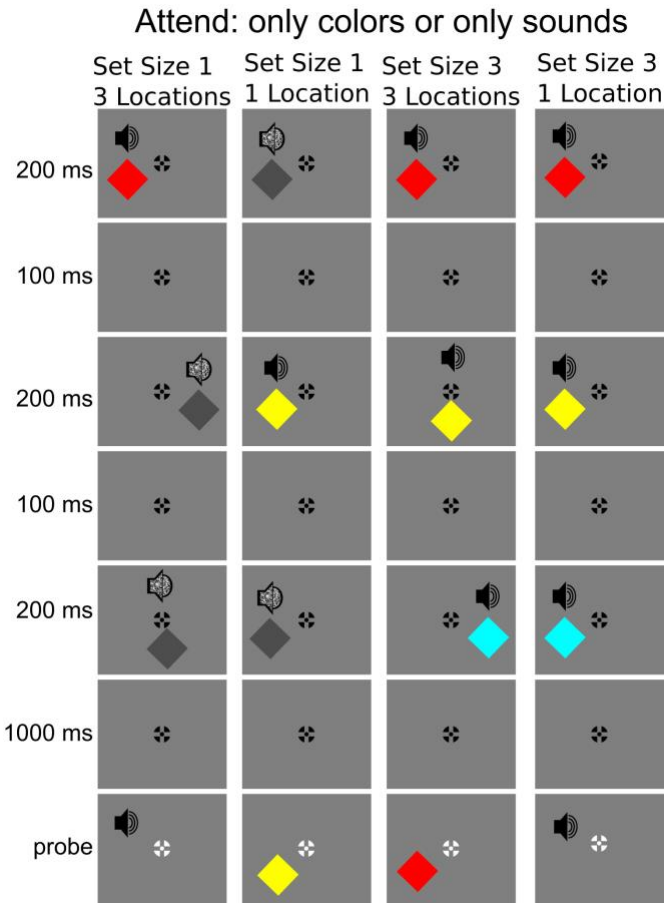

Figure 29. Experiment 2 task. Subjects attended to either only colors or only sounds. Each trial could have either one or three relevant stimuli (set size 1 trials had the same placeholders as experiment 1). Stimuli could appear in the same location for all presentations or in 3 unique locations.

We had behavioral data for 15 out of 16 subjects. Mean accuracy across subjects ranged from a minimum of 89.3% (auditory set size 3, different locations) to a maximum of 98.4% (visual set size 1, same locations). A repeated-measures ANOVA revealed a non-significant effect of attended modality ($F_{(1,14)} = 2.85$, $p = 0.11$), a significant effect of WM load ($F_{(1,14)} = 35.0$, $p = 3.8 \times 10^{-5}$), and a (although very close to threshold) significant effect of the number of attended

locations ($F_{(1,14)} = 4.68$, $p = 0.048$). There was a non-significant interaction between modality and load ($F_{(1,14)} = 0.83$, $p = 0.38$), but significant interactions between modality and number of locations ($F_{(1,14)} = 19.07$, $p = 6.44 \times 10^{-4}$), load and number of locations ($F_{(1,14)} = 6.03$, $p = 0.028$), and three-way interaction ($F_{(1,14)} = 14.78$, $p = 0.0018$).

We used RSA to examine the predictive power of three regressors: (1) pointer load model—predicts similarity based on the number of *items* stored, regardless of sensory modality. (2) Spatial attention model—predicts similarity based on the total number of *locations attended* regardless of sensory modality and pointer load. (3) Sensory modality model—predicts similarity based on the stored sensory modality.

Modality-independent pointers and spatial attention reflect distinct selection processes

All models exhibited statistically significant correlations (spatial attention: $W = 136$, $p = 1.5 \times 10^{-5}$; attended modality: $W = 128$, $p = 3.8 \times 10^{-4}$; pointer load: $W = 134$, $p = 4.6 \times 10^{-5}$) that were sustained throughout both stimulus presentation and delay-period phases of the trial (Figure 30B). The spatial attention model explained robust variance, starting immediately after the second stimulus presentation when the number of attended positions could be disambiguated. Critically, the pointer load model explained unique variance throughout both stimulus presentation and the delay, showing that this load signal cannot be explained by variations in the spatial extent of attention or by modality-specific activity. We also replicated these RSA results while swapping the sensory modality model for two models: (1) visual load model—predicted similarity based on number of colors stored and (2) auditory load model—predicted similarity based on the number of sounds stored. This change did not affect the outcomes for the pointer load and spatial attention models (Figure 30C, pointer load: $W = 122$, $p = 1.68 \times 10^{-3}$; spatial

attention: W = 136, $p = 1.5 \times 10^{-5}$). Both the visual (W = 119, $p = 3.14 \times 10^{-3}$) and auditory (W = 106, $p = 0.025$) load models explained reliable variance, although this was transient and early for the auditory load model. We note that when ranking values prior to computing correlations, as done in prior work (Kiat et a., 2022), the auditory load model no longer had a significant semipartial correlation when averaging across the delay period. However, no other result for this analysis (crucially pointer load) changed substantially. These findings bolster the conclusion from experiment 1 that modality-specific neural activity was tracking sensory activity related to color and sound storage rather than a shift between discrete task sets. Furthermore, this modality-specific activity is independent from the pointer-related activity.

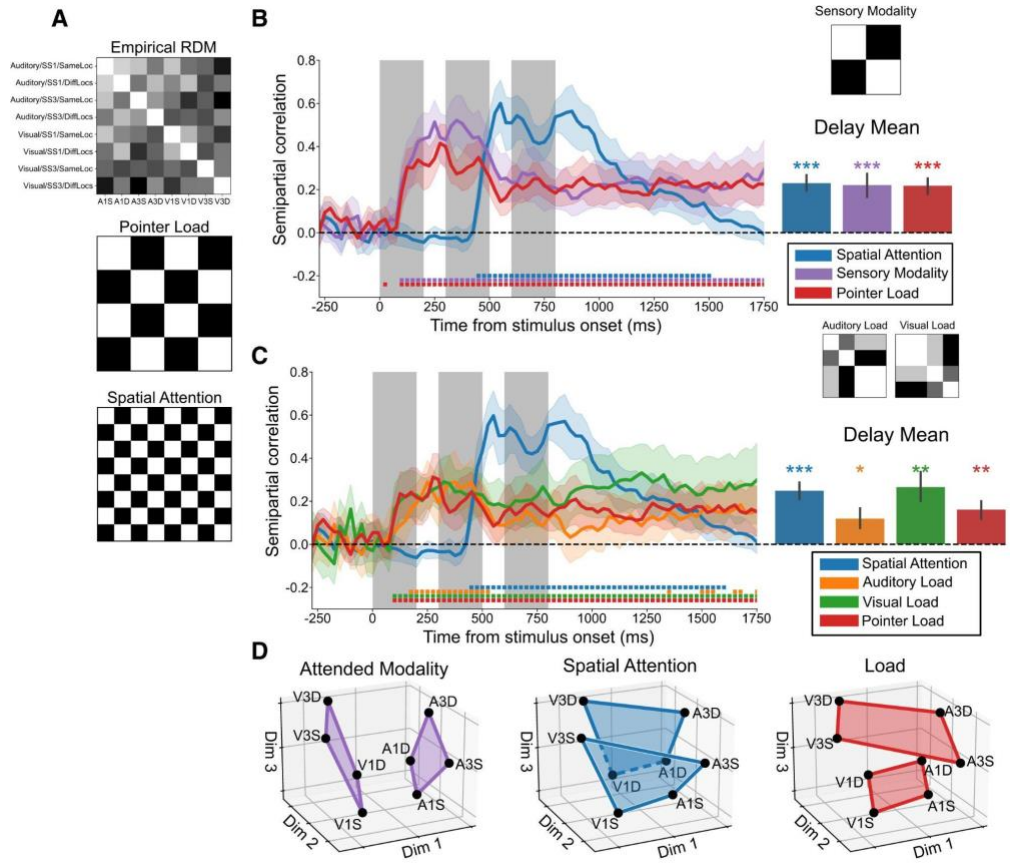

Figure 30. RSA results for experiment 2 (n = 16). Condition labels (as shown in the empirical RDM) are such that the first letter (A/V) denotes modality, the number (1/3) denotes set size, and

*the final portion (Same/Diff) denotes stimuli appearing in the same or different spatial locations. (A) RDMs used in all comparisons (representative empirical RDM, pointer load, and spatial attention). (B) RSA results with the sensory modality model (auditory or visual, shown in top right). Left: semipartial correlations of each factor over time. The stim period is denoted by the gray-shaded region, and significant time points (Wilcoxon signed-rank test, FDR corrected) are denoted by colored boxes under the graph. Right: average correlation over delay period (t > 800 ms). ***p < 0.001, **p < 0.01, *p < 0.05 (Wilcoxon signed-rank test, uncorrected). (C) RSA results with auditory and visual modality-specific load models (in top right). (D) 3D MDS projection of delay-period empirical RDM, points grouped by sensory modality, spatial attention demand, and WM load (points within the same colored plane are consistent within that feature). Tick marks are equal in distance. See also Figures A2 and A4 for replications of this effect, an alternate spatial attention regressor, and a pupil size regressor.*

Interestingly, our spatial attention model, which assumed that attention was allocated to both targets and distractors, explained substantial variance. However, one might predict that spatial attention would eventually be restricted to locations containing targets. To test this, we reran RSA with a spatial attention model based solely on the number of relevant target locations (Figure A2). In fact, this model began to explain some variance, particularly toward the end of the delay (Figure A2B). However, it did not explain unique variance at any time point from that explained by the spatial attention model, which allocated attention to both targets and distractors (Figure A2C), although the variance explained was significant in the delay average (W = 103, p = 0.037). Regardless of which model for spatial attention was used, the pointer load and attended modality models explained robust variance throughout the stimulus and delay periods. These

results suggest an interesting distinction: while signals tracking WM are shaped by relevance early after stimulus onset, signals tracking spatial attention exhibit spatial selectivity in a delayed fashion (Hakim et al., 2019).

We considered the possibility that our pointer load measure may reflect a more generalized signature of cognitive effort. While it has been difficult to find precise measures of effort (Westbrook & Braver, 2015), some have argued that it plays a substantial role in the neural substrates of WM (Master, Li & Curtis, 2024; Kitzbichler et al., 2011). However, many studies that have looked at effort and WM load simultaneously, instead of using one as a proxy for the other, have identified them as unique signals that are both present in nearly all cognitive tasks (Jones, Thyer et al., 2024; Kardan et al., 2020). We note that the tasks performed were not particularly difficult (mean accuracies of 95% and 94% in experiments 1 and 2, respectively), and there were only modest declines in accuracy in the higher load conditions (2% and 7% differences in experiments 1 and 2, respectively). Nevertheless, we carried out additional analyses to address this alternative explanation. Pupil size has often been used as an objective measure of effort, particularly when the sensory energy of the display is matched across putative variations in effort (van der Wel & van Steenbergen, 2018; Keene et al., 2022). Indeed, in prior work, we have directly tested whether voltage-based decoding could be explained by variations in pupil size (Jones, Thyer et al., 2024), and we found no evidence that variations in pupil size could explain voltage-based measures of WM load. These findings notwithstanding, we examined the impact of effort by replicating all RSA analyses with an additional RDM representing each individual's average baselined change in pupil diameter per experimental condition.

The key question was whether or not including pupil data undermined our evidence for pointers. The pupil size model never explained unique variance, and including it did not have a consistent impact across the two experiments. In experiment 1's analysis (Figure A3), inclusion of the pupil size regressor reduced the variance explained of the pointer load model in the delay period (51/58 to 18/58 time points significant), although the model remained significant in the aggregate (W = 139, $p = 0.0399$). In contrast, in experiment 2, including pupil size had no reliable effect on the variance explained by pointer and sensory modality factors (Figure A4A: 67/70 significant time points both with and without pupil model, delay aggregate W = 114, $p = 4.27 * 10^{-4}$), and it showed only a slight reduction in the analysis where modality-specific load factors were included (Figure A4B: 67/70 to 46/70 significant time points, delay aggregate W = 99, $p = 0.0128$). The inconsistent impact of the pupil regressor suggests that cognitive effort is not a robust source of the modality-general pointer signal.

Finally, an MDS projection of these experimental conditions into 3D space reveals a pleasingly interpretable cubic configuration (Figure 30D). Set size 1 and set size 3 define the top and bottom of the cube. Visual and auditory conditions define the left and right sides of the cube. And finally, the number of attended locations defines the front and back faces of the cube. Thus, the MDS plot implies separability between modality-independent pointers, feature-selective neural signals, and spatial attention. To summarize, experiment 2 replicated the initial demonstration of a dissociation between load-sensitive and feature-specific neural activity while also showing that the number of attended positions explained unique variance in EEG activity. These findings corroborate recent claims that the deployment of spatial attention and the gating of items into WM represent distinct aspects of attentional control (Jones, Diaz et al., 2024; Awh

& Vogel, 2025; Hakim et al., 2019). Moreover, these results show that spatial attention cannot explain the modality-independent neural activity that tracked the number of items stored in WM.

Discussion

WM is a cornerstone for broader intellectual function (Cowan et al., 2005; Unsworth et al., 2014; Fukuda et al., 2010), motivating the search for a clear taxonomy of its neural components. Our findings reveal a modality-general signature of WM storage that tracks the number of individuated items stored, independent of the sensory modality of those items. Multivariate decoding models trained on one sensory modality enabled precise decoding of WM load in the other sensory modality. Critically, RSA analyses corroborated the presence of this modality-general load signal while also showing that *distinct variance* in ongoing EEG activity was explained by the stored sensory modality and the number of attended positions in the display. Thus, parallel but separate neural signals track the number of individuated items stored and the featural content of those items (Xu & Chun, 2006; Xu & Chun, 2009). Moreover, we reinforce recent work that has argued for a dissociation between deployments of spatial attention and the selective encoding of items into WM (Jones, Diaz et al., 2024; Hakim et al., 36; Günseli et al., 2019).

What is the computational role of this modality-general load activity? Our working hypothesis is that it reflects the deployment of spatiotemporal "pointers" that bind the selected items to the surrounding event context (Thyet et al., 2022; Awh & Vogel, 2025), an operation that has been highlighted in major models of WM (Baddeley, 2000; Hedayati, Donnell & Wyble, 2022; Oberauer, 2019; Oberauer & Lin, 2017) and in prominent theories of dynamic visual cognition (Pylyshyn, 2009; Kahneman, Treisman & Gibbs, 1992; Kanwisher, 1987). This literature

elucidates how perception and action in dynamic environments require the observer to track attended items through space and time, despite changes in appearance or position (Pylyshyn, 2009; Kahneman, Treisman & Gibbs, 1992). Critically, the insight that perceptual inputs typically evolve across an unfolding event has motivated the idea that spatiotemporal tracking—sometimes referred to as *tokenization* (Kanwisher, 1987)—is distinct from the maintenance of the featural details of tracked items. This separation affords continuous spatiotemporal tracking even when the featural details of the item are unstable. Thus, this contextual binding process provides an attractive explanation for a modality-independent, item-based signature of the number of relevant items stored in WM. Moreover, our findings show that these modality-general pointers operate in parallel with modality-specific signals that index the specific constellation of features held in WM (i.e., visual, auditory, or both). Thus, pointers may support item-based contextual binding while parallel processes support the maintenance of the attended featural details. This said, we are not presently able to rule out other processes that could generate this signal. For example, subitizing and judging numerosity is one process that seems to be limited by the number of individuated objects (Halberda, Sires & Feigenson, 2006). Further work is therefore necessary to directly link spatiotemporal tracking to this process of content-independent storage.

Our results are in line with past arguments for supramodal aspects of attentional networks mediating WM storage.[55] For example, top-down modulation by the dorsal attention network (DAN) has been heavily implicated in WM storage (Gazzaley & Nobre, 2012; Majerus et al., 2012; Naghavi & Nyberg, 2005; Todd & Marois, 2004). Majerus et al. (2016) observed strong generalizability between fMRI patterns of WM for visual and verbal stimuli, particularly in posterior intraparietal sulcus, as well as across the DAN more broadly, during encoding and

maintenance periods. Similarly, Rizza et al. (2024) found strong correspondence between activation patterns for visual or auditory presentations of a spatial mapping task in several brain regions associated with the DAN, including the frontal eye fields and superior parietal lobule. While we see a clear connection with these ideas, our findings also show that WM encoding generates load signals that are dissociable from those tracking the deployment of spatial attention and cognition more broadly (Jones, Diaz et al., 2024; Diaz et al., 2021; Hakim et al., 2019; Bae & Luck, 2018).

Importantly, there are extant behavioral findings that, at first glance, appear to conflict with our proposal that WM storage depends on a modality-general pointer system. Various studies have found that WM performance is enhanced when the memory set includes mixed stimulus types (e.g., visual and verbal or visual and auditory stimuli) (Cocchini et al., 2002; Baddeley & Hitch, 1974; Fougnie & Marois, 2011; Henderson, 1972). In some cases, little or no interference is observed when information from separate modalities is concurrently stored, implying that each modality has its own storage capacity. Results like this have motivated models that explain behavioral response accuracy based entirely on the similarity between (noisy) competing memories (Schurgin, Wixted & Brady, 2020) or due to interference during retrieval instead of limits on storage capacity per se (Oberauer & Lin, 2017). Thus, while past work has provided both behavioral and neural evidence of item limits on WM storage (Adam, Vogel & Awh, 2017; Ngiam et al., 2023), it is nonetheless possible that these models are addressing distinct stages of processing with distinct limiting factors. Although past work suggests that item-based load signals are capacity-limited and predictive of individual WM ability (Thyer et al., 2022; Luria et al., 2016, Adam, Vogel & Awh, 2020), interference that scales up with inter-item similarity may have a strong effect on decision stages of processing.

Thus, interference during decision stages could explain the impact of inter-item similarity, even if a common pointer operation supports distinct sensory modalities.

In summary, we present electrophysiological evidence for a neural signature of WM storage that generalizes across sensory modalities, supporting a broad class of models that distinguish between abstract control processes that enable the gating and manipulation of specific thoughts and the stimulus-specific representations that determine the content of those thoughts (Awh & Vogel, 2025; Lundqvist et al., 2022; O'Reilly, Ranganath & Russin, 2022). We propose that these modality-general load signals reflect a content-independent operation for the contextual binding of items to the surrounding event, an essential component of cognition in virtually all perceptually guided tasks. Indeed, given the prominence of event codes in theories of long-term memory encoding and access, we hypothesize that the assignment of spatiotemporal pointers may serve as the initial step in the formation of durable episodic memories (Awh & Vogel, 2025; Yonelinas, 2024; Fukuda & Vogel, 2019).

Materials & Methods

Experimental model and study participant details

Participants were recruited from the greater University of Chicago and Hyde Park community in exchange for payment ($20 / hour). Twenty-eight participated in experiment 1 (4 excluded), and twenty in experiment 2 (4 excluded). Subjects were excluded from the final sample if fewer than 150 trials per stimulus condition were present after rejection of trials containing artifacts. Participants were between the age of 18-35, reported normal or corrected-to-normal color vision and hearing, and no history of stroke or neurological disorder. All participants gave informed

consent according to procedures approved by the University of Chicago Institutional Review Board.

Our target sample size for experiment 1 was 24 participants, a conservative estimate based on prior studies in the lab (Thyer et al., 2022; Jones, Thyer et al., 2024). Twenty-eight subjects (mean age 24.7, 13 male, 13 female, 1 declined to state, 1 data lost) participated. Four subjects were excluded from the final sample due to excessive eye movements and noise in EEG recordings (see artifact rejection). All pupil size analyses were run on 19 subjects with usable pupil data.

Our target sample size for experiment 2 was 16 participants. We assumed that increasing the difference between set sizes would improve overall SNR, and prior pilot studies still found reliable load effects at 16 subjects. A total of 20 participants completed the experiment (12 female, 6 male, 1 declined to state, 1 data not saved, mean age, 24.6, SD 4.3), with 16 remaining after artifact rejection. All pupil size analyses were run on 14 subjects with usable pupil data.

We recorded self-identified sex for all participants, with an option to decline to specify. We did not record ancestry, race, ethnicity, or socioeconomic status. However, we do not believe that these factors should substantially affect the results. All analyses were performed within subjects.

Experimental Procedures

Both experiments were single-probe sequential change detection tasks. We used bimodal stimuli, in which a colored diamond was presented with each auditory stimulus, in a spatially colocalized manner. We used a set of 5 recordings of environmental sounds used in prior studies (Uddin et al., 2018a; Uddin et al., 2018b), clips each of which lasted 200ms. The possible sounds were of a

camera shutter, crow, car horn, bell, or zipper, in addition to a distractor white noise sound. Stimuli were matched on intensity (70 dB SPL), sampling rate (44100 Hz), and duration. All stimuli had a bias towards the right (90°), left (-90°) or center (30°), and were synthesized by filtering recordings through a head-related transfer function (Martin, 2021; Gardner & Martin, 1995) based on volume balance. Visual stimuli were colored red #FF0000, yellow #FFFF00, green #00FF00, cyan #00FFFF, blue #0000FF, or gray (distractor) #555555 diamonds with a 1 degree visual angle radius, lateralized to the left, right (both by 6 degrees), or center of fixation. Luminances of all stimuli were, respectively, 25.2, 106.9, 83.84, 92.61, 9.50, or 17.90 cd/m^3. Stimuli were jittered by up to 2 degrees from their center point. Stimuli presented at the center location could not be less than 1.5 degrees from fixation. We predicted that spatial and temporal colocalization would act as grouping cues which would cause the stimuli would be viewed as bound audiovisual objects; this aligned with our results and with subjects' subjective experience (upon being asked after data collection).

In experiment 1, two pairs of visual and auditory stimuli were presented for 200ms, with a 100ms ISI. In set size 2 trials, 2 target stimulus pair were presented. In set size 1 trials, placeholder pairs (gray squares and white noise) were presented at one timepoint in order to match stimulus energy across set sizes. The order of target pairs and placeholder pairs was random across trials. We manipulated the information subjects stored in WM by directing subjects to only remember auditory items, visual items, or all items (conjunction) for a given block. Therefore, experiment 1 had a 3 (attended-modality conditions) x 2 (set sizes) design. Subjects completed 24 blocks of 56 trials, cycling between sensory modality conditions, for a total of 1,344 trials and 224 trials per cell of the design. After a 1s delay following the second

stimulus pair presentation, a single probe was presented according to the attended modality. Subjects answered whether this was one of the previously presented objects.

The goal of experiment 2 was to determine whether the modality general load signal seen in experiment 1 was driven by spatial attention, or if it persisted even when spatial attention was directly manipulated. We replicated experiment 1's design with the following modifications. First, we varied the number of attended spatial locations. In "same location" conditions, all stimuli appeared in the same location after each other, randomly selected from the options of left, right, or center, ensuring that all trials had equal demands to spatial attention regardless of working memory load. "Different location" conditions were identical to those in experiment 1, with stimulus pairs appearing at distinct locations within a presentation sequence. Second, we removed the conjunction condition to reduce the experiment run time. Finally, we changed the set sizes to 1 and 3 to increase the magnitude of the possible spatial attention confound, as well as the magnitude of the putative WM load signal. This design allows us to test the generalization of load signals while controlling the spatial information between conditions. Therefore, experiment 3 had a 2 (attended-modalities) x 2 (set sizes) x 2 (spatial conditions) design. In total, subjects completed 26 blocks of 56 trials, for a total of 1456 trials and 182 trials per cell of the design. All other details, including stimuli and placeholders, were identical to experiment 1.

Apparati & Data Acquisition

Subjects were tested in a dimly lit, electrically shielded chamber. Stimuli were presented on a neutral gray background (RGB: 127,127,127). Subjects performed the experiment on a gamma-corrected 24-in. LCD monitor, with a 1920x1080 resolution and 120Hz refresh rate, at a distance of 75 cm. Subjects gave their response by a keyboard press (z = same, / = different), and rested

their heads on a foam chin rest for the duration of trial blocks. Auditory stimuli were presented via in-ear earphones (Neurospec ER3C), with disposable foam tips. Experiments were designed using PsychoPy3 (Peirce et al., 2019). During all trials, subjects were expected to fixate on a cross in the center; we used the shape identified by Thaler et al. (2013) as optimal for fixation. We applied a real-time eye-tracking rejection procedure during the experiment. In this, any eye movements (determined by eye-tracking) of more than 1.5 degrees visual angle from the center of fixation would cause the trial to be immediately aborted and a trial of the same condition repeated at the end of the experiment.

We recorded EEG activity from 30 active Ag/AgCl electrodes (Brain Products actiCHamp) at International 10-20 system sites Fp1/2, Fz, F3/4, F7/8, FC1/2, FC5/6, Cz, C3/4, CP1/2, CP5/6, Pz, P3/4, P7/8, PO3/4, PO7/8, Oz, and O1/2. Prior to starting recording, impedances for all electrodes were below 10 kΩ. A ground electrode was placed at position FPz, and two references were placed on the left and right mastoids. All electrodes were referenced online to the right mastoid, and rereferenced offline to the algebraic average of the left and right mastoids. Additionally, to track eye movements, we recorded EOG data using passive electrodes. We placed a ground electrode on the left cheek, two electrodes to track horizontal eye movements ~ 1 cm from the horizontal canthus of each eye, and two to track vertical eye movements above and below the right eye. Data were recorded at 1000 Hz and filtered online (high cutoff: 80Hz, low cutoff: 0.01 Hz, slope: 12 dB / octave) using BrainVision Recorder on a Windows PC. Eye-tracking data was collected using a desk-mounted infrared EyeLink 1000 Plus system (SR Research) at 1000 Hz, that was calibrated between blocks and after any breaks. We used eye-tracking data preferentially for artifact rejection, and EOG when eye-tracking was not available or malfunctioning.In experiment 1, data were segmented offline to 1800 ms epochs

time-locked to the onset of the first stimulus (-300 ms to +1500 ms), and baseline-corrected with the 300 ms immediately prior to stimulus onset. In experiment 2, 2100 ms epochs were used (-300 ms to +1800 ms).

Quantification and statistical analysis

Overview

For each analysis, sample size (number of participants) and the statistical tests used are present in the figure legends. Sample size was uniform for all analyses in the same figure. Where asterisk notation was used to denote significance, we used the following thresholds: *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, n.s = $p > 0.05$. P values and Bayes Factors for statistical tests, where applicable, are presented in the results text. When evaluating performance over time, significant time windows are marked with a colored box, corresponding to a FDR corrected p-value (see Statistical Tests subsection) below 0.05. Bar heights and plot lines represent means, and error bars (included shaded regions) denote SEM.

Artifact Rejection and Subject Exclusions

We applied an automated artifact rejection procedure to identify trials contaminated by ocular or muscular artifacts, and then visually inspected all trials to manually reject significantly noisy trials or remove false positives from rejection (these primarily occurred due to drift in the EOG or brief lapses in eye-tracking recording. Trials were rejected if they met any of the following criteria:

-Eyetracking: Eye movements of over 1.5 degrees visual angle away from fixation.

-EOG: Any trial where the absolute EOG value was greater than 75 μV from pre-trial baseline. This was only used when eye-tracking was not available for a subject.

-Blinks: Trials where the eye-tracker lost focus on the eye for any portion. Several of these were false positives, and were re-included in the final sample if vertical EOG and frontal EEG electrodes did not show a deflection.

-Saturated Electrodes: Trials where any channel has flatlined or exhibits a step function

-Additional EEG artifacts: skin potentials, muscle artifacts, and excessive noise. In experiment 1, we excluded 1 electrode from analysis for 1 subject due to high noise. However, we still had sufficient data to successfully run all analyses.

In experiment 1, we rejected between 5.6% (conjunction ss1) to 6.5% (auditory ss1) of trials. In experiment 2, we rejected between 3.5% (auditory ss3 different locations) to 4.3% (auditory ss1 1 location) of trials. Subjects were excluded from the final sample if they had fewer than 150 trials per load condition remaining after artifact rejection.

Decoding analyses

We applied a multivariate binned-trial classification procedure (mvLoad) to classify working memory load within subjects (Adam et al., 2020). When attempting to use cross-decoding to evaluate the generalizability of neural patterns across feature dimensions, we used a parallel but higher-powered alternative metric to decoding accuracy which we called "hyperplane contrast." The goal of this approach is to draw a hyperplane in n-dimensional space at the midpoint of the logit function as determined by the training set. We then measure the average signed distance of each condition to this hyperplane (scikit-learn: decision_function) and take the difference between the two distances as the contrast. This is done both for held-out

samples from the training set, as well as a test group that the classifier is entirely naïve to when appropriate. This is analogous to estimating the magnitude of the multivariate difference vector separating each pair of conditions, and is insensitive to the types of additive shifts that changes in other signals may cause (Jones, Thyer et al., 2024). These distances are additionally easier to interpret than classification accuracy, as they are not affected by ceiling effects and apply in a continuous space as opposed to the average of several binary classifications.

We used the following pipeline for decoding analyses. On a given iteration, we first formed "bins" by randomly sampling trials from the same condition into groups of 20, and averaging across the trials in each group. This was done without replacement, and any trials that were not assigned to a bin (due to the trial counts not cleanly dividing by 20) were dropped for that iteration. We then split these binned trials into a training (70%) and testing (30%) set using the train_test_split function from scikit-learn (Pedregosa et al., 2011). Training sets were always balanced via random dropping to ensure that an equal number of trial bins from each training condition were used. We averaged these time series using a sliding window (50ms, 25ms step). Training data were standardized for each time window using the StandardScaler function, and test data were standardized to the mean and standard deviation of the training data. These data were then fed into a regularized logistic regression classifier (scikit-learn LogisticRegression), and used to predict appropriate labels and a confidence value for the test set. This procedure was repeated over 1,000 iterations, including the initial random binning, for each subject and each time window.

We chose a bin size of 20 a priori based on prior work (Thyer et al., 2022; Jones, Thyer et al., 2024). However, to ensure our results were invariant to the specific bin size used we performed a downsampling analysis while varying the varying the bin size, using trial bins of

size 1, 5, 10, 20, and 30 (Figure A5). SNR generally monotonically increased with greater bin size, particularly when crosstraining. We additionally observed sustained crosstraining performance throughout the delay period when using a bin size of 10 or greater; the 1 trial and 5 trial bins were more sporadic.

Statistical Tests

Hyperplane contrasts were evaluated using a one-tailed paired t-test, as we would not expect these to be meaningly lower than zero. We compared the mean contrast of each subject at each timepoint. Because we conducted independent significance tests at multiple timepoints, all p-values were corrected according to the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995), with a false discovery rate set to 0.05. When we were comparing the values of two contrasts against each other, we used a combination of one-tailed and two-tailed tests based on our alternative hypotheses.

For all parametric tests, we calculated the Bayes factor (BF) using the pingouin package (Vallat, 2018), which compares the relative degree of evidence for the null and alternative hypotheses. Calculation of Bayes factors addresses one of the primary weaknesses of null-hypothesis significance testing, which is its difficulty in interpreting nonsignificant results. Instead of providing a likelihood statistic for the null hypothesis, the Bayes factor represents the likelihood ratio of the null and alternative hypotheses. Generally, a BF of 1 represents equal evidence for both (absence of evidence), while BFs of 1/3 and 1/10 represent weak and strong, respectively, evidence for the null hypothesis, and BFs of 3 and 10 represent weak and strong evidence for the alternative. All other statistical results were calculated using Scipy (Virtanen et al., 2020).

To ensure robustness of our correlations in Figure 28, we removed subjects with outlier values from regressions. For each regression, we calculated Cook's distance for each datapoint to identify high-leverage outliers. We defined influential subjects as those with a Cook's distance greater than 0.5. Two high-leverage subjects, with Cook's distances of 0.54 and 0.95 were excluded. We replicated all other analyses for experiment 1 with these two subjects excluded, and did not observe any qualitative differences.

Representational Similarity Analysis

We additionally used representational similarity analysis (Kriegeskorte et al., 2008), a form of multivariate pattern analysis which analyzes the similarity of activation patterns in the brain across conditions in order to more effectively model the effects of load and stimulus modality simultaneously. Unlike evaluating classifier discriminability, it allows us to examine the effects of one factor (say, set size) when the effects of others are controlled for.

We performed RSA analyses by computing a distance metric between each possible pair of conditions, and then combining them into a single matrix of dissimilarities (representational dissimilarity matrix: RDM). The distance metric used here was a crossvalidated estimate of the Mahalanobis distance, also known as the linear discriminant contrast (LDC), which produces a distance metric that is largely unbiased by noise (Diedrichsen, Provost & Zareamoghaddam, 2016; Nili et al., 2014; Walther et al., 2016). We note that the LDC is conceptually analogous to the hyperplane contrast presented above, just based specifically on a Linear Discriminant Analysis (LDA), rather than a Logistic Regression model. This distance metric is particularly ideal for inference tests, as unlike non-cross-validated distance measures, which have a lower bound of 0, the LDC can take on negative values, such that if two samples vary solely due to

chance, the mean of the distribution of their LDC values across iterations will be zero. This procedure used a similar pipeline as the decoding analyses, although we did not bin trials prior to classification. Training and testing trials were obtained using the StratifiedShuffleSplit function from scikit-learn using a 50/50 split. Distances were calculated over subjects and timepoints using a 50 ms sliding window (25 ms step), over 10,000 iterations, and subsequently averaged over each iteration. We computed LDC between two condition pairs as

$$d^2_{LDC}(i,j) = (m_i - m_j)_{Train} * \Sigma^{-1}_{train} * (m_i - m_j)_{Test}$$

Where m_i and m_j are the average EEG voltage values for a given condition at the given timepoint (i.e. a vector of the same length as the number of electrodes), computed for both the training and testing half, respectively, and Σ_{Train} is the covariance matrix of the training set. We calculated this by first demeaning trials by subtracting the condition average voltage, computing the covariance matrix across conditions, and then regularizing by the Ledoit-Wolf procedure (Walther et al., 2016; Ledoit & Wolf, 2004).

In order to evaluate models, we designed RDMs which were intended to describe the predictions of different theoretical factors which may be present in the data. Prior work (Jones, Thyer et al., 2024; Jones, Diaz et al., 2024; Kiat et al., 2022) has rank transformed RDMs prior to calculating correlations. We did not do so to improve interpretability, but note that the results of our models are nearly qualitatively identical with or without the rank transformation; the only noticeable effect was that the auditory load model in Figure 30C no longer became significant after ranking. We then applied a linear regression model with the empirical RDM as the independent variable, and each theoretical factor model as a predictor. Then, we calculated the semipartial correlation (a measure of the unique variance explained by each regressor) of each theoretical model to the empirical RDM.

In Experiment 1, we tested the following theoretical factors. First, we included a pointer load model, assumed a distance of 0 between all conditions of the same set size regardless of attended modality (e.g. visual-1 and auditory-1) and a distance of 1 between all conditions of the other set size (e.g. visual-1 and visual-2). Second, we included an alternative, feature based model, which coded each condition by the number of feature values being maintained. In this model, for example, conjunction set size 2 would have 4 features (2 colors and 2 sounds), while visual set size 1 would have 1 feature, meaning the difference between them is 3. We also included two models designed to code for how the different modality conditions may be represented. In the Graded model, the conjunction condition fell between the visual and auditory conditions (e.g. auditory=-1, conjunction=0, visual=1), reflecting a single attended-feature axis. In the Discrete model, each condition (auditory, conjunction, visual) was equidistant from the other 2 (a distance of 1 to each other), reflecting an equilateral triangle.

Experiment 2 included the following theoretical factors. First, the pointer load model was identical to that of experiment 1, given there were only 2 set sizes to compare against. Next, a spatial attention factor was coded based on the number of presented locations (3 vs 1), with a distance of 0 between the conditions with the same number of locations and a distance of 2 between conditions with a different number of locations. We also examined a version of this spatial attention factor that was coded based on the number of relevant target locations (i.e., all set size 1 conditions, as well as all set size 3 with repeated locations, only had 1 target location), but that model did not explain as much unique variance as the pure spatial attention factor, nor did it change the results relating to the other factors. We also included 2 different modeling approaches to capture modality-specific signals. In one model, we coded only for the sensory modality, such that all conditions of a given modality had a distance of 0 between them, and all

condition pairs that spanned the 2 modalities had a distance of 1. This model was analogous to the attended feature models used in experiment 1. In contrast, we also tested a model which examined modality-specific load signals. This consisted of 2 RDMs, on per modality, using the following logic. Set size 1 trials for a given modality (e.g. visual) would have a modality-specific load of 1, set size 2 trials of that modality would have a modality-specific load of 2, and any set size of the other modality (auditory set size 1 or 2 in this example) would have a modality-specific load of 0.

Prior to running regression analyses, we calculated the variance inflation factor (VIF) for each empirical model to ensure that estimates of each theoretical model's individual contribution were reliable and not potentially deflated; in all cases, each theoretical RDM's VIF was <3, indicating low multicollinearity between variables. For each factor, the semipartial correlation was calculated as the square root of the difference between the full model's R^2 and the R^2 value of a sub-model with this factor removed, multiplied by the sign of the factor's β in the full model. This was done for each subject. We used Wilcoxon tests to evaluate whether semipartial correlations were significant at the group level, while applying the Benjamini-Hochberg procedure when testing over multiple timepoints. The presence of a statistically significant semipartial correlation can provide evidence for the unique contribution of a given theoretical factor to the neural signal.

To visualize the representational structure of the conditions, we applied metric multidimensional scaling (MDS) to the distance matrices calculated for RSA. We used both 2-D and 3-D projections based on experimental conditions (2-D for experiment 2's 2-level design and 3-D for experiment 3's 3-level design). We first averaged the RDM over all subjects and delay

period timepoints, and then used the metric MDS class implemented by scikit-learn to calculate the appropriate projection.

Chapter 3: Neural signatures of working memory storage track capacity limits once isolated from spatial attention

Abstract

Recent work has identified a multivariate EEG signal that scales with the number of items in working memory (WM) regardless of the specific content and which is separable from spatial attention. While this suggests that a core mechanism of WM may be consistent across featural information, it has not yet been examined whether this putative WM load signal tracks behavioral capacity limits. Here, we recorded EEG from participants in two studies (N_1=26; N_2=30) while they completed a traditional change detection task across a range of set sizes that spanned capacity limits. Rather than equate spatial attention directly, participants also performed a task using the same displays, in which they had to spatially attend the locations of the initial items in preparation for the presentation of brief, rare targets. Multivariate decoding found that the tasks contained a spatial attention signal that scaled with set size, and that the change detection task contained an additional, WM-storage specific signal. To isolate how this WM-specific signal changed across set sizes, we developed a novel analytical approach combining hierarchical Bayesian modeling with representational similarity analysis to estimate how the underlying functions change across set sizes simultaneously. We find that the WM load function peaks or plateaus around set size 2 or 3, consistent with behavioral capacity estimates. Thus, the findings from this work are twofold. First, spatial attention can be allocated in the absence of WM, even when information (location) must be actively maintained to succeed at a task, arguing for a refined definition of WM storage. Second, the WM load function tracks behavioral capacity limits, supporting the argument that it reflects a core, content-independent component of WM.

Introduction

Whether it's remembering your dual-factor authentication code while you log into your institutional account or remembering the name of someone you were just introduced to, we must actively hold onto information to help us navigate everyday life. Our ability to actively maintain and manipulate information - working memory (WM) - is a core cognitive process. This is underscored by the fact that individual differences in WM capacity are highly related to broad measures of intellectual function, such as attentional control and fluid intelligence (Cowan et al., 2005; Vogel et al., 2005; Fukuda et al., 2010). Despite WM's ubiquity and necessity in everyday cognition, it is a severely capacity limited system, with most people only able to store about 3-4 items at once (Cowan, 2001). A fundamental, open question in the field of WM is from where this capacity limit arises.

One way of answering this question is to identify how information in WM is represented in the brain. While studies of WM representations originally focused on identifying the exact content being stored in WM (e.g., Fuster & Alexander, 1971; Fuster & Jervey, 1981; Funahashi et al., 1989; Goldman-Rakic, 1995; Harrison and Tong, 2009; Serences et al., 2009; D'Esposito and Postle, 2015; Rademaker et al., 2019), an alternative approach is to examine signals which track the amount of information in WM, rather than the specific content (Todd & Marois, 2004; Xu & Chun, 2006; Vogel & Machizawa, 2004; Luria et al., 2016). This latter approach provides a lens into the common WM processes that are engaged regardless of the specific content being stored.

Recent work examining EEG data has identified a multivariate signal that appears to track the number of items in WM, which we will refer to as WM load (Adam et al., 2020; Thyer et al., 2022). This signal scales independently of the exact content being maintained, including

colors, orientations and motion coherence levels, but also sounds and words (Jones, Diaz, et al., 2024, Suplica et al., 2025, Chang et al., 2026). Critically, this signal can be dissociated from signals tracking spatial attention to relevant locations. For example, the signal persists even after controlling for spatial attention either through overlapping stimuli or sequential presentations at repeated location (Jones, Thyer et al., 2024, Suplica et al., 2025; Chang et al., 2026). These studies also found that manipulations of the relevant areas produce a signal that is orthogonal to the WM load signal. Moreover, previous EEG work found that spatial attention, as measured by lateralized alpha power, can be deployed to an item's location without producing an apparent WM load signal for that item (Hakim et al., 2019).

We have proposed that the multivariate WM load signal reflects a core component of WM storage. For example, it may reflect the capacity-limited assignment of indices that enable the selective access and manipulation of items, as well as the binding between those items to representations of the current context or event (Thyer et al., 2022; Awh & Vogel., 2025). While such a mechanism was originally proposed to explain human's ability to track items over time (Pylyshyn et al., 1979, Kahneman et al., 1992), it has recently been incorporated into models of WM (e.g., Oberauer & Lin, 2017; Hedayati et al., 2022). As behavioral work has found that people can only hold around 3-4 items in WM at once, it may be that we can only allocate 3-4 indices simultaneously.

If the multivariate WM load signal does reflect a core, capacity-limited component of WM, then it should reflect these capacity limits. However, no work has yet looked at how this multivariate load signal behaves when participants are asked to remember supracapacity arrays. If the signal reflects this core component of WM, then signal should peak around behavioral capacity limits, after which it may plateau or even decrease (Fukuda, Woodman & Vogel, 2015).

In contrast, if the signal reflects a more task-general state like effort or arousal, we should expect it to continue increasing with set size (Kardan et al., 2020).

The current work aims to test the prediction that the multivariate WM load signal should peak at behavioral capacity limits by examining how the signal varies from subcapacity to supracapacity arrays, while controlling for spatial attention. Rather than equate spatial attention, which would be difficult at higher set sizes, we sought to simultaneously estimate how spatial attention changes across set sizes using a previously established paradigm (Hakim et al., 2019), combined with modern multivariate decoding approaches. Here, participants are given the same displays as in a standard change detection task, but are merely required to attend to the locations of the initially presents items in preparation for processing rare, briefly presented targets. Importantly, this paradigm still requires the active maintenance of information (location) to succeed.

Together, this work tackles two questions. First, can spatial attention be allocated without engaging WM? Second, how does the WM load signal behave at supracapacity set sizes? To answer these questions, we collected two datasets (N_1=26, N_2=30) of participants completing both a change detection task and this target detection task across a range of set sizes. Experiment 1 included set sizes 2, 4, 6, and 8, and an additional manipulation of the presence of placeholder stimuli to estimate whether changes in filtering demands across set sizes impacted our estimates of the WM load signal. As Experiment 1 indicated that filtering demands did not meaningfully impact WM or spatial attention signals in this paradigm, Experiment 2 does not include this manipulation of placeholder presence, and instead included a larger range of set sizes (0, 1, 2, 3, 4, 6, and 8) to more precisely estimate whether and at what point the WM load signal function plateaus.

To anticipate the results, we found the following. First, spatial attention can be engaged without WM storage, but tasks requiring WM storage engage both WM and spatial attention. As a result, classifier-based predictions of WM load will reflect a mixture of both WM and spatial attention signals. To isolate the changes in WM load across set sizes from those of spatial attention, we developed a novel analytic approach that combined representational similarity analysis with hierarchical Bayesian modeling to estimate the underlying signals. This approach revealed that the WM load signal peaked around set size 2 or 3, consistent with previous behavioral and neural estimates (Balaban, Fukuda & Luria, 2019; Vogel et al., 2004; Fukuda, Woodman & Vogel, 2015). This contrasts with reports of effort, which were consistent across the two tasks, and increased significantly across every change in set size. Overall, these results are consistent with the theory that the WM load signal tracks a core component of WM.

Materials & Methods

Participants

Experiments 1 and 2 included 34 and 33 volunteers, respectively, participating for monetary compensation ($20 per hour). Participants were between the ages of 18 and 35 years old, reported normal or corrected-to-normal visual acuity, and provided informed consent according to procedures approved by the University of Chicago Institutional Review Board. Participants were recruited from the University of Chicago and the surrounding community via online advertisements and fliers posted on the university campus. Participation in Experiment 1 was not an exclusion criterion for Experiment 2.

Experiment 1. Our target sample in Experiment 1 was 25 participants. While previous studies have reliably decoded WM load with few participants (e.g., Jones, Diaz et al., 2024), the SNR between set sizes is reduced at higher set sizes (Thyer et al., 2022), and none have attempted to decode WM load between very high set sizes (6 and 8). Thus, we chose to collect a larger sample to improve our power. 34 volunteers participated in Experiment 1 (18 female, 15 male, 1 non-binary; mean age = 25.91 years, SD = 4.37). 8 participants were excluded from the final dataset due to insufficient trials after artifact rejection (n=7), or behavioral performance (n=1). The final sample was 26 participants (15 female, 10 male, 1 non-binary; mean age = 26.57 years, SD = 4.41).

Experiment 2. Our target sample in Experiment 2 was 30 participants. 33 volunteers participated in Experiment 210 (22 female, 9 male, 2 non-binary; mean age = 24.18 years, SD = 4.11). 3 participants were excluded from the final dataset due to insufficient trials after artifact rejection, leaving a final sample of 30 participants (22 female, 7 male, 1 non-binary; mean age = 24.4 years, SD = 4.15). One participant applied the instructions for set size 0 trials in the digits task to the set size 0 condition of the change detection task for the first four blocks of the session (see below), leading to high behavioral error rates in this condition; this participant returned for a second session, and the two datasets were combined, dropping the set size 0 trials of the change detection task from those blocks of session 1.

Apparatus

We tested the subjects in a dimly lit, electrically shielded chamber. Stimuli were generated using PsychoPy (Peirce et al., 2019). Subjects viewed the stimuli on a gamma-corrected 24 in LCD

monitor (refresh rate = 120 Hz; resolution = 1,080 × 1,920 pixels) with their chins on a padded chin rest at a viewing distance of approximately 82 cm.

Task Procedures

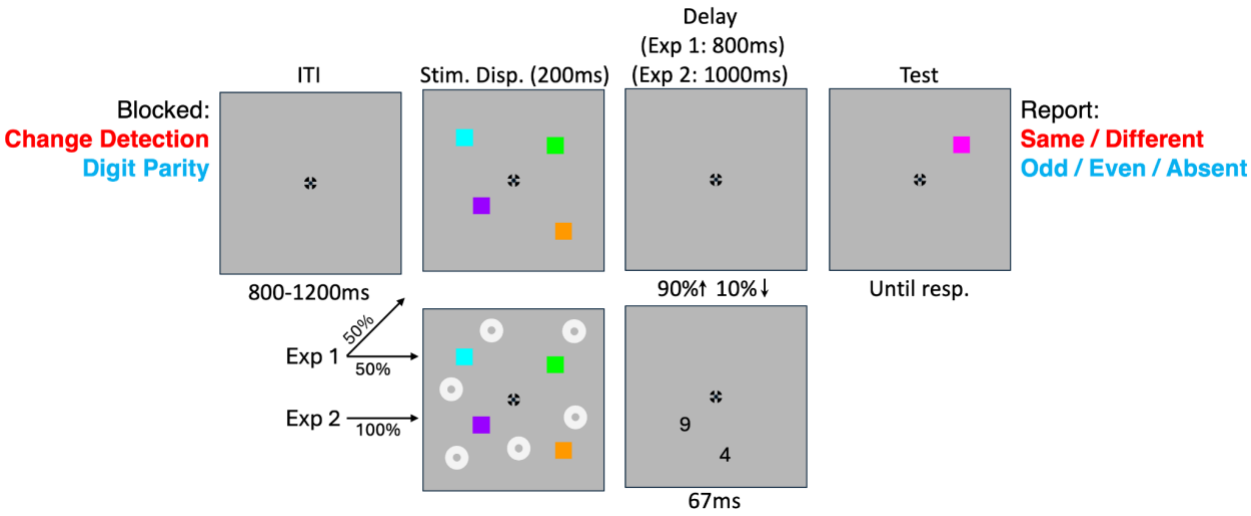

Figure 31. Experimental Task Design. Trials consisted of initial arrays of colored squares, potentially with placeholders, followed by a delay period, and finally a single colored square probe. In the change detection task, participants judge whether the probe square was the same color as before or different. In the Digit parity task, participants would attend to locations of the colored squares in preparation for the presentation of rare digits. If digits did appear, participants judged whether the digit at the attended location was odd or even.

Luminance-balanced displays. For each experiment, the luminance of the target color set was measured using an LS-150 Luminance Meter. Next, the luminance meter was used to identify an RGB value that produced a gray matching the average luminance of the target set in each

experiment. Placeholders were colored with this gray and set to cover the same area as the targets, controlling the total area and luminance across set sizes.

Stimulus Displays. All tasks in both experiments used the same sequence of displays on each trial (Figure 31). Stimuli were presented on a dim, neutral gray background (RGB values = [95.625, 95.625, 95.625]). A central fixation point was displayed at all times. Following the recommendations of Thaler et al. (2013), it consisted of a black circle with a radius of 0.25 degrees of visual angle (DVA) across the diameter with a fixation cross matching the background gray overlaid on top with a width of 0.16 DVA, and a central black dot with a radius of 0.075 DVA (i.e., a diameter of 0.15 DVA).

After an inter-trial interval (ITI), an initial display was presented. It consisted of a number of colored squares defined by the set size of the trial (2,4,6, or 8 in Experiment 1; 1,2,3,4,6, or 8 in Experiment 2). Squares had side lengths of 1 DVA and were placed within a grid consisting of 3 rows and 5 columns. Each row spanned 4 DVA, and each column spanned 3 DVA. The central cell containing the fixation point was not available for placement, leaving 14 possible cells. Within each cell, squares were jittered with the requirement that they could not be within 0.5 DVA of a cell edge.

Colors were drawn from the following set of 9 options: Red (Hex = #AF0000), Blue (Hex = #0000FF), Green (Hex = #008C00), Cyan (Hex = #00FFFF), Orange (Hex = #FF8000), Pink (Hex = #FF00FF), Yellow (Hex = #FFFF00), Black (Hex = #00000), and White (Hex = #FFFFFF). In half of the trials in Experiment 1, and all trials in Experiment 2, placeholder stimuli were included in the initial array. Placeholders consisted of luminance-matched gray (Hex = #929292) donuts with an inner radius of 0.1 DVA and an outer radius computed to match

the same area of the square (1 DVA^2): 0.573 DVA. The number of placeholders was set on a given trial to ensure that 10 items were presented, and placeholders were placed in the remaining available grid cells. Placeholders were jittered with the same constraints as squares.

Following the initial display, a delay period consisting of the fixation cross was presented. On some trials (see below), 2 digits were briefly presented during the delay period. The possible set of digits was 1, 2, 3, 4, 5, 6, 7, and 8. One digit would appear at the location previously occupied by a colored square, and the other would appear at another location not occupied by a colored square, though it may have been occupied by a placeholder. One digit was always odd while the other was always even.

Following the delay period, a final colored square was presented. Its location was chosen among the set of square locations presented at the start of the trial. The square would stay on the screen until participants made a button press, which initiated the next trial. See *Experiment 1* and *Experiment 2* sections below for a description of the task paradigms and exact display timings.

Experiment 1. Experiment 1 consisted of 2 tasks: a whole-field change detection task and a rare-target task (Hakim et al., 2019). Both tasks used the same sequence of displays on each trial, described in the previous section. ITIs lasted between 800 and 1200 ms. The initial display of colored squares and placeholder donuts lasted 200 ms and was followed by an 800 ms delay period. Digits were presented at some point during the delay period on 10% of all trials. Their presentation time was selected from a uniform distribution across the delay period, excluding the first and last 100 ms, and they were displayed for 67 ms. Following the delay period, a single probe square was displayed until participants made a button press. Participants were encouraged to blink once the probe square appeared, right before or as they made the button press, on the

majority of trials, to reduce strain and the probability of eye movements or blinks during the initial stimulus presentation and delay period.

In the change detection task, participants were presented with either 2, 4, 6, or 8 colored squares to remember and, after a delay period, had to judge whether a single probe square had changed in color from its original presentation at the same location. Changes occurred on 50% of the trials, and participants responded with a 'same color' judgement using the 'z' key, and a 'different color' judgement using the '/' key. Each color in the initial display was unique, and on change trials, the probe color was drawn from the set of colors not used on that trial.

To assess whether the presence of placeholder stimuli induced an additional filtering demand signal that scales with WM load, placeholders were only included on half of all trials. When they were included, the number of placeholders displayed was chosen to set the total number of items on the screen to 10. For example, for the maximum set size of 8 colored squares, 2 additional placeholders were presented.

The digits task was designed to identify how spatial attention signals changed with set size while reducing WM demands, following the logic of Hakim et al. (2019). Participants were told to simply attend to the locations of the colored squares in preparation for the possible presentation of 2 digits during the delay period. If digits appeared, which occurred on 10% of all trials, only one would appear at an attended location, and participants would note whether that digit was odd using the 'z' key or even using the '/' key. Participants withheld their response until the probe colored square appeared at the end of the trial. If no digits appeared, then participants would press the spacebar to end the trial and begin the next one. Note that digits also appeared in the WM task, though participants were told to ignore them.

Before the EEG recording, participants first practiced a version of the digits task in which digits appeared on every trial in blocks consisting of 16 trials. Once participants were above chance and could verbally describe the task instructions, they completed a practice block that matched the trial proportions that they would experience in the main session. Once participants completed practice of the digits task, they would perform practice blocks of the change detection task. Once participants were above chance and could verbally describe the task, they advanced to the main session.

Participants completed 10 blocks of each task with 80 trials per block, for a total of 20 blocks and 1600 trials, alternating between the 2 tasks every 5 blocks. After each set of 5 blocks, the experimenter opened the EEG booth to confirm with the participant that they were switching to the next task and to provide a longer break. The initial order of the tasks was set randomly for each participant. Within each block, participants performed 10 trials for each combination of set size and placeholder condition, which were randomly intermixed. 1 trial per condition contained digits. In total, participants completed 100 trials for each of 16 conditions formed from 2 tasks, 2 placeholder levels, and 4 set sizes.

Participants received real-time feedback for eye movements made either immediately preceding trial onset (300ms before stimulus presentation), during the encoding period, or during the delay period of the trial (see section '*Eye tracking*' below). In cases in which an eye movement was detected, defined as the average distance of the two eyes from the fixation point being greater than 1.5°, the trial was aborted. Participants were presented with the fixation point, red text indicating that an eye movement had been detected, and a red dot (radius: 8 pixels) indicating the location at which the feedback software first detected that the eyes had left fixation. Participants pressed the spacebar to advance, and a version of the trial (containing the

same high-level information, but with the exact colors and locations randomly generated) was appended to a set of makeup blocks at the end of the session. Participants were allowed to self-time their breaks between blocks, and were offered coffee or tea, which they could accept at any time during the session. Acquisition sessions lasted approximately 4 hours, including preparation.

Experiment 2. Experiment 2's task followed the same structure as Experiment 1, with the following modifications. First, the delay period was extended to 1000ms. Second, as the results of Experiment 1 suggested that placeholders had minimal impact on the WM load and spatial attention signals, placeholders were included on every trial. Third, to reduce possible confusion during the change detection task, digits were no longer presented on any trials of that task. Fourth, the presentation of digits was made fully probabilistic so that participants could not predict the number of digit trials that they would experience within a block. Finally, the set of possible set sizes was expanded to 0, 1, 2, 3, 4, 6, and 8.

Set size 0 trials were modified versions of the main tasks. Both tasks began with trials consisting of only placeholders, followed by a blank delay period. In the digits task, the test display of 10% of trials would consist of 2 digits, one of which would be in a colored frame. When that occurred, participants had to judge the parity of the framed digit. On 90% of trials, participants would see a colored square, and would need to press the spacebar to indicate that no digit appeared. In the working memory task, test displays would consist of 2 colored squares, and participants would judge whether the colors of the 2 squares were the same as one another or different. For example, two green squares would be a 'same' response, while a green square and a blue square would be a 'different' response.

As mentioned above, one participant confused the two set size 0 conditions, and would press the space bar for every set size 0 trial in the change detection task for the first four blocks before catching their mistake. This participant returned for a second session, and the session was combined after dropping the set size 0 trials from the problematic blocks. In addition, two participants did not switch to the other task after the first five blocks, until partway through the sixth block. For one participant, the sixth block was dropped and the participant was included, but the other participant did not have enough trials after the block was excluded and was removed from the dataset.

Participants completed 10 blocks of each task with 84 trials per block, for a total of 20 blocks and 1680 trials, alternating between the 2 tasks every 5 blocks. The initial order of the tasks was set randomly for each participant. Within each block, participants performed 12 trials of each set size, which were randomly intermixed. In total, participants completed 120 trials for each of 14 conditions formed from 2 tasks and 7 set sizes.

Participants received real-time feedback for eye movements made during the encoding and delay periods of the trial (see section '*Eye tracking*' below). In cases in which an eye movement was detected, defined as the average distance of the two eyes from the fixation point being greater than 1.5°, the trial was aborted, and a version of the trial (containing the same high-level information, but with the exact colors and locations randomly generated) was appended to the end of the block. Makeup trials were randomly ordered at the end of a block.

As in Experiment 1, participants were allowed to self-time their breaks between blocks, and were offered coffee or tea, which they could accept at any time during the session. Acquisition sessions lasted approximately 4 hours, including preparation.

Effort Survey. Following the main data collection of Experiment 2, participants completed 2 additional tasks. First, they completed a survey in which they indicated how much mental effort they were applying to perform the various conditions. Participants were presented with a black bar at the bottom of the screen indicating how much effort they applied. The far left was labelled "no effort," while the right was labelled as "max possible effort." All conditions were presented simultaneously with image icons and titles, stacked vertically so that they would not obstruct one another. From each image, a vertical line descended to the slider bar so that participants had an accurate estimate of where each condition fell along the effort slider. Participants moved conditions along the slider by clicking and dragging the icons. Conditions were ordered by set size, with set size 0 at the bottom, and set size 8 at the top. The vertical order of the tasks was randomized across participants. Image icons were also outlined either "green" or "gold", using the default values of PsychoPy. The color-to-task mapping were also randomized across participants.

The slider line was 15 DVA wide and 0.2 DVA tall, presented 10 DVA below fixation. Icons were 0.7 DVA, and the text labels were 0.2 DVA. There was an additional buffer of 0.1 DVA between each image. The connector from each icon to the slider line was 0.1 DVA wide.

EEG Acquisition & Preprocessing

We recorded EEG activity from 30 active Ag/AgCl electrodes mounted in an elastic cap (actiCHamp, Brain Products). We recorded from international 10-20 sites Fp1, Fp2, F7, F3, Fz, F4, F8, FT9, FC5, FC1, FC2, FC6, FT10, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1, Oz, and O2. Two additional electrodes were affixed with stickers to the left and right mastoids, and a ground electrode was placed in the elastic cap at position Fpz. Impedance values

were brought below 7 kΩ at the beginning of the session using BD PrecisionGlide Blunt Fill Needles (18-gauge,1-1/2 inches). All sites were recorded with a right-mastoid reference and were rereferenced offline to the algebraic average of the left and right mastoids. Data were filtered online (low cutoff = 0.01 Hz, high cutoff = 80 Hz, slope from low to high cutoff = 12 dB/octave) and were digitized at 500 Hz using BrainVision Recorder running on a PC. Offline, data were low-pass filtered with a high cutoff of 60 Hz using MNE (Hamming window with 0.0194 passband ripple, 53 dB stopband attenuation, upper transition bandwidth: 15Hz (-6 dB cutoff frequency: 67.50Hz). Following this, the EEG data were segmented into epochs time-locked to the onset of the initial array from -450ms before the initial stimulus to 1400ms after. Data were baseline-corrected using the 300ms preceding stimulus onset in Experiment 1, and the 350ms preceding stimulus onset in Experiment 2. Trials were examined for artifacts from the beginning of the baseline period to the end of the delay period. Given excessive noise in Fp1 and Fp2, these electrodes were ignored during the artifact detection process and dropped from all analyses.

Eye tracking

We monitored gaze position using a desk-mounted EyeLink 1000 Plus infrared eye-tracking camera (SR Research). The gaze position of both eyes was sampled at 1000 Hz. We calibrated the eye tracker before each test block and between trials during the blocks, if necessary. For the real-time feedback, the position data from the eye tracker was sampled every 10ms, and the positions of the 2 eyes were averaged together. To ensure that the real-time eye tracking was accurate for fair feedback, we ran a drift-correction procedure every five trials. After collection, the data was downsampled to 500 Hz and concatenated with the EEG data. We drift corrected

the eye-tracking data for each trial by subtracting the mean gaze position measured during the same baseline window used for the EEG data.

For 1 subject in experiment 1, the sampling rate was erroneously switched to 250 Hz partway through the session. This data was split, and the latter section was upsampled to 500 before being recombined with the first section and concatenated with the EEG data.

Artifact Rejection

Eye movements, blinks, blocking, drift, and muscle artifacts were detected by applying automatic criteria, and we discarded any epochs contaminated by artifacts between the beginning of the baseline period and the end of the delay. Artifacts were detected using a combination of MNE and custom Python scripts (Gramfort et al., 2013). Trials containing rare digits were dropped. Trials contaminated by EEG artifacts were excluded from both EEG analysis and behavioral analysis. All subjects included in the final sample had at least two-thirds of the original trial counts of each condition, excluding digit presentations (60 in Experiment 1; 72 in Experiment 2). In Experiment 1, 7 participants were rejected, and in Experiment 2, 3 participants were rejected.

Drift and muscle artifacts. We checked for drift (e.g., skin potentials) and muscle artifacts. We excluded trials where EEG activity was >100 μV or less than −100 μV. We excluded trials with peak-to-peak activity >100 μV within a 200 ms window with 100 ms steps. We also excluded trials with a step change of 60 μV between the first and last halves of a 250ms window, with 20ms steps. Lastly, we excluded epochs in which any electrodes flatlined for at least 200ms, and epochs in which any electrode had a linear drift with a slope of at least 75 μV/S and an r^2 of at

least 0.3. The epoched data were manually inspected after initial labeling; any trials flagged due to large alpha in a subset of channels were reincluded if no other artifacts were detected.

Eye movements and blinks. Trials were flagged as containing a blink if the eye tracker could not detect the pupil at any time during the rejection window. Eye movements were defined as any moment when the gaze location was greater than 1° from the baseline period fixation point. Saccades were defined by windows in which the eye gaze shifted at least 0.5° between the first and last halves of an 80ms window with 10ms steps. An artifact was only flagged if it was present in both eyes on a given trial.

Behavioral Analysis

For the change-detection task, performance was computed at each set size as accuracy and Cowan's K, an estimate of the number of items successfully maintained in working memory (Cowan, 2001). In Experiment 1, we collapsed across placeholder levels. Note that because computing Cowan's K at a given set size includes multiplication by that set size value, the K estimate for set size 0 conditions in Experiment 2 is uninformative, and it is excluded from any behavioral analysis.

For the digits task, we computed the accuracy of participants on digit-present trials, excluding trials in which participants pressed the spacebar. We also examined the omission error rate (probability of pressing the spacebar on digit-present trials) and the commission error rate (probability of pressing 'z' or '/' on digit-absent trials). One participant was excluded in Experiment 1 for a high omission rate (average of 58.4%). Generally, rates for omission errors (Experiment 1 mean=1.28%, SD=1.6%; Experiment 2 mean = 0.42%, SD =0.93%) and

commission errors (Experiment 1 mean=0.41%, SD=0.85%; Experiment 2 mean = 0.16%, SD = 0.32%) were quite low, suggesting that participants were appropriately engaged in the task.

Changes in performance across set sizes were computed using a repeated-measures analysis of variance (rmANOVA) and post-hoc t-tests implemented in JASP, version 0.96 (JASP Team, 2026). ANOVA models were compared against null models that only contained random intercepts. For post-hoc t-tests, p-values were Bonferroni corrected, and posterior odds were corrected by setting the probability that the null hypothesis holds across all comparisons to 0.5 (Westfall, Johnson, & Utts, 1997). We took the same approach to examine participants effort ratings across two factors: set size and task.

Multivariate Classification & Generalization Analysis

For both tasks, we applied machine learning classification to test whether we could identify multivariate signals that scaled with set size. The analysis was implemented using Scikit-learn (Pedregosa et al., 2011) and custom Python scripts; Scikit-learn functions are noted in italics. We trained a linear discriminant analysis (LDA; *LinearDiscriminantAnalysis*) classifier using baselined EEG data on 2 set size conditions separately for each task. In Experiment 1, we trained on set size 2 and set size 4; in Experiment 2, we trained on set size 1 and set size 3. Note that for Experiment 1, all decoding analyses relied on trials with placeholders presented to control for stimulus energy. To increase SNR, we temporally smoothed the EEG data by applying a sliding window average with a width of 50ms and a step of 25ms. Following this, we applied a k-fold procedure with 5 folds (*StratifiedKFold)*. The data were randomly split into 5 folds, 4 of which were used for training while the last was held out as test data. Before fitting the model, only the 2 training conditions were selected from the training data, and they were randomly downsampled

to match the number of trials between the two. Then, at each time point and for each electrode, both the training and testing data were standardized using the mean and standard deviation of the training data (*StandardScaler*). The classifier was trained to discriminate between the conditions of interest. Then, we applied the classifier to every trial of the test set, including those from other conditions, and recorded the distance from the classifier's decision boundary. This was repeated 5 times, once for each fold to be used as the test data, and the whole procedure was repeated 1000 times.

To convert these trial-level predictions into a measure of decodability, we computed the linear discriminant contrast (LDC; Walther et al., 2016; Jones, Thyer et al., 2024; Suplica et al., 2025; Chang et al., 2026). We averaged the trial-level predictions across iterations and then averaged trials within each condition. Next, we subtracted the average predictions of the lower set size condition in our training set (set size 2 in Experiment 1 and set size 1 in Experiment 2) from the average predictions of the higher set size condition in our training set (set size 4 in Experiment 1 and set size 3 in Experiment 2). This measure is unbiased, and is centered on 0 when two conditions can't be decoded from one another, and greater than 0 otherwise (Walther et al., 2016). To test whether decodability is greater than chance, we ran a cluster-based permutation test using a one-tailed t-statistic at every time point and the default cluster inclusion threshold of alpha=.05 using MNE's procedure (Maris & Oostenveld, 2007; Gramfort et al., 2013).

To assess whether the signals decoded in one task generalized to the other task, we ran two additional analyses. First, we also computed the hyperplane contrast for the other task. For example, if we trained on set sizes 1 and 3 of the working memory task, then we examined the hyperplane contrast for the predictions of set sizes 1 and 3 of the digits task. We assessed

whether the hyperplane contrast of the held-out task was above chance following the method described above, and we assessed whether the hyperplane contrast for the trained-on task differed from that of the held-out task at each time point using a two-tailed cluster-based permutation test, again using the default cluster inclusion threshold of alpha=.05.

Second, we examined the predictions of the trained classifier on all set sizes for both tasks. To increase power, we averaged across the delay period, excluding the first and last 100 ms, when digits could never be presented. We then ran an rmANOVA with 2 factors: set size and task condition. This allowed us to examine whether the signal identified in one task remained consistent in the other task, even for set sizes that the model was never trained on. As for the behavior, ANOVAs and post-hoc t-tests were implemented in JASP, version 0.96 (Jasp Team, 2026).

For Experiment 1, we also ran the above analyses within task, examining generalization from placeholder-present to placeholder-absent trials. Given the marginal effects detected, we averaged across predictions of both placeholder levels for the main analyses.

Interpreting Decodability via Haufe Transform

To aid in interpreting the decoding results, we converted the weight matrices (W) of each model into the change in voltage patterns across electrodes (A), following Haufe et al. (2014):

$$A = \sum X * W$$

Where $\sum X$ is the covariance matrix between features (e.g., electrodes), computed automatically when fitting an LDA model. Note that when using LDA, this approach effectively resolves to

mass-univariate analysis, though the units are in terms of standard deviations, as the data is standardized before model fitting (Haufe et al., 2014). We averaged these patterns across folds and repetitions. To improve SNR while retaining the ability to gauge the temporal dynamics of the signals, we binned the subsequent timeseries in 200ms windows, starting with the onset of the trial (i.e., 0-200ms, 200-400ms, etc.). Within each time window, we computed whether any clusters of electrodes were significantly different from 0, using the default spatial adjacencies and cluster inclusions thresholds of MNE. When decoding using eye features (see below), the features don't have a spatial relationship. Therefore, we Bonferroni corrected the p-values across features in this case to correct for multiple comparisons in this case.

Eye Analysis and EEG Comparison

To assess whether spatial attention or working memory signals are present in eye signals, we repeated the above analysis on features extracted from the eyes. We began with each eye's x-pixel position, y-pixel position, and pupil size, baselined as described above. From the x and y positions, we computed the Euclidean distance from the baselined fixation point. Lastly, given the sluggish nature of the pupil and the relative brevity of our trials, we computed the temporal derivative of the pupil (Koevoet, Jones et al., 2026), as well as the other 3 features. This resulted in 8 features per eye and 16 features across both eyes.

Using these extracted eye features, we applied the exact same analysis as above, using the same folds, downsamplings, and repetitions so that trial-level predictions would be based on the exact same training set as the EEG data. We then ran the exact same analyses as above.

We next assessed whether trial-level predictions from eye signals were associated with those of EEG signals. Using the held-out trials of the training conditions (Experiment 1: set size

2 vs 4 with placeholders; Experiment 2: 1 vs 3) we fit Bayesian hierarchical models describing the relationship between the two signals, along with a contrast factor for the 2 training conditions:

$$EEGpred_{t,s} \sim Normal(P_{t,s}, Sigma)$$

$$P_{t,s} = B_{eye,s} * EYEpred_{t,s} + B_{cond,s}*Cond + B_{int,s}$$

$$Sigma \sim HalfNormal(1)$$

Where $EEGpred_{t,s}$ is the classifier's prediction from the EEG data for trial *t* of subject *s*, $EYEpred_{t,s}$ is the complementary prediction from the eye signals, Cond is a condition contrast marker with -0.5 for the low set size condition and +0.5 for the high set size condition, and B are the subject-level betas describing how the eye predictions, condition contrast, and intercept impact the EEG-based predictions. Subject-level betas are drawn from population distributions with hyperpriors:

$$B_{eye,s} \sim Normal(B_{eyes}, Sigma_{eyes})$$

$$B_{eyes} \sim Normal(0, 1)$$

$$Sigma_{eyes} \sim HalfNormal(1)$$

$$B_{cond,s} \sim Normal(B_{cond}, Sigma_{cond})$$

$$B_{cond} \sim Normal(0, 1)$$

$$Sigma_{cond} \sim HalfNormal(1)$$

$B_{int,s} \sim Normal(B_{int}, Sigma_{int})$

$B_{int} \sim Normal(0, 1)$

$Sigma_{int} \sim HalfNormal(1)$

Models were implemented in PyMC (Abril-Pla et al., 2023). We examined the population-level posteriors for each parameter to assess whether a factor reliably contributed to the EEG-based predictions. For each model, we also fit 2 alternate versions: one that excluded the eye-based predictions and only included the condition contrast as a factor, and one that excluded the condition contrast and only included the eye-based predictions as a factor. We compared the models using leave-one-out cross-validation (LOOCV), and examined both the expected log pointwise predictive density (ELPD) and the model weight as our measures of model selection. Results are presented in Appendix 3.

Representational Similarity Analysis & Bayesian Modeling

The multivariate classification analysis revealed that EEG data from the change detection task contained multiple signals that scaled with set size. Specifically, there seemed to be both a spatial attention-based signal and an additional working memory-specific signal. To estimate how the working memory-specific signal changed across set sizes, while controlling for changes in the spatial attention signal, we applied a novel analysis that combined representational similarity analysis (RSA) and Bayesian modeling.

At a high level, this analysis consisted of 2 steps. First, we computed the neural dissimilarity between every pair of conditions (16 conditions in Experiment 1, 12 conditions in Experiment 2 after excluding set size 0). Second, we fit a hierarchical Bayesian model that

assumed that the neural dissimilarity between 2 conditions was related to the sum of differences between the conditions across our factors of interest: spatial attention and working memory. Critically, we did not a priori define how either signal should change across set sizes, which we will term the signal set size function. Instead, the model was given priors on possible shapes the signal set size function could take, and estimated the posterior of the shape at the population and individual level.

To compute the neural dissimilarity between every pair of conditions, we followed previously described methods (Jones, Diaz et al., 2024; Jones, Thyer et al., 2024; Suplica et al., 2025; Chang et al., 2026), applying the procedure at each time point as in the decoding analysis. We used the cross-validated Mahalanobis distance, which is identical to the linear discriminant contrast that we used in our classification analysis (Walther et al., 2016). To compute the LDC efficiently, the data were split into 2 halves, one for training and one for testing. For each condition, we computed the mean pattern of voltage separately for the training and testing sets. To correct for the covariance between electrodes, we first demeaned each trial of the training set by its condition's mean pattern. We computed the covariance across electrodes using the demeaned training trials, regularized using the Ledoit-Wolf procedure (Ledoit & Wolf, 2004; Walther et al., 2016). The distance between each pair of conditions (i and j) was computed as

$$D^2_{LDC}(i, j) = (m_i - m_j)_{train} * \Sigma^{-1}_{train} * (m_i - m_j)_{test}$$

Where m_i and m_j are the mean voltage patterns for conditions i and j in the train or test sets, and Σ^{-1}_{train} is the inverse of the regularized covariance matrix. Note that while this example focuses on computing the difference between 2 conditions, the actual estimation is performed on all

conditions simultaneously using a covariance matrix estimated across all conditions in the training set, as implemented in rsatoolbox (Nili et al., 2014).

We then converted from squared distances to linear distances, which stabilized model fitting:

$$D_{LDC}(i, j) = Sign[D^2_{LDC}(i, j)] * (D^2_{LDC}(i, j))^{-½}$$

The Sign[] operator extracts the sign of the original distance estimate. Note that because the distance estimate is cross-validated, it is unbiased, and can be negative.

To produce stable distance estimates, we repeated this procedure 1,000 times with different train-test splits and averaged the resulting RDMs for each participant. To improve SNR for model fitting, we then average RDMs across the delay period, excluding the first and last 100ms when digits could never appear.

We next applied a hierarchical Bayesian analysis. The model assumed that the dissimilarity between 2 conditions was drawn from a normal distribution, centered on the sum of dissimilarities across our signals of interest:

$$D_{LDC}(i, j)_{sub} \sim Normal(D_{Attention}(i,j)_{sub} + D_{WM}(i,j)_{sub}, Sigma_S)$$

$$Sigma_{sub} \sim Exp(scale=Sigma_{pop})$$

$$Sigma_{pop} \sim Exp(scale=0.2)$$

For each signal, we set shape constraints for how the signal changed from one set size to the next, i.e. the signal set size function. For spatial attention, we assumed that the signal increased monotonically. We implemented this by defining the relative steps from one set size to the next

as being drawn from a Dirichlet function, which divides the interval between 0 and 1 into a series of discrete steps. Because the sum of the steps is constrained to be 1 by the Dirichlet function, the function was also scaled. Thus, the function signal (F) at a given step size (s) was defined for each subject as:

$$F_{attention}(s) = B * \Sigma^{i=s}_{min\text{-}ss} Steps_{attention}(i)$$

$$Steps_{attention} \sim Dirichlet(alpha_{pop})$$

$$B \sim Exp(scale=B_{pop})$$

$$B_{pop} \sim Exp(scale=0.2)$$

Where alpha defines the priors on where the steps should occur. In equation X, min-ss refers to the smallest set size for the given experiment (2 in Experiment 1, 1 in Experiment 2). Because the observed data are distances, the signal value for the smallest set size is arbitrary and we therefore set it at 0.

The alpha and beta population parameters were drawn from distributions with their own hyperpriors. The alpha constrains the priors on subject-specific steps in two dimensions: the relative values set the average locations of steps, while the overall magnitude sets the variability of draws relative to those means, with lower overall values increasing variability and larger overall values reducing variability. Thus, alpha is defined by its own Dirichlet distribution and scalar, and this scalar is distinct from the scalar that scales individual functions.

$$alpha_{unscaled} \sim Dirichlet(\mathbf{1})$$

$beta_{alpha} \sim Exp(scale=0.2)$

$alpha_{pop} = beta_{alpha} * alpha_{unscaled}$

Where **1** reflects a vector of ones, i.e., that the steps are drawn assuming they are evenly spaced across the [0,1] interval.

The distance between conditions i and j along the spatial attention signal was defined as:

$$D_{Attention}(i,j) = | F_{attention}(setsize(i)) - F_{attention}(setsize(j)) |$$

For Experiment 1, we included an additional function for sensory energy based on the number of items on the screen. This followed the same structure as the spatial attention function, with the same hyperpriors, but it included an additional level because the presence of placeholders on half of the trials meant that the total set of values was [2, 4, 6, 8, 10].

While spatial attention and sensory energy are assumed to increase monotonically (see *Limitations* section in the discussion), we did not want to assume that the WM signal would necessarily continue increasing at higher set sizes. Indeed, behavioral capacity limits are often around 3 items, even when participants are presented with higher set sizes (Fukuda, Woodman & Vogel, 2015; Balaban, Fukuda & Luria, 2019), and previous work examining how the contralateral delay activity (CDA), another putative measure of working memory load, found that the signal appears to plateau or even decrease at supracapacity set sizes (Vogel & Machizawa, 2004; Fukuda and Woodman). Therefore, if the multivariate load signal tracked the number of items in working memory, we might expect it to plateau or decrease after a certain set size. In contrast, if the signal reflected some other cognitive operation that was content-

independent, such as effort or arousal, it may continue to increase across set sizes (Kardan et al., 2020).

To avoid constraining the shape of the WM set size function, we simply assumed that the WM load signal for any set size in the WM task is higher than the average load signal across all of the digits task set sizes, which we treat as the lowest load value. This is supported by the results of Hakim et al. (2019), as well as the decoding results in both experiments, which suggest that the WM task engages an additional process that is not present in the digits task. We pinned the WM load value in the digits task to 0, as it is the lowest WM load condition, and assert that the WM load signal for each set size of the WM task is greater than 0. To do this, we model the mean values of the WM function at being drawn from a truncated normal distribution, with a lower bound at 0:

$$Fln_{WM}(attentionTask) = 0$$

$$Fln_{WM}(s) \sim Truncated\text{-}Normal(0.5, 0.25, min=0)$$

This set of equations defined the population-level WM set size function. Individual functions were also sampled from a bounded normal distribution with a minimum of 0, centered on the population function:

$$F_{WM,sub}(s) \sim Truncated\text{-}Normal(F_{WM}(s), sigma_{WM}(s), min=0)$$

$$Sigma_{WM}(s) \sim Exp(Sigma_{pop})$$

$$Sigma_{pop} \sim Exp(scale=0.25)$$

And the difference in WM load signal between conditions i and j for a given subject was defined as:

$$D_{WM}(i,j) = | F_{WM}(setsize(i)) - F_{WM}(setsize(j)) |$$

In addition to this model, we considered alternative models in which the WM signal plateaued after a certain set size. We did this by truncating the population-level random walk to that set size in the model, and then replicating the final value across the missing set sizes. For example, if a model contained a plateau at set size 4, then the value of population walk at set size 4 would be duplicated for set size 6 and 8. Individuals were still allowed to vary at each set size. Because the observed variables are differences computed across consistent underlying conditions, the individual datapoints are not independent (Diedrichsen et al., 2020). Therefore, common model comparison approaches like LOOCV, which rely on estimating the likelihood of individual points under the model while assuming independence, are not appropriate. Thus, we compared models by estimating the Bayes factors of the different plateau points. To compare a model which plateaus at a given set size s to a model which plateaus at the preceding set size, we asked whether the posterior difference between s and the preceding set size was reliably different from 0, for the model that plateaued at s. We computed the Bayes Factor using the savage dickey ratio, i.e. by comparing the density of the posterior at the value of 0 to the density of the prior at the value of 0. This provides bayes factors comparing the point of plateaus between adjacent set sizes, e.g. BF_{21} or BF_{86}. To compute the Bayes factors between non-adjacent set sizes, we multiplied the Bayes factors including the intervening set sizes, as in the following equations:

$$BF_{10} = P(\Theta|M_1) / P(\Theta|M_0)$$

$$BF_{21} = P(\Theta|M_2) / P(\Theta|M_1)$$

$$BF_{20} = P(\Theta|M_2) / P(\Theta|M_0) = [P(\Theta|M_2) / P(\Theta|M_1)] * [P(\Theta|M_1) / P(\Theta|M_0)] = BF_{21}*BF_{10}$$

Where Θ is the data, M_s is a model which plateaus at set size s, and $P(\Theta|M_s)$ is the probability of the data given M_s. Because a model which plateaus at set size s is used to compute the Bayes factor between a model which plateaus at set size s and the preceding set size, we fit models for all possible plateau points except for the lowest. That is, in Experiment 1 we fit models with plateau points at set size 4, 6, and 8, and in Experiment 2 we fit models with plateau points at set size 2, 3, 4, 6, and 8.

Results

Behavior performance and reported effort is impacted by set size in both tasks.

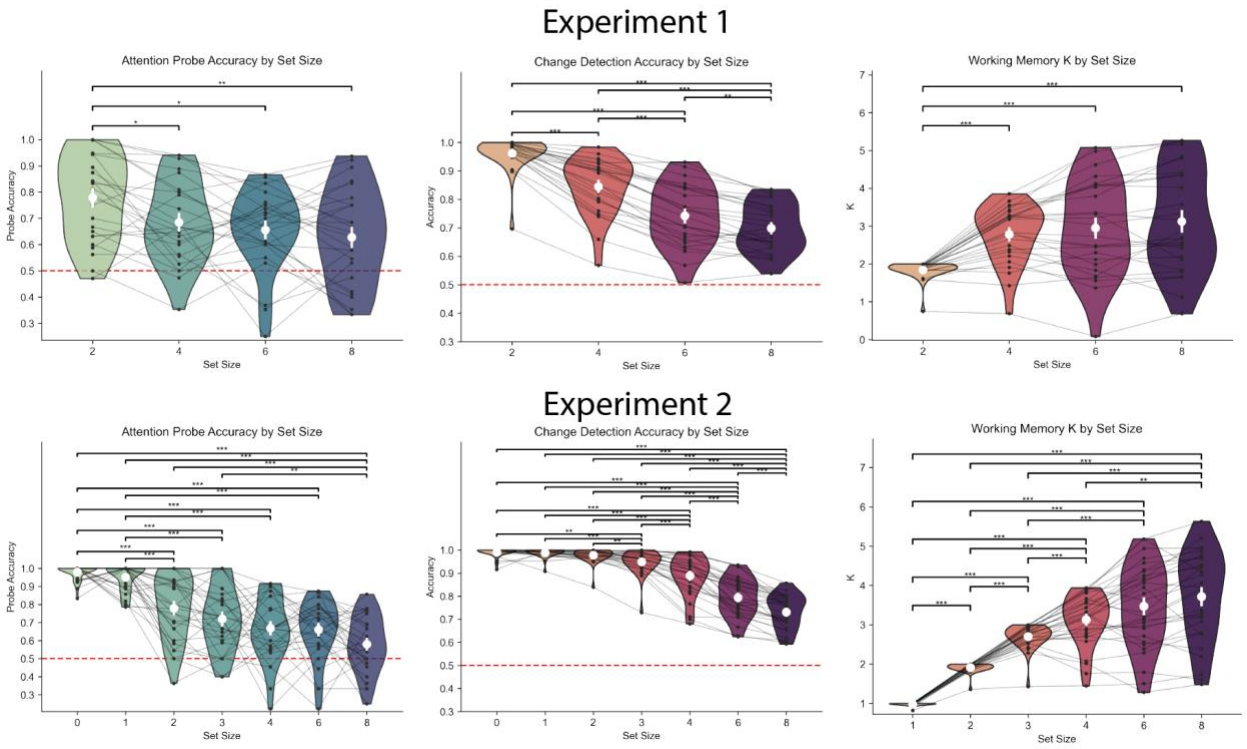

*Figure 32. Behavioral Performance. Left column shows the change in accuracy across set sizes for probe-present trials in the digits task. Middle column shows the change in accuracy across set sizes for the change detection task. Right column shows the change in estimated working memory capacity (Cowan's k) across set sizes in the change detection task. Black dots are individuals. White dots and bars show the mean and its standard error at each set size. Asterisks refer to the significance threshold crossed by the post-hoc t-test, after Bonferroni correction: *=.05, **=.01, ***=.001.*

We first asked whether set size had a reliable impact on performance in both tasks (Figure 32). For the change detection task, we examined both accuracy and Cowan's K estimates.

Experiment 1. A repeated-measures ANOVA on accuracy in the change detection task found a main effect of set size (F=191.0, p<.001, BF_m=1e32). Post-hoc t-tests revealed that each set size had significantly lower performance than the previous (alls ts >4.415, all ps <= .001, all Posterior Odds >= 68.45).

Analysis of Cowan's K, which estimates the number of items that an individual maintained in working memory, again found a main effect of set size (F=23.14, p<.001, BF_m=7e7). However, in contrast to accuracy, K estimates were relatively stable for set sizes above the estimated capacity limit. K estimates for set size 2 were significantly different from those of set size 4, 6, and 8 (all ts > 5.306, all ps < .001, all Posterior Odds >= 326.4), but K estimates did not significantly differ between set sizes 4, 6, and 8 (all ts <= 2.19, all ps >=.227, all Posterior Odds < .653).

Finally, we examined performance on the digits task, specifically on the probe-present trials. A repeated measures ANOVA found a main effect of set size (F=7.187, p<.001, BF_m=110.9). Post-hoc tests revealed that accuracy was also stable for set sizes above 2: set size 2 different significantly from set size 4, 6, and 8 (alls ts>3.241, all Ps<=.02, all Posterior Odds >=4.956), but performance did not significantly differ between set sizes 4, 6, and 8 (all ts < 1.531, all ps>=0.83, all Posterior Odds <=0.241).

Experiment 2. A repeated-measures ANOVA on accuracy found a main effect of set size (F=171.4, p<.001, BF_m=4e69). Unlike in Experiment 1, which only had a single set size within standard estimates of capacity limits, post-hoc t-tests revealed that these differences were largely driven by changes in accuracy at the higher set sizes. Accuracy on set size 0 trials was not significantly different from set size 1 (t(29)=0.409, P_{Bonf}=1, Posterior Odds = 0.046), nor set size

2 (t(29)=2.205, p_{Bonf}=.747, Posterior Odds=0.344). In addition, there was only a trend towards accuracy differing between set size 1 and set size 2 conditions (t(29)=3.285, P_{Bonf}=.056, Posterior Odds = 3.07). In contrast, all other comparisons were significantly different, as accuracy significantly decreased from one set size to the next (all ts >=4.379, all P_{Bonf} <=.003, all Posterior Odds >=41.43).

Analysis of Cowan's K, which estimates the number of items that an individual maintained in working memory, again found a main effect of set size (F=113.7, p<.001, BF_m=5e47). As in Experiment 1, K estimates were relatively stable for set sizes above the estimated capacity limit, though in Experiment 2 the estimates did increase slightly. Specifically, there was only a trend towards K differing between set size 4 and set size 6 (t(29)=2.991, P_{Bonf}=.084, Posterior Odds = 1.910) and between set size 6 and set size 8 (t(29)=2.748, P_{Bonf}=.153, Posterior Odds = 1.149), but there was a significant difference between set size 4 and set size 8 (t(29)=4.591, P_{Bonf}=.001, Posterior Odds = 83.50). All other comparisons showed that K significantly differed across set sizes (all ts >=4.864, all P_{Bonf} <.001, all Posterior Odds >=638.7).

Last, we examined performance on the probe-present trials of the digits task. A repeated measures ANOVA found a main effect of set size (F=38.33, p<.001, BF_m=3e28). Post-hoc tests revealed that performance on set size 0 was not significantly different from set size 1 (t=1.610, p_{Bonf}=1.0, Posterior Odds=0.135), though it was different from all other set size (all ts>=6.156, all ps <.001, all Posterior Odds >= 3730). Set size 1 performance was also significantly different from all higher set sizes (all ts>5.466, all ps<.001, all Posterior Odds >= 645). Set Size 2 was not significantly different from set size 3 or set size 4 (ts <= 2.411, ps>.472, Posterior Odds <0.501). In comparing set size 3 to higher set sizes, it did not significantly differ from set size 4 or 6 (ts

<= 2.47, ps>=.406, Posterior Odds <=.106), though it was significantly different from set size 8 (t=4.094, p=.007, Posterior Odds = 20.56). As in Experiment 1, there was no significant difference in performance between set size 4, 6 and 8 (all ts < 2.477, all ps >.406, all Posterior Odds <=.568).

Summary. To summarize, we found that performance in the change-detection task, as measured by accuracy, decreased across higher set sizes. In contrast, when we used Cowan's K to estimate the number of items that participants held in mind, the estimate was largely consistent across set sizes of 4 and greater, though it did significantly increase from set size 4 to 8 in Experiment 2. This difference between Experiment 1 and 2 might be driven by the increased sample size in Experiment 2, both in the number of trials and number of subjects. Finally, we found that performance on probe-present trials of the digits task decreased across intermediate set sizes, before stabilizing at set sizes of 4 and greater.

Experiment 2 Effort Ratings. For each task and every set size, we asked participants to report how much effort they allocated to each condition on a slider bar (see Materials & Methods). The range spanned from "no effort" to "max effort possible". We tested whether effort differed across tasks and across set sizes using a repeated measures ANOVA (Figure 33). We found a main effect of set size (F=167, p<.001), as well as a main effect of task (F=5.156, p=.03), and an interaction between set size and task (F=2.735, p=.042). Post-hoc tests found that the main effect of set size reflected an increase in reported effort between every pair of set sizes (all ts >=6.935, all p_{Bonf}s<.001, all Cohen's d's>=3.77, all Posterior Odds>1.3e7), The main effect of task was driven by effort reports for the digits task actually being slightly higher than the effort ratings for

the change detection task. Despite this, post-hoc tests revealed that no effort ratings significantly differed between tasks for any particular set size (all ts < 3.0, all p_{Bonf}s > 0.465, all Cohen's D < 0.455). Regarding the interaction between set size and task, this appears to be driven by a few comparisons between tasks and adjacent set sizes that were not reliable, namely set size 4 in the digits task and set size 6 in the change detection task (t=3.0, p_{Bonf}=.462, Cohen's d = .515), and set size 6 in the change detection task and set size 8 in the digits task (t=3.69, p=.073, Cohen's d=5.84). All other comparisons, excluding within the same set size, were significantly different (all ts>=5.14, all ps<.001, all Cohen's ds >= .640).

In summary, while we find that the digits task produced higher effort ratings on average, these ratings were largely quite similar at each set size. In contrast, effort ratings significantly varied across every pair of set sizes.

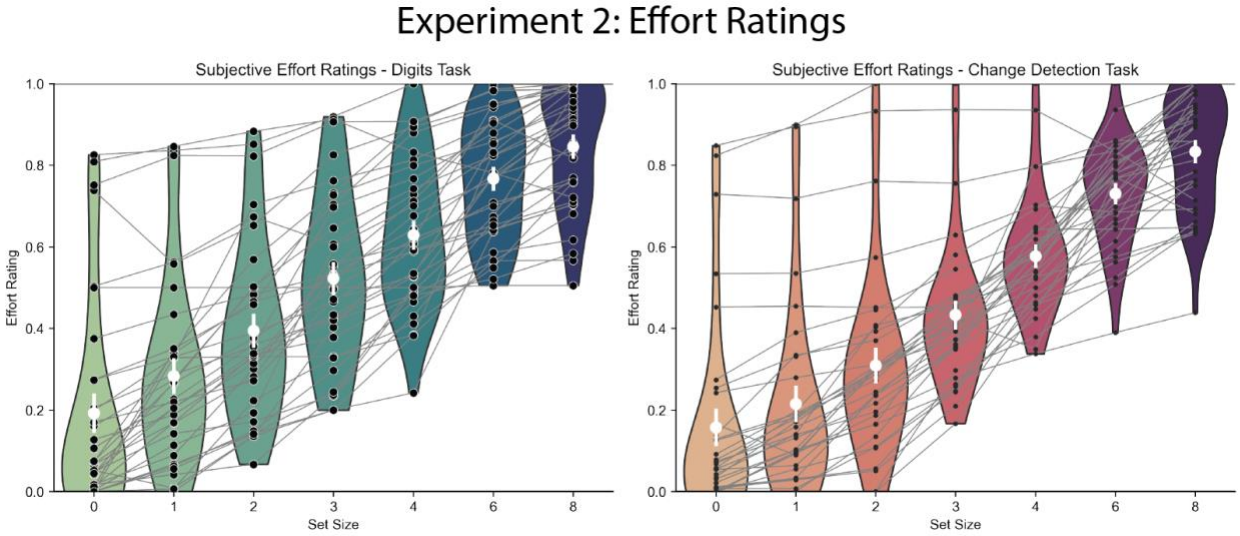

Figure 33. Experiment 2 subjective effort ratings. Left column shows the effort ratings for the digits task; right column shows the effort ratings for the change detection task. Black dots are individuals. White dots and bars show the mean and its standard error at each set size. As all

pairwise comparisons across set sizes were significant within each task, markers are excluded for visual clarity.

Placeholders induce minimal set-size-dependent filtering signals

In Experiment 1, we assessed whether the presence of placeholders induced filtering demands that changed across set sizes. To test this, we trained models decoding set size 2 against set size 4, either in the digits task or the change detection task. For each trained model, we recorded its predictions on all 16 conditions, labeled by set size, task, and placeholder condition. To increase power, we averaged the model predictions across the delay period, excluding the first and last 100ms when digits could never be presented. We then ran a repeated measures ANOVA to assess the impact of placeholders on model predictions. While set size and task are factors in the model, our primary interest here is in the main effect of placeholder presence and its interaction with these other factors

We first assessed the predictions of the model trained on set sizes 2 and 4 of the digits task. A rmANOVA revealed no significant main effect of placeholder presence ($F=0.453$, $p=.507$), nor any two-way interactions (Set Size * Placeholder: $F=.053$, $p=.953$; Task*Placeholder: $F=2.079$, $p=.162$), though there was a trend towards a three-way interaction between set size, task, and placeholder presence ($F=2.422$, $p=.079$). However, Bayesian model comparison revealed that the most likely model of the data was a model containing only set size as a factor ($BF_M=14.30$).

We next assessed the predictions of the model trained on set sizes 2 and 4 of the change detection task. A rmANOVA revealed no significant main effect of placeholder presence ($F=2.16$, $p=.154$), nor a significant interaction with set size ($F=1.217$, $p=.308$), nor a three-way

interaction between set size, task, and placeholder presence ($F=0.954$, $p=.411$). There was a significant interaction between task and placeholder presence ($F=7.532$, $p=.011$). Post-hoc tests revealed that this was driven by a trend towards different predictions between placeholder present and placeholder absent conditions of the digits task ($t=2.853$, $p_{Bonf}=.051$, Cohen's $d=0.25$), with no other differences between the other conditions (all $ts<1.437$, all $ps>.979$). Despite the presence of this small effect, Bayesian model comparison revealed that the most likely model of the data was a model containing a main effect of set size, task, and an interaction between set size and task, with no factors related to placeholder presence ($BF_M=23.51$).

Thus, it appears that the presence of placeholders did not reliably impact the predictions of the models, such that a model trained with placeholders present made nearly the same predictions for trials without placeholders. For the remaining decoding results, though we only train on placeholder-present trials, we average across placeholder presence for improved power and clearer comparison with Experiment 2.

Spatial attention signals are consistent across tasks

Figure 34. Hyperplane contrasts for decoding of set size in both experiments. Left column shows the contrasts obtained by training on the digits task and testing on either task. Right column shows contrasts obtained by training the model on the change detection task and testing on either task. Shaded regions show the standard error of the mean. Bottom squares indicate significant clusters of time points in which the contrasts are greater than chance. Top dots indicate significant clusters of time points in which the contrast for the non-trained-on task is different from the contrast of the trained-on task. Below each timeseries plot is the Haufe pattern

identified by the model in 200ms time windows. White electrodes are members of a cluster that significantly differs from 0.

We asked whether any signal scales with set size in the digits task, and whether that signal was also present in the change detection task (Figures 34-35, left columns). We found that in both experiments, we could decode set size within the digits task, and that a model trained on the digits task could generalize to the change-detection task without a drop in performance.[4] There were no clusters of time points in which the decodability of set size significantly differed between the tasks.

Next, we asked whether the digit model's predictions are consistent across all conditions (Figure 35). To increase power, we averaged across the delay period and performed a two-way repeated measures ANOVA with factors for set size and for task. For both experiments, rmANOVAs revealed a main effect of set size (Experiment 1: $F=24.01$, $p<.001$; Experiment 2: $F=32.34$, $p<.001$), but not a main effect of task (Experiment 1: $F=2.814$, $p=.106$; Experiment 2: $F=0.416$, $p=.524$), nor an interaction effect (Experiment 1: $F=1.685$, $p=.182$; Experiment 2: $F=0.710$, $p=.612$). Bayesian model comparison found that a model containing only set size as a factor was the most likely model for both Experiments (Experiment 1: $BF_M=3.939$; Experiment 2: $BF_M=14.94$).

[4] The decodability of set size in the digits task was not sustained through the delay period in Experiment 1, though it was in Experiment 2. This could be explained by a few factors impacting the SNR of the decoding analysis. First, Experiment 2 contains more trials per condition and more participants. Second, Experiment 2 contrasts set size 1 and set size 3, while Experiment 1 contrasts set size 2 and set size 4. Given that the classifier predictions shown in Figure 35 appear to show a saturating signal at high set sizes, it is likely that the difference between set size 2 and set size 4 is smaller than the difference between set size 1 and set size 3. Regardless of the difference in overall timing, the core results are the same.

Taken together, both experiments indicate that the processes that scale with set size in the digits task are preserved in the change detection task.

Working memory demands induce additional processes above and beyond spatial attention

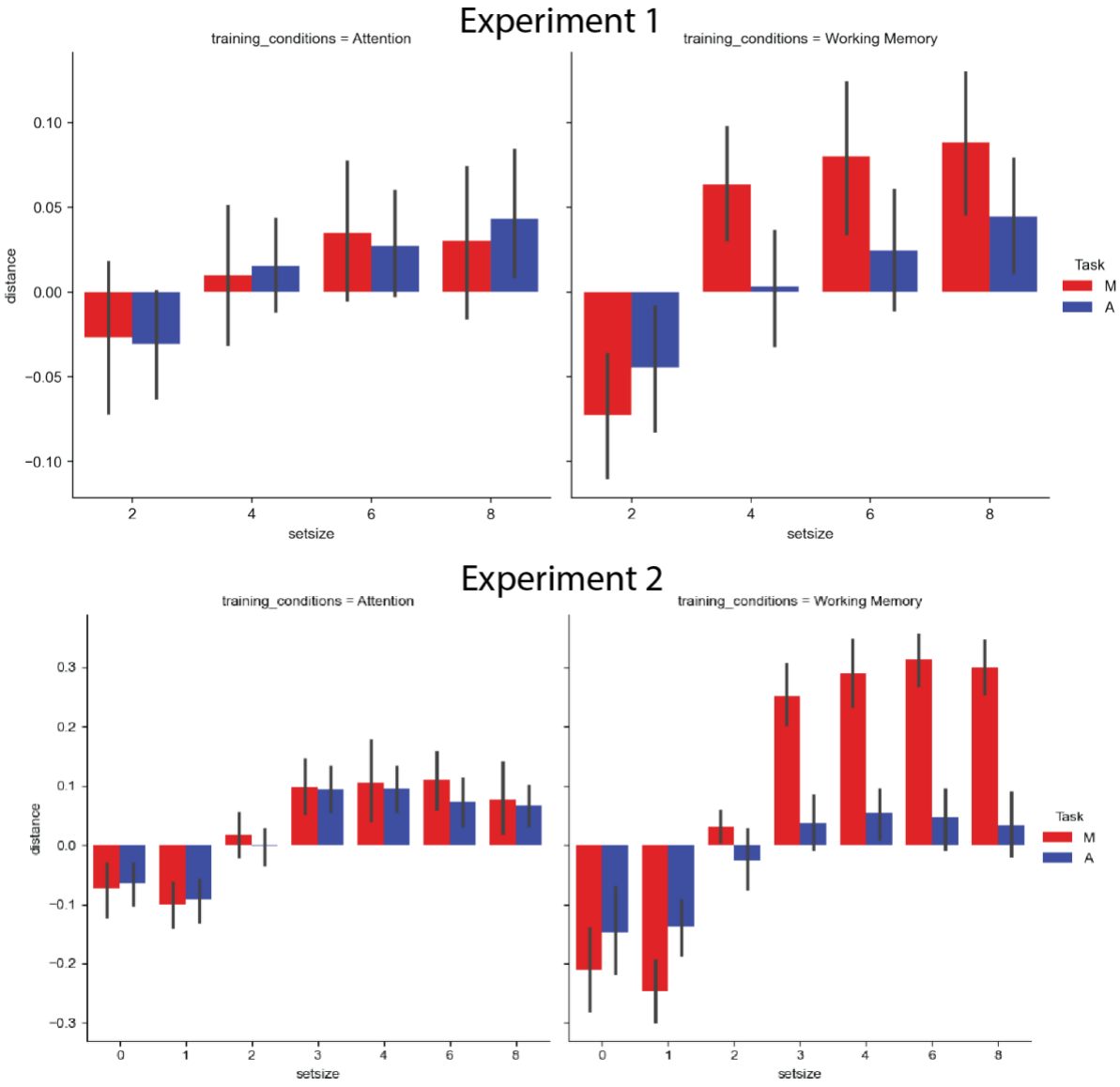

Figure 35. Condition predictions for decoding of set size in both experiments, averaged across the delay period. Left column shows the predictions obtained by training on the digits task. Right column shows predictions obtained by training the model on the change detection task.

The previous analyses found that the signals that scaled with set size in the digits task are preserved in the change detection task. This likely reflects the consist allocation of spatial attention in both tasks (Hakim et al., 2019). We next asked whether additional set size dependent signals were present in the change detection task, consistent with the selective engagement of WM storage. To test this possibility, we trained a classifier to decode set size in the change-detection task, and assessed its performance on both tasks. If change detection includes additional processes that scale with set size, such as WM storage, then the decodability of set size in the change detection task should be much larger than that of the digits task. In contrast, If the same processes are engaged in both tasks, then we should see the same generalization that we saw when training on the digits task.

Our results support the additional storage hypothesis. Namely, the hyperplane contrast for set size in the change detection task was significantly larger than the contrast for set size when tested on the digits task across the delay period in both Experiments (Figure 34, right column).

We also examined the model's predictions across all conditions, averaged across the delay period (Figure 35). In both experiments, there was a main effect of set size (Experiment 1: $F=54.24$, $p<.001$; Experiment 2: $F=86.03$, $p<.001$). In Experiment 1 there was no significant main effect of task ($F=0.24$, $p=.629$), though the main effect was significant in Experiment 2 ($F=30.80$, $p<.001$), driven by more negative predictions in the digits task than in the change detection task overall ($t=5.550$, $p<.001$). Critically, in both experiments, there was a significant interaction between set size and task (Experiment 1: $F=23.96$, $p<.001$; Experiment 2: $F=36.92$, $p<.001$), and Bayesian model comparison selected models including the interaction between set size and task as the most likely (Experiment 1: $BF_M=5e8$; Experiment 2: $BF_M>1e76$). This interaction reflects an overall increased expansion away from the hyperplane for the change

detection conditions, relative to the digit conditions. In Experiment 1, predictions of set size 2 in the change detection task were significantly more negative than those in the digits task ($t=3.764$, $p_{Bonf}=.025$), while for set sizes 4, 6, and 8, change detection predictions were numerically more positive for change detection conditions than digit conditions, though not significantly so (alls ts $<=3.1$, all p_{Bonf}s $>= .133$). Experiment 2 held a similar pattern. While Set size 0 predictions did not differ significantly between tasks ($t=2.645$, $p_{Bonf}=1.0$), set size 1 predictions were significantly more negative in the change detection task than in the digits task ($t=4.652$, $p_{Bonf}<.001$). In contrast, while set size 2 predictions were not significantly different ($t=2.269$, $p_{Bonf}=1.0$), predictions for set sizes 3, 4, 6, and 8 were all significantly more positive in the change detection task than in the digits task (all ts $>= 5.762$, all p_{Bonf}s $< .001$).

Aligned with these decoding results, we find that the voltage patterns, estimated via Haufe transformation, (Haufe et al., 2014) additionally differ across the two tasks. In the digits task, both experiments show a significant negative-going pattern in a central and posterior cluster of electrodes. In Experiment 1, this cluster shifts from a midline and right-lateralized cluster to a posterior cluster from 0-200ms to 200-400ms, while in Experiment 2, the posterior cluster is sustained across both time windows. In contrast to this, the change detection task appeared to largely contain a superset of clusters. Specifically, in addition to the negative-going posterior cluster, the change detection task also included negative-going frontoparietal electrodes in the 400-600ms window, which flipped to a positive-going pattern in the 800-1000 ms window, and in the 1000-1200 ms window of experiment 2. These frontal electrodes did not produce any significant changes in any time windows for the digits task.

Taken together, the model predictions support the theory that additional WM storage processes are engaged in the change detection task that are not present in the digits task (Hakim et al., 2019).[5]

Isolated working memory load signals plateau at behavioral capacity limits

The second major aim of the current work was to assess whether multivariate working memory load signals reflect behavioral capacity limits at supracapacity set sizes. However, the decoding results above indicate that both spatial attention and WM storage signals change with set size in the change detection task. Thus, classifier predictions alone would not provide a pure assessment of whether signals specific to WM storage are linked to prior estimates of item limits in WM capacity.

Previous research has leveraged representational similarity analysis (RSA) to estimate the individual contributions of factors to the EEG signal (Kiat & Luck 2018; Kikumoto & Mayr, 2020; Jones, Diaz et al., 2024; Jones, Thyer et al., 2024; Suplica et al., 2025; Chang et al., 2026). However, these approaches rely on explicit predictions of how these factors of interest should vary across the conditions to construct the theoretical dissimilarity matrices that are subsequently compared to the empirical dissimilarity matrix. In our case, we do not have clear predictions for how the WM load signal should behave. For example, if it tracks a core, capacity limited process specific to WM storage, then it may plateau or even decrease after a certain point (Fukuda, Woodman & Vogel, 2015). However, if it reflects a more general process like effort, then it may continue to increase across set sizes (Kardan et al., 2020).

[5] We also examined whether our decoding results could be driven by changes in eye features. Overall, we found that these eye features could not explain our results (Appendix 3).

To estimate how the WM load function changes across set sizes in a data-driven manner while isolating it from changes in spatial attention, we developed a novel analytical approach applying hierarchical Bayesian modeling to representational similarity analysis. We modeled the neural dissimilarities between 2 conditions as a function of the dissimilarities between two underlying signals: spatial attention and working memory load. For spatial attention, we assumed that its signal increased monotonically with the number of attended areas. For the working memory function, we simply assumed that working memory load was higher in the change detection task than the spatial attention task. In Experiment 1, we also included a factor for the total sensory energy that also increases monotonically. See *Materials & Methods* for details.[6]

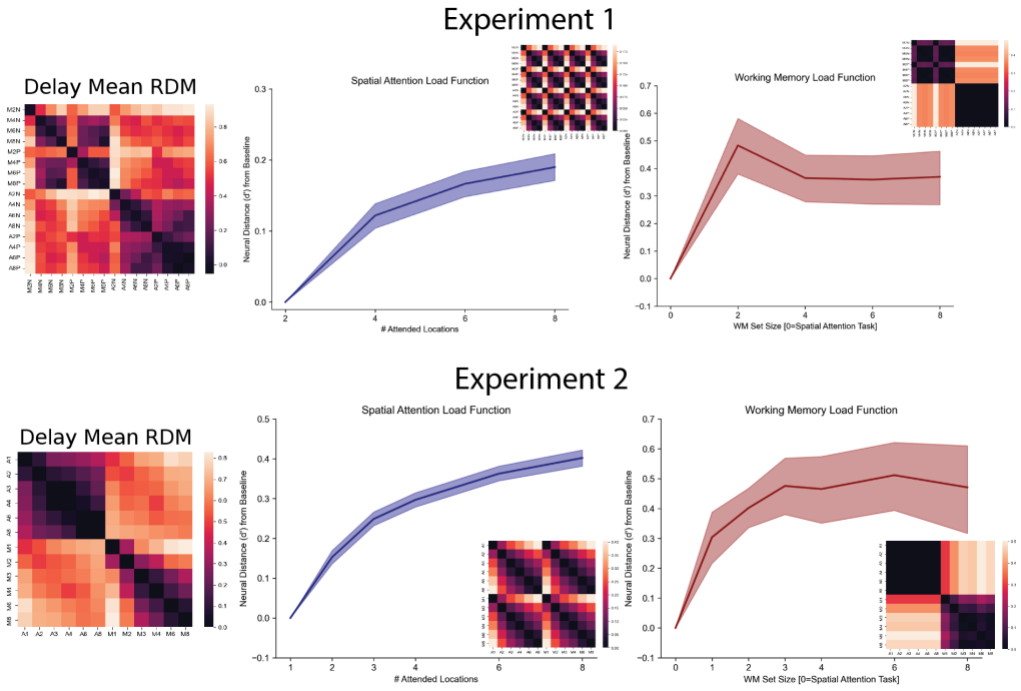

Figure 36. RSA and Bayesian modeling results. The left column shows the representation dissimilarity matrices for the 2 experiments, averaged across the delay period. The Middle

[6] Experiment 2 included a set size 0 condition as an alternate lowest set size that would also allow us to isolate baseline differences between the tasks. However, we found that SS0 largely behaved like SS1 in both tasks in the decoding results, making it difficult to identify plausible assumptions for its relationship to the other conditions. Therefore, we excluded SS0 conditions from this analysis.

column shows the posterior of the population-level spatial attention signal set size function. The Right column shows the posterior of the population-level working memory signal set size function. Insets show the expected dissimilarity matrices of each factor's set size function, based on the average over the posterior. Dark lines show the average value across the posterior, while shaded regions reflect the 94% highest density interval.

The results are presented in Figure 36. We find that spatial attention increases with a relatively log-like function. In contrast, the WM load function appears to plateau between set size 2 and 4. In Experiment 1, the WM load function peaks at set size 2, before actually decreasing to set size 4 and stabilizing. In Experiment 2, the WM load function increases up to set size 3, after which it is relatively stable.

To assess whether WM load signals plateaued at a certain set size, we computed the Bayes factors between models assuming plateaus at each possible set size. In Experiment 1, we found that a model with a plateau at set size 2 was the most likely overall, though the evidence for it being more likely than set size 4 was ambiguous (BF_{24}=1.206). Both models were more likely than models which plateaued at set size 6 or 8 (BFs>6.5, see Table 2). The results of Experiment 2 were highly similar (Table 3). We found that a model with a plateau at set size 3 was the most likely overall, though evidence for it being more likely than set size 2 was ambiguous (BF_{32}=1.219). Both models were more likely than models which plateaued at any other set size (BFs>4.88, see table X).

Experiment 1 WM Function Model Comparison		
Plateau Point (set size)	$BF_{Best,M}$	$BF_{M,next}$
2	**1.00**	**1.206**
4	1.206	6.533
6	7.882	5.274
8	34.46	-

Table 2. Experiment 1 plateau model comparison. The left column defines the plateau point of the model M. The middle column compares the bayes factors between the most likely model and the row's model. The right column shows the bayes factor between the row's model and the next most likely model (the row below).

Experiment 2 WM Function Model Comparison		
Plateau Point (set size)	$BF_{Best,M}$	$BF_{M,next}$
3	**1.00**	**1.219**
2	1.219	4.884
4	5.955	4.353
6	25.923	1.578
1	40.911	2.387
8	97.681	-

Table 3. Experiment 2 plateau model comparison. The left column defines the plateau point of the model M. The middle column compares the bayes factors between the most likely model and the row's model. The right column shows the bayes factor between the row's model and the next most likely model (the row below).

Taken together, the model comparison finds evidence for WM to plateau around a set size of 2 to 3 items. This aligns with previous large scale studies estimating behavioral capacity limits across the population, which suggest that on average participants maintain around 2.67 items (Balaban, Fukuda & Luria, 2019).

Discussion

This work found that the multivariate WM load signal, once isolated from confounding spatial attention signals, tracked behavioral capacity limits (Balaban, Fukuda & Luria, 2019). Following set size of 2 or 3, the WM load signal effectively plateaued. This supports the theory that this load signal reflects a core component of WM storage. This core component may be an indexing operation that enables selective access and manipulation of individual items in working memory (Oberauer & Lin, 2017; Hedayati et al., 2022; Awh & Vogel, 2025). Future work can leverage the multivariate WM load signal to perform hypothesis testing by manipulating potential storage or indexing demands and seeing how the WM load signal behaves. Importantly, any test of changes in the WM load signal will need to rely on careful consideration of spatial attention and how it may be controlled for to isolate WM specific signals.

The isolation of the WM load signal from spatial attention signals was only possible because we found that the digits task did not require WM storage. This is notable because the task still requires the active maintenance of information, the items' locations, without external cues to succeed at the task, for participants to successfully identify and judge the parity of the target digit. Indeed, we found that these locations were likely being actively tracked over the

delay period, through the allocation of spatial attention. These requirements match standard definitions of working memory. For example, Foster, Vogel & Awh (2022) define working memory as "an online memory system that allows us to hold information 'in mind' in service of ongoing cognitive processing" and Bays et al. (2024), state of the field of WM that "[t]he dominant notion was that visual WM holds internal copies of visual objects or features, which can be directly accessed for judgement or decision-making at a later point in time." Thus, spatial attention could be considered a form of working memory for locations. However, the additional, capacity-limited signal associated specifically with storage in the change detection task argues that spatial attention is distinct from the standard forms of WM storage that are commonly studied.

This motivates a refinement of our definition of WM storage to be more specific than merely the active maintenance of information. Instead, we propose defining WM as the severely capacity-limited cognitive system that enables active maintenance of information at the item level. This definition emphasizes two critical components of WM. First, it highlights the stark capacity limits seen in the WM load signal here, as well as in previous behavioral and neural paradigms (Balaban, Fukuda & Luria, 2019; Cowan et al., 2001; Oberauer et al., 2018), and for which individual differences are associated with differences in critical life outcomes (Mashburn et al., 2023; Fukuda et al., 2010; Cowan et al., 2005). Second, aligned with previous work showing behavioral and neural consistency across visual features including multifeature items (Luck & Vogel, 1997; Vogel, Woodman & Luck, 2001; Woodman & Vogel, 2008; Luria & Vogel, 2011; Rajsic, Burton & Woodman, 2019; Thyer et al., 2022; Jones, Thyer et al., 2024; Suplica et al., 2025; Chang et al., 2026), we propose that WM mechanisms apply at the item, or "chunk" level, and are therefore consistent across features. This contrasts with spatial attention,

which can only be used to support the maintenance of location information. Indeed, to the extent that spatial attention aids working memory (e.g., Williams et al., 2013), it may do so in a manner akin to subvocal rehearsal by enabling feature-specific maintenance in a manner that increases the overall effective capacity of the system (Baddeley & Hitch, 1974; Chen & Cowan, 2009).

A lingering question is whether this WM load signal simply reflects a more task general process that should scale with set size, such as effort or arousal. The current results argue against this possibility in a few ways. First, we found that reported effort was relatively consistent across the tasks, and that if anything it was higher in the digits task. While this may explain the spatial attention signals shared across both tasks, it cannot explain the presence of the additional, WM-specific signal. Furthermore, the Bayesian modeling approach allows for any consistent signals between the two tasks to be lumped into the spatial attention signal function, meaning that, by definition, any change in the estimated WM load function must reflect changes that are specific to the change detection task. Second, the recovered WM function plateaued at low set sizes, around behavioral capacity limits, while participants' reported effort increased with each subsequent set size, aligned with previously identified putative effort signals (Kardan et al., 2020). Thus, it appears that the WM load signal identified here cannot be explained by more general processes.

Limitations

Our design relies on two different tasks to isolate WM load signals from spatial attention signals. A signal tracking the overall task set would increase the overall difference between digit conditions and change detection conditions. Because our modeling approach did not include a

factor for task set, any signal tracking the difference between the tasks would be absorbed into the WM function. This would impact the average value of the estimated WM function across set sizes. As a result, the overall signal strength is not directly interpretable as it may reflect the combination of WM load and task set differences. This additionally limits the model's ability to identify individual differences in the WM function, as both the WM function and the strength of the task set signal may vary across individuals.

We attempted to resolve this by including a set size 0 condition in Experiment 2 so that we could directly model the contrast between task sets. However, our decoding results revealed that the set size 0 condition largely behaved like set size 1. As a result, we chose not to include set size 0 in the model so that we could keep the modeling assumptions simple. Despite the fact that the average value of the WM function is not easily interpretable due to the potential presence of a task set signal, the changes from one set size to the next, and therefore the overall shape, is. In both experiments, we find evidence that the WM load function plateaus around set size 2-3, consisted with previous estimates of behavioral capacity limits.

Separately, we chose to model the spatial attention set size function as monotonic. While this is a relaxation from an assumption of linear increases, and we are not aware of any work suggesting that spatial attention may peak or begin to compress at the range of locations tested in this work, it's still more constrained than the WM function. Unfortunately, we found that models in which spatial attention could vary in the same manner as working memory could not reliably converge, at least given the size of our dataset and the structure of our conditions. Future work is needed to find the minimum assumption set that supports the reliable identification of function shapes through this new method.

Chapter 4: Electroencephalogram decoding suggests separate indexing mechanisms for attentional tracking of featureless objects and working memory storage

Abstract

Recently, multivariate decoding of EEG data has identified a signal that scales with the number of items in working memory (WM), regardless of the specific content being maintained or the number of spatially attended locations. One possibility is that this signal reflects an abstract indexing process that binds the content of items to their context in space and time for maintenance and accessibility. To explore this possibility, we examined whether a similar load signal exists for "featureless objects", which can only be described by their coordinates in space and time. On each trial, participants viewed a dense grid of crosses of random orientations. A moving object was implemented by having a single cross change from one random orientation to another, with changes occurring between adjacent crosses across frames. These transients yielded a persisting trackable object, even though (a) there is no constant surface feature across frames, and (b) it is impossible to identify objects on any static frame. In 2 EEG datasets (N_1=25; N_2=25), participants completed a task in which they tracked either 1 or 2 cued featureless objects, and a WM task in which they remembered either colors or the shape and location of "dot cloud" stimuli while spatial attention was controlled for. In both experiments, EEG decoding found a stable signal that scaled with the number of tracked featureless objects. However, we found no consistent evidence that the tracking load signal and the WM load signal generalized to one another. Although planned studies will examine whether this conclusion generalizes to

traditional tracking tasks that may rely on other mechanisms (Lu and Sperling, 1996), these results suggest that distinct mechanisms may guide WM storage and attentional tracking.

Introduction

A critical question in the field of working memory (WM) is what mechanisms support the selective maintenance of and access to items in WM. One means of answering this question is to identify the neural signatures of working memory storage, and how these signatures change across task demands. Following this approach, previous studies examining EEG data have identified a multivariate signal that scales with the number of items in working memory, regardless of their exact content (Thyer et al., 2022; Jones, Thyer et al., 2024, Suplica et al., 2025; Chang et al., 2026). Furthermore, this WM load signal is separable from signals tracking spatial attention (Jones, Diaz et al., 2024; Suplica et al., 2025; Chang et al., 2026), and mimics behavioral capacity limits, plateauing around 2-3 items (Chapter 3), supporting the interpretation that it reflects a specific, key mechanism of WM storage.

The content-independent nature of this load signal is consistent with a class of theories suggesting that WM storage relies on the allocation of abstract indexing or binding operations to items (Swan & Wyble, 2014; Hedayati et al., 2022; Oberauer & Lin, 2017; Oberauer & Lin, 2024; Awh & Vogel, 2025). These operations provide an address for individual items maintained in working memory, allowing rapid and selective access to the items and their component features. In line with these theories, we have proposed that the WM load signal reflects the spatiotemporal indexing of items in working memory (Awh & Vogel, 2025). This theory was adapted from work by Pylyshyn (1989; 2009) and Kahneman (Kahneman, Treisman & Gibbs,

1992) that aimed to explain how people can track items over dynamic scenes, even as low level features such as luminance and shape may drastically change. We propose that these spatiotemporal indices, or "pointers" provide a means of preserving item identity while updating over low-level features.

A shared mechanism between WM storage and attentional tracking would explain why both paradigms produce lateralized posterior signals that scale with the number of targets (Drew et al., 2004; Drew et al., 2011). This shared signal may reflect the number of pointers allocated to succeed in the task. However, an alternative explanation of the above results is that participants are simultaneously encoding the tracked items into WM, producing a storage signal.

The aim of this work is to test whether WM storage and tracking rely on the same spatiotemporal indexing operation. To test this, we had participants perform a tracking task relying on the recently developed "featureless object paradigm" (Bai & Scholl, in prep). On each trial, participants viewed a dense grid of crosses of random orientations. A moving object was implemented by having a single cross change from one random orientation to another, with changes occurring between adjacent crosses across frames. These transients yielded a persisting trackable object, even though (a) there is no constant surface feature across frames, and (b) it is impossible to identify objects on any static frame. We assessed whether we could decode the number of featureless objects that participants tracked, and whether this featureless object load signal generalized to the WM load signal

To isolate WM load from spatial attention, we employed 2 working memory tasks designed to independently manipulate these two factors (Jones, Diaz et al., 2024). Experiment 1 used a change detection task with sequential presentation of the targets, allowing the targets to overlap in space. Experiment 2 used "dot cloud" stimuli which could change in size to match in

overall area across set sizes. Both WM tasks afforded independent manipulations of WM load and spatial attention, allowing us to assess whether either of these components generalized to the tracking task.

To anticipate the results, we found sustained WM load signals that were separable from spatial attention, particularly in Experiment 2. Furthermore, we found that we could decode the number of featureless objects tracked. Critically, the WM load signals did not generalize to the tracking task, suggesting that they rely on different mechanisms, rather than a shared spatiotemporal indexing operation.

Materials & Methods

Participants

Experiments 1 and 2 included 30 and 25 volunteers, respectively, participating for monetary compensation ($20 per hour). Participants were between the ages of 18 and 35 years old, reported normal or corrected-to-normal visual acuity, and provided informed consent according to procedures approved by the University of Chicago Institutional Review Board. Participants were recruited from the University of Chicago and the surrounding community via online advertisements and fliers posted on the university campus. Participants in Experiment 2 did not participate in Experiment 1.

Experiment 1. Our target sample in Experiment 1 was 25 participants. This was based on the sample size of a previous dataset using the same sequential presentation working memory task to

decode working memory load while controlling for spatial attention, which had a sample size of 20 participants (Experiment 1b of Jones et al., 2024x). However, given the aim of generalizing between this established signal and a signal from a novel task in a different paradigm, we increased the sample size. 30 volunteers participated in Experiment 1 (16 female, 14 male; mean age = 25.53 years, SD = 4.07). 5 participants were excluded from the final dataset due to insufficient trials after artifact rejection. The final sample was 25 participants (13 female, 12 male; mean age = 25.44 years, SD = 3.91).

Experiment 2. Our target sample in Experiment 2 was 25 participants. 25 volunteers participated in Experiment 2 (18 female, 5 male, 2 nonbinary; mean age = 24.64 years, SD = 4.09). All were included after artifact rejection.

Apparatus

We tested the subjects in a dimly lit, electrically shielded chamber. Stimuli were generated using PsychoPy (Peirce et al., 2019). Subjects viewed the stimuli on a gamma-corrected 24 in LCD monitor (refresh rate = 120 Hz; resolution = 1,080 × 1,920 pixels) with their chins on a padded chin rest at a viewing distance of approximately 80 cm.

Task Procedures

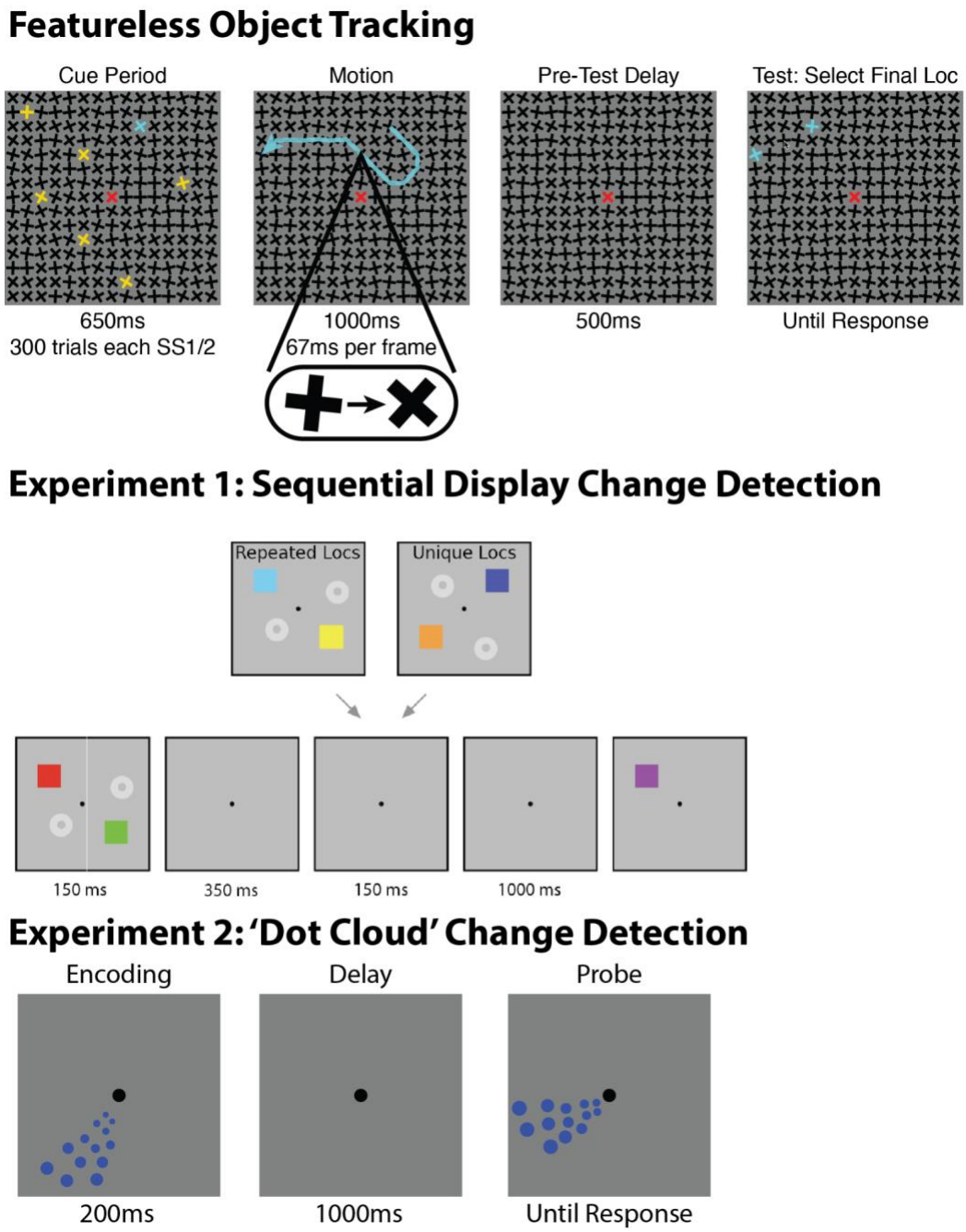

Figure 37. Task Displays. Participants in both experiments performed a tracking task using the featureless objects paradigm. In Experiment 1, participants performed a change detection task with sequentially presented targets. Participants saw 2 initial screens, each of which contained 1 or 2 targets along with placeholders. The second screen contained targets that either repeated the locations of the first screen, or which swapped the targets and placeholders. At test,

participants had to judge whether the color of the probe item matched any color that had been presented at the location previously. In Experiment 2, participants performed a spatial change detection on 'dot cloud' stimuli. Participants saw 1 or 2 clouds, and then after a delay, had to judge whether the represented cloud had changed in size or location. Note that dots clouds are enlarged for clarity in this schematic.

Luminance-balanced displays. For the working memory task in Experiment 1, the luminance of each color in the target set was recorded using an LS-150 Luminance Meter, and the luminance meter was used to identify an RGB value that produced a gray matching the average luminance of the target color set. Placeholders in this experiment were colored this gray and set to cover the same area as the targets, controlling the total area and luminance across set sizes. For the tracking task, the luminance of the two cueing colors was measured, and the colors were refined until they matched in luminance. The same process was applied for the two colors in the working memory task in Experiment 2.

Featureless object tracking task. Both experiments used the same featureless object tracking task, in which participants tracked a sequence of spatiotemporally adjacent changes (Figure 37, top row). Displays consisted of a grid of randomly oriented black crosses (RGB: 0, 0, 0) against a medium gray background (RGB: 127.5, 127.5, 127.5). The grid consisted of 23 rows and 25 columns, each of which was 0.525 degrees of visual angle (DVA) long. Crosses were presented at the center of each grid, and were 0.5 DVA long. The central cross was colored red (RGB: 255, 0, 0) to serve as the fixation point.

At the start of each trial, participants were cued to the initial positions of these featureless objects for 650ms. Two colors were used to cue the target objects as well as the distractor objects: Cyan (RGB: 0, 255, 255) and a luminance-matched gold (RGB: 239, 215, 0). Across blocks, the target-cueing color and the distractor-cueing color would swap, and the initial order was randomized across participants. During this cueing period, one or two crosses would be colored with the target cue, and five or six would be colored with the distractor cue, respectively, for a total of seven objects. Crosses were placed with the constraints that they must be at least 10 cells apart as measured in city-block units and could not appear at the fixation cross.

Following the cueing period, the crosses would turn black again, and a 1000 ms tracking period would begin. Here, featureless objects were induced as a sequence of spatiotemporally adjacent changes. For each object, a cross would randomly change orientations every 67ms, starting with the initial cued crosses. The new orientation was 30-60 degrees away from the previous orientation. The next change would appear at a cross in one of the 8 positions surrounding the previous cross.

The changes propagated across the crosses to reveal the featureless objects, with the following constraints. First, objects changed directions every 5 to 10 updates. When objects changed directions, they would turn by 45 degrees. For example, if an object was traveling upward, then it could turn to head upward and to the right or to the left. Second, objects "bounce" off the edges of the grid. For example, if a cross was traveling down and to the left when it reached the left edge of the grid, it would continue traveling down and to the right. Third, to avoid objects from passing through one another and inducing swaps, objects were repelled from one another. If objects were closer than 10 cells via city-block distance, then a sequential process would occur in which one object would compute the next position of the other

object within its 10-cell radius, and choose the available direction (straight, 45 degrees clockwise, or 45 degrees counterclockwise) that would maximize its distance from that position. The order of this process across objects was randomized from frame to frame so that no subset of objects was more responsive than another overall. This process could occur over multiple sequential frames, causing the objects to turn away from each other relatively quickly. Objects could cross through the fixation point, in which case the central cross would randomly rotate, but preserve its color. After the tracking period, to end the perception of motion, there was a 500ms static period. All crosses were still across frames, and the display did not change in any way.

Finally, a test display was presented. One of the target objects was randomly selected to be probed, and its final location's cross was colored with the target color, along with a nearby lure cross. The lure was randomly placed 6 cells away from the true location via city-block units, with the constraint that in set size 2 trials, it had to be farther away from the non-probed target than from the probed target. The cursor appeared directly between the two options, and participants used the mouse to click on the true final location. Responses were saved out as the mouse's position, and trial-level accuracy was computed as whether the mouse was closer to the correct location or the lure location.

Responding initiated the next trial with an 800-1200ms intertrial interval (ITI). All crosses reset to their original colors, but their orientations were preserved and presented across the ITI. Participants were encouraged to blink once the test display appeared, right before or as they made the mouse press, on the majority of trials, to reduce strain and the probability of eye movements or blinks during the initial stimulus presentation and delay period. Trial counts are provided for each experiment below.

Experiment 1 working memory task. Participants in Experiment 1 completed a version of a working memory task with sequential target presentations (Figure 37, middle row) that has previously been shown to produce WM load signals while controlling for spatial attention (Experiment 1B of Jones et al., 2024x).

Stimuli were presented on a dim, neutral gray background (RGB: 90, 90, 90). A central fixation point was displayed at all times. Following the recommendations of Thaler et al. (2013), it consisted of a black circle with a radius of 0.25 degrees of visual angle (DVA) across the diameter with a fixation cross matching the background gray overlaid on top with a width of 0.16 DVA, and a central black dot with a radius of 0.075 DVA (i.e., a diameter of 0.15 DVA).

Participants were sequentially presented with two screens containing targets to maintain in working memory. Each screen contained one or two colored targets, either squares or donuts, along with placeholders of the other shape. Squares had a side length of 2 DVA, while donuts had an inner radius of 0.3 DVA, and an outer radius that was computed to match the total area with that of the colored squares (4 DVA^2): 1.168 DVA. Across blocks, the target shape and the placeholder shape would swap, and the initial order was randomized across participants.

Colors were drawn from the following set of six colors: red (RGB: 255, 0, 0), green (RGB: 0, 255, 0), blue (0, 0, 255), yellow (255, 255, 0), magenta (255, 0, 255), and cyan (0, 255, 255). In addition to the colored squares, luminance-matched gray (RGB: 135, 135, 135) placeholders were presented to set the total number of objects at 4 on every screen. The stimuli were presented as in Experiment 1b of Jones et al., 2024x. Items were presented within an area spanning 12.5° vertically and horizontally, with a 3° gap around the fixation point. The area was divided into a 4 x 4 grid, each cell of which would contain one object, which was jittered within (~0.25º).

Within the task, there were two set sizes (2 and 4) and two main location conditions (same and different). Participants were given either set size 2 or set size 4 trials, with the targets split evenly across the two displays. Each display lasted 150ms, with a 350ms interstimulus interval. For both location conditions, the positions for the first array were randomly chosen from the 16 possible locations without replacement. In the same-location condition, the same locations were used for the second array. Otherwise, the target locations were randomly swapped with the placeholder locations. In addition to these 4 conditions, there was also a set size 2, simultaneous condition. In this condition, two colored squares and two placeholders were presented on the first display, while four placeholders were presented on the second display, at the same location. This condition was included to reduce differences in the expected set size between the displays, and is not included in the main analyses.

Following the initial displays, there was a 1000ms delay period. After this, a colored target was presented at one of the original target locations, and participants were asked to judge whether the color matched *any* colors that were previously presented at that location. Changes occurred on 50% of the trials, and participants responded with a 'matching color' judgement using the 'z' key, and a 'different color' judgement using the '/' key. Each color in the initial display was unique, and on change trials, the probe color was drawn from the set of colors not used on that trial.

As in the featureless object tracking task, responding initiated the next trial with an 800-1200ms intertrial interval (ITI). Participants were encouraged to blink once the test display appeared, right before or as they made the button press, on the majority of trials, to reduce strain and the probability of eye movements or blinks during the initial stimulus presentation and delay period.

Experiment 1 EEG Session. Before the EEG session, participants practiced the working memory task, then the tracking task. Participants completed blocks with 5 trials per condition (25 total trials for the WM task and 10 trials for the tracking task). Participants practiced a task until they were reliably above chance, and could verbally explain the instructions to the experimenter. After both tasks were sufficiently practiced, they advanced to the main recording session.

Participants completed 10 blocks of each task, for a total of 20. Participants alternated between the two tasks every block; the initial order was randomized across participants. Tracking blocks consisted of 30 trials per set size for a total of 60, while working memory blocks consisted of 20 trials per condition, for a total of 100. Together, participants completed 1600 trials.

Participants received real-time feedback for eye movements in a period beginning immediately preceding trial onset (300 ms before stimulus presentation) and ending at the onset of the test display (see section '*Eye tracking*' below). In cases in which an eye movement was detected, defined as the average distance of the two eyes from the fixation point being greater than 1.25°, the trial was aborted. Participants were presented with the fixation point, red text indicating that an eye movement had been detected, and a red dot (radius: 8 pixels) indicating the location at which the feedback software first detected that the eyes had left fixation. Participants pressed the spacebar to advance, and a version of the trial (containing the same high-level information, but with the details randomly generated) was appended to a set of makeup blocks at the end of the session. Participants were allowed to self-time their breaks between blocks, and were offered coffee or tea, which they could accept at any time during the session. Acquisition sessions lasted approximately 4 hours, including preparation.

Experiment 1 Effort Survey. Following the main data collection of Experiment 1, participants completed a survey in which they indicated how much mental effort they were applying to perform the various conditions. Participants were presented with a black bar at the bottom of the screen indicating how much effort they applied. The far left was labelled "no effort," while the right was labelled as "max possible effort." All conditions were presented simultaneously with image icons and titles, stacked vertically so that they would not obstruct one another. From each image, a vertical line descended to the slider bar so that participants had an accurate estimate of where each condition fell along the effort slider. Participants moved conditions along the slider by clicking and dragging the icons. Conditions consisted of set size 1 and 2 of the tracking task, and five conditions for the working memory task: 2 same-locations, 2 different-locations, 2 simultaneous, 4 same-locations, and 4 different-locations. The vertical order of the conditions was randomized across participants.

The slider line was 15 DVA wide and 0.2 DVA tall, presented 6 DVA below fixation. Icons were 1.25 DVA, and the text labels were 0.5 DVA. There was an additional buffer of 0.25 DVA between each image. The connector from each icon to the slider line was 0.1 DVA wide.

Experiment 2 working memory task. Participants in Experiment 2 completed a version of a working memory task (Figure 37, bottom row) that has been previously shown to independently manipulate WM load and spatial attention signals without the need for sequential presentations or additional placeholders (Experiment 2 of Jones et al., 2024x).

Stimuli were presented on a dim, neutral gray background (RGB: 100, 100, 100). A central fixation point was displayed at all times. Following the recommendations of Thaler et al.

(2013), it consisted of a black circle with a radius of 0.25 degrees of visual angle (DVA) across the diameter with a fixation cross matching the background gray overlaid on top with a width of 0.16 DVA, and a central black dot with a radius of 0.075 DVA (i.e., a diameter of 0.15 DVA).

Trials began with a 200ms stimulus display. Participants were presented with one or two sets of colored dots, termed dot clouds. Clouds were defined by their color, which could be either a darkish green (RGB: 0, 178, 0), or a luminance-matched purple (RGB: 255, 75, 255). Clouds could be presented in 1 of 6 equally sized circular wedges around the fixation point. The inner radius of the wedges was 2 DVA from fixation, and the outer radius was 6 DVA. Each wedge spanned 60 degrees around the circle. These wedges were further divided into a grid consisting of 8 columns and 15 rows. Each cell could only contain a single dot. Dots increased linearly in size with distance from fixation, starting at 0.175 DVA in the closest circular row and finishing at 0.225 DVA in the last circular row. Individual dots were jittered within their cells with the constraint that they could not pass any edge of the cells. The overall grid was randomly rotated between 0 and 60 degrees on every trial to prevent the appearance of consistent edges that participants could use to guide decisions.

Beginning with set size 2 trials, Individual cloud sizes were defined to span either half of a wedge (4 columns, 30 degrees), or a whole wedge (8 columns, 60 degrees). Dot assignment within this area was pseudorandomized. Each column was required to hold at least 2 dots to prevent obvious gaps or a reduction in the span of a cloud, while the remaining dots were assigned randomly to the remaining locations. A cloud could have a minimum of 20 dots. To reduce participants' reliance on density for making size judgments, the total number of dots varied from trial to trial in a uniform sample with a minimum of 40 and a maximum of 56. The locations of the two clouds were drawn such that they could not appear in the same wedge, to

avoid any additional processing related to disentangling the overlap between clouds (Jones et al., 2024x). In addition, the locations of the clouds were constrained to lie along the edge of the wedge that was closest to the other cloud. Therefore, if the clouds appeared in adjacent wedges, they would appear as adjacent to one another, even if one or both only spanned half of a wedge. Because there were 6 wedges total, this occurred on 40% of trials. Of the remaining 60%, 40% included clouds with 1 wedge between them, and 20% included clouds with two wedges between them in either direction.

Half of set size 1 trials were designed to match the general statistics of set size 2 trials. That is, these set size 1 trials were generated following the same structure of set size 2 trials with adjacent clouds, except that the dots were all a single color. This cloud spanned 8, 12, or 16 columns, and had a total of 40 to 56 dots per trial. This design matched set size 2 adjacent trials in terms of the total relevant area as well as the number and density of dots.

On the other half of set size 1 trials, a “small” cloud would be displayed. This cloud mixed the statistics of the “large” set size 1 cloud and individual set size 2 clouds. Namely, while it still contained 40 to 56 dots, it only spanned 4 or 8 columns. Note that both size conditions for set size 1 include an 8-column cloud, such that there is overlap between the two size distributions. However, the “large” set size 1 trials match the size distribution of set size 2 trials, and is the condition used to decode WM load.

Following the 200ms stimulus presentation, there was a 1000ms delay period. After this, participants were presented with a single dot cloud and had to judge whether it had changed in its overall spatial configuration. On every trial, dot clouds were redrawn with a new sample of dots to discourage participants from simply memorizing a subset of dots. Participants were asked to judge whether the dot cloud had changed in its overall size or location. Location changes were

generated by shifting the underlying grid between 45 and 60 degrees and redrawing the cloud. Size changes were generated by redrawing a cloud that was an alternate size option. For set size 2 clouds and set size 1 "small" clouds, clouds would change from spanning 8 columns to spanning 4 and vice versa. For set size 1 "large" clouds, clouds could change to any adjacent size option. For example, if the cloud spanned 12 columns, then it could change to either 8 or 16 columns. Changes occurred on 50% of the trials, and participants responded with a 'same' judgement using the 'z' key, and a 'different' judgement using the '/' key. On change trials, only one change type occurred.

Experiment 2 EEG Session. Before the EEG session, participants practiced the working memory task, then the tracking task. Participants completed blocks with 5 trials per condition (25 total trials for the WM task and 10 trials for the tracking task). Participants practiced a task until they were reliably above chance, and could verbally explain the instructions to the experimenter. After both tasks were sufficiently practiced, they advanced to the main recording session.

Participants completed 9 blocks of each task, for a total of 18. Participants alternated between the two tasks every block; the initial order was randomized across participants. Tracking blocks consisted of 30 trials per set size for a total of 60, while working memory blocks consisted of 120 trials per block to account for every combination of location, set size, and set size 1 size condition. Together, participants completed 1620 trials.

Participants received real-time feedback for eye movements in a period beginning immediately preceding trial onset (300ms before stimulus presentation) and ending at the onset of the test display (see section '*Eye tracking*' below). In cases in which an eye movement was detected, defined as the average distance of the two eyes from the fixation point being greater

than 1.25°, the trial was aborted. Participants were presented with the fixation point, red text indicating that an eye movement had been detected, and a red dot (radius: 8 pixels) indicating the location at which the feedback software first detected that the eyes had left fixation. Participants pressed the spacebar to advance, and a version of the trial (containing the same high-level information, but with the details randomly generated) was appended to a set of makeup blocks at the end of the session. Participants were allowed to self-time their breaks between blocks, and were offered coffee or tea, which they could accept at any time during the session. Acquisition sessions lasted approximately 4 hours, including preparation.

Experiment 2 Effort Survey. Following the main data collection of Experiment 1, participants completed a survey in which they indicated how much mental effort they were applying to perform the various conditions. Participants were presented with a black bar at the bottom of the screen indicating how much effort they applied. The far left was labelled "no effort," while the right was labelled as "max possible effort." All conditions were presented simultaneously with image icons and titles, stacked vertically so that they would not obstruct one another. From each image, a vertical line descended to the slider bar so that participants had an accurate estimate of where each condition fell along the effort slider. Participants moved conditions along the slider by clicking and dragging the icons. Conditions consisted of set size 1 and 2 of the tracking task, and five conditions for the working memory task. The WM task conditions were divided either the general size of the cloud in set size 1, or by the gap between clouds in set size 2. These conditions were labelled as the following: "1 small-to-medium", "1 large-to-extra-large", "2 adjacent", "2 near", and "2 far". The vertical order of the conditions was randomized across participants.

The slider line was 15 DVA wide and 0.2 DVA tall, presented 8 DVA below fixation. Icons were 1.25 DVA, and the text labels were 0.5 DVA. There was an additional buffer of 0.25 DVA between each image. The connector from each icon to the slider line was 0.1 DVA wide.

EEG Acquisition & Preprocessing

We recorded EEG activity from 30 active Ag/AgCl electrodes mounted in an elastic cap (actiCHamp, Brain Products). We recorded from international 10-20 sites Fp1, Fp2, F7, F3, Fz, F4, F8, FT9, FC5, FC1, FC2, FC6, FT10, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1, Oz, and O2. Two additional electrodes were affixed with stickers to the left and right mastoids, and a ground electrode was placed in the elastic cap at position Fpz. Impedance values were brought below 7 kΩ at the beginning of the session using BD PrecisionGlide Blunt Fill Needles (18-gauge,1-1/2 inches). All sites were recorded with a right-mastoid reference and were rereferenced offline to the algebraic average of the left and right mastoids. Data were filtered online (low cutoff = 0.01 Hz, high cutoff = 80 Hz, slope from low to high cutoff = 12 dB/octave) and were digitized at 500 Hz using BrainVision Recorder running on a PC. Offline, data were low-pass filtered with a high cutoff of 60 Hz using MNE (Hamming window with 0.0194 passband ripple, 53 dB stopband attenuation, upper transition bandwidth: 15Hz (-6 dB cutoff frequency: 67.50Hz). Following this, the EEG data were segmented into epochs time-locked to the onset of the initial array from -300ms before the initial stimulus to 2700ms after. Data were baseline-corrected using the 300ms preceding stimulus onset. Trials were examined for artifacts from the beginning of the baseline period to the onset of the test display. Fp1 and Fp2 were highly noisy in a subset of participants. Those electrodes were ignored for those participants and

subsequently excluded from the main analyses. Including the electrodes for the participants without high noise levels produced qualitatively identical results.

Eye tracking

We monitored gaze position using a desk-mounted EyeLink 1000 Plus infrared eye-tracking camera (SR Research). The gaze position of both eyes was sampled at 1000 Hz. We calibrated the eye tracker before each test block and between trials during the blocks, if necessary. For the real-time feedback, the position data from the eyetracker was sampled every 10ms, and the positions of the 2 eyes were averaged together. To ensure that the real-time eyetracking was accurate for fair feedback, we ran a drift-correction procedure every five trials. After collection, the data was downsampled to 500 Hz and concatenated with the EEG data. We drift corrected the eye-tracking data for each trial by subtracting the mean gaze position measured during the same baseline window used for the EEG data.

Artifact Rejection

Eye movements, blinks, blocking, drift, and muscle artifacts were detected by applying automatic criteria, and we discarded any epochs contaminated by artifacts between the beginning of the baseline period and the end of the delay. Artifacts were detected using a combination of MNE and custom Python scripts (MNE citation). Trials containing digits were dropped. Trials contaminated by EEG artifacts were excluded from both EEG analysis and behavioral analysis. All subjects included in the final sample had at least 100 trials per condition.

Drift and muscle artifacts. We checked for drift (e.g., skin potentials) and muscle artifacts. We excluded trials where EEG activity was >100 μV or less than −100 μV. We excluded trials with peak-to-peak activity >100 μV within a 200 ms window with 100 ms steps. We also excluded trials with a step change of 60 μV between the first and last halves of a 250ms window, with 20ms steps. Lastly, we excluded epochs in which any electrodes flatlined for at least 200ms, and epochs in which any electrode had a linear drift with a slope of at least 75 μV/S and an r^2 of at least 0.3. The epoched data were manually inspected after initial labeling; any trials flagged due to large alpha in a subset of channels were reincluded if no other artifacts were flagged.

Eye movements and blinks. Trials were flagged as containing a blink if the eye tracker could not detect the pupil at any time during the rejection window. Eye movements were defined as any moment when the gaze location was greater than 1° from the baseline period fixation point. Saccades were defined by windows in which the eye gaze shifted at least 0.5° between the first and last halves of an 80ms window with 10ms steps. An artifact was only flagged if it was present in both eyes on a given trial.

Behavioral Analysis

Accuracy was computed for every condition. For the tracking task, we ran a repeated measures t-test was run to test whether performance changed from set size 1 to set size 2. For the working memory task in Experiment 1, we ran a repeated measures analysis of variance (rmANOVA) with two factors: set size (2 vs 4) and location condition ('same' vs 'different'). Set size 2 with simultaneous presentation produced comparable performance to the other set size 2 conditions. For the working memory task in Experiment 2, we first ran a repeated measures t-test to compare

set size 1 and set size 2 while collapsing across sub-conditions. We next assessed whether performance varied within sub-conditions. Within set size 1, we ran a repeated measures t-test comparing performance on 'small' cloud trials and 'large' cloud trials. Within set size 2, we ran a repeated measures ANOVA based on the distance between clouds, either adjacent, 1 wedge separated ('near'), or 2 wedge separated ('far').

We repeated all of the above analyses on the effort ratings that participants provided at the end of the EEG sessions. T-tests were computed using the pingouin package in python. rmANOVAs and post-hoc t-tests were implemented in JASP, version 0.96 (Jasp Team, 2026). For post-hoc t-tests, p-values were Bonferroni corrected, and posterior odds were corrected by setting the probability that the null hypothesis holds across all comparisons to 0.5 (Westfall, Johnson, & Utts, 1997).

Multivariate Classification & Generalization Analyses

For all tasks, we applied machine learning classification to test whether we could identify multivariate signals that scaled with set size or relevant area. The analysis was implemented using Scikit-learn (Pedregosa et al., 2011) and custom Python scripts; Scikit-learn functions are noted in italics. We trained a linear discriminant analysis (LDA; *LinearDiscriminantAnalysis*) classifier using EEG data. For the tracking task in both experiments, we compared set size 1 vs set size 2.

In both WM tasks, we trained a set of models that either isolated WM load or spatial attention. For the WM task in Experiment 1, we trained two models. The first compared 2-different and 4-same to contrast WM load while controlling for the total number of relevant locations. The second compared 4-same and 4-different to contract spatial attention while

controlling for set size. Training on 2-same vs 2-different produced similar but weaker results. For the WM task in Experiment 2, we trained 3 models. The first compared set size 1 'large' trials and set size 2 adjacent trials to contrast WM load while controlling the attended area. The next two attempted to isolate changes in spatial attention in different ways. The first compared set size 1 "small" and set size 1 "large" trials, to contrast the breadth of spatial attention while controlling for set size. Trials in which the cloud spanned 1 wedge (8 columns) were dropped, since they were present in both conditions. Finally, we compared set size 2 "adjacent" trials to set size 2 "near" trials (1 wedge between the clouds), to contrast the potential spread of spatial attention to multiple locations, while controlling for set size.

To increase SNR, we temporally smoothed the EEG data by applying a sliding window average with a width of 50ms and a step of 25ms. Following this, we applied a k-fold procedure with 5 folds (*StratifiedKFold)*. The data were randomly split into 5 folds, 4 of which were used for training while the last was held out as test data. Before fitting the model, only the 2 training conditions were selected from the training data, and they were randomly downsampled to match the number of trials between the two. Then, at each time point and for each electrode, both the training and testing data were standardized using the mean and standard deviation of the training data (*StandardScaler*). The classifier was trained to discriminate between the conditions of interest. Then, we applied the classifier to every trial of the test set and recorded the distance from the classifier's decision boundary. This was repeated 5 times, once for each fold to be used as the test data, and the whole procedure was repeated 1000 times.

To convert these trial-level predictions into a measure of decodability, we computed the linear discriminant contrast (LDC; Walther et al., 2016; Jones et al., 2024; Suplica et al., 2025; Chang et al., 2026). We averaged the trial-level predictions across iterations and then averaged

trials within each condition. Next, we subtracted the average predictions of the lower set size or smaller relevant area condition in our training set from the average predictions of the higher set size or larger relevant area condition in our training set. This measure is unbiased and is centered on 0 when two conditions can't be decoded from one another, and greater than 0 otherwise (Walther et al., 2016). To test whether decodability is greater than chance, we ran a cluster-based permutation test using a one-tailed t-statistic at every time point and the default cluster inclusion threshold of alpha=.05 using MNE's procedure (Maris & Oostenveld, 2007; Gramfort et al., 2013).

To assess whether the different WM models successfully isolated their unique signals, we applied the trained model to the held out conditions within the WM task. For each held out condition, we tested whether it was significantly different from each training condition, using a two-tailed cluster-based permutation test with the same parameters as above.

To assess whether the signals isolated in one WM classifier model generalized to the tracking task and vice versa, we applied the trained model to the held-out conditions of the other task. However, the tasks have different temporal structures, and signals might change over time. To resolve this, we ran a temporal generalization analysis in which we used the model trained on each time point to test every other time point. We also applied this approach within the test trials of the training condition for each model to estimate how dynamic these signals were across time. Significant clusters of time points were again assessed through permutation test implemented in MNE with the same parameters.

Eye Analysis and EEG Comparison

To assess whether any of the signals identified through the above analyses are associated with changes in the eyes, repeated the above multivariate classification analyses on features extracted from the eyes. We began with each eye's x-pixel position, y-pixel position, and pupil size, baselined as described above. From the x and y positions, we computed the Euclidean distance from the baselined fixation point. Lastly, given the sluggish nature of the pupil and the relative brevity of our trials, we computed the temporal derivative of the pupil (Koevoet et al., 2025), as well as the other 3 features. This resulted in 8 features per eye and 16 features for participants with both eyes recorded. For Experiment 1, all participants had both eyes recorded. In Experiment 2, 23 of the 25 participants had both eyes recorded while the remaining 2 had one eye recorded.

Using these extracted eye features, we applied the exact same decoding analysis as above, using the same folds, downsamplings, and repetitions so that trial-level predictions would be based on the exact same training set as the EEG data. We then assessed whether trial-level predictions from eye signals were associated with those of EEG signals. To increase SNR, we averaged across the delay period for the WM tasks, and across the tracking period and pre-test static period for the tracking tasks. We fit Bayesian hierarchical models describing the relationship between the two signals, along with a contrast factor for the 2 training conditions:

$$EEGpred_{t,s} \sim Normal(P_{t,s}, Sigma)$$

$$P_{t,s} = B_{eye,s} * EYEpred_{t,s} + B_{cond,s}*Cond + B_{int,s}$$

$$Sigma \sim HalfNormal(1)$$

Where $EEGpred_{t,s}$ is the classifier's prediction from the EEG data for trial *t* of subject *s*, $EYEpred_{t,s}$ is the complementary prediction from the eye signals, Cond is a condition contrast marker with -0.5 for the low set size or smaller spatial attention condition and +0.5 for the high set size or larger spatial attention condition, and B are the subject-level betas describing how the eye predictions, condition contrast, and intercept impact the EEG-based predictions. Subject-level betas are drawn from population distributions with hyperpriors:

$$B_{eye,s} \sim Normal(B_{eyes}, Sigma_{eyes})$$

$$B_{eyes} \sim Normal(0, 1)$$

$$Sigma_{eyes} \sim HalfNormal(1)$$

$$B_{cond,s} \sim Normal(B_{cond}, Sigma_{cond})$$

$$B_{cond} \sim Normal(0, 1)$$

$$Sigma_{cond} \sim HalfNormal(1)$$

$$B_{int,s} \sim Normal(B_{int}, Sigma_{int})$$

$$B_{int} \sim Normal(0, 1)$$

$$Sigma_{int} \sim HalfNormal(1)$$

Models were implemented in PyMC (Abril-Pla et al., 2023). We examined the population-level posteriors for each parameter to assess whether a factor reliably contributed to the EEG-based predictions. For each model, we also fit 2 alternate versions: one that excluded the eye-based

predictions and only included the condition contrast as a factor, and one that excluded the condition contrast and only included the eye-based predictions as a factor. We compared the models using leave-one-out cross-validation (LOOCV), and examined both the expected log pointwise predictive density (ELPD) and the model weight as our measures of model selection.

Results: Experiment 1

Behavioral Analyses

Experiment 1. We first assessed the impact of set size on performance in the tracking task. We found that performance on set size 2 trials (Mean (SD): 88.99% (7.71%)) was significantly lower than performance on set size 1 trials (Mean (SD): 95.59% (3.69%); t=6.68, p<.001, Cohen's d=1.09, BF_{10} = 2.646e4). Along with this, participants reported allocating more effort for set size 2 trials than for set size 1 trials (t=4.952, p<.001, Cohen's d=0.775, BF_{10}=531).

We next asked whether our manipulations of set size and location impacted performance and effort on the WM task. A repeated measures ANOVA on accuracy found a main effect of set size (F=188.2, p<.001) and a main effect of location (F=7.897, p=.01), though no significant interaction (F=2.371, p=.137). Bayesian model comparison revealed that a model with only main effects of set size and location was the most likely model (BF_M=2.982). Post-hoc t-tests found a significant drop in performance from set size 2 to set size 4 (mean difference: 11.6%, t=13.72, p<.001, Cohen's d=2.332, BF_{10}=9e18), as well as a small drop in performance on same location trials, relative to different location trials (mean difference = 1.3%, t=2.810, p=.01, Cohen's d=.253, BF_{10}=3.866). Finally, we examined effort ratings. As in performance, a repeated measures ANOVA found a main effect of set size (F=48.15, p<.001), and a main effect of

location (F=6.744, p=.016), though no significant interaction (F=1.937, p=.177). Bayesian model comparison revealed that a model with only main effects of set size and location was the most likely model (BF_M=3.570). Unsurprisingly, participants reported allocating more effort for set size 4 conditions than for set size 2 conditions (t=6.939, p<.001, Cohen's D=1.474, BF_{10}=1.59e9). However, in contrast to the difference in performance across location conditions, participants actually reported allocating more effort in the different location condition, relative to the same location condition (t=2.597, p=.016, Cohen's d =.272, BF_{10}=4.113). Note that for both performance and effort ratings, the effect of location is quite small, with Bayes factors suggesting only weak evidence in favor of the effect over the null.

Experiment 2. Performance on the tracking task was highly consistent with that of Experiment 1. We found that performance on set size 2 trials (Mean (SD): 88.41% (7.01%)) was significantly lower than performance on set size 1 trials (Mean (SD): 95.48% (4.37%); t=8.357, p<.001, Cohen's d=1.19, BF_{10} = 9.143e5). Along with this, participants reported allocating more effort for set size 2 trials than for set size 1 trials (t=7.12, p<.001, Cohen's d=1.15, BF_{10}=4.316e4).

We next examined how performance and effort varied across the WM task. We found that performance was significantly lower for set size 2 (Mean (SD): 84.58% (8.58%)) than for set size 1 (Mean (SD): 90.12% (5.91%); t=5.821, p<.001, Cohen's d=.743, BF_{10}=3864.738), and that this drop in performance across set sizes was met with a corresponding increase in reported effort (t=4.22, p<.001, Cohen's d=1.16, BF_{10}=90.384). Finally, we examined how performance and effort changed within each set size. Within set size 1, we compared small clouds to large clouds. We found that performance for large clouds (Mean (SD): 86.07% (6.21%)) was significantly lower than performance for small clouds (Mean (SD): 94.10% (6.29%)). This could

be explained by a scaling of the just noticeable difference for changes with the size of the clouds. For large clouds, both size and location changes are relatively smaller than they are for large clouds. Despite the difference in performance, participants did not reliably report allocating more effort into trials with larger clouds ($t=1.31$, $p=.204$, Cohen's $d=0.258$, $BF_{10}=.465$).

Within set size 2, we examined whether the distance between clouds impacted performance. Conditions were labeled by whether the clouds were in adjacent wedges (SS2 Adj), separated by 1 wedge (60 degrees) at the closest point (SS2 Near), or separated by 2 wedges (120 degrees) in both directions (SS2 Far). A repeated measures ANOVA revealed a main effect of distance ($F=5.425$, $p=.009$), supported by Bayesian model comparison ($BF_M=5.596$). Post-hoc tests revealed that this difference was largely driven by a difference in SS2 Far trials relative to the other conditions. Specifically, there was a trend for accuracy on far trials to be better than SS2 Adj trials ($t=2.54$, $p_{Bonf}=.05$, Cohen's $d=.301$, $BF_{10}=2.294$), and a significant improvement in accuracy on SS2 Far trials relative to SS2 Near trials ($t=3.54$, $p_{Bonf}=.005$, Cohen's $d = .331$, $BF_{10}=22.36$). There was no difference between adjacent trials and near trials ($t=.254$, $p_{Bonf}=1$, Cohen's $d=.03$, $BF_{10}=.217$). Participants did not report a significant difference in effort allocation across distances ($F=1.58$, $p=.221$); Bayesian model comparison supported the null model over a model including a factor for distance ($BF_M=2.491$).

WM load and spatial attention signals are separable

We next asked whether we could decode the number of items in WM separately from the relevant areas to attend. Because our tasks have different temporal sequences, and our signals of interest might be dynamic across time, we also examined the temporal stability of our signals through temporal generalization analysis.

Experiment 1. In Experiment 1 we tested two classifiers that were designed to separate signals tracking WM load from those tracking spatial attention. First, we trained a classifier to decode set size 2 trials with different locations (SS2 Diff) from set size 4 trials with the same locations (SS4 Same; Figure 38, top row). Both conditions have the same number of relevant locations (2), so they should differ only in WM load over the course of the delay period. We found that we could significantly classify set size in these conditions, starting ~250ms after the first stimulus presentation and extending through the end of the delay period. Note that each stimulus display differs in the number of targets (1 vs 2), meaning that early classification can be driven by transient differences in WM load or spatial attention.

To assess whether the classifier reliably differentiated between WM load regardless of spatial attention in any time window, we examined the classifier's predictions on the held-out conditions: set size 2 with same locations, and set size 4 with the different locations. For each held out condition, we tested whether its predictions were significantly different from those of the two training conditions, separately. We found that the two held out conditions matched the same set size training conditions after the first stimulus presentations, when the displays were matched. However, after the second stimulus presentation, both held-out conditions were significantly deflected towards the classifier's decision boundary from ~700ms until ~1200ms. Notably, the directions of the deflections were not consistent with a spatial attention account defined by the number of relevant locations. First, SS4 Diff, which has the same WM load but *more* relevant locations than SS4 Same, was predicted to be more like SS2 Diff, rather than further from it, relative to SS4 Same. In parallel, SS2 Same, which has the same WM load but *fewer* relevant locations than SS2 Diff, was predicted to be more like SS4 Same, rather than

further from is, relative to SS2 Diff. An alternative hypothesis is that this deflection tracks a *reorienting* of attention in the different locations, rather than the total number of attended areas. Under this hypothesis, our training conditions can be defined by their WM load and whether attention was reoriented, with SS2 Diff having a low set size and reorientation, while SS4 Same has a higher set size but no reorientation. Both held out conditions contain a flipped mixture of these components: SS2 Same has a low set size but no reorientation, while SS4 Diff has a high set size but reorientation.

The above results indicate that the WM load classifier did not isolate WM load through the first period of the delay, and may also capture the reorientation of attention. Following this period, we found that starting ~1300ms after the trial onset, the held-out conditions were no longer significantly different, suggesting that at this point in the trial, the signal differentiating our two training conditions largely isolated differences in WM load.

We next trained a model to distinguish between conditions that were matched on WM set size while differing in the number of relevant locations: SS4 Same and SS4 Diff (Figure 38, second row). We found that we could transiently decode these conditions starting after the second stimulus, when the two conditions became differentiated, but decodability waned over the delay period and was no longer significant ~1450ms after the trial onset. We next examined the classifier's predictions for the held-out conditions SS2 Same and SS2 Diff. The results are aligned with a reorienting explanation. Specifically, SS2 Same, which would not require reorienting, was never significantly different from SS4 same, which also would not require reorienting, but it was significantly different from SS4 Diff across the delay period. In contrast, SS2 Diff, which would require reorienting, was significantly different from SS4 Same across the first half of the delay period, though its predictions were overall closer to the decision boundary

than SS4 Diff, and significantly so in the middle of the delay period. This orderly difference may be explained as reflecting the number of locations reoriented to, rather than the total number of relevant locations that participants may be attending to.

Finally, we examined the stability of these signals across time (Figure 38, bottom row). Aligned with our condition generalization results above, we found that the WM load signal appeared to consist of two phases (green boxes); one set of signals appeared following stimulus presentations, while another emerged late in the delay period. In contrast, we found that the spatial attention signal was quite dynamic, with minimal generalization between spread-out time points.

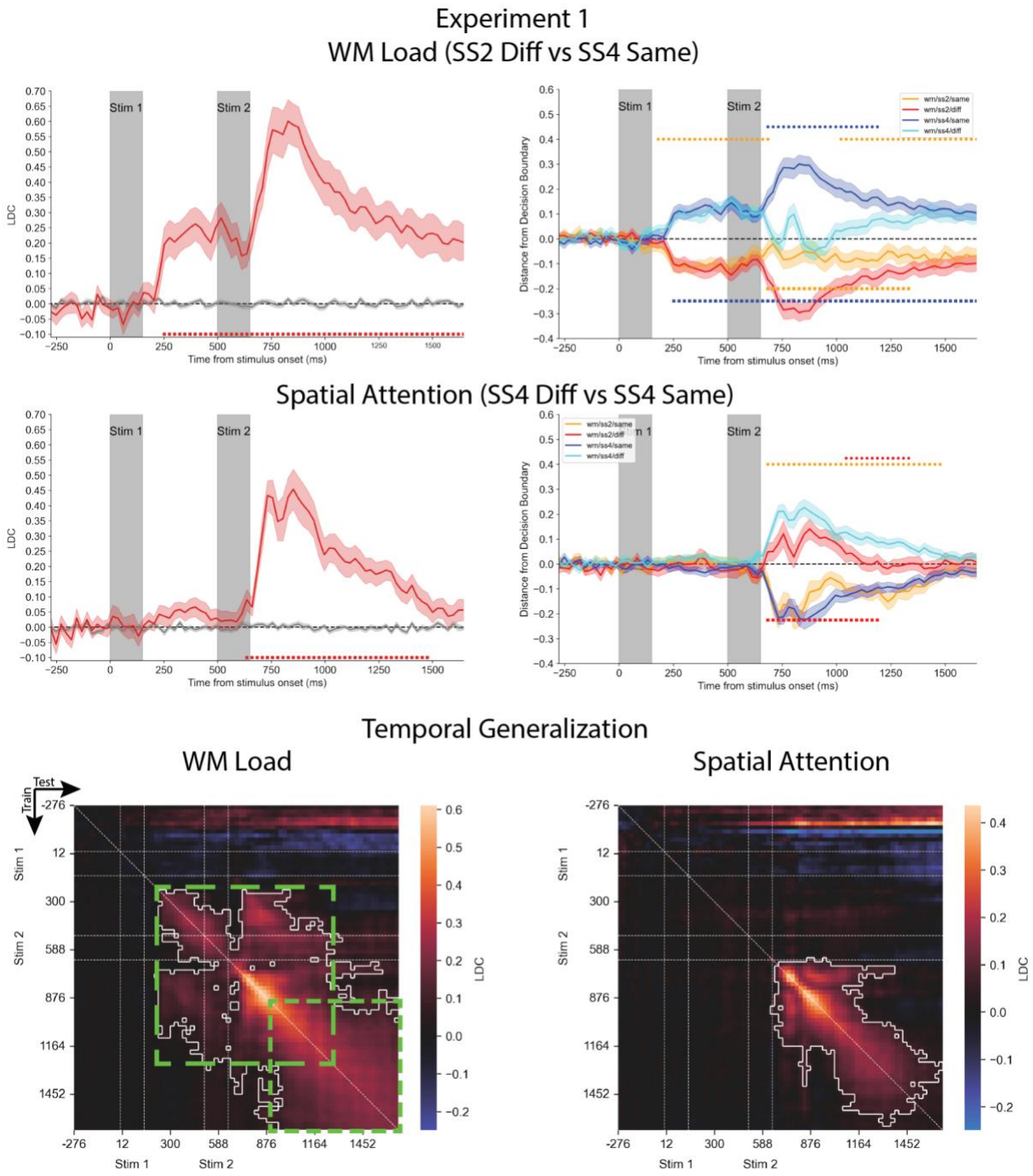

Figure 38. Experiment 1 Change Detection Task Decoding. Top row presents the decoding timeseries for conditions contrasting WM load while controlling for spatial attention (SS2 Diff, SS4 Same). The second row presents the decoding timeseries for conditions contrasting spatial attention while controlling WM Load (SS4 Same, SS4 Diff). For each of these row, the left

timeseries shows the overall decodability of the two conditions. Red squares indicate significant clusters of timepoints for which decoding is greater than chance. The right timeseries show the predictions of the models on all 4 conditions. Colored squares in the lower end of the plot indicate significant clusters of timepoints for which the condition is different from the lower training condition, and colored squares in the upper end of the plot indicate significant clusters of timepoints for which the condition is different from the higher training condition. Bottom row shows the temporal generalization results for the two models. White outlines show clusters of timepoint pairs which are significantly above chance.

Experiment 2. In Experiment 1, we found that our conditions could only isolate WM load signals from other signals late in the delay period. This may be due to a combination of the difference in the number of targets across displays, along with the difference in the need to reorient. In Experiment 2, we leveraged the dot clouds WM task, which controls for these potential differences.

We trained and tested 3 models to compare potential WM load signals and spatial attention signals. First, we trained a model to discriminate between a pair of conditions that differed in set size, but were matched on spatially relevant statistics: set size 1 trials with large clouds (SS1 Large), and set size 2 trials with adjacent clouds (SS2 Adj). These conditions are matched in the area, density, and luminance of the individual dots presented. Results are presented in the top row of Figure 39. We find that we can decode set size starting at the end of the stimulus display, extending through the end of the delay period. We then examined the model's predictions on held-out conditions: set size 1 with small clouds (SS1 Small), set size 2

clouds that were 60 degrees apart (SS2 Near), and set size 2 clouds that were 120 degrees apart (SS2 Far). The predictions for each of these conditions matched those of the training condition with the same set size. SS1 Small predictions were significantly different from SS2 Adj predictions across the delay period, but not those of SS1 Large, and SS2 Near and SS2 Far predictions were significantly different from SS1 Large predictions across the delay period, but not those of SS2 Adj. Thus, the WM load model seems selectively sensitive to the number of clouds being maintained in WM, not their size or relative positions.

We next examined whether spatial attention signals were impacted by the size of clouds or the distance between them. To examine signals tracking the size of the clouds, we trained a model contrasting SS1 Small and SS1 Large clouds (Figure 39, second row). As in Experiment 1, we could decode this putative spatial attention signal starting during the stimulus presentation, though it waned across the delay period until it was no longer significantly decodable, ~700ms after the trial onset. In addition, the classifier's predictions for the held-out SS2 conditions (SS2 Adj, SS2 Near, SS2 Far) largely matched those of SS1 large, with a few exceptions. During the stimulus period, SS2 Near and SS2 Far were predicted to be significantly further from the decision boundary than SS1 Large, and during the delay period, there was a transient cluster of time points for which SS2 Far was significantly further from the decision boundary than SS1 Large again. However, SS2 Adj predictions were not significantly different from SS1 Large predictions at any time point.

To examine signals tracking the distance between the clouds, we trained a model contrasting SS2 Adj and SS2 Near. As above, we could transiently decode the distance between clouds starting from the stimulus period until ~750ms. Furthermore, the classifier’s predictions for the held-out conditions (SS2 Far, SS1 Small, SS1 Large) largely tracked the gap between

relevant areas, with one exception. SS1 Small predictions were significantly further from the decision boundary during the stimulus period than SS2 Adj, but they were not significantly different from one another during the delay period, and SS1 Large predictions were never significantly different from those of SS2 Adj at any time point. Interestingly, SS2 Far predictions were significantly further from the decision boundary than those of SS2 Near in two clusters of timepoints, one midway through the stimulus period into the early delay, and another in the middle of the delay. This suggests that this signal parametrically tracks the distance between relevant locations, rather than a binary property that reflects whether attention is at a single location or multiple locations.

Taken together, we find that these two forms of spatial attention are largely separable in the dot clouds paradigm. There does appear to be some overlap of the signals during the stimulus period, when predictions for both models seem to be impacted by the size and distance of the clouds. This may reflect the initial allocation and spread of spatial attention. Following this initial period, these two signals seem distinct. Critically, we found that our WM load signal of interest appeared to be isolated from these changes in spatial attention. A model trained on SS1 Large and SS2 Adj trials was not reliably impacted by differences in the other conditions, and predictions for SS1 Large and SS2 Adj trials were consistent across both spatial attention models.

Lastly, we examined the temporal stability of these signals (Figure 39, bottom row). As above, we found that the WM load signal was fairly dynamic, though it appeared to be stabilizing late in the delay period, ~800ms. In contrast, the cloud size signal was highly dynamic. Finally, the cloud gap signal appeared to be dynamic in the earliest moments before stabilizing during the early delay period, as it waned.

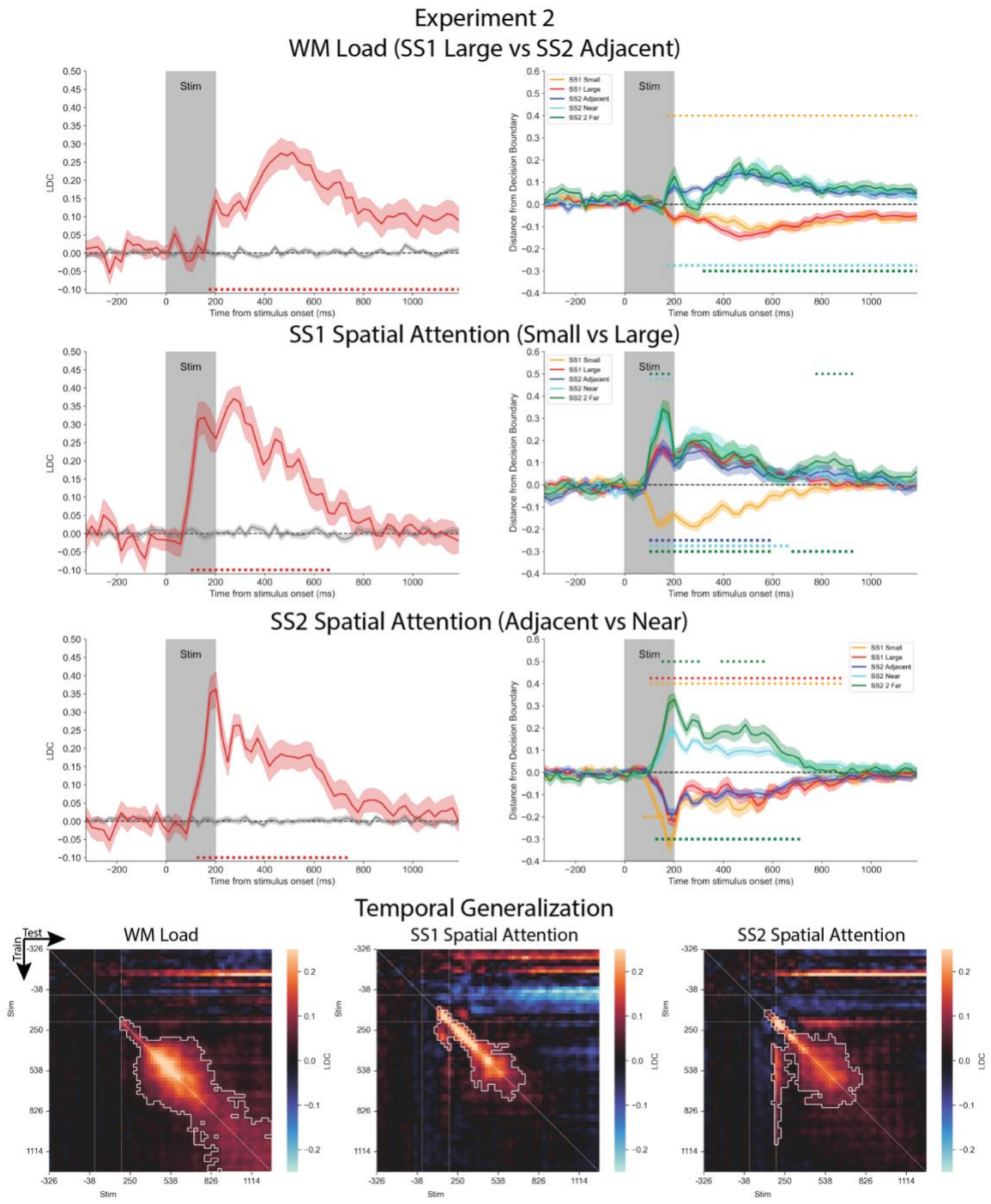

Figure 39. Experiment 2 Change Detection Task Decoding. Top row presents the decoding timeseries for conditions contrasting WM load while controlling for spatial attention (SS1 Large, SS2 Adj). The second row presents the decoding timeseries for conditions the manipulate overall cloud size while controlling for WM load (SS1 Small, SS1 Large). The third row presents the

decoding timeseries for conditions that manipulate the gap between clouds while controlling for WM load (SS2 Adj, SS2 Near). For each of these row, the left timeseries shows the overall decodability of the two conditions. Red squares indicate significant clusters of timepoints for which decoding is greater than chance. The right timeseries show the predictions of the models on all 4 conditions. Colored squares in the lower end of the plot indicate significant clusters of timepoints for which the condition is different from the lower training condition, and colored squares in the upper end of the plot indicate significant clusters of timepoints for which the condition is different from the higher training condition. Bottom row shows the temporal generalization results for the three models. White outlines show clusters of timepoint pairs which are significantly above chance.

Multivariate signals scale with the number of featureless objects tracked

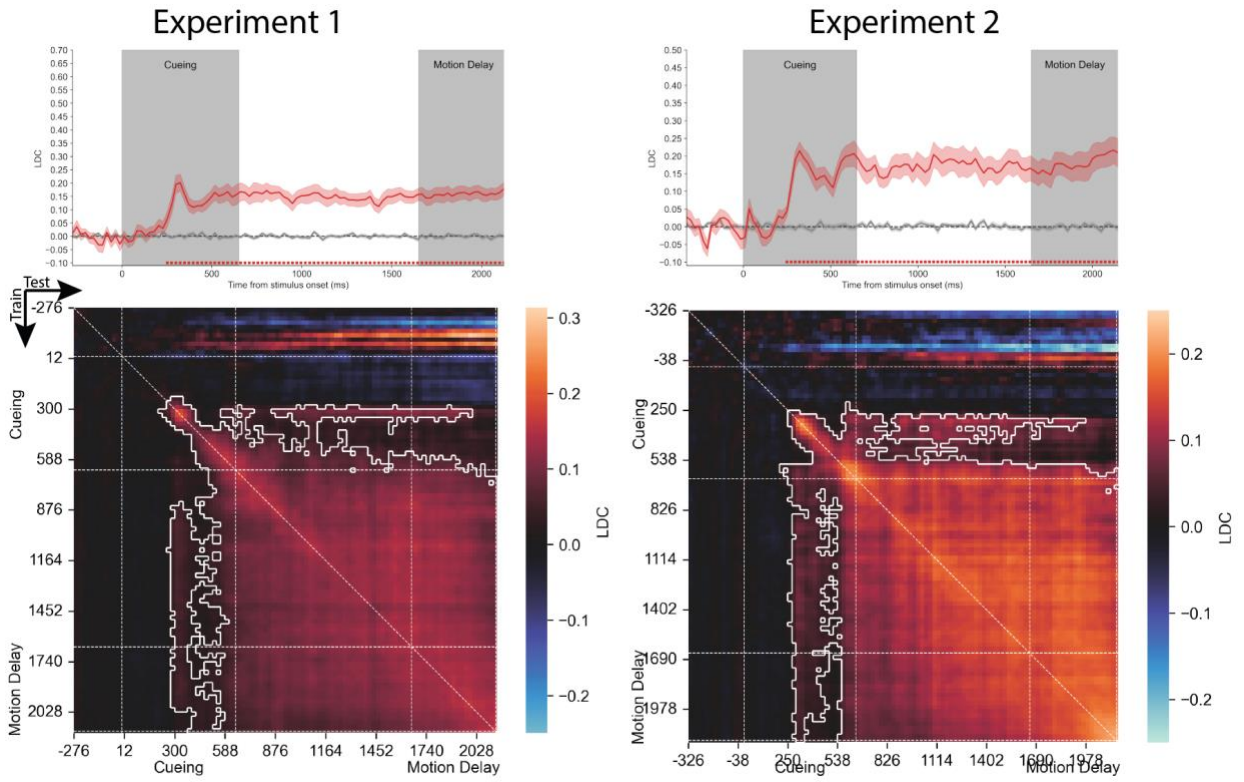

Figure 40. Decoding results for featureless object tracking load. The left and right columns show the results for Experiments 1 and 2, respectively. The top plot shows the decoding timeseries. Red squares indicate significant clusters timepoints that are different from chance. The bottom plots show the temporal generalization results. White outlines show clusters of timepoint pairs which are significantly above chance.

We next asked whether we could decode the number of featureless objects being tracked in our novel tracking task (Figure 40). We trained a classifier to decode set size in the featureless object task. The classifier could reliably distinguish between set size 1 and set size 2 trials from ~200ms into the cue period through the tracking period and the pre-test delay period. Temporal generalization analysis revealed that the signal was somewhat dynamic through the cueing period, before stabilizing through the rest of the trial. Timepoints from the end of the cueing

period and the beginning of the tracking period significantly generalized to the end of the pre-test delay, and vice versa, indicating that the signal was highly stable over time.

The results from Experiment 2 replicated Experiment 1 nearly identically. We found the same time window of decoding, starting ~200ms in the cue period and extending through the rest of the trial, with a highly stable signal beginning around the beginning of the tracking period. Thus, the tracking signal seems to be highly reliable and stable, even across distinct groups of participants.

EEG signals cannot be explained by changes in the eyes

Before assessing whether the signals identified in the tracking task were consistent with those in the WM tasks, we asked whether the signals were driven by changes in the eyes. For each model trained on EEG data described above, we trained complementary models using the x-position, y-position, Euclidean distance from the initial fixation, pupil size, and the temporal derivatives of these factors for each eye available. We then compared the trial-level EEG-based predictions against those of the eye-based predictions using a hierarchical linear regression approach. We modeled EEG-based predictions on each trial as being drawn from a normal distribution whose mean was defined by a linear sum of an intercept, a trial condition contrast, and the eye-based predictions. To assess whether the EEG-based predictions could be explained solely from the eyes, we compared these models to two alternatives that excluded either the condition contrast or the eye-based predictions.

Experiment 1. We found that for every classifier, a model containing condition information was necessary to explain the EEG-based predictions (Table 4). For the WM load classifer, a model containing both condition labels and eye-based predictions was the highest performing (weight=62.54%), while in the spatial attention and tracking classifiers, models containing only the condition label were the highest performing (weight: 70.44-98.74%). Note that in all cases, the second-highest performing model was comparable in ELPD to the highest performing, meaning that for every classifier, it was ambiguous whether a model including eye-based predictors was better than a model using only the condition labels. Regardless, models using only the eye-based predictions to explain the EEG-based predictions were always the lowest-ranked, with much lower ELPD than the top two, and low weight (1.26-4.54%).

Experiment 1 EEG & Eye prediction models comparison				
Decoding Model	Rank	Name	ELPD difference from best model: Mean (SD)	Weight (%)
WM Load (2 Diff vs 4 Same)	**1**	**condition&eyes**	**-**	**62.54**
	2	condition-only	2.195 (3.321)	32.93
	3	eyes-only	331.291 (25.901)	4.54
Spatial Attention (4 Same vs 4 Diff)	**1**	**condition-only**	**-**	**98.74**
	2	condition&eyes	1.539 (1.044)	0
	3	eyes-only	126.540 (15.764)	1.26

Table 4. Model comparison results explaining EEG-based predictions from eye-based predictions in Experiment 1. The column explanations are as follows. Decoding Model: the classifier whose predictions are being tested, defined by the conditions it was trained on. Rank:

the rank order of the 3 models being tested for that classifier's predictions. Name: the model being ranked. ELPD difference: the difference in expected log pointwise density, a measure of the likelihood of the data under the model, estimated via efficient leave-one-out-cross-validation (LOOCV). Weight: the relative weighting of the models when combined to explain the data. Roughly interpreted as the probability that the model captures the true data generating process.

Decoding Model	Rank	Name	ELPD difference from best model: Mean (SD)	Weight (%)
Tracking Load	**1**	**condition-only**	**-**	**70.44**
	2	condition&eyes	0.717 (1.876)	27.61
	3	eyes-only	80.641 (12.93)	1.95

Table 4, continued.

Experiment 2. The results of Experiment 2 largely replicated those of Experiment 1 (Table 5). In Experiment 2, a model containing both the condition label and eye-based prediction was the winning model for the tracking load EEG predictions (weight = 77.33%), in comparison to Experiment 1, in which a model only containing condition labels was the top performer. However, as in Experiment 1, the ELPD between the top two models was quite comparable, suggesting ambiguity over the reliability of the relationship between EEG-based predictions and eye-based predictions. For the 3 classifiers defined on conditions in the WM task, the models including only condition labels were the top performers (weight: 71.64-98.29%), though the ELPD was again comparable to the models including the eye-based predictions. In all cases, the models including only eye-based predictions were the lowest performers, with much lower ELPD and consistently low weight (1.71-7.52%).

Taken together with the results of Experiment 1, it appears that while eye-based signals may be related to EEG-based signals, one does not explain the other.

Experiment 2 EEG & Eye prediction models comparison				
Decoding Model	Rank	Name	ELPD difference from best model: Mean (SD)	Weight (%)
WM Load (1 Large vs 2 Adj)	**1**	**condition-only**	**-**	**71.64**
	2	condition&eyes	0.795 (2.040)	23.04
	3	eyes-only	142.223 (17.684)	5.32
SS 1 Spatial Attention (1 Small vs 1 Large)	**1**	**condition-only**	**-**	**98.29**
	2	condition&eyes	1.439 (1.130)	0
	3	eyes-only	44.112 (9.474)	1.71
SS 2 Spatial Attention (2 Adj vs 2 Near)	**1**	**condition-only**	**-**	**92.48**
	2	condition&eyes	1.533 (1.351)	0
	3	eyes-only	35.261 (9.065)	7.52
Tracking Load	**1**	**condition&eyes**	**-**	**77.33**
	2	condition-only	5.137 (4.011)	18.36
	3	eyes-only	104.760 (14.910)	4.31

Table 5. Model comparison results explaining EEG-based predictions from eye-based predictions in Experiment 2. The column explanations are as follows. Decoding Model: the classifier whose predictions are being tested, defined by the conditions it was trained on. Rank: the rank order of the 3 models being tested for that classifier's predictions. Name: the model

being ranked. ELPD difference: the difference in expected log pointwise density, a measure of the likelihood of the data under the model, between the ranked model and the best model. Weight: the relative weighting of the models when combined to explain the data. Roughly interpreted as the probability that the model captures the true data generating process.

Tracking load signals are independent from WM load signals

Finally, we assessed whether the WM load and spatial attention signals identified in the WM tasks generalized to the load signal in the featureless object tracking task. Because of the difference in task timings as well as the dynamic signals that we identified above, we performed a temporal generalization analysis in which we trained on every time point and tested on every other time point. For all analyses, we trained classifiers on the WM conditions and tested on the tracking conditions, but the results were qualitatively identical when training on the tracking conditions and testing on the WM conditions.

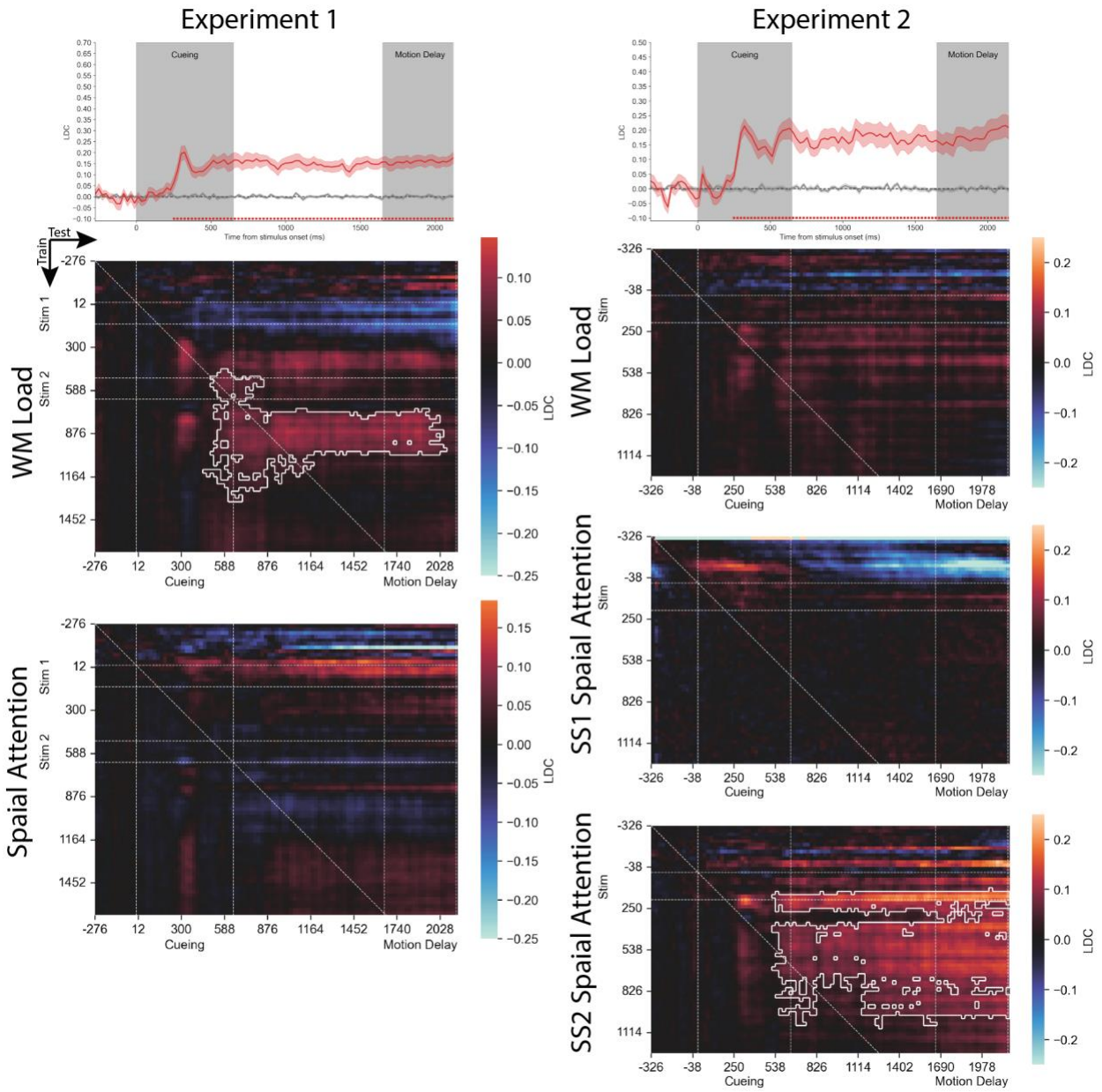

Figure 41. WM and Tracking Generalization Results. The left and right columns show the results for Experiments 1 and 2, respectively. Each plot shows the temporal generalization results when training on some subset of conditions from the WM tasks and testing on SS1 and SS2 of the tracking task. White outlines show clusters of timepoint pairs which are significantly above chance.

Experiment 1. To compare WM load to tracking load, we trained a model on SS2 Diff and SS4 Same, and tested it on SS1 and SS2 of the tracking task (Figure 41, left column). We found a transient window, starting with the second presentation through the first half of the delay period, that generalized to the stable period of the tracking task, the tracking period and the pre-test delay period. This is the period in which the WM load model seemed to reflect both WM load and attentional reorienting. However, we found that a model that was trained to decode SS4 Same from SS4 Diff, which might isolate this reorienting, did not generalize to any time point of the tracking task.

Critically, the late delay period of the WM load model did not generalize to the Tracking task. Overall, this pattern of results suggests that some transient signal which scales with set size in the WM task, but is not related to attentional reorienting or WM load, generalizes to the tracking task.

Experiment 2. Experiment 2 allowed us to compare a set of models that more clearly isolated WM load and two separate spatial attention signals against the load signal in the tracking task. Critically, we found that a model trained to isolate WM load using SS1 Large and SS2 Adj trials did not generalize to the tracking task at any time point (Figure 41, right column). This was also the case for a model trained to decode the size of the clouds using SS1 Small and SS1 Large trials. However, a model trained to decode the *distance* between clouds using SS2 Adj and SS2 Near trials did generalize to the tracking task. This generalization began around the end of the WM stimulus presentation and extended fairly far into the WM delay, reflecting the times for which decoding of distance was significantly above chance. These timepoints generalized to the stable period of the tracking task, namely the tracking period and the pre-test delay.

Discussion

Here, we tested the hypothesis that the content-independent WM load signal reflected the allocation of spatiotemporal indices that enable the selective maintenance of and access to items in WM. To begin, we found that we could decode both the number of items stored in WM (WM load) and the number of featureless objects tracked (tracking load) in a sustained manner. Despite this, we did not find consistent evidence that the WM load signal generalized to the tracking load. While Experiment 1 revealed a transient period in which the WM load signal generalized to the tracking task, examination of the model predictions for the other WM conditions revealed that the WM load signal was not isolated in this time window.

Examination of classifier predictions in Experiment 2 revealed a clear separation between apparent WM load signals and two separate spatial attention signals. These two spatial signals appeared to be initially interlinked, before separating into signals tracking the total relevant area and the distance between relevant areas, respectively. Surprisingly, we found that the WM load signal isolated in Experiment 2 did not generalize to the tracking load signal. Instead, only the signal tracking the distance between the clouds. We provide a post-hoc interpretation of this result, as well as the transient generalization seen in Experiment 1, below.

The current work does not provide evidence that WM relies on spatiotemporal indexing. However, it is possible that our tasks failed to tap into the key constructs of interest. Specifically, the tracking task may rely on separate mechanisms from traditional tracking paradigms, preventing the allocation of spatiotemporal indices. Lu & Sperling (1995; 1996) previously identified 3 forms of motion detection that could enable tracking. The first two orders, enabling tracking of luminance differences and textures, are fast and monocular, while the third is slower,

binocular, and applicable to broader forms of detectable motion. They propose that this third order of motion detection can be used to infer motion in complex stimuli that would not be detectable by the first two, such as binocular stereograms, and changes in orthogonal motions patterns, such horizontal shifts in the spatial positions of columns defined by transient dots shifting up or down in alternation. For these latter paradigms, they propose that participants must rely on top down control to prioritize certain features, producing locations on a salience map that evolve over time, generating a sense of motion. It may be that traditional tracking tasks can leverage the first two forms of motion detection, enabling the allocation of spatiotemporal indices.

The salience map hypothesis may also explain the pattern of generalization that we see between our experiments. In Experiment 2, we see generalization between the tracking load signal and a signal tracking the distance between clouds. A salience map marking the locations of clouds could serve as a mnemonic aid (Williams et al., 2013), particularly in this task which probes memory for spatial locations. It may be that in this case and in the case of the tracking load signal, the classifiers are picking up on changes in the spatial characteristics of the saliency map, such as the distance between salient locations.

How might changes in the spatial characteristics of the saliency map explain the pattern of generalization we see in Experiment 1? We see a transient generalization to the tracking load signal from classifiers trained on set size 2 trials with different locations and set size 4 trials with same locations, but not from classifiers trained on same location trials and different location trials within set size 4. First, the saliency map effects may have been lower overall in Experiment 1, where location was less important, and the repetition of locations may discourage the use of spatial information as a mnemonic aid. Second it is possible that the difference in saliency maps

compounds across presentations. Under this assumption, the set size 4 same location and set size 4 different location trials have only begin to differ in the second presentation, while set size 2 different trials differ from set size 4 same trials at each presentation. These differences compounding may produce a strong enough signal to be detectable, and generalize to the tracking task. In line with this, there was a small cluster of timepoints following the first presentation of the stimuli in the 2 Diff vs 4 Same model which did generalize to the tracking task, but the significance of the cluster was only trending ($p=.076$). Thus, it may be that the tracking task relies exclusively on motion inferred from the saliency map, linking it to certain aspects of spatial attention in our working memory tasks. While we acknowledge that this interpretation is post-hoc, future work can compare the tracking load signals generated by the featureless object tracking task to signals generated by more traditional tracking tasks and to salience maps produced by the different conditions of our working memory tasks (Harrison, Thayer & Sprague, 2026) to assess whether they rely on different mechanisms.

Another noteworthy feature of this work is the dissociation between the load signals identified and changes in difficulty or effort across conditions. Participants' performances dropped with increases in set size in both tasks, and their reported effort increased. Therefore, if these load signals reflected effort, we would predict generalization between the WM load signal and the tracking load signal, which is not what we found. Instead, the strongest generalization from the tracking load signal was to the signal separating SS2 Adj and SS2 Near cloud conditions in Experiment 2, for which participants did not significantly differ in performance or effort ratings. This, combined with the lack of generalization between tasks, suggests that the load signals reflect the specific increase in certain cognitive processes employed in service of performing the tasks at hand, rather than a more domain general process.

Limitations

In addition to the non-traditional tracking task, this work also employed two non-traditional WM tasks. As noted above in the context of the featureless object tracking task, this introduces the possibility that the tasks are not relying on the same processes as in the traditional paradigms. For example, in the sequential WM task of Experiment 1, participants may recode items to reduce interference at repeated locations, relying on a different mode of storage than those typically used in change detection tasks. In line with this, we found that the putative WM load signal was fairly dynamic, and did not appear to actually isolate set size specific signals until late in the delay period. While these aspects are true for the sequential task, this explanation seems less likely to drive the lack of generalization seen with the dot cloud task in Experiment 2. There, the WM load signal seemed selectively sensitive to the number of clouds presented, regardless of their spatial arrangement, across the delay period. In addition, unpublished work has provided initial evidence that a WM load signal identified via SS1 Large and SS2 Adj conditions generalizes to predict set size in a more traditional change detection task, similar to version used in Thyer et al., (2022) and Jones, Thyer et al., (2024). Thus it appears likely that, at least in Experiment 2, our WM task was engaging the same mechanisms as traditional change detection tasks and serving as a reference point with which to compare WM mechanisms to those engaged in our tracking task.

One additional concern is that the displays between tasks were extremely different. It is possible that additional, sensory-specific signals scale with set size. For example, the act of allocating attention may modulate the gain of neural activity coding for low level stimulus features, producing an interaction between set size and sensory-specific signals (Martínez et al.,

2006; Müller et al., 2006). If there are sensory-specific signals that also scale with set size, these might dwarf content-independent load signals in classifiers trained to decode load, reducing our sensitivity to generalization. Despite this, we still see some generalization across displays. In particular, the generalization between Experiment 2's cloud distance model and the tracking task seems to largely track the timepoints for which both classifiers find significant decoding. This indicates that potential sensory differences aren't enough to dwarf all signals.

Conclusions

This work collected two datasets comparing the tracking of featureless objects to the storage of items in working memory, while controlling for spatial attention. Across the datasets, we did not find reliable generalization between our WM load signals and the load signal for tracking featureless objects, suggesting that they don't rely on the same mechanisms. While it's possible that WM storage doesn't rely on spatiotemporal indexing, an alternative explanation is that the featureless object tracking task relies on a distinct process that infers motion over changes in salience maps, preventing our ability to test the spatiotemporal indexing hypothesis. Future work that contrasts featureless object tracking against traditional tracking tasks while potentially manipulating salience of locations, could provide insight into the underlying mechanisms of attentional tracking and WM storage, and whether they rely on spatiotemporal indices or other mechanisms to enable selective access and manipulation.

General Discussion

This work provides two major findings: First, voluntarily storing items in working memory is distinct from voluntarily attending to the original locations of the items (Chapters 1-2). Indeed, this form of spatial attention can be sustained without anything being stored in working memory at all (Chapter 3), prompting a refinement of our definition of working memory above and beyond the simple maintenance of information. Second, WM storage mechanisms are applied across vastly different types of information and track behavioral capacity limits (Chapters 2 and 3). This supports theories suggesting that WM storage relies on the abstract binding of items to a representation of the current event to enable selective maintenance and access to these items. I discuss each of these in turn.

How does the relationship between spatial attention and working memory impact our definition of working memory?

Previous work has found evidence of a tight link between spatial attention and working memory. For example, spatial attention is sustained at the location of items encoded into WM even when location is irrelevant to the task (Foster et al., 2017), and disruption of spatial attention also disrupts working memory performance (Awh et al., 1998; Williams et al., 2013). These results are consistent with theories that spatial attention can act as part of the rehearsal mechanism for working memory (Awh & Jonides, 2001) or that visual working memory may be the internally directed form of visual attention (Chun, 2011).

In contrast, Chapters 1 and 2 found that the WM load signal was dissociable from signals tracking the allocation of spatial attention, even in spatial working memory tasks such as Experiment 2 of Chapter 1. This is aligned with previous dissociations between neural signals

associated with spatial attention and working memory storage (Fukuda et al., 2015; Bae & Luck, 2018; Günseli et al., 2019; Hakim et al., 2021; Diaz et al., 2019). Thus, we must consider the distinct roles that spatial attention and working memory may serve in the service of selecting and maintaining access to pertinent information.

In the discussion of Chapter 1, I considered 3 potential, not mutually exclusive roles for spatial attention. First, space may act as a key organizing dimension in the building and subsequent maintenance of object representations, serving as a prerequisite (Abrahamse et al., 2014; Pertzov & Husain, 2014; Schneegans & Bays, 2017; Treisman & Zhang, 2006). Second, the allocation of spatial attention may improve the signal-to-noise ratio of early sensory processing (Martínez et al., 2006), shaping the fidelity of representations available for working memory storage. Third, spatial attention may be a specific instance of feature-based attention, particularly in tasks where spatial location and envelope are the task-relevant features, such as Experiment 2 of Chapter 1 (Serences et al., 2009; Woodman & Vogel, 2008). As noted above, these possibilities are not mutually exclusive, and future work is required to isolate each possibility's contribution to performance on WM and attentional tasks.

While Chapters 1 and 2 found that the allocation of spatial attention can be independently manipulated in WM paradigms, Chapter 3 found that spatial attention can be allocated in the *absence* of WM storage. This result replicates Hakim et al. (2019) while relying on more sensitive multivariate measures and extending the range of set sizes to include supracapacity arrays. This is noteworthy because the spatial attention task in Chapter 3 requires the active maintenance of information, location, in the absence of external cues, for participants to successfully identify and make a judgment about the relevant digit. These requirements match standard definitions of working memory. For example, Foster, Vogel & Awh (2022) define

working memory as “an online memory system that allows us to hold information ‘in mind’ in service of ongoing cognitive processing” and Bays et al. (2024), state of the field of WM that “[t]he dominant notion was that visual WM holds internal copies of visual objects or features, which can be directly accessed for judgement or decision-making at a later point in time.” Thus, spatial attention could be considered a form of working memory for locations. However, the additional signal associated specifically with WM storage in the change detection tasks of Chapter 3 argues that spatial attention is distinct from the standard forms of WM storage commonly studied.

Here, I propose defining WM as the severely capacity-limited cognitive system that enables active maintenance of information at the item level. This definition emphasizes two critical components of WM. First, it highlights the stark capacity limits seen in behavioral and neural paradigms (Balaban, Fukuda & Luria, 2019; Cowan et al., 2001; Oberauer et al., 2018; Chapter 3), and for which individual differences are associated with differences in critical life outcomes (Mashburn et al., 2023; Fukuda et al., 2010; Cowan et al., 2005). Second, aligned with the results of Chapter 2 and previous work showing behavioral and neural consistency across visual features including multifeature items (Luck & Vogel, 1997; Vogel, Woodman & Luck, 2001; Woodman & Vogel, 2008; Luria & Vogel, 2011; Rajsic, Burton & Woodman, 2019; Thyer et al., 2022; Chang et al., 2026), I propose that WM mechanisms apply at the item, or “chunk” level, and are therefore consistent across features. This contrasts with spatial attention, which can only be used to support the maintenance of location information. Indeed, to the extent that spatial attention aids working memory (e.g., Williams et al., 2013), it may do so in a manner akin to subvocal rehearsal by enabling feature-specific maintenance in a manner that increases the overall effective capacity of the system (Baddeley & Hitch, 1974; Chen & Cowan, 2009).

What mechanisms underlie storage in working memory?

Chapters 1-3 identify the content-independent, capacity-limited nature of WM storage, emphasized in our refined definition of working memory above. However, what exact mechanisms underlie WM storage have not been determined. Here, I consider possible mechanisms, including the spatiotemporal indexing account assessed in Chapter 4.

As mentioned throughout the preceding chapters, a large class of theories proposes that a major component of WM is its ability to bind items to the current episode to maintain selective access to those items (Oberauer & Lin, 2017; Hedayati et al., 2022; Awh & Vogel, 2025). These theories are driven by links between long-term memory research, which focuses on item-context bindings as the mechanism underlying episodic memory (Eacott and Gaffan, 2005; Davachi, 2006; Ranganath, 2010; Knierim et al., 2006; Libby et al., 2020), and the attentional tracking literature, which relies on indexing operations to preserve item identity across space and time (Kahneman, Treisman & Gibbs, 1992; Pylyshyn, 1989). While the exact implementation of these ideas varies from theory to theory, the key insight is that having an index divorced from the actual content itself allows observers to maintain a coherent sense of the object, even as low-level features consistently change throughout the ongoing scene. Theories of working memory that incorporate these indexing mechanisms often leverage them as a bottleneck on the system, explaining both the content-independent and capacity-limited nature of the WM storage signal.

Chapter 4 was designed as a specific test of the "pointer theory," an instance of this indexing or binding hypothesis (Thyer et al., 2022; Awh & Vogel, 2025). The pointer theory largely builds off the work of Pylyshyn (1989; 2009), and proposes that items in working memory are selectively tagged with spatiotemporal indices. These indices enable selective access

to the items, allowing participants to compare them to newly presented items or perform other manipulations like updating their feature representations, depending on the task demands. In Chapter 4, I attempted to isolate this spatiotemporal indexing process through the newly developed featureless object paradigm (Bai & Scholl, in prep). I found that the signal produced in this paradigm was distinct from WM storage, providing initial evidence against spatiotemporal indices as the mechanism underlying the WM storage signal. However, we also found evidence that some aspects of spatial attention may have been shared between the tasks, suggesting that participants may not have needed spatiotemporal indices to perform the featureless object tracking task. Instead, participants may rely on "third-order" motion detection mechanisms (Lu & Sperling, 1995; Lu & Sperling, 1996). Under this hypothesis, participants might be able to infer motion by attending to certain features to mark locations on a salience map and subsequently tracking the evolution of these locations across the map. Future work is planned to compare the WM storage signal to more traditional tracking paradigms.

In addition to the pointer theory, other indexing or binding theories have been proposed that instantiate the core mechanism in distinct ways, requiring additional experimentation. For example, Wyble and colleagues (Swan & Wyble, 2014; Hedayati et al., 2022) have proposed the presence of abstract *tokens* that are associated with specific feature values, such as color and orientation, through an intermediate set of neural units (the binding pool) via coactivation. At test, activating a given token leads to a cascade of activity from the token, back to the relevant feature values, enabling comparison or other manipulations. A third theory is Oberauer & Lin's (2017; 2024) interference theory, which relies on binding context and feature information to create more representations that are more discriminable from one another, facilitating accurate retrieval. Notably, the interference model does not meaningfully distinguish between the

representation of context and the representation of additional features, though it often treats location specifically as the context. Instead, objects are effectively defined as the concatenation of vectors representing different feature dimensions, including a context representation vector. If an object lacks a feature value, that section of its representation vector is simply set to some null value. While neural signatures of these proposed mechanisms have not been directly examined with respect to these theories, work doing so would enable a more direct comparison with the WM storage signal. For example, experimental paradigms that modulate the probability of successful item-context binding may produce signals overlapping with those that change with the number of items held in working memory.

The above theories are largely motivated by behavioral paradigms. As a result, it can be difficult to operationalize some of the key mechanisms in a manner that allows for the study of their neural signatures. In contrast, neurally inspired mechanisms have also been proposed that may underlie a seemingly content-independent, capacity-limited storage mechanism. One possibility, which would appear to conflict with the apparent content-independent nature of the storage signal, is that the signal is tracking the number of compressed representations being concurrently represented in the prefrontal cortex (Christophel et al., 2017; Bancroft et al., 2014). Under this theory, the PFC can be engaged to represent simplified versions of the maintained stimulus, particularly for tasks where simplified representations are sufficient (Bancroft et al., 2014). In support of this theory, Seabra et al. (2026) found that representations of color were primarily aligned with categorical boundaries in the superior precentral sulcus, while early visual regions tracked more continuous changes across hue values. All of our studies examining the content-independence of the WM storage signal used stimulus sets with few items, potentially enabling the use of simplified representations in the PFC. Computational models of working

memory have also implicated the representations PFC as a core, flexible substrate supporting WM maintenance. For example, Bouchacourt & Buschman (2019) found that completely random connections between simulated PFC neurons and content-specific units produced both the inheritance of weak neural tuning curves and capacity-limiting interference between representations within the PFC cells. Relatedly, O'Reilly, Frank, and colleagues (O'Reilly & Frank, 2006; Hazy et al., 2007; Soni & Frank, 2025) have proposed a model of WM storage that relies on the duplication of representations in PFC "stripes" to sustain representations over time and enable selective access. Though these theories motivate the presence of WM representations in the PFC through different mechanisms, they are all consistent with the possibility that stimuli in our experiments are also represented in the PFC, regardless of the specific content being maintained. Future work must examine whether stimuli of the kind studied here are reliably represented in frontal regions, and if so, whether they are presented in overlapping neural subspaces, or whether some separation is preserved.

Duplicated representations across the cortex are also consistent with a recently proposed "spatial computing" theory of working memory dynamics (Lundqvist et al., 2023). This theory proposes that item representations are moved across the cortex, existing at different spatially meaningful locations depending on factors such as the order of encoding. Such an organization would allow for top-down control by targeting spatially meaningful regions (such as the position associated with the first item in a sequence) to select a specific item, rather than relying on precise neuron-to-neuron mapping. This account predicts consistent regions of cortex to be engaged across content types, which could also produce consistent storage signals as well, potentially at both the sources and the sites of control. However, this theory is still largely

untested, and the details of how representations are duplicated into the initial spatially meaningful positions is underspecified.

Finally, there are theories of WM storage that focus on the synchrony of feature representations, entrained to ongoing neural oscillations (e.g., Lisman & Idiart, 1995; Mongillo et al., 2008; Taher, Torcini & Olmi, 2020). Specifically, these theories propose that an item's features are selectively reactivated over range of phases of some underlying, low-frequency oscillation (e.g., theta), keeping the item relatively active and accessible. This mechanism may be content-independent in that multiple features of a given item could be entrained in the same phase subspace, and it may be capacity-limited, assuming that items need some minimum amount of the phase subspace to remain coherent and accessible. While these models are largely untested in humans (but see Bahramisharif et al., 2018), future work can assess whether high-level features of this model, such as the overall frequency or the size of phase subspace spanned by individual items, change systematically with the number of items stored, consistent with the WM storage signal.

In summary, there is a set of cognitive and neural mechanisms that have been proposed that may align with the content-independent, capacity-limited signal that I have observed in this work. The broad class of cognitively motivated theories focuses on abstract indexing or binding operations, while the neural theories are more mechanistically variable, but have common themes such as duplicated representations in the PFC. It's worth stating that certain instantiations of these neural mechanisms might serve as instances of the binding or indexing operations. For example, the phase subspaces of the oscillation models, or the PFC stripes of O'Reilly and Frank's model, may serve as the flexible tokens in Wyble's binding pool model (Swan & Wyble, 2014; Hedayati et al., 2022). In addition, the act of synchronization of features within a phase

subspace in the oscillation models may act as the binding mechanism of the interference model (Oberauer & Lin, 2017). Integration of these different classes of theories and their models may provide additional insights, along with a testing bed to directly compare their various predictions and the likelihood that they underlie WM storage.

Conclusions

Working memory is a critical cognitive process. The ability to actively hold information in mind, even in the absence of external inputs, allows us to make flexible decisions and adapt to ever-changing situations. In this work, I leveraged a neural signature of working memory storage as a lens into its core components. The attributes of working memory storage revealed through this approach - that working memory storage relies on content-independent, yet capacity-limited mechanisms - allow us to refine our understanding of working memory and focus on testing the class of theories that are compatible with these attributes. Ultimately, researchers will likely need to integrate cognitive and neural theories to account for behavioral and neural signatures of working memory, including the storage signal, and to allow us to understand humans' ability to successfully navigate everyday life.

Appendix 1: Chapter 1 Supplementary Figures

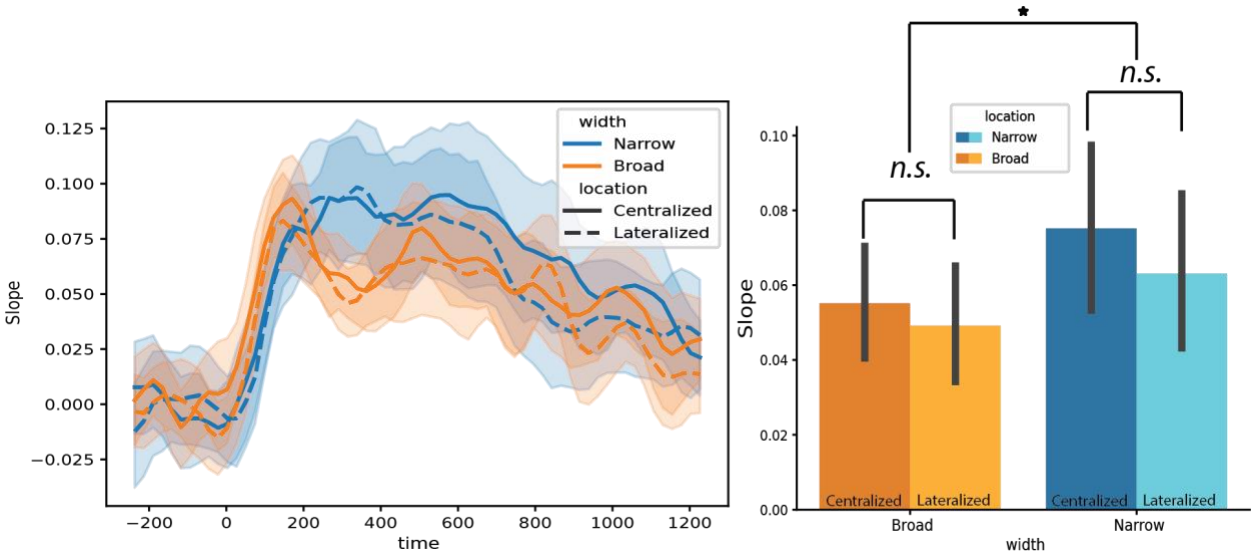

Figure A1. Examining the slope of Set Size 1 Narrow and Broad clouds, split by the location of the clouds. While narrow clouds produced significantly higher slopes than broad clouds, slopes were not significantly impacted by the location of the clouds, nor was there a significant interaction. This indicates that the differences between narrow and broad clouds were not driven by broad clouds straddling visual hemifields in some locations.

Appendix 2: Chapter 2 Supplementary Figures

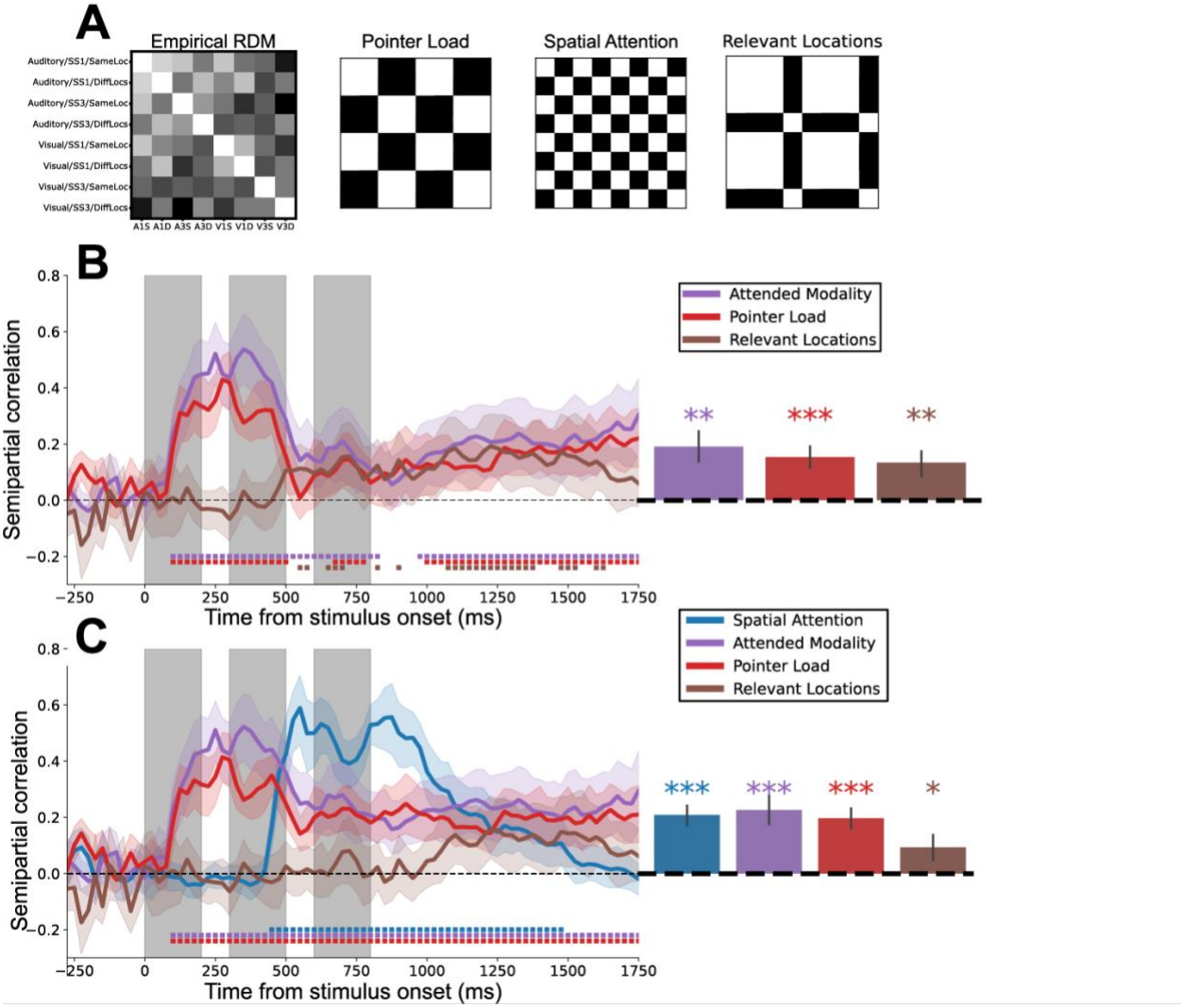

Figure A2. Experiment 2 RSA with inclusion of relevant locations model (n=14), Related to Figure 30. The relevant locations model predicted that spatial attention would only be allocated to targets, not distractors — i.e. all conditions except set size 3 different location trials have only 1 location. A: All RDMs used in this figure. B: RSA results with only relevant locations model. Left: Semipartial correlations of each factor over time. Stim period denoted by gray shaded region, significant timepoints (wilcoxcon signed-rank test, FDR corrected), denoted by colored boxes under the graph. Right: average correlation over delay period (t>800 ms). Attended

Modality: W=119, p = 3.1 x 10^{-3}; Pointer Load: W = 126, p = 6.6 x 10^{-4}; Relevant Locations: W = 117, p = 4.5 x 10^{-3} . C: RSA results with both relevant locations and target/distractor agnostic "spatial attention" model. Delay average p-values: Spatial Attention: W=135, p = 3.1 x 10^{-5} Attended Modality: W=128, p = 3.8 x 10^{-4}; Pointer Load: W = 133, p = 7.6 x 10^{-5}; Relevant Locations: W = 105, p = 0.029.

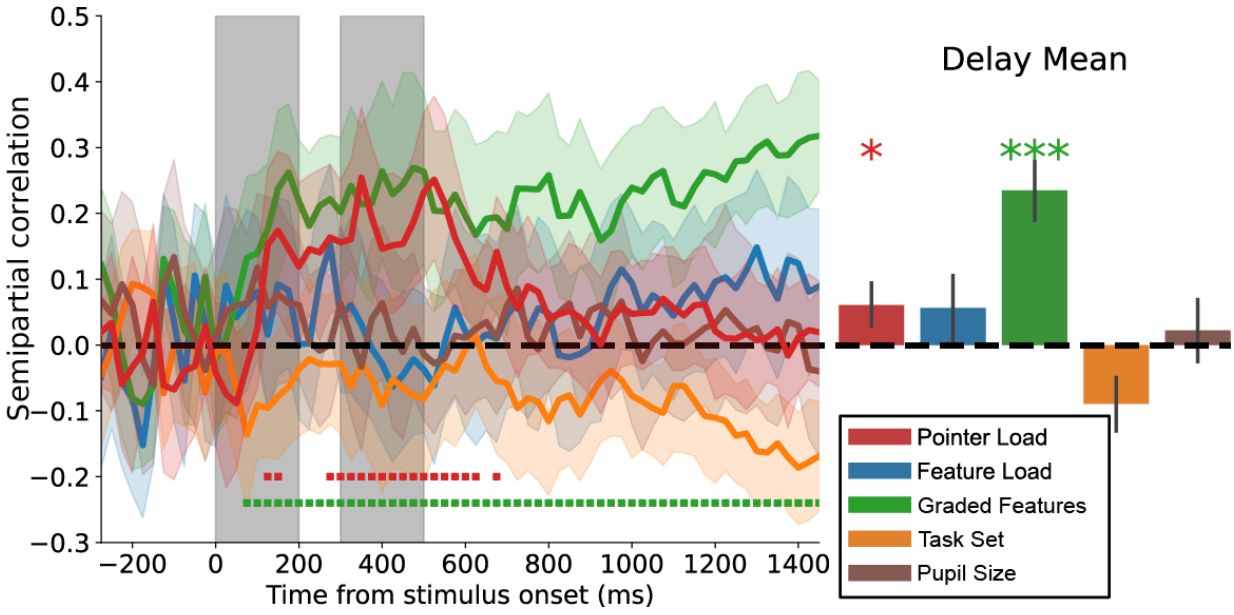

*Figure A3. RSA results for experiment 1 with inclusion of pupil regressor (n=19), Related to Figure 28. Pupil size model was the mean pupil size over the delay period per subject per condition. Line graph: Semipartial correlations of each factor to the empirical RDM over time. Stim period denoted by gray shaded region, significant timepoints (wilcoxcon signed-rank test, FDR corrected), denoted by colored boxes under the graph. Bar Graph: Semipartial correlations averaged across the delay period (t>500 ms). *** = p < 0.001, ** = p < 0.01, * = p < 0.05 (wilcoxcon signed-rank test, uncorrected).*

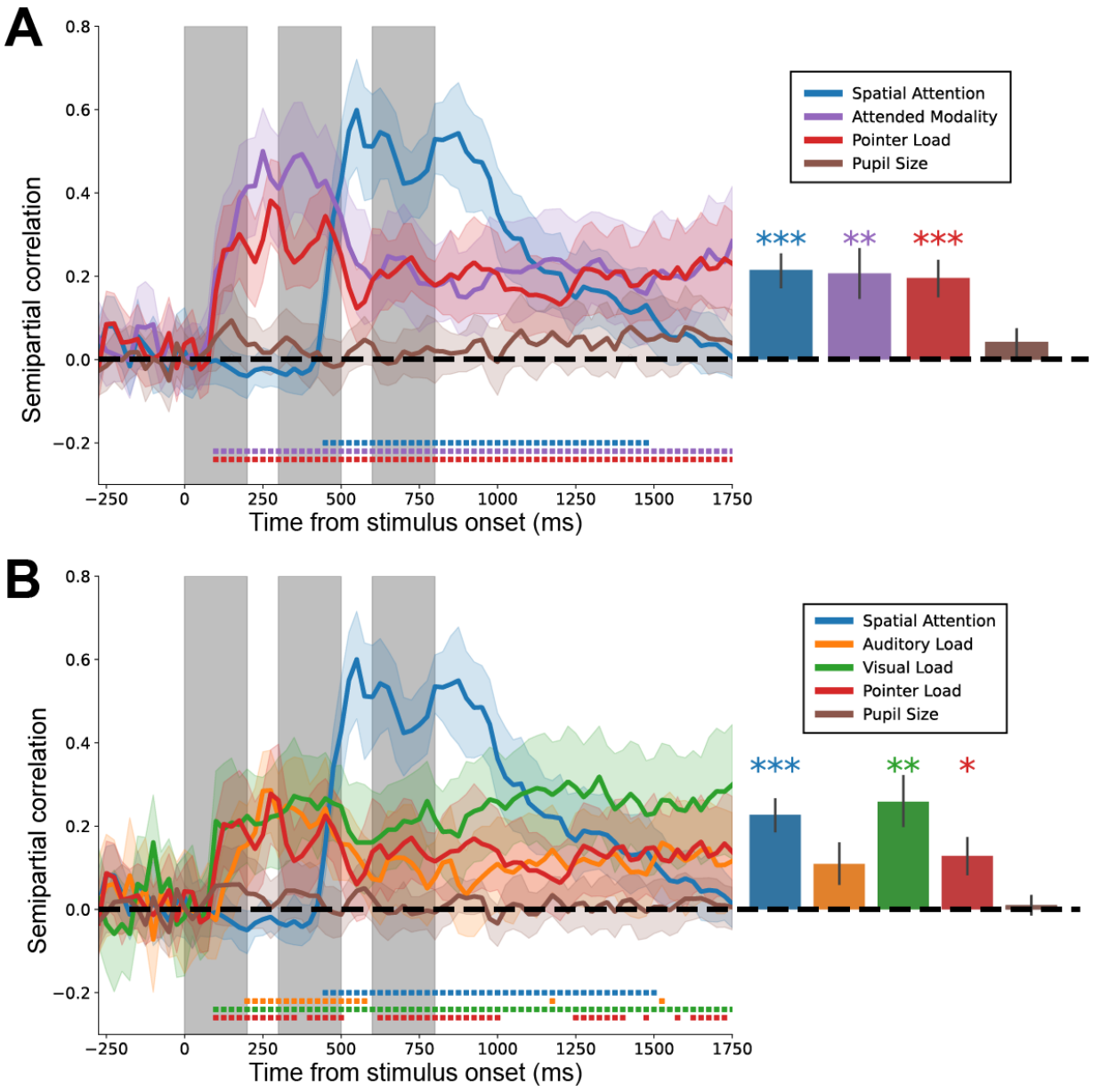

Figure A4. RSA results for experiment 2 with inclusion of pupil regressor (n=14), Related to Figure 30. Pupil size model was the mean pupil size over the delay period per subject per condition. Top 2: with spatial attention, attended modality, and pointer load (corresponds to Figure 30B) Bottom 2: with spatial attention, auditory/visual load, and pointer load (corresponds to Figure 30C) Line graphs: Semipartial correlations of each factor to the

*empirical RDM over time. Stim period denoted by gray shaded region, significant timepoints (wilcoxcon signed-rank test, FDR corrected), denoted by colored boxes under the graph. Bar graphs: Semipartial correlations averaged across the delay period (t>500 ms). *** = p < 0.001, ** = p < 0.01, * = p < 0.05 (wilcoxcon signed-rank test, uncorrected).*

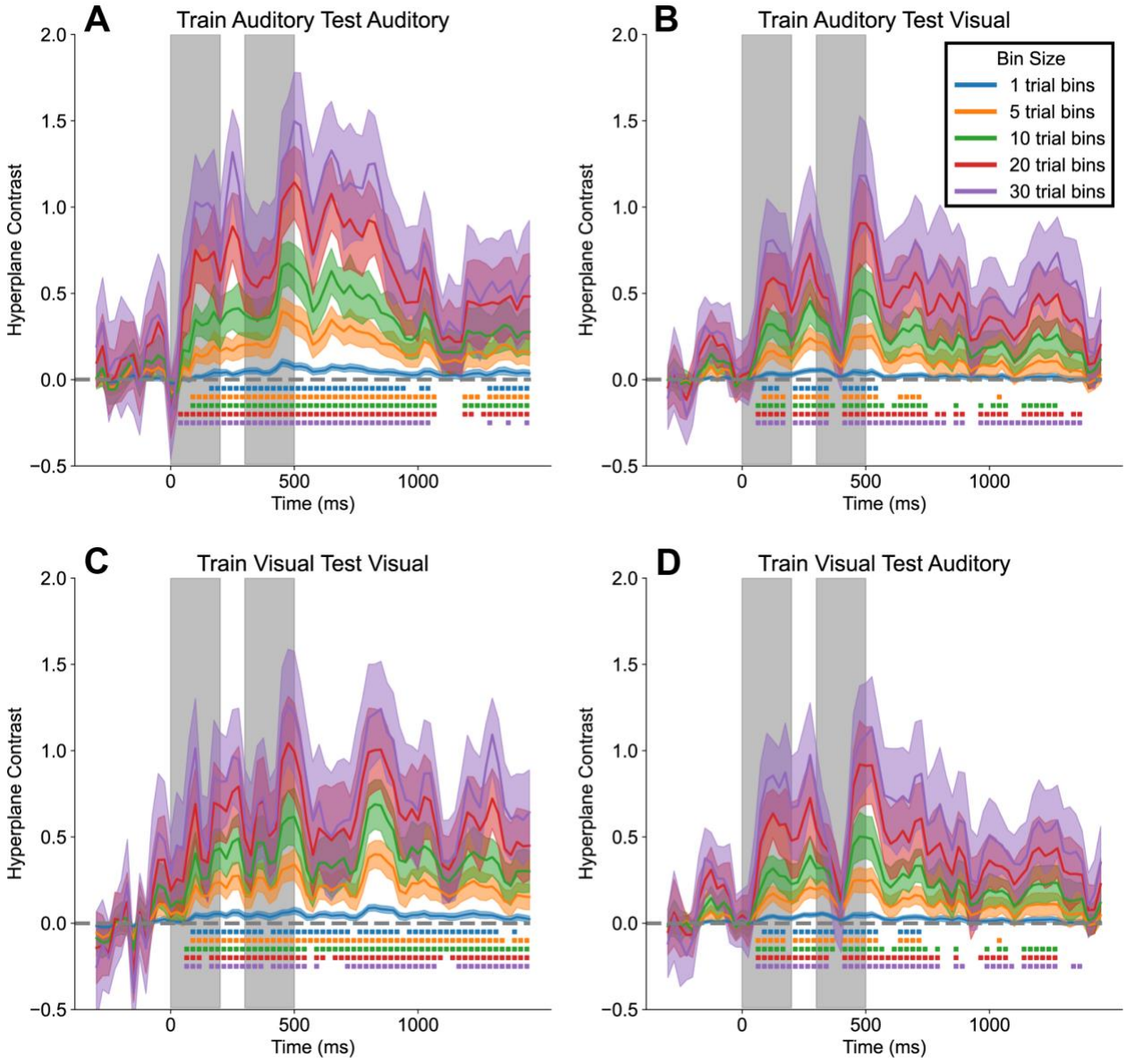

Figure A5. Downsampling results for comparisons in 26A-B with different "miniblock" trial bin sizes, Related to Methods (Decoding analyses) and Figure 26. A-B correspond to blue (train/test

within modality) and orange (train/test across modality) curves in Figure 26A, respectively. C-D are same but for Figure 26B. As in Figure 26, stimulus period is shown in by gray vertical bars, lines denote mean hyperplane contrast, shaded region denotes SEM, and squares under each timepoint indicate contrast values greater than zero (FDR-corrected one-tailed t-test).

Appendix 3: Chapter 3 Eye Feature Analysis

Eye features track spatial attention and working memory, but do not explain EEG signals

The previous results found that we can decode set size from EEG data in both the digits task and the change detection task. We next asked whether this was driven by changes in eye features such as movements and pupil size. Previous work has found that both spatial attention and working memory impact pupil size independently (Koevoet, Jones et al., 2026). Thus, we wondered whether the decoding results above were driven by subtle changes in the eyes that produced changes in the electrical signals detected by EEG.

We first reran the above analyses on the 16 eye features (X position, Y position, Euclidean distance, pupil size, and temporal derivatives for each eye). We found similar results, with a few notable differences (Figure A6).

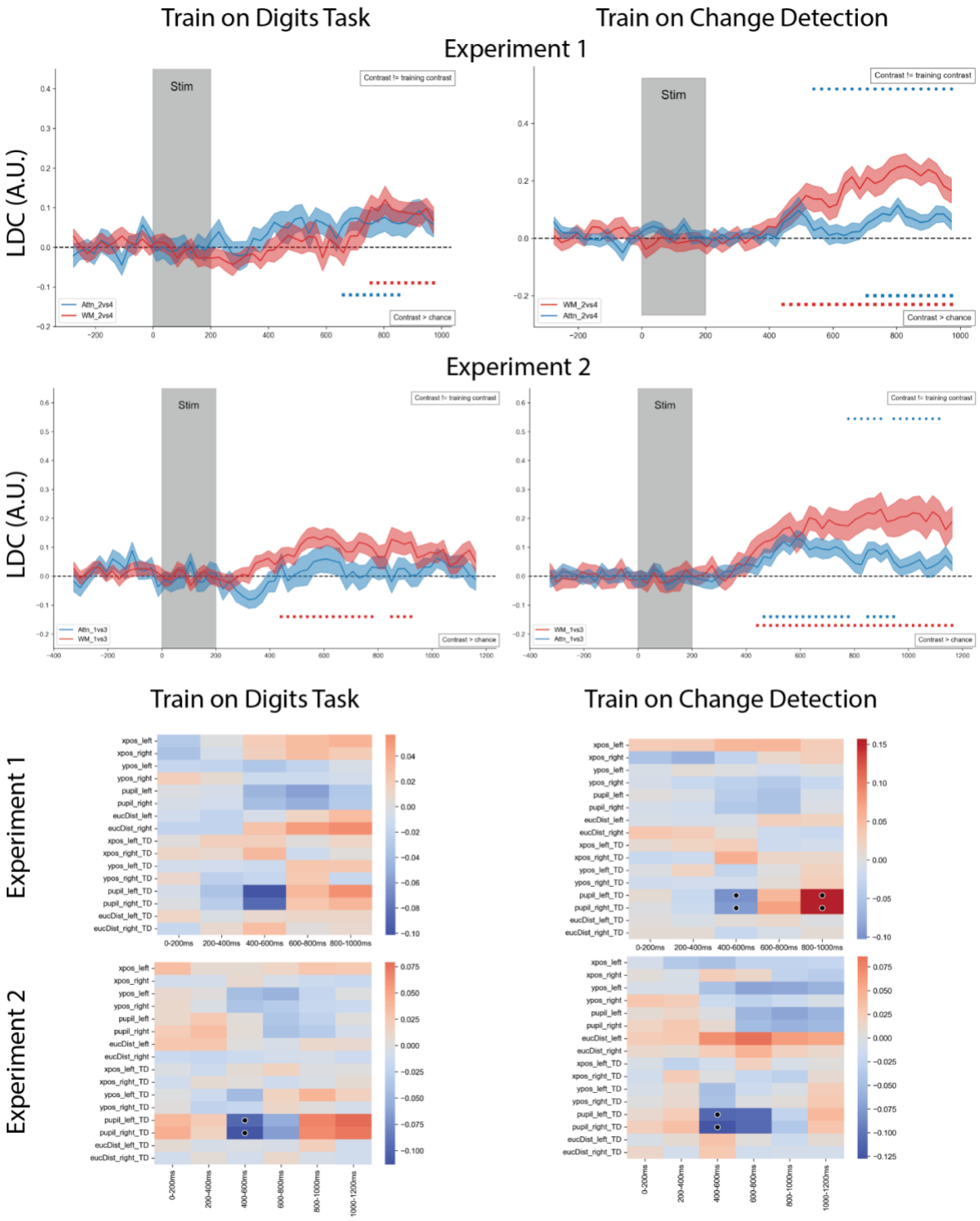

Figure A6. Hyperplane contrasts for decoding of set size in both experiments based on eye features. Left column shows the contrasts obtained by training on the digits task and testing on either task. Right column shows contrasts obtained by training the model on the change

detection task and testing on either task. Shaded regions show the standard error of the mean. Bottom squares indicate significant clusters of time points in which the contrasts are greater than chance. Top dots indicate significant clusters of time points in which the contrast for the non-trained-on task is different from the contrast of the trained-on task. Below the timeseries plots are the Haufe pattern identified by the model in 200ms time windows. Black and white dots indicate time windows that are significantly different from 0, following Bonferroni correction.

In Experiment 1, when the model was trained on the digits task, the only reliable timepoints were late in the delay period, with a single significant cluster between ~650 and ~850ms for data from the digits task, and between ~750 and ~950ms in the change detection task, and no reliable differences between the two task's decodability timeseries. In contrast, when a model was trained on data from the change detection task, it identified timepoints of significant decoding in both tasks in larger clusters (digits: ~700-1000ms; change detection: ~450-1000ms), and there was a significant cluster of timepoints for which decodability was higher for the change detection task than the digits task (~500-1000ms).

In Experiment 2, the patterns were similar, though the overall time points with significant decoding appeared earlier in the delay period. Importantly, when training on the digits task, the classifier could not decode set size in the digits task, but it could decode set size in the change detection task. When training on the change detection task, the classifier could decode set size in both tasks, and the hyperplane contrast was significantly larger in the change detection task again.

While these patterns of results, particularly for Experiment 1, are similar to the decoding results seen in EEG, the larger cluster sizes when testing the change detection task compare to the digits task, regardless of the training conditions, is also consistent with asymmetries in the signal-to-noise ratio (SNR) between the 2 tasks. Previous work has found that the impact of SNR varies depending on the overall signal strength and the ratio of the differences (van den Hurk and Op de Beeck, 2019; Koevoet, Voet et al., 2026), but a common result is that training on the high SNR improves the model's sensitivity, leading to an improvement in decoding for all conditions that appears to be a rescaling of decoding timecourses (Koevoet, Voet et al., 2026). This account assumes that the same features are changing with set size in both tasks, but to different extents.

Aligned with this difference in SNR, we find that for both task models, the only significant deflections are in the temporal derivative of the pupils (Figure A6). In three of the four models, we see negative values reflecting constricting with spatial attention and WM load in the 400-600ms window. The fourth model, trained on the digits task in Experiment 1, also shows this constriction, but it is not significant. For the change detection task Experiment 1, we also see dilation later in the trial period, in the 800-1000ms, though this isn't replicated in Experiment 2. Thus, it appears that while the EEG-based decoding results above were driven by spatially separable signals, both spatial attention and WM load laregly impact a single factor, pupil size.

Finally, we assessed whether EEG signals were linked to eye-based signals at the trial level. For each model, we assessed whether the EEG-based trial predictions of the training conditions could be explained by the EYE-based predictions, above and beyond a simple condition label differentiating the 2 set sizes. To improve SNR and to include the possibility that EEG signals reflect motor preparation before eye movements, we averaged over time. For EEG predictions, we averaged across the whole delay period, excluding the first and last 100ms, and

for eye predictions, we averaged across the second half of the delay period, also excluding the last 100ms. We chose the second half of the delay period for the eye predictions as this was the window in which eye decoding was most consistently significant across all of the models in both experiments.

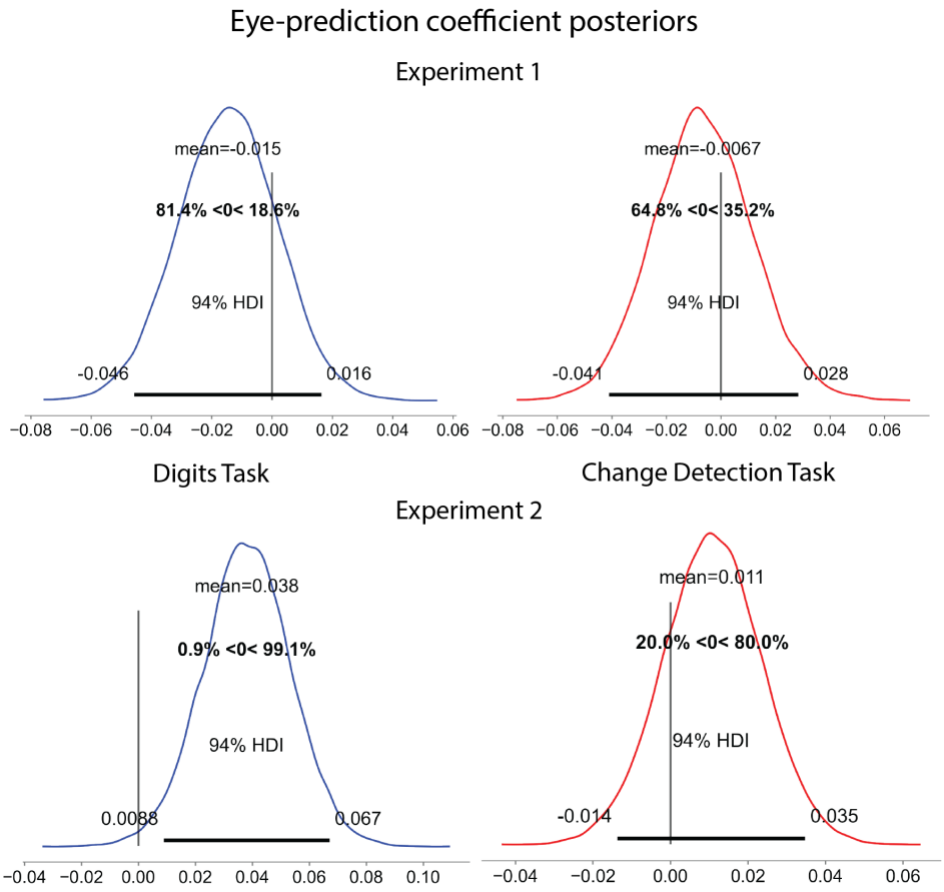

Figure A7. Posteriors of coefficients linking EEG-based predictions and eye-based predictions. The left column shows the results for models trained on the digits task, while the right column shows the results for models trained on the change detection task. Vertical lines indicate a value 0, while horizontal lines reflect the 94% highest-density interval for the posterior.

Our results did not provide evidence that trial-level eye movements were reliably associated with trial-level EEG signals (Figure A7). We tested 4 models, a digits-task model and

a change-detection-task model in both experiments. Only the digits-task model in Experiment 2 produced a posterior estimate of the association between EEG predictions and eye predictions whose 94% highest density interval (HDI) excluded 0 (mean=0.038, HDI=[.0075, .071]). The posterior for all other models included 0 (Experiment 2 change detection: mean=.0092, HDI=[-.016, .033]; Experiment 1 digits: mean=-.083, HDI=[-.04, .024]; Experiment 1 change detection: mean=-0.11, HDI=[-.045, .024]). Indeed, the means of the posteriors for Experiment 1 were actually negative, suggesting that if anything, eye-based predictions of higher set sizes were associated with EEG predictions of lower set sizes.

Experiment	Conditions	Rank	Name	ELPD difference from best model: Mean (SD)	Weight (%)
1	Digits: 2vs4	**1**	**condition-only**	**-**	**89.74**
		2	condition&eyes	1.29 (0.99)	0
		3	eyes-only	3.865 (3.11)	10.26
	Change Detection: 2vs4	**1**	**condition-only**	**-**	**97.78**
		2	condition&eyes	1.01 (1.31)	0
		3	eyes-only	72.78 (12.096)	2.22
2	Digits: 1vs3	**1**	**condition&eyes**	**-**	**72.15**
		2	condition-only	2.62 (3.11)	22.84
		3	eyes-only	79.23 (13.23)	5.01
	Change Detection: 1vs3	**1**	**condition-only**	**-**	**97.28**
		2	condition&eyes	1.58 (1.25)	0
		3	eyes-only	440.552 (28.99)	2.72

Table A1. Model comparison results for explaining EEG-based predictions. The first two columns specify the experiment and the conditions of the trials being modeled. The third and fourth columns show the rank of the 3 models, and their model names. "Condition-only" refers to a model that only contains a condition contrast as a factor for explaining the EEG-

predictions. "Eyes-only" refers to a model that contains only the eye-based trial level predictions for explaining the EEG-predictions. "condition&eyes" refers to a model which contains both factors. The fifth columns show the mean difference and standard deviation in expected pointwise predictive density (ELPD) between the best model and the row's model. The sixth column shows the relative weight for each model.

We also performed model comparison, comparing two variants of the full model: only with only a predictor based on eyes ("eyes-only"), and a model which only had a predictor based on the condition labels ("condition-only"). See Table A1 for a summary of the results For every task in both experiments, a model including the condition contrast was the highest ranked via leave-one-out-cross-validation (LOOCV), and the highest weights (72.15-97.28%). For both tasks in Experiment 1, as well as the change detection task in Experiment 2, the model containing only a condition factor was the highest ranked and weighted (89.74-97.78%). For the digits task in Experiment 2, the highest ranked and weighted (72.15%) model included both the condition label and the eye-based predictions. Finally, for every task in both experiments, the model using only eye-based predictions as a factor was the lowest ranked, and consistently received very low weightings (2.22-10.26%).

Taken together, it appears unlikely that eye movements were the major driver of EEG-based signals or that the EEG signals reflected preparation of subsequent eye movements.

References

Abrahamse, E., van Dijck, J.-P., Majerus, S., & Fias, W. (2014). Finding the answer in space: The mental white-board hypothesis on serial order in working memory. *Frontiers in Human Neuroscience*, *8*, Article 932. https:// doi.org/10.3389/fnhum.2014.00932

Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C. J., Kochurov, M., ... & Zinkov, R. (2023). PyMC: a modern, and comprehensive probabilistic programming framework in Python. PeerJ Computer Science, 9, e1516.

Adam, K.C.S., Vogel, E.K., and Awh, E. (2017). Clear evidence for item limits in visual working memory. Cogn. Psychol. 97, 79–97.

Adam, K. C. S., Vogel, E. K., & Awh, E. (2020). Multivariate analysis reveals a generalizable human electrophysiological signature of working memory load. *Psychophysiology*, *57*(12), Article e13691.

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in Cognitive Sciences*, *5*(3), 119–126.

Awh, E., Jonides, J., & Reuter-Lorenz, P. A. (1998). Rehearsal in spatial working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 780.

Awh, E., Vogel, E. K., & Oh, S. H. (2006). Interactions between attention and working memory. *Neuroscience*, *139*(1), 201–208.

Awh, E., and Vogel, E.K. (2025). Working memory needs pointers. Trends Cogn. Sci. 29, 230–241.

Baddeley, A. D. (1986). Working memory oxford. England: Oxford Uni.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? Trends Cogn. Sci. 4, 417–423.

Baddeley, A.D., and Hitch, G. (1974). Working Memory. In Psychology of Learning and Motivation (Elsevier), pp. 47–89.

Bae, G. Y., & Luck, S. J. (2018). Dissociable decoding of spatial attention and working memory from EEG oscillations and sustained potentials. *Journal of Neuroscience*, *38*(2), 409–422.

Bahramisharif, A., Jensen, O., Jacobs, J., & Lisman, J. (2018). Serial representation of items during working memory maintenance at letter-selective cortical sites. PLoS biology, 16(8), e2003805.

Balaban H, Drew T, Luria R (2019) Neural evidence for an object-based pointer system underlying working memory. Cortex 119:362–372.

Balaban, H., Fukuda, K., & Luria, R. (2019). What can half a million change detection trials tell us about visual working memory?. Cognition, 191, 103984.

Bancroft, T. D., Hockley, W. E., & Servos, P. (2014). Does stimulus complexity determine whether working memory storage relies on prefrontal or sensory cortex?. Attention, Perception, & Psychophysics, 76(7), 1954-1961.

Bays, P. M., Schneegans, S., Ma, W. J., & Brady, T. F. (2024). Representation and computation in visual working memory. Nature human behaviour, 8(6), 1016-1034.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. B Methodol. 57, 289–300.

Bisley, J. W., & Goldberg, M. E. (2003). Neuronal activity in the lateral intraparietal area and spatial attention. *Science*, *299*(5603), 81–86.

Bouchacourt F, Buschman TJ (2019) A flexible model of working memory. Neuron 103:147–160.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/ 156856897X00357

Chang W, Epstein W, Jones, HM, Ngiam WXQ, Awh, E; Content-Independent Pointers Mediate Working Memory Storage for Both Visual and Verbal Stimuli. *J Cogn Neurosci* 2026; doi: https://doi.org/10.1162/JOCN.a.2522

Chen, Z., & Cowan, N. (2009). Core verbal working-memory capacity: The limit in words retained without covert articulation. Quarterly journal of experimental psychology, 62(7), 1420-1429.

Chen H, Wyble B (2016) Attribute amnesia reflects a lack of memory consolidation for attended information. J Exp Psychol Hum Percept Perform 42: 225–234.

Chota S, Leto C, van Zantwijk L, Van der Stigchel S (2022) Attention rhythmically samples multi-feature objects in working memory. Sci Rep 12: 14703.

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J. D. (2017). The distributed nature of working memory. Trends in cognitive sciences, 21(2), 111-124.

Chun, M. M. (2011). Visual working memory as visual attention sustained internally over time. *Neuropsychologia*, *49*(6), 1407–1409.

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 109–127.

Cocchini, G., Logie, R.H., Della Sala, S.D., MacPherson, S.E., and Baddeley, A.D. (2002). Concurrent performance of two memory tasks: Evidence for domain-specific working memory systems. Mem. Cognit. 30, 1086–1095.

Cowan, N. (1999). An Embedded-Processes Model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and brain sciences, 24(1), 87-114.

Cowan, N., Elliott, E.M., Scott Saults, J., Morey, C.C., Mattox, S., Hismjatullina, A., and Conway, A.R.A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. Cogn. Psychol. 51, 42–100.

Cowan, N., Saults, J.S., and Blume, C.L. (2014). Central and peripheral components of working memory storage. J. Exp. Psychol. Gen. 143, 1806–1836.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21.

D'Esposito M, Postle BR (2015) The cognitive neuroscience of working memory. Annu Rev Psychol 66:115–142.

Diaz, G. K., Vogel, E. K., & Awh, E. (2021). Perceptual grouping reveals distinct roles for sustained slow wave activity and alpha oscillations in working memory. *Journal of Cognitive Neuroscience*, *33*(7), 1354–1364.

Diedrichsen, J., Berlot, E., Mur, M., Schütt, H. H., Shahbazi, M., & Kriegeskorte, N. (2020). Comparing representational geometries using whitened unbiased-distance-matrix similarity. arXiv preprint arXiv:2007.02789.

Diedrichsen, J., Provost, S., and Zareamoghaddam, H. (2016). On the distribution of cross-validated Mahalanobis distances. Preprint at arXiv.

Drew, T., Horowitz, T. S., Wolfe, J. M., & Vogel, E. K. (2012). Neural measures of dynamic changes in attentive tracking load. Journal of Cognitive Neuroscience, 24(2), 440-450.

Drew, T., Horowitz, T. S., Wolfe, J. M., & Vogel, E. K. (2011). Delineating the neural signatures of tracking spatial position and working memory during attentive tracking. Journal of Neuroscience, 31(2), 659-668.

Emrich SM, Riggall AC, Larocque JJ, Postle BR (2013) Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. J Neurosci 33:6516–6523.

Ester EF, Anderson DE, Serences JT, Awh E (2013) A neural measure of precision in visual working memory. J Cogn Neurosci 25:754–761.

Feldmann-Wüstefeld, T., & Awh, E. (2020). Alpha-band activity tracks the zoom lens of attention. *Journal of Cognitive Neuroscience*, *32*(2), 272–282.

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1:1–47.

Fiebelkorn IC, Saalmann YB, Kastner S (2013) Rhythmic sampling within and between objects despite sustained attention at a cued location. Curr Biol 23:2553–2558.

Fiebelkorn, I. C., Pinsk, M. A., & Kastner, S. (2019). The medio-dorsal pulvinar coordinates the macaque fronto-parietal network during rhythmic spatial attention. *Nature Communications*, *10*(1), Article 215.

Foster, J. J., Sutterer, D. W., Serences, J. T., Vogel, E. K., & Awh, E. (2017). Alpha-band oscillations enable spatially and temporally resolved tracking of covert spatial attention. *Psychological Science*, *28*(7), 929–941.

Foster, J. J., Vogel, E. K., & Awh, E. (2024). 13 Working memory as persistent neural activity. In The Oxford Handbook of Human Memory, Two Volume Pack: Foundations and Applications, online edn (pp. 353-370). Academic, Oxford.

Fougnie, D., and Marois, R. (2011). What limits working memory capacity? Evidence for modality-specific sources to the simultaneous storage of visual and auditory arrays. J. Exp. Psychol. Learn. Mem. Cogn. 37, 1329–1341.

Fougnie, D., Zughni, S., Godwin, D., and Marois, R. (2015). Working memory storage is intrinsically domain specific. J. Exp. Psychol. Gen. 144, 30–47.

Fukuda, K., Mance, I., & Vogel, E. K. (2015). α power modulation and event-related slow wave provide dissociable correlates of visual working memory. *Journal of Neuroscience*, *35*(41), 14009–14016.

Fukuda, K., and Vogel, E.K. (2019). Visual short-term memory capacity predicts the ‘‘bandwidth’’ of visual long-term memory encoding. Mem. Cognit. 47, 1481–1497.

Fukuda K, Vogel E, Mayr U, Awh E (2010) Quantity, not quality: the relationship between fluid intelligence and working memory capacity. Psychon Bull Rev 17:673–679.

Fukuda, K., Woodman, G. F., & Vogel, E. K. (2015). Individual differences in visual working memory capacity: Contributions of attentional control to storage. Mechanisms of sensory working memory: Attention and performance XXV, 105-119.

Funahashi S, Bruce CJ, Goldman-Rakic PS (1989) Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. J Neurophysiol 61: 331–349.

Fuster JM, Alexander GE (1971) Neuron activity related to short-term memory. Science 173:652–654.

Fuster, J.M., and Jervey, J.P. (1981). Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. Science 212, 952–955.

Gardner, W.G., and Martin, K.D. (1995). HRTF measurements of a KEMAR. J. Acoust. Soc. Am. 97, 3907–3908

Gaspelin, N., & Luck, S. J. (2018). The role of inhibition in avoiding distraction by salient stimuli. *Trends in Cognitive Sciences*, *22*(1), 79–92.

Gazzaley, A., and Nobre, A.C. (2012). Top-down modulation: bridging selective attention and working memory. Trends Cogn. Sci. 16, 129–135.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., … Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*, Article 267. https://doi.org/10.3389/ fnins.2013.00267

Goldman-Rakic PS (1995) Cellular basis of working memory. Neuron 14: 477–485.

Guimond, S., Vachon, F., Nolden, S., Lefebvre, C., Grimault, S., and Jolicoeur, P. (2011). Electrophysiological correlates of the maintenance of the representation of pitch objects in acoustic short-term memory. Psychophysiology 48, 1500–1509.

Günseli, E., Fahrenfort, J. J., van Moorselaar, D., Daoultzis, K. C., Meeter, M., & Olivers, C. N. (2019). EEG dynamics reveal a dissociation between storage and selective attention within working memory. *Scientific Reports*, *9*(1), Article 13499. https://doi.org/10.1038/s41598-019-49577-0.

Günseli, E., Foster, J., Sutterer, D., Todorova, L., Vogel, E. K., & Awh, E. (2022). Overlapping neural representations for dynamic visual imagery and stationary storage in spatial working memory. *BioRxiv*, 2022-09. https://doi.org/10.1101/2022.09.24.509255

Hakim, N., Adam, K. C. S., Gunseli, E., Awh, E., & Vogel, E. K. (2019). Dissecting the neural focus of attention reveals distinct processes for spatial attention and object-based storage in visual working memory. *Psychological Science*, *30*(4), 526–540.

Hakim, N., Feldmann-Wüstefeld, T., Awh, E., & Vogel, E. K. (2021). Controlling the flow of distracting information in working memory. *Cerebral Cortex*, *31*(7), 3323–3337.

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. Psychological Science, 17(7), 572–576.

Harrison SA, Tong F (2009) Decoding reveals the contents of visual working memory in early visual areas. Nature 458:632–635.

Harrison, A. H., Thayer, D. D., & Sprague, T. C. (2026). Neural priority maps encode behavioral relevance independently across multiple attended locations. bioRxiv, 2026-04.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage, 87, 96-110.

Hedayati S, O'Donnell RE, Wyble B (2022) A model of working memory for latent representations. Nat Hum Behav 6:709–719.

Henderson, L. (1972). Visual and verbal codes: Spatial information survives the icon. Q. J. Exp. Psychol. 24, 439–447.

Hillyard, S. A., Vogel, E. K., & Luck, S. J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: Electrophysiological and neuroimaging evidence. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *353*(1373), 1257–1270.

Holcombe AO, Chen W-Y (2013) Splitting attention reduces temporal resolution from 7Hz for tracking one object to <3Hz when tracking three. J Vis 13:12.

Iman, R. L., & Conover, W. J. (1979). The use of the rank transform in regression. *Technometrics*, *21*(4), 499–509.

Jones HM, Diaz GK, Ngiam WXQ, Awh E (2024) Electroencephalogram decoding reveals distinct processes for directing spatial attention and encoding into working memory. Psychol Sci 35:1108–1138.

Jones, H.M., Thyer, W.S., Suplica, D., and Awh, E. (2024). Cortically Disparate Visual Features Evoke Content-Independent Load Signals during Storage in Working Memory. J. Neurosci. 44, e0448242024.

Kahneman D, Treisman A, Gibbs BJ (1992) The reviewing of object files: object-specific integration of information. Cogn Psychol 24: 175–219.

Kanwisher, N.G. (1987). Repetition blindness: Type recognition without token individuation. Cognition 27, 117–143.Kanwisher, N.G. (1987). Repetition blindness: Type recognition without token individuation. Cognition 27, 117–143.

Kardan, O., Adam, K.C.S., Mance, I., Churchill, N.W., Vogel, E.K., and Berman, M.G. (2020). Distinguishing cognitive effort and working memory load using scale-invariance and alpha suppression in EEG. Neuroimage 211, 116622.

Karnath, H. O., & Rorden, C. (2012). The anatomy of spatial neglect. *Neuropsychologia*, *50*(6), 1010–1017.

Keene, P.A., deBettencourt, M.T., Awh, E., and Vogel, E.K. (2022). Pupillometry signatures of sustained attention and working memory. Atten. Percept. Psychophys. 84, 2472–2482.

Kiat, J. E., Hayes, T. R., Henderson, J. M., & Luck, S. J. (2022). Rapid extraction of the spatial distribution of physical saliency and semantic informativeness from natural scenes in the human brain. *Journal of Neuroscience*, *42*(1), 97–108.

Kikumoto, A., and Mayr, U. (2020). Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. Proc. Natl. Acad. Sci. USA 117, 10603–10608.

Kitzbichler, M.G., Henson, R.N.A., Smith, M.L., Nathan, P.J., and Bullmore, E.T. (2011). Cognitive Effort Drives Workspace Configuration of Human Brain Functional Networks. J. Neurosci. 31, 8259–8270.

Koevoet, D., Jones, H. M., Van der Stigchel, S., & Awh, E. (2026). Dissociating Spatial Attention and Working Memory Storage with Pupillometry. Journal of Cognitive Neuroscience, 38(4), 656-668.

Koevoet D, Strauch C, Van der Stigchel S, Mathôt S, Naber M (2023) Revealing visual working memory operations with pupillometry: encoding, maintenance, and prioritization. Wiley Interdiscip Rev Cogn Sci 15:e1668.

Koevoet, D., Voet, V., Jones, H. M., Awh, E., Strauch, C., & Van der Stigchel, S. (2026). Population-Level Activity Dissociates Preparatory Overt from Covert Attention. Journal of Neuroscience, 46(2).

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, Article 4. doi: 10.3389/neuro.06.004.2008 Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411.

Landau AN, Fries P (2012) Attention samples stimuli rhythmically. Curr Biol 22:1000–1004.

Ledoit O, Wolf M (2004) A well-conditioned estimator for large-dimensional covariance matrices. J Multivar Anal 88:365–411.

Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., and Postle, B.R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. J. Cogn. Neurosci. 24, 61–79.

Libby, A., and Buschman, T.J. (2021). Rotational dynamics reduce interference between sensory and memory representations. Nat. Neurosci. 24, 715–726.

Lisman, J. E., & Idiart, M. A. (1995). Storage of 7±2 short-term memories in oscillatory subcycles. Science, 267(5203), 1512-1515.

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, Article 213. https://doi.org/10.3389/fnhum.2014.00213

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Wild, A., Knight, P., Rouder, J. N., Morey, R. D., & Wagenmakers, E. J. (2019). JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software*, *88*, 1–17.

Lu, Z. L., & Sperling, G. (1995). The functional architecture of human visual motion perception. Vision research, 35(19), 2697-2722.

Lu, Z. L., & Sperling, G. (1996). Three systems for visual motion perception. Current Directions in Psychological Science, 5(2), 44-53.

Luck, S. J., Vogel, E. K., & Shapiro, K. L. (1996). Word meanings can be accessed but not reported during the attentional blink. *Nature*, *383*(6601), 616–618.

Luck SJ, Vogel EK (1997) The capacity of visual working memory for features and conjunctions. Nature 390:279–281.

Lundqvist, M., Brincat, S. L., Rose, J., Warden, M. R., Buschman, T. J., Miller, E. K., & Herman, P. (2023). Working memory control dynamics follow principles of spatial computing. Nature Communications, 14(1), 1429.

Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience & Biobehavioral Reviews*, *62*, 100–108.

Luria, R., & Vogel, E. K. (2011). Shape and color conjunction stimuli are represented as bound objects in visual working memory. Neuropsychologia, 49(6), 1632-1639.

Majerus, S., Attout, L., D'Argembeau, A., Degueldre, C., Fias, W., Maquet, P., Martinez Perez, T., Stawarczyk, D., Salmon, E., Van der Linden, M., et al. (2012). Attention Supports Verbal Short-Term Memory via Competition between Dorsal and Ventral Attention Networks. Cereb. Cortex 22, 1086–1097.

Majerus, S., Cowan, N., Pe´ ters, F., Van Calster, L., Phillips, C., and Schrouff, J. (2016). Cross-Modal Decoding of Neural Patterns Associated with Working Memory: Evidence for Attention-Based Accounts of Working Memory. Cereb. Cortex 26, 166–179.

Mangun, G. R., & Hillyard, S. A. (1991). Modulations of sensory-evoked brain potentials indicate changes in per-ceptual processing during visual-spatial priming. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(4), 1057–1074.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

Martin, R.A. (2021). PyPortfolioOpt: portfolio optimization in Python. J. Open Source Software 6, 3066.

Martínez, A., Anllo-Vento, L., Sereno, M. I., Frank, L. R., Buxton, R. B., Dubowitz, D. J., Wong, E. C., Hinrichs, H., Heinze, H. J. & Hillyard, S. A. (1999). Involvement of striate and extrastriate visual cortical areas in spatial attention. *Nature Neuroscience*, *2*(4), 364–369.

Martínez, A., Teder-Sälejärvi, W., Vazquez, M., Molholm, S., Foxe, J. J., Javitt, D. C., Di Russo, F., Worden, M. S., & Hillyard, S. A. (2006). Objects are highlighted by spatial attention. *Journal of Cognitive Neuroscience*, *18*(2), 298–310.

Mashburn, C. A., Burgoyne, A. P., & Engle, R. W. (2023). Working memory, intelligence, and life success. Memory in science for society: There is nothing as practical as a good theory, 149-184.

Master, S.L., Li, S., and Curtis, C.E. (2024). Trying Harder: How Cognitive Effort Sculpts Neural Representations during Working Memory. J. Neurosci. 44, e0060242024.

Mongillo G, Barak O, Tsodyks M (2008) Synaptic theory of working memory. Science 319:1543–1546.

Müller, M. M., Andersen, S., Trujillo, N. J., Valdes-Sosa, P., Malinowski, P., & Hillyard, S. A. (2006). Feature-selective attention enhances color signals in early visual areas of the human brain. *Proceedings of the National Academy of Sciences, USA*, *103*(38), 14250–14254.

Naber M, Frässle S, Rutishauser U, Einhäuser W (2013) Pupil size signals novelty and predicts later retrieval success for declarative memories of natural scenes. J Vis 13:11.

Naghavi, H.R., and Nyberg, L. (2005). Common fronto-parietal activity in attention, memory, and consciousness: Shared demands on integration? Conscious. Cogn. 14, 390–425.

Ngiam, W.X.Q., Foster, J.J., Adam, K.C.S., and Awh, E. (2023). Distinguishing guesses from fuzzy memories: Further evidence for item limits in visual working memory. Atten. Percept. Psychophys. 85, 1695–1709.

Ngiam WX, Loetscher KB, Awh E (2023) Object-based encoding constrains storage in visual working memory. J Exp Psychol Gen 153:86–101.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. PLoS Comp. Biol. 10, e1003553.

Oberauer K (2019) Working memory capacity limits memory for bindings. J Cogn 2:40.

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D., Conway, A., Cowan, N., ... & Ward, G. (2018). Benchmarks for models of short-term and working memory. Psychological bulletin, 144(9), 885.

Oberauer K, Lin HY (2017) An interference model of visual working memory. Psychol Rev 124:21.

Oberauer, K., and Lin, H.-Y. (2024). An interference model for visual and verbal working memory. J. Exp. Psychol. Learn. Mem. Cogn. 50, 858–888.

Olson IR, Jiang Y (2002) Is visual short-term memory object based? Rejection of the "strong-object" hypothesis. Percept Psychophys 64:1055–1067.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 1–9.

O'Reilly, R.C., Ranganath, C., and Russin, J.L. (2022). The Structure of Systematicity in the Brain. Curr. Dir. Psychol. Sci. 31, 124–130.

Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature*, *592*(7855), 601–605.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*, 195–203.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Pertzov, Y., & Husain, M. (2014). The privileged role of loca-tion in visual working memory. *Attention, Perception, & Psychophysics*, *76*, 1914–1924.

Pylyshyn Z (1989) The role of location indexes in spatial perception: a sketch of the FINST spatial-index model. Cognition 32:65–97.

Pylyshyn ZW (2009) Perception, representation, and the world: The FINST that binds. In: Computation, cognition, and Pylyshyn (Dedrick D, Trick L, eds), pp 3–48. Cambridge, MA: MIT Press.

Rademaker RL, Chunharas C, Serences JT (2019) Coexisting representations of sensory and mnemonic information in human visual cortex. Nat Neurosci 22:1336–1344.

Rajan, A., Meyyappan, S., Liu, Y., Samuel, I.B.H., Nandi, B., Mangun, G.R., and Ding, M. (2021). The Microstructure of Attentional Control in the Dorsal Attention Network. J. Cogn. Neurosci. 33, 965–983.

Rajsic, J., Burton, J.A., and Woodman, G.F. (2019). Contralateral delay activity tracks the storage of visually presented letters and words. Psychophysiology 56, e13282.

Rizza, A., Pedale, T., Mastroberardino, S., Olivetti Belardinelli, M., Van der Lubbe, R.H.J., Spence, C., and Santangelo, V. (2024). Working Memory Maintenance of Visual and Auditory Spatial Information Relies on Supramodal Neural Codes in the Dorsal Frontoparietal Cortex. Brain Sci. 14, 123.

Sandhaeger F, Siegel M (2023) Testing the generalization of neural representations. Neuroimage 278:120258.

Saults, J.S., and Cowan, N. (2007). A central capacity limit to the simulta-neous storage of visual and auditory arrays in working memory. J. Exp. Psychol. Gen. 136, 663–684.

Schneegans, S., & Bays, P. M. (2017). Neural architecture for feature binding in visual working memory. *Journal of Neuroscience*, *37*, 3913–3925.

Schurgin, M.W., Wixted, J.T., and Brady, T.F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. Nat. Hum. Behav. 4, 1156–1172.

Seabra, J. P., Gui, A. M., Chopurian, V., Souza, A. S., Allefeld, C., & Christophel, T. B. (2025). Visuospatial working memory is cortically enabled through veridical, categorical and semantic representations. bioRxiv, 2025-08.

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science*, *20*(2), 207–214.

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring latent visual working memory representations in human cortex. *Neuron*, *91*(3), 694–707.

Sutterer, D. W., Foster, J. J., Adam, K. C. S., Vogel, E. K., & Awh, E. (2019). Item-specific delay activity demonstrates concurrent storage of multiple active neural representations in working memory. *PLOS Biology*, *17*(4), Article e3000239. https://doi.org/10.1371/journal.pbio.3000239

Swan G, Wyble B (2014) The binding pool: a model of shared neural resources for distinct items in visual working memory. Atten Percept Psychophys 76:2136–2157.

Taher, H., Torcini, A., & Olmi, S. (2020). Exact neural mass model for synaptic-based working memory. PLOS Computational Biology, 16(12), e1008533.

Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research*, *76*, 31–42.

Todd JJ, Marois R (2004) Capacity limit of visual short-term memory in human posterior parietal cortex. Nature 428:751–754.

Treisman, A., & Zhang, W. (2006). Location and binding in visual working memory. *Memory & Cognition*, *34*, 1704–1719.

Thyer, W., Adam, K. C. S., Diaz, G. K., Velázquez Sánchez,I. N., Vogel, E. K., & Awh, E. (2022). Storage in visual working memory recruits a content-independent pointer system. *Psychological Science*, *33*(10), 1680–1694.

Uddin, S., Heald, S.L.M., Van Hedger, S.C., Klos, S., and Nusbaum, H.C. (2018). Understanding environmental sounds in sentence context. Cognition 172, 134–143.

Uddin, S., Heald, S.L.M., Van Hedger, S.C., and Nusbaum, H.C. (2018). Hearing sounds as words: Neural responses to environmental sounds in the context of fluent speech. Brain Lang. 179, 51–61.

Unsworth, N., Fukuda, K., Awh, E., and Vogel, E.K. (2014). Working mem-ory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. Cogn. Psychol. 71, 1–26.

van den Hurk, J., & Op de Beeck, H. P. (2019). Generalization asymmetry in multivariate cross-classification: When representation A generalizes better to representation B than B to A. BioRxiv, 592410.

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E. J. (2020). Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's r. *Journal of Applied Statistics*, *47*(16), 2984–3006.

Vaina LM (1994) Functional segregation of color and motion processing in the human visual cortex: clinical evidence. Cereb Cortex 4:555–572.

Vallat, R. (2018). Pingouin: statistics in Python. J. Open Source Software 3, 1026.

van der Wel P, van Steenbergen H (2018) Pupil dilation as an index of effort in cognitive control tasks: a review. Psychon Bull Rev 25:2005–2015.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272.

Vo, V.A., Sutterer, D.W., Foster, J.J., Sprague, T.C., Awh, E., and Serences, J.T. (2022). Shared Representational Formats for Information Maintained in Working Memory and Information Retrieved from Long-Term Memory. Cereb. Cortex 32, 1077–1092.

Vogel, E. K., Luck, S. J., & Shapiro, K. L. (1998). Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(6), 1656. https://doi.org/10.1037//0096-1523.24.6.1656

Vogel, E. K., & Luck, S. J. (2002). Delayed working memory consolidation during the attentional blink. *Psychonomic Bulletin & Review*, *9*(4), 739–743.

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), 748–751.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. Journal of experimental psychology: human perception and performance, 27(1), 92.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200.

Wang, B., van Driel, J., Ort, E., & Theeuwes, J. (2019). Anticipatory distractor suppression elicited by statistical regularities in visual search. *Journal of Cognitive Neuroscience*, *31*(10), 1535–1548.

Westbrook, A., and Braver, T.S. (2015). Cognitive effort: A neuroeconomic approach. Cogn. Affect. Behav. Neurosci. 15, 395–415.

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. Biometrika, 84(2), 419-427.

Williams, M., Pouget, P., Boucher, L., & Woodman, G. F. (2013). Visual–spatial attention aids the maintenance of object representations in visual working memory. *Memory & Cognition*, *41*, 698–715.

Wimmer K, Nykamp DQ, Constantinidis C, Compte A (2014) Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat Neurosci 17:431–439.

Woodman, G. F., & Vogel, E. K. (2008). Selective storage and maintenance of an object's features in visual working memory. *Psychonomic Bulletin & Review*, *15*(1), 223–229.

Woodman, G. F., Wang, S., Sutterer, D. W., Reinhart, R. M., & Fukuda, K. (2022). Alpha suppression indexes a spotlight of visual-spatial attention that can shine on both perceptual and memory representations. *Psychonomic Bulletin & Review*, *29*(3), 681–698.

Xu Y, Chun MM (2006) Dissociable neural mechanisms supporting visual short-term memory for objects. Nature 440:91–95.

Yonelinas AP (2024) The role of recollection and familiarity in visual working memory: a mixture of threshold and signal detection processes. Psychol Rev 131:321–348.

Yu, Q., and Shim, W.M. (2017). Occipital, parietal, and frontal cortices selectively maintain task-relevant features of multi-feature objects in visual working memory. NeuroImage 157, 97–107.

Zeki SM (1978) Functional specialisation in the visual cortex of the rhesus monkey. Nature 274:423–428.

Zhao, Y., Kuai, S., Zanto, T.P., and Ku, Y. (2020). Neural Correlates Underlying the Precision of Visual Working Memory. Neuroscience 425, 301–311.