Flexible gaze reinstatement during working memory for natural scenes

Yueying Dong*, Yun-chen Hung, Lana Gaspariani, and Anastasia Kiyonaga

Department of Cognitive Science, University of California, San Diego

yud070@ucsd.edu

Abstract

Working memory (WM) representations may be more action-oriented and anatomically widespread than previously assumed. For instance, oculomotor signatures like gaze biases can reflect spatial WM content. However, the specificity and functional relevance of such signatures is unclear. Here, we tracked eye gaze during WM for naturalistic images, and manipulated which image features were most task-relevant (visual or semantic). In two experiments, we found that the eyes carried out item-specific spatiotemporal gaze sequences during WM maintenance, retracing the scanpath from stimulus encoding. Therefore, gaze patterns during WM track the identity of complex, natural images held in mind. Moreover, the degree of such WM gaze reinstatement was stronger when the task prioritized visual detail, and ramped up in anticipation of the test probe. Oculomotor WM signals are therefore malleable to when and how the content will be used, indicating that distributed WM representations are prospectively-oriented and functionally flexible to situational needs.

Keywords: visual working memory; visual attention; eye movements; gaze reinstatement; sensorimotor recruitment

Research transparency statement

General Disclosures

Conflicts of interest: All authors declare no conflicts of interest.
Funding: This work was partially supported by the Air Force Office of Scientific Research (AFOSR) under award number FA9550-22-1-0230 as well as NIH award 1R01EY036843-01
Artificial intelligence: No artificial intelligence assisted technologies were used in this research or the creation of this article.
Ethics: This research received approval from the local Office of IRB Administration (OIA)

Study Disclosures

Preregistration: No aspects of the study were preregistered.
Materials: Some study materials are publicly available (https://github.com/YueyingDong/dong_etal_paper_code/tree/main)
Data: Provisional datasets can be accessed here: https://osf.io/PYEBZ/.
Analysis scripts: Provisional code can be accessed here: https://github.com/YueyingDong/dong_etal_paper_code/tree/main

Introduction

Flexible goal-directed cognition relies on the ability to guide behavior based on task-relevant information that is held 'in mind.' Likewise, decades of research have sought to unveil the hidden mental representations that support this working memory (WM) function (Christophel et al., 2017; Leavitt et al., 2017; Sreenivasan et al., 2014; Xu, 2017). Prominent WM models assert that feature-specific WM content is stored in sensorimotor cortical activations (D'Esposito & Postle, 2015; Serences, 2016), but newer evidence suggests that more peripheral signals outside the brain can also carry WM information. For instance, during a WM delay, small gaze biases veer toward memorized locations, and pupil size tracks the remembered brightness of visual WM stimuli (van Ede et al., 2019; Zokaei et al., 2019). Such ocular WM signatures may provide a valuable window into internal goal states, and into the linkage between WM and motor systems. However, the specific information contained in oculomotor inflections, and their functional sensitivity to task demands, remain unclear.

A growing body of work describes the WM-related information that is reflected in eye movements. During a WM delay, micro-saccadic movements are biased toward the spatial location where a task-relevant visual stimulus was encoded (Liu et al., 2022; Vries & Ede, 2023). This location-specific saccade bias is considered a signature of internal attention focusing, via spatial selection among items held in mind (van Ede et al., 2021). Beyond simple spatial coding, gaze and saccade biases can also track imagery for geometric shapes (Laeng et al., 2014), and the remembered angle of objects in continuous orientation space (Linde-Domingo & Spitzer, 2024). Yet, such tasks may still be considered fundamentally spatial, leaving it unclear whether oculomotor WM signatures reflect mere spatial tagging, or carry more complex feature information.

Long-term memory (LTM) research provides a clue as to the complexity the eyes might convey during WM. Namely, during LTM retrieval, eye movements have been found to recapitulate the spatiotemporal sequences that occurred during encoding (Johansson & Johansson, 2014; Wynn et al., 2019). When recalling a complex natural scene, the eyes may retrace the steps they made when initially viewing that scene. This phenomenon is termed 'gaze reinstatement', and some evidence suggests that it may support retrieval success (Holm & Mäntylä, 2007; Johansson & Johansson, 2014; Wynn et al., 2020). Although the correspondence between WM and LTM remains contested, both may operate over similar visual representations (Cowan, 1999; Fukuda & Woodman, 2017; Schurgin, 2018). In that case, WM may also engage eye movements that represent rich visual detail in their spatio-temporal patterns, rather than time-invariant spatial tagging.

However, the granularity of the oculomotor WM representation may depend on the task at hand. WM is tightly intertwined with action (Boettcher et al., 2021; Miller et al.,

2020; Olivers & Roelfsema, 2020), and the same information may be maintained with different neural representations depending on behavioral demands (Nobre & Stokes, 2019; Trentin et al., 2023, 2024). Indeed, different brain regions or feature representations can be recruited during WM depending on how WM will be tested. For instance, WM maintenance for precise visual features can preferentially engage posterior regions like extrastriate cortex, whereas more categorical maintenance preferentially recruits prefrontal cortex instead (Lee et al., 2013). WM delay activity for the same visuo-spatial sample can also show a more motor-like pattern when an upcoming manual response is known, but a more sensory-like pattern when the response is unknown (Bae & Chen, 2024; Henderson et al., 2022). Thus, if WM is geared to guide near future behavior, the WM representation should match the context of what the task requires. For instance, we might expect richer information to be expressed in the eyes during WM when the task requires precise visual detail vs. a coarse categorical label.

Here, we address two primary questions about WM signatures in peripheral oculomotor activity. First, what information do they convey? Specifically, we ask whether gaze patterns carry item-specific information for natural scene images during WM. Second, are they flexible to prospective task demands? We ask whether WM gaze patterns retrospectively follow the physical properties of encoded stimuli, or whether they prospectively adapt to the relevant features for the task context. To do this, we manipulate (1) whether visual detail or semantic stimulus properties are emphasized at test, and (2) whether that test condition is known before or after stimulus encoding. Analytically, we assess gaze reinstatement by testing whether spatio-temporal eye movement sequences recapitulate from WM encoding to delay. We also use time-resolved representational similarity analyses (RSA) to gauge the stimulus information in gaze patterns across different WM task epochs. The work reveals a sensitive coupling between WM maintenance and motor systems, where the expression of mental content in gaze is strongly yoked to expected task demands.

Methods Overview

Across two eye-tracking experiments, we manipulated how WM content would be used for a subsequent match-to-sample task. In both experiments, participants were asked to remember a naturalistic image over a delay of several seconds, and they were cued as to whether the visual or semantic image properties would be most task relevant (**Figure 1A**). On each trial, a color cue indicated whether the WM probe would test on either the visual detail or the category of the sample image. In **Experiment 1**, the color cue appeared before the WM sample stimulus (**precue** task, n = 41), whereas in **Experiment 2** the cue appeared after the WM sample encoding period (**retrocue** task,

n = 41). In both tasks, a probe image appeared after a blank 3sec delay, and participants judged whether it was a match or non-match to the sample image. On 'visual' trials (green cue), participants saw either a small segment from the sample image (match), or a segment of a different image from the same category (non-match). On 'semantic' trials (purple cue), participants saw either a different image from the same category as the sample (match) or an image from a different category altogether (non-match). For example, when remembering an image of a bear, on a visual trial a match would be cropped from the exact same bear image, while on a semantic trial a match would be an image of a different bear (**Figure 1B**). Thus, to successfully perform the task required either maintenance of precise visual detail or the broader semantic category. Each sample image appeared four times in both the visual and semantic conditions, so that we could compare metrics for the same stimuli across trials and task conditions. We conducted eye-tracking during both experiments, to test whether gaze patterns convey image-specific information about WM content, and whether they are flexible to the WM task demands.

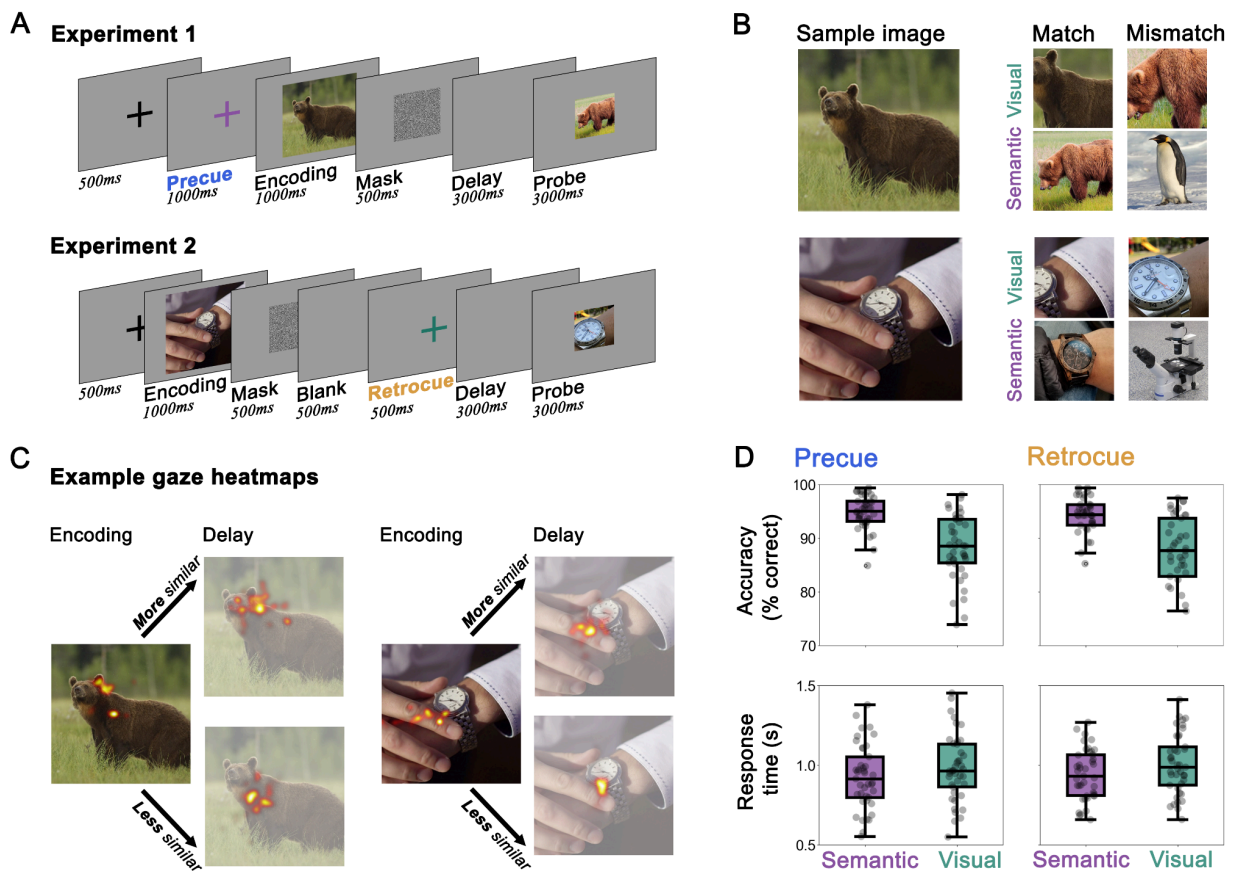

Figure 1. Experiment design and memory performance for both experiments. (**A**) Task schematics for Experiment 1 (precue) and Experiment 2 (retrocue). In both experiments, a color cue indicated whether the probe would test for the visual detail (green) or semantic category

(purple) of the WM sample image. However, in the precue task the cue appeared before encoding, whereas in the retrocue task it appeared after encoding. (**B**) Example WM sample stimuli and memory probes under the different trial conditions. (**C**) Example gaze heatmaps from encoding and delay epochs during WM for select images. The delay epoch examples show trials where the gaze heatmaps would be considered either more similar (top) or less similar (bottom) to the gaze during encoding of the same image. Yellower shades represent a longer dwell time. (**D**) Working memory recognition performance for both experiments and trial conditions. Accuracy reflects % correct and response time reflects the duration between probe onset and button press. Data points reflect individual subject means, the box reflects the interquartile range, and the horizontal line reflects the group median.

Results

Working memory recognition performance

Overall, participants performed the tasks well (mean accuracy: 90.49% correct). But performance was somewhat better on semantic trials (vs. visual; **Figure 1D**). When participants were probed on the semantic category of the WM sample instead of the visual detail, responses were both more accurate (precue: $t(40) = -8.78$, $p < 0.001$, *cohen's d* = -1.38; retrocue: $t(40) = -5.69$, $p < 0.001$, *cohen's d* = -0.86) and faster (precue: $t(40) = 4.78$, $p < 0.001$, *cohen's d* = 0.27; retrocue: $t(40) = 4.89$, $p < 0.001$, *cohen's d* = 0.36), in both pre- and retrocue experiments. However, pre- and retrocue task performance did not differ from each other (*Accuracy*: $t(80) = 1.64$, $p = 0.11$, *cohen's d* = 0.36; *Response time*: $t(80) = -0.48$, $p = 0.63$, *cohen's d* = -0.11). Consistent with previous findings, we observed a semantic task advantage (Brown & and Wesley, 2013; Lee et al., 2013), however, the two different experiments were comparable to each other.

Gaze follows an image-specific scanpath during working memory delays

Our first goal was to assess whether the gaze patterns during a WM delay recapitulated the encoding patterns for the same image. **Figure 1C** illustrates some example gaze heatmaps that vary in similarity to each other from encoding to delay. However, there are many ways to quantify these patterns. To capture both the spatial and temporal properties of the gaze signals, here we first examined scanpaths — or sequences of successive eye movements — between encoding and delay epochs (**Figure 2A** and **B**).

Two prevalent approaches to scanpath analysis can be loosely classified as string-based and vector-based (Coutrot et al., 2018; Fahimi & Bruce, 2021).The

string-based method represents scanpaths as sequences of letters, each corresponding to a region on the monitor screen. Similarity is then quantified between two strings using sequence alignment algorithms (Cristino et al., 2010; Fahimi & Bruce, 2021). This yields a single similarity score for each comparison. The vector-based method measures the similarity across multiple vector geometry features, yielding multiple distinct scores (Dewhurst et al., 2012). Both approaches assess the spatio-temporal information in eye-movement patterns, but they do so using different underlying metrics.

Our first analyses use the string-based approach, assessing similarity between two scanpaths via the **ScanMatch** algorithm (Coutrot et al., 2018). We quantified the similarity between encoding and delay period scanpaths, both within and between every trial, producing a similarity matrix for every participant (**Figure 2C**). We then compared the strength of encoding-delay similarity for three scanpath comparison conditions: within the same trial ('within trial'), between different trials when the same image was remembered ('across trial'), and between different trials when a different image was remembered ('different image'; **Figure 2B** and **C**). If gaze patterns during WM convey image-specific spatio-temporal information, then encoding-delay scanpath similarity should be greater across trials when the same image is remembered, as compared to trials when different images are remembered.

We ran a one-way repeated measures ANOVA on ScanMatch similarity scores, with a 3-level factor of trial scanpath comparison: "within trial"; "across trial"; "different image". In both experiments, we found that scanpaths during the WM delay resembled encoding scanpaths for the same images (**Figure 2D**). Encoding and delay scanpaths were most similar to each other within the same trial, then across trials with the same sample image, followed by different trials with different sample images (precue: $F(2,80) = 46.15$, $p < 0.001$, $\eta_p^2 = 0.54$; retrocue: $F(2,80) = 92.75$, $p < 0.001$, $\eta_p^2 = 0.70$). Follow-up *t*-tests confirmed that 'within trial' similarity scores were greater than 'across trial' (precue: $t(40) = 5.49$, $p < 0.001$, *cohen's d* = 1.17; retrocue: $t(40) = 8.95$, $p < 0.001$, *cohen's d* = 1.93), and 'across trial' scores were greater than 'different image' scores as well (precue: $t(40) = 9.82$, $p < 0.001$, *cohen's d* = 2.22; retrocue: $t(40) = 2.92$, $p = 0.006$, *cohen's d* = 0.59). Thus, shared gaze patterns during WM are not limited to a particular trial, but can carry image-specific information that generalizes over different trials, if the same information is being remembered.

Gaze patterns during a blank WM maintenance period may (at least partially) retrace the unique, image-specific paths that the eyes took during encoding. We will refer to this encoding-delay similarity as 'WM gaze reinstatement', akin to the encoding-retrieval gaze patterns that have been observed during long-term memory (Johansson & Johansson, 2014; Wynn et al., 2019).

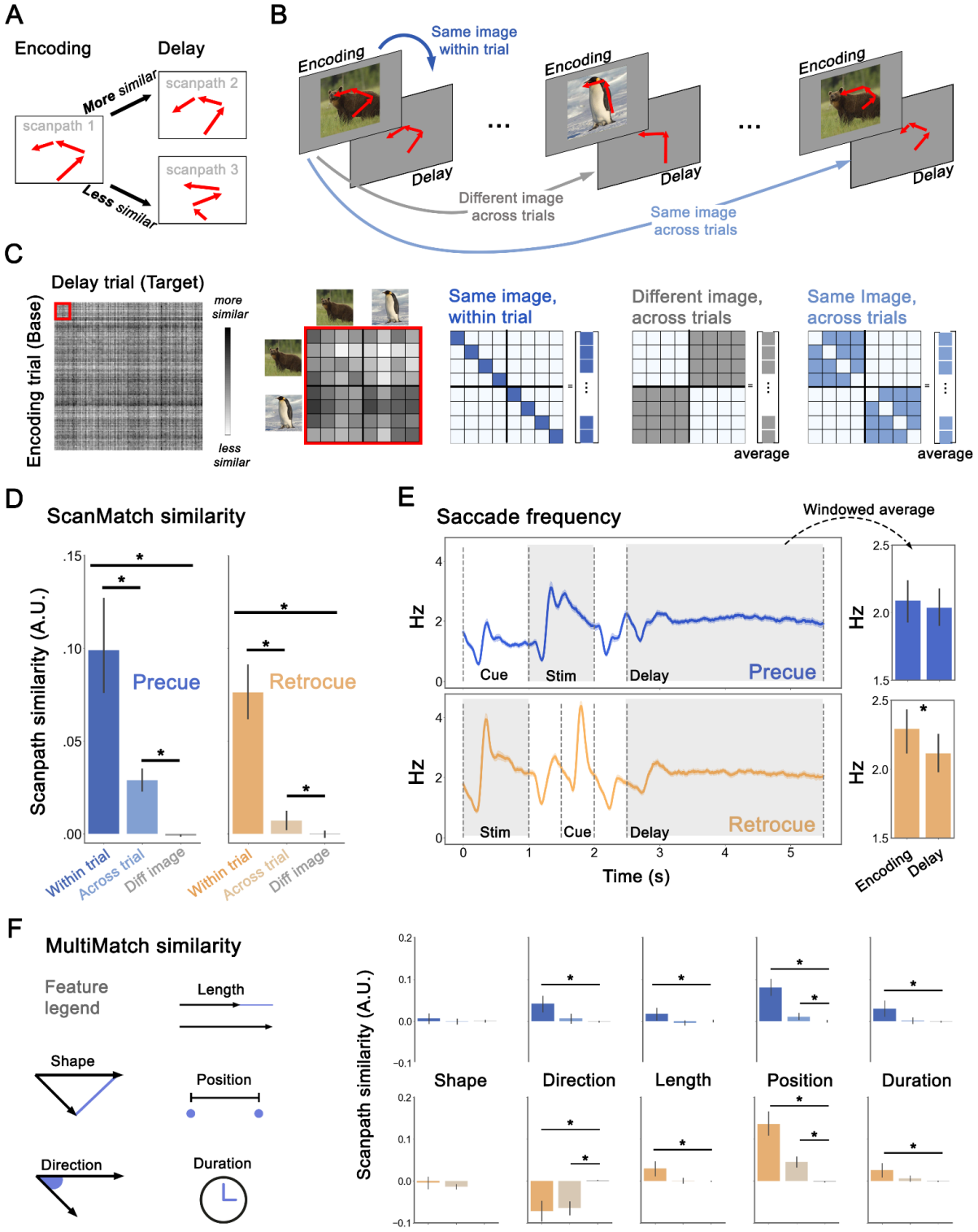

Figure 2. Scanpath method and results for both experiments. (**A**) An example illustration of similar vs. dissimilar scanpaths from encoding to delay. (**B**) Schematics illustrating the three types of encoding-delay scanpath similarity that were compared: within the same trial ("within

trial”), across different trials where different images were remembered (“different image”), and across different trials where the same image was remembered (“across trial”). (**C**) Left: The subject-averaged scanpath similarity matrix (ScanMatch algorithm) from 320 trials in the precue experiment. Diagonal values represent within-trial comparisons between encoding and delay scanpaths. Darker shades represent more similarity between encoding-delay scanpath pairs. Middle: A subset of the similarity matrix for two images, in just the visual condition. Each image appeared 4 times in each of the visual and semantic trial conditions (8 times total per experiment). Right: Masks applied to the scanpath similarity matrix to extract the three scanpath comparison conditions (“within trial”,“different image”, and “across trial”). Shaded regions represent 1, and white space represents 0. (**D**) The mean of individual subject scanpath similarity scores (ScanMatch algorithm) across comparison conditions and experiments. Note the similarity matrices were normalized within subject to account for baseline variability. Horizontal lines across the top of the plot indicate a main effect of comparison condition, whereas shorter lines accompany significant post-hoc pairwise comparisons. (**E**) Left: Average saccade frequency, over the time course of the trial, in both experiments. Right: Windowed average of saccade frequency for the encoding and delay epochs, in both experiments (areas shaded in grey on the frequency timecourse). (**F**) Scanpath similarity scores for five geometric dimensions (MultiMatch algorithm), across comparison conditions. Shape and direction measure the vector distance and angular distance between aligned saccade vectors, length measures saccade amplitudes, position measures the Euclidean distance between aligned fixations of the saccade vectors, and duration measures the fixation dwell time. Across all panels, error bars reflect 95% CI, and shaded error bands reflect ± 1SEM.

WM gaze reinstatement is stronger when the task goal is known before encoding

Here, participants were cued as to the task-relevant feature either before or after encoding the WM sample image (i.e., precue vs. retrocue). We reasoned that this manipulation might alter encoding-delay gaze similarities. For instance, participants might adopt different strategies in the different tasks. Participants might take a more focused encoding approach in the precue experiment, to extract only the most relevant information. However, they may need to encode more dimensions of information in the retrocue experiment, because the probe demands are not yet known. This could yield stronger encoding-delay gaze similarity in the precue experiment.

Indeed, encoding-delay scanpath similarity scores appeared overall lower during the retrocue task (**Figure 2D**). To investigate this difference, and ensure that eye movements weren’t overall lower during the retrocue task, we examined additional gaze metrics. We first examined saccade frequencies, across the whole trial epoch, in both experiments (**Figure 2E**). Saccade frequencies were descriptively elevated overall in the retrocue task (vs. precue), but especially during the WM sample encoding period. Saccade frequency was higher during encoding (as compared to delay) in only the retrocue task (precue: $t(40) = 0.75$, $p = 0.46$, *cohen’s d* = 0.11; retrocue: $t(40) = 2.56$, p

= 0.01, *cohen's d* = 0.38). Rather than a reduction in eye movements that could explain dampened encoding-delay similarity, participants may take a different approach to encoding the WM sample during a retrocue experiment.

To examine this further, we also tested for 'WM gaze reinstatement' using vector-based metrics that quantify multiple distinct scanpath properties. This complements the string-based metrics in the previous section, which yield a single distance measure for each scanpath comparison. In this section, the vector-based approach, implemented with the **MultiMatch** algorithm (Dewhurst et al., 2012; Jarodzka et al., 2010; Wagner et al., 2019), evaluates eye movement sequences along five geometric dimensions: shape, direction, length, position, and duration (**Figure 2F**).

In the precue experiment, all metrics except for shape exhibited WM gaze reinstatement within trials (**Figure 2F**; see **Supplementary Table 1** for statistics). However, the image-specific gaze reinstatement across trials was only present in the position metric (see **Supplementary Table 2** for statistics). This indicates that gaze tended to visit the same coordinate positions whenever the same image was remembered. The retrocue experiment similarly showed within-trial gaze reinstatement for all metrics except shape, and between trials for only the position metric. However, the direction metric showed that scanpaths were inversely related to each other from encoding to delay. That is, rather than the positive relationship that was observed between encoding-delay scanpaths in the precue experiment, the scanpaths showed a negative relationship in the retrocue experiment. Despite gaze reinstatement in position during the delay, the angular direction of the saccade vectors were more like the opposite of those during encoding. This could arise if gaze revisited the same coordinate positions as during encoding, but in a different order.

In sum, the timing of task cues may influence WM encoding strategies, and the nature of encoding-delay reinstatement in gaze. Saccades were overall more frequent during retrocue task encoding, but gaze position showed image-specific encoding-delay similarity in both experiments. However, the direction gaze metric showed sharply different patterns between the pre- and retrocue experiments, suggesting that eye movements may have followed a reversed temporal sequence when the task condition was cued after encoding.

WM gaze reinstatement is stronger when the task requires visual detail

A primary goal of this study was to test whether WM gaze patterns would be flexible to the expected demands of an upcoming WM test, even when the remembered stimuli were physically identical. We therefore examined whether WM gaze

reinstatement – as indexed by encoding-delay scanpath similarity – differed when the task emphasized visual vs. semantic maintenance strategies.

We first assessed encoding-delay similarity within and between trials (as in **Figure 2D**), but separately for visual and semantic trial conditions (**Figure 3A**). Here, we examined scanpath similarity only across trials of the same type (visual or semantic). For both experiments, we ran a repeated-measure ANOVA on ScanMatch similarity scores with factors of scanpath comparison ("within trial", "across trial", "different image") and trial feature (visual vs. semantic). Like earlier analyses, both experiments showed a main effect of scanpath comparison, whereby encoding-delay similarity was greatest within trial, then across trials with the same image, then across trials with different images (precue: $F(2,80) = 45.45$, $p < 0.001$, $\eta_p^2 = 0.53$; retrocue: $F(2,80) = 91.10$, $p < 0.001$, $\eta_p^2 = 0.70$). Additionally, the precue experiment showed an effect of trial feature ($F(1,40) = 13.72$, $p = 0.001$, $\eta_p^2 = 0.26$), and an interaction between factors ($F(2,80) = 4.03$, $p = 0.02$, $\eta_p^2 = 0.01$; **Figure 3B)**. Encoding-delay scanpath similarity was overall stronger for visual trials (vs. semantic) and the difference between visual and semantic strategies was most pronounced within trials. Therefore, WM gaze reinstatement was relatively amplified when the WM probe encouraged a precise visual memory strategy. However, in the retrocue experiment, there was no effect of trial feature, nor an interaction between factors (trial feature: $F(1,40) = 0.46$, $p = 0.50$, $\eta_p^2 = 0.01$; interaction: $F(2,80) = 1.76$, $p = 0.18$, $\eta_p^2 = 0.04$). Therefore, the gaze reinstatement difference between visual and semantic trials only emerged when the cue allowed a task-specific strategy during encoding.

The vector-based MultiMatch results also followed this pattern (**Figure 3C**; **Supplementary Figure 1** shows all scanpaths comparison conditions), and the pattern from earlier analyses (as in **Figure 2F**). The difference between visual and semantic trial gaze reinstatement was evident only in the position metric for the precue task (interaction: $F(2,80) = 3.81$, $p = 0.03$, $\eta_p^2 = 0.09$), and the retrocue task showed a converse pattern in the direction metric from encoding to delay (interaction: $F(2,80) = 3.68$, $p = 0.03$, $\eta_p^2 = 0.08$; **Supplementary Table 3** and **4** show all conditions and statistics). The difference between visual and semantic trials was not explained by differences in overall rates of eye movement, as saccade frequencies did not differ between trial features (**Supplementary Figure 2**). Therefore, the extent and type of WM gaze reinstatement was flexible to both task- and trial-level strategy.

The effect of trial feature on encoding-delay scanpath similarity (i.e., **Figure 3B**), suggests that scanpaths only show generalizable image-specific information across trials when the task encourages a visual strategy. However, this effect could arise because a visual encoding strategy is always more generalizable to other trials, or instead because only visual trials show strong alignment from encoding to delay.

Therefore, we also tested encoding-delay scanpath similarity across the two trial features – for instance, pairing a visual encoding trial with a semantic delay trial (**Figure 3D**). We did this cross-trial feature analysis for only the 'across trial' comparisons, to gauge which conditions yielded the most generalizable image-specific gaze patterns across trials (**Figure 3E**; See **Supplementary Figure 3** for "different image" comparisons). For only the precue task, we found that any combination of trials that included the visual strategy (as either the encoding or delay trial) produced superior scanpath similarity scores (**Supplementary Table 5**). Scanpath similarity was strongest when comparing encoding from visual trials to delay from visual trials, but encoding-delay comparisons across trial features (visual-semantic or semantic-visual) were stronger than comparisons between only semantic trials. Therefore, image-specific information in gaze patterns is minimal when the task encourages a semantic strategy.

The timecourse of pupil dilation also corroborated task- and trial-level strategy modulations (**Figure 3F**). Pupils can be expected to constrict at the onset of a new visual stimulus (Mathôt et al., 2018; Oster et al., 2022), and during fine-grained visual discrimination (Ebitz & Moore, 2019; Einhäuser, 2017). Here, we found that stimulus-induced constriction was also modulated by task context. In the precue task, the constriction during encoding was greater on visual trials, and pupil size also remained smaller on visual trials throughout the WM delay. In the retrocue task, pupil size was identical throughout encoding (as expected when the trial feature is not yet known), but visual and semantic trials diverged shortly after the retrocue offset. Both WM maintenance (precue) and internal attentional selection (retrocue) showed relative pupil constriction, consistent with visual discrimination processes, when the upcoming probe would require more fine detail.

Together, gaze and pupil metrics indicate that the same WM content may be subject to different encoding and maintenance strategies when the task context and probe demands differ. WM delay period gaze patterns held closer to encoding patterns when visual details were emphasized – but only when the task strategy was cued before encoding. Upcoming WM task demands can modulate the information content in image-specific eye movement patterns, during a blank WM delay.

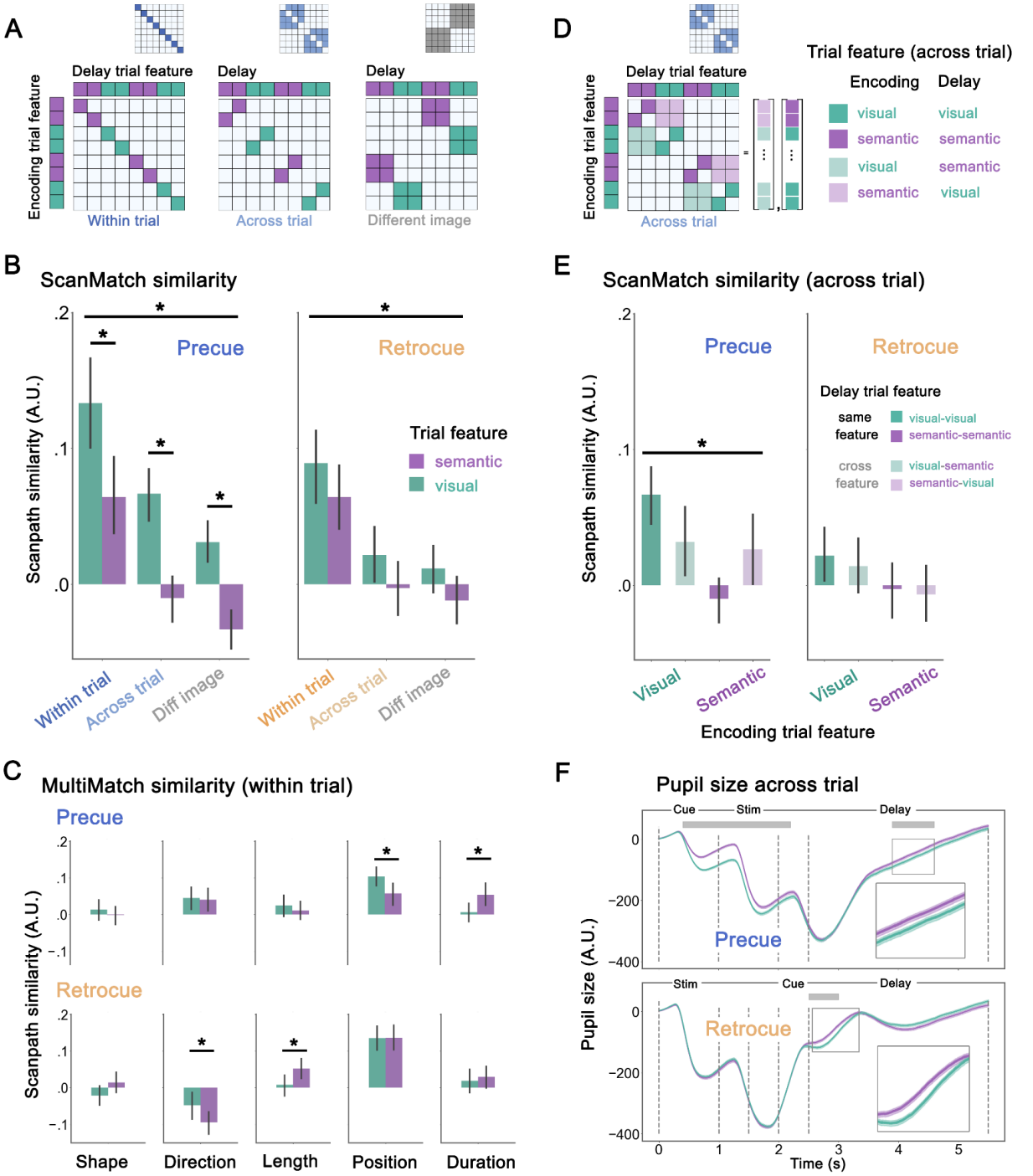

Figure 3. Scanpath similarity between encoding and delay epochs, across visual and semantic trial feature conditions. (**A**) A similar matrix schematic as in Figure 2, now illustrating visual (green) and semantic (purple) conditions, for two different images. The first 4 rows/columns all reflect trials with the same sample image, while the second half reflect a different sample image. Scanpath similarity scores for each comparison condition ("within trial", "across trial", "different image") were partitioned into visual or semantic trial feature conditions. (**B**) The mean of

individual subject scanpath similarity scores (ScanMatch algorithm) across comparison conditions and trial features, in precue and retrocue experiments. Note the similarity matrices were normalized within subject to account for baseline variability. Horizontal lines across the top of the plot indicate a main effect of comparison condition, whereas shorter lines accompany significant post-hoc pairwise comparisons. Note that the precue experiment also showed both a main effect of trial feature as well as an interaction between factors. (**C**) Scanpath similarity scores for visual and semantic trials, across five geometric dimensions (MultiMatch algorithm), for "within trial" scanpath comparison. (**D**) Schematic for cross-feature comparisons within the "across trial" condition. Darker shades represent congruent encoding and delay trial features (visual-visual, semantic-semantic), and lighter shades represent incongruent (visual-semantic, semantic-visual) (**E**) Encoding-delay scanpath similarity scores across trials of either the same type (visual or semantic), or across trial types. Note that the "same feature" condition depicts the same data as in the "across trial" condition of panel B, but here it is contrasted with scanpath similarity between trials that emphasized different features. Significance bar represents an interaction between encoding trial feature and delay trial congruence. (**F**) The timecourse of average pupil size across the trial, for visual and semantic trial features, in precue and retrocue experiments. Grey bands at top of plots indicate periods where permutation testing yielded a significant difference between visual (green) and semantic (semantic) conditions.

Gaze tracks the similarity relationships between naturalistic images during WM

The prior analyses have shown that gaze patterns during the WM delay share properties with the gaze patterns made during stimulus encoding, and that this similarity is accentuated when the task encourages a more visual encoding approach. However, even in cases when eye movement patterns differ from encoding to delay, they may still convey stimulus-specific information. For instance, when stimulus locations change from encoding to probe, location-specific WM gaze biases can shift to prospectively reflect expected future locations (Boettcher et al., 2021; Liu et al., 2024). To examine the timecourse of stimulus representation in the gaze patterns – and when visual vs. semantic differences might arise – these final analyses take a representational similarity analysis (RSA) approach.

RSA is often applied to multi-voxel fMRI activity patterns to characterize the stimulus information carried in physiological signals (Kriegeskorte et al., 2008). RSA is geared to gauge the representational information in lower-dimensional structure, even when specific activity patterns vary. For instance, two similar stimuli might be expected to show activity patterns that remain similarly correlated over time, even if the underlying activity for both stimuli changes. This approach is therefore well-suited to extract stable representational information over the course of a WM task, when activity dynamically fluctuates. This approach has also been applied to patterns of microsaccade bias during WM, to show that they reflect the relational structure between

stimuli of different orientations (Linde-Domingo & Spitzer, 2024). That is, remembering orientations that are more nearby to each other in feature space is associated with gaze patterns that are more similar to each other.

However, the similarity between our complex naturalistic stimuli is less straightforward to quantify. While a single feature like orientation varies along just one dimension, natural scenes can vary along myriad dimensions. We therefore used VGG-16, a convolutional neural network specialized for image recognition, to glean the similarity structure of our stimulus space (Simonyan & Zisserman, 2015). Briefly, we computed the pairwise correlations for every image pair, based on their activity in each layer of the model. We then computed the correlation between this stimulus similarity structure and the gaze similarity structure (Euclidean distance in gaze position between images), at every time point across the trial (**Figure 4A**). When the two structures are correlated (**Figure 4B**), we can say that the gaze pattern conveys stimulus-specific information. We found that, for both pre- and retrocue tasks, the last layer of the model showed the strongest correlation to the gaze data (**Figure 4C**), so the remaining analyses use the similarity structure from that final layer.

We partitioned the trial pairs into three categories, based on the corresponding trial feature comparisons: correlations between visual-to-visual, semantic-to-semantic, or semantic-to-visual trials (the Euclidean distance structure between visual to semantic trials is identical to that between semantic to visual trials). We separately ran the RSA for these three comparisons and found that representational information was strongly present during encoding, in both the pre- and retrocue experiments (**Figure 4C**). In other words, the correlations between trials in the gaze structure resembled the model-derived similarity between stimuli. More similar WM sample stimuli showed more similar gaze patterns. Permutation testing confirmed that this correlation was above chance in all three scanpath comparisons (visual-visual: 1.2s to 3s, $p < 0.001$; semantic-semantic: 1.2 s to 2.9 s, $p < 0.001$; semantic-visual: 0.9s to 3s, $p < 0.001$). Image-specific gaze patterns during WM encoding therefore reflected the relational structure between these naturalistic images.

Stimulus information tapered off after encoding, but ramped back up again toward the end of the delay. In both experiments, WM stimulus information in the gaze pattern – at least of the type that is captured by the VGG-16 similarity structure – emerged late in the WM delay, only as the probe approached. In the precue experiment, representational information was present in the delay for all three conditions (visual-visual: 4.9s to 5.4s, $p = 0.04$; semantic-semantic: 3.8s to 4.4s, $p = 0.03$; 4.6s to 5.2s, $p = 0.03$; semantic-visual: 4.4s to 5.4s, $p = 0.02$). However, in the retrocue experiment, this late-delay escalation occurred only in the visual-visual condition (4.8s to 5.4s, $p = 0.02$). No significant clusters were detected in the semantic-semantic condition, although one cluster approached significance in the semantic-visual condition

(5.5 s to 5.9 s, $p = 0.06$). An additional permutation analysis tested the difference between trial feature comparisons (visual-visual, semantic-semantic, semantic-visual) and found an effect in only the retrocue experiment (5.4 s to 5.9 s $p = 0.04$). Therefore, stimulus-specific information emerged during the retrocue WM delay for only visual trials.

RSA showed that gaze patterns reflect image-specific information during WM, even for complex naturalistic images. At various points across a WM trial, stimuli that share more similar image properties also evoke more similar gaze patterns. Notably, we found that the expression of stimulus information in the gaze rose in anticipation of the WM probe, consistent with the idea that WM may engage oculomotor systems in a prospective manner.

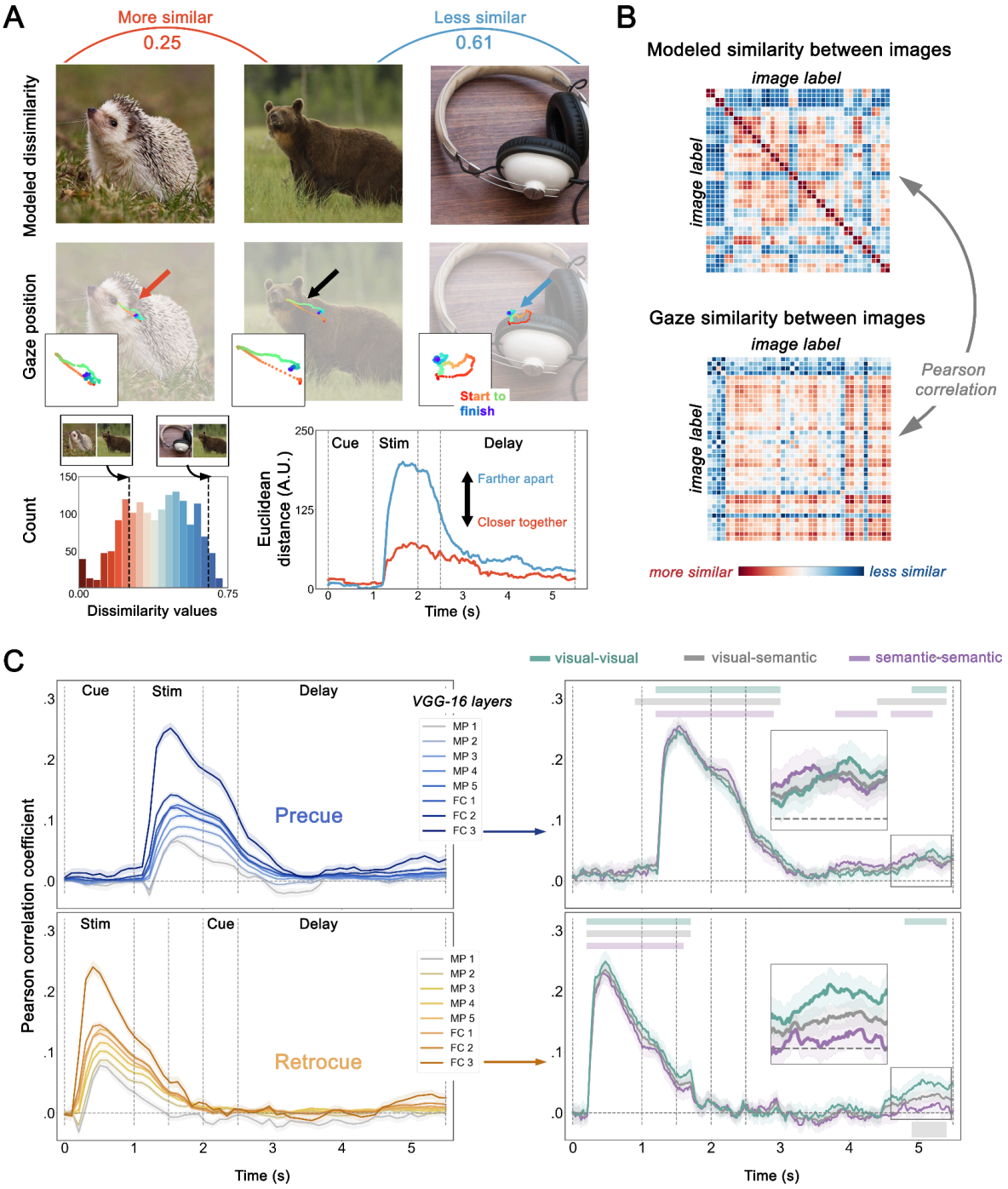

Figure 4. Gaze RSA methods and results. (**A**) Evaluating stimulus similarity and gaze similarity for representational similarity analyses (RSA). Top panel: the similarity values between example image pairs, as assessed via correlation in the final layer of VGG-16. The first pair [bear, hedgehog] was rated as relatively more similar, and the second [bear, earphone] as less similar, according to their activity patterns in the network. Below that are example gaze coordinates during WM for the corresponding image. Cooler colors represent later time points. Lower panel: The distribution of scores, corresponding to the similarities between all unique image pairs (left),

and an example time course of Euclidean distances in gaze between those image pairs. (**B**) Top: matrix depicting pairwise similarity scores for every image pair in the stimulus set, assessed via correlations between VGG-16 layer activity. Bottom: average gaze similarity (Euclidean distance) for respective stimulus image pairs, collapsed across time. (**C**) Pearson correlation between gaze similarity and modeled similarity, across every trial time point, evaluated for key max pooling (MC) and fully connected (FC) VGG-16 layers. Darker lines represent later layers of the model. (**D**) Pearson correlation between gaze similarity and modeled similarity (from activity in layer FC3), split by trial feature conditions (visual-visual, visual-semantic, and semantic-semantic). Thin shaded bands on top of plot show periods when correlation values were greater than chance, for each trial feature condition. Thicker band at bottom of plot shows periods when trial feature conditions differed from each other.

Discussion

Across two experiments, we asked participants to remember real-world images, and we manipulated whether the visual details or semantic labels were most relevant for the WM test. We assessed WM gaze reinstatement, operationalized as the similarity in spatio-temporal eye movement patterns from stimulus encoding to WM delay. Additionally, we used deep convolutional neural network modeling to guide RSA of the evolving stimulus information in gaze patterns across a trial. We found that eye movement patterns tracked specific images held in mind. Notably, WM gaze reinstatement was most pronounced when the task required greater visual detail, and when the task strategy was known before stimulus encoding. The findings reveal precise signatures of mnemonic content in oculomotor inflections, which are flexible to the most task-relevant stimulus features.

Distributed working memory representations across the nervous system

The representational basis for WM has been the subject of ongoing debate (Leavitt et al., 2017; Sreenivasan et al., 2014; Xu, 2017). Most human neuroimaging supports the idea that WM maintenance engages shared neural representations with sensory perception (D'Esposito & Postle, 2015; Serences et al., 2009), but it is becoming increasingly clear that WM-related information is widely distributed across the brain (Christophel et al., 2017; Courtney et al., 1997; Dotson et al., 2018). Beyond the cortex, WM content information may also be read out from peripheral sensorimotor apparati (Dong et al., 2025; Dong & Kiyonaga, 2024; van Ede et al., 2019; Zokaei et al., 2019). Here, we find that precise and flexible WM content information is present in the eyes. More than simple location tagging, gaze patterns during WM retrace spatio-temporal sequences associated with specific naturalistic images. This WM gaze reinstatement is also more than a lingering sensory-evoked trace, or reflexive response to visual stimulation, as it was acutely sensitive to which features were most task-relevant. WM may therefore

recruit whatever relevant sensory and motor structures are available across the nervous system to support current needs (Postle et al., 2006).

Working memory encoding strategy varies with task context

A hallmark of WM is its flexibility. Humans can strategically shift and spread attention among items held in mind, and can flexibly use mental content toward a variety of goals (Kiyonaga & Egner, 2013; Nobre & Stokes, 2019; Souza & Oberauer, 2016). However, the mechanisms that support flexible internal representation have been contested (Oberauer & Awh, 2022; Stokes, 2015). Here, we varied contextual WM demands, at both the level of the overarching task (precue vs. retrocue) and at the trial level within tasks (visual vs. semantic), to examine how oculomotor WM adapts to these situational factors.

We found that task-level context impacted stimulus encoding and maintenance strategies, as indexed by eye movement patterns. A dynamic and uncertain environment would encourage exploration behavior (Cohen et al., 2007). In our case, when the relevant stimulus features were unknown during encoding (i.e., retrocue task), the task may favor a more holistic, exploratory sampling strategy. Likewise, we observed relatively increased saccade frequency during retrocue task encoding. However, when the relevant feature is then cued after encoding, the information may undergo a representational transformation during maintenance (Souza & Oberauer, 2016). Likewise, we found that some gaze metrics showed anti-correlated patterns from encoding-delay in the retrocue task. During the precue task, on the other hand, encoding can be proactively targeted to a particular task strategy. WM maintenance may then sustain that strategy, yielding more similar gaze patterns from encoding to delay.

Collectively, oculomotor WM signatures reflect flexibility in encoding-delay WM strategies. When the task mode is known before encoding, the eyes may sample information selectively, but maintain a stable representation during the delay. Conversely, when the task-relevant features are unknown at encoding, the eyes may sample more holistically, but then undergo a representational transformation during the delay.

Working memory maintenance signatures anticipate task-relevant features

In addition to the strategic differences prompted by task-level cue timing, oculomotor WM signatures also varied with the trial-level cued feature. When the WM test required more precise visual memory, delay period gaze resembled the encoding pattern more strongly. That is, given the same encoded stimulus, the delay-spanning eye movement patterns depended on the expected upcoming task demands. This aligns with neuroimaging findings that visual cortical regions are preferentially engaged

during WM for more visual (vs. semantic) features (Lee et al., 2013). Likewise, we speculate that WM flexibly recruits distributed sensorimotor apparati, whose representations are tailored to the task demand. Rather than reflectively represent encoded stimulus features, sensorimotor WM representations adapt to prioritize the type of representation required for successful performance (Ball et al., 2013; McAteer et al., 2023).

Further corroborating the flexibility in WM maintenance strategies, we also found that pupil size was relatively smaller during the WM delay when visual detail was emphasized. Smaller pupil size has been shown to benefit fine-grained discrimination, while larger pupil size benefits detection (Eberhardt et al., 2022; Ebitz & Moore, 2019; Mathôt et al., 2018; Mathôt & Ivanov, 2019). Our findings suggest that such pupillary mechanisms for perception are also engaged during 'internal' visual discrimination. Together, these results show that various ocular inflections during WM are modulated by the stimulus features that are most task-relevant. Humans may remember the same information differently depending on what they need to do with it.

Multiple routes to successful working memory

It is unclear whether eye movements during WM play any causal role in maintenance. There is some evidence that interfering with eye movements can interfere with WM and LTM (Postle et al., 2006; Johansson & Johansson, 2014, 2020), but the link is tenuous (Liu et al., 2022; Loaiza & Souza, 2022; Walcher et al., 2024). Here, pre- and retrocue tasks exhibited distinct oculomotor signatures despite comparable behavior, suggesting that the eye movements are not strictly essential. Instead, there can be multiple pathways to the same behavioral outcome. Indeed, people with different preferred maintenance strategies (e.g., visual vs. verbal) tend to perform equally well on most WM tasks, even if their ocular engagement varies (Keogh & Pearson, 2018; Korda et al., 2024). Oculomotor signatures are likely one of many distributed WM representations.

The current findings have both practical and theoretical implications. This work suggests that gaze may be a feasible tool to gauge the integrity of different underlying representations (e.g., sensory vs. linguistic). The findings also suggest that the representations underlying WM may differ depending on the resolution required for the task at hand. Therefore, our understanding of visual WM storage based on lab tasks that test for precise visual features may not apply to 'good enough' WM that can be achieved with a semantic label or category code.

Concluding remarks

Information that we hold in WM can be detected from our eyes, the first stop for visual perception. We found that, during WM maintenance, eye movements recapitulate

the spatiotemporal sequences made during encoding. This WM gaze reinstatement carries not only information about simple visual stimuli such as orientation or shapes, but also image-specific patterns for complex natural scenes. Most critically, we see that the WM oculomotor signal is not merely a reflexive sensory reinstatement, but also varies with WM strategy and prospective task demands. In this sense, the eyes make an informative, accessible readout of otherwise hidden mental representations. WM is not confined to the central nervous system but flexibly and prospectively engages the sensorimotor periphery.

Detailed Methods

Participants

The experiments were approved by the local Office of IRB Administration (OIA). Participants were adults (over 18 years of age) with normal or corrected to normal (via contact lens) vision, recruited via UCSD's SONA platform. Participants completed informed consent and received course credit as compensation. The study took 60 - 80 minutes to complete.

We recruited 42 participants for Experiment 1 (precue), and 41 for Experiment 2 (retrocue). We had no precedent to estimate expected effect sizes for this type of analysis, so we based our sample size on related studies that find reliable effects when testing similar multivariate analyses of eye movement data (Linde-Domingo & Spitzer, 2024). One participant was excluded from the precue experiment for data quality reasons (see *Preprocessing* for Eyetracking QA details), and no participants were excluded from the retrocue experiment. After exclusions, 41 participants were included in analyses for the precue experiment (*27 females, 12 males, 1 declined to report, 1 other, mean age = 22.84 ± 3.85*) and 41 for the retrocue experiment (*34 females, 5 males, 1 declined to report, 1 other, mean age = 20.31 ± 2.55*).

Stimuli and task procedure

The experiments were programmed and presented in Psychopy, v2023.2.3. Participants completed the task in a dimly lit environment on a 1920 x 1080 monitor (24 inch, 60.96cm) with 60hz refresh rate. Participants sat 60cm away from the screen center, their heads stabilized by a chinrest.

The WM sample stimuli were 40 unique naturalistic images from the *Things* database (Hebart & Stoinski, 2019).Two images were taken from each of 20 randomly-selected categories (e.g., bears, watches, etc). The image set was the same for all participants, to obtain comparable gaze measures across participants.

In both experiments, each trial started with a 500ms fixation, where a black cross appeared at the center of the screen (128 x 128 pixels, subtending 3.5º visual angle). In the **precue experiment** (**Figure 1A**, top panel), the color of the cross then changed to either green or purple, representing the trial condition cue. The color of the cross cued which stimulus feature would be tested at the end of the trial. This precue remained onscreen for 1000ms. If the precue (the cross) was green, the WM probe would test for visual detail of the sample image, but if the precue was purple, the probe would test for the semantic category of the sample stimulus. The color cues were designed to be equiluminant to each other and to the neutral grey background (green: RGB 0, 128, 0; purple: RGB 128, 0, 128). Participants were instructed and trained, before the main experiment, to learn the meaning of the cues.

Following the precue, a WM sample image (540 x 540 pixels; subtending 14º visual angle) was shown for 1000ms, centered on the display screen. After the sample image, a noise mask appeared for 500ms, and was then followed by a 3000ms blank memory delay. Lastly, at the end of each trial, a probe image appeared for 3000ms. Probe images subtended 7º visual angle (270 x 270 pixels). In the visual condition (i.e., green cue), the probe was a cropped portion of either the sample image or a different image from the same category. The crop was 20% of the original image size (then scaled to 270 x 270 pixels), and the crop center was drawn randomly within a circular range of 270 pixels from the image center (see **Figure 1B** for example probes). This technique, of showing participants only a segment of the full image, was designed to encourage a more visual maintenance strategy for fine image details. In the semantic condition (i.e., purple cue), the probe image was never an exact match to the sample image, but was drawn from either the same or different category as the sample. The semantic probe images were uncropped, but were scaled down to 270 x 270 pixels to match the size of the visual probes. In this condition, only a semantic label was necessary to distinguish whether the probe was a (category) match, and visual details would be little help since the category exemplar image was always different from the sample. Non-match probe stimuli were taken from outside the sample stimulus set, drawn from the full image repository. In both conditions, participants made a keyboard button press to indicate whether the probe was a match or non-match to the sample – either a visual match to the original image, or a different image from the matching category (Y for match; N for mismatch). Match and non-match probes, as well as visual vs. semantic trials, occurred equally often and in random order.

The trial structure of the retrocue experiment (**Figure 1A**, bottom panel) broadly followed that of the precue experiment, except for one critical difference: instead of being told before encoding what stimulus feature would be tested, the participant was retroactively cued after the sample stimulus had been encoded. Each trial started with a 500ms fixation, followed by a 1000ms stimuli encoding period. Next the mask was

shown for 500ms, and subsequently, a 500ms blank screen occurred. Following the blank screen, the retrocue appeared as a color cross. As in the precue experiment, the color (purple or green) indicated the task relevant feature (visual or semantic). Lastly after a 3000ms memory delay, participants completed the same aforementioned probe task.

Both experiments consisted of 320 trials, divided into 16 blocks of 20 trials each. Each unique image was presented 8 times in total – 4 times in each the visual and semantic conditions. Within that constraint, images were drawn from the stimulus set in random order for each trial. One complete session of either experiment contained 160 visual and 160 semantic trials.

Eye-tracking analysis

We implemented continuous eye tracking throughout both experiments via a desk-mounted Eyelink Portable Duo, recording at 1000hz (SR Research Ltd.). We performed a 13-point calibration using the built-in calibration module, and performed a drift check at the beginning of every block (every 20 trials) to ensure that central fixation was maintained. Data was exported using the EyeLink Data Viewer software package (Version 4.3.1).

Preprocessing. The eye-tracking preprocessing procedure was similar to previous work, where blink artifacts were detected as outliers in the velocity space of pupillary change (Dong & Kiyonaga, 2024). The data were first transformed into *Trial x Time* format, and for each trial, we obtained a smoothed derivative array, $\vec{V}$, from the original pupil size, $\vec{X}$

$$\vec{V}_n = \frac{\vec{X}_{n+2} + \vec{X}_{n+1} - \vec{X}_{n-1} - \vec{X}_{n-2}}{6\Delta t}$$

Next, we defined blinks as any time point that exceeded a velocity threshold

Threshold = Median $(\vec{V})$ *+ 3.5 * Median absolute deviation*$(\vec{V})$

The timepoints exceeding this threshold were marked as blinks, and each was padded with a 50ms epoch on each side. Then, we masked the identified artifacts from the gaze data, removing any data during these periods. The missing gaze data were reconstructed, depending on the availability of data surrounding each gap, using linear,

nearest-neighbor, or no interpolation. After interpolation, a trial was discarded if it had more than 20% data missing from the main epoch of interest, that is, the 5.5s from fixation offset to the end of the memory delay. Participants were excluded if they had more than 40% trials rejected, either due to poor eye-tracking quality or missing probe responses. After excluding one subject for poor data quality, trial-wise data cleaning retained 91.3% of data in the precue experiment (203 - 320 trials per participant), and 93.5% of data in the retrocue experiment (255 - 319 trials per participant).

Saccade detection. Similar to detecting blinks from the pupil array, saccadic activity could be detected from gaze as outliers in velocity space. For this procedure, we adopted the approach outlined in **Engbert & Kliegl 2003** (Engbert & Kliegl, 2003). To reduce aliasing for saccade detection, we smoothed the gaze data using a rolling window average of size 11ms, then calculated the gaze velocity on the horizontal and vertical axis $\vec{V}_x$, $\vec{V}_y$. The 2d velocity was subsequently calculated as

$$\vec{V}_{2d} = \sqrt{(\vec{V}_x)^2 + (\vec{V}_y)^2},$$

Saccades were outliers on this 2d velocity space that exceeded the threshold of median($\vec{V}_{2d}$) + 6 * median absolute deviation($\vec{V}_{2d}$). Additionally, we set a minimum shift distance of 8 pixels for a movement to be considered saccadic, equating to 0.4% change relative to the screen size and 0.21º visual angle. These thresholds followed those determined by prior work (Dong & Kiyonaga, 2024). We also calculated the frequency of saccades, for control analysis (**Figure 2E**), by applying a rolling window filter. Here, F_N represents the saccade frequency at Nth window.

$$F_N(\#saccades/s) = \frac{\#\ of\ saccades\ at\ Nth\ window}{100ms} * 1000ms/s$$

Scanpath analysis. A scanpath is a series of fixations. To transform the continuous gaze array into scanpaths, we first epoched the identified saccades to the two events of interest: the stimulus encoding and memory delay periods. We then extracted fixation locations (the average eye position on the horizontal and vertical axes between each detected saccade), as well as the fixation duration (dwell time in between each saccade (Coutrot et al., 2018; Cristino et al., 2010; Jarodzka et al., 2010)). This procedure was applied to every trial, resulting in an encoding and a delay scanpath for each. These scanpaths were then used as input to the analyses described below.

The primary analyses compare the similarity between encoding and delay scanpaths (**Figures 2** and **3**). To quantify scanpath similarity, we used both string-based and vector-based approaches. For the string-based approach, we used a MATLAB implementation of the ScanMatch algorithm (Cristino et al., 2010). In brief, the string-based analyses discretize the continuous gaze recording into a string of letters, each representing a region on the monitor screen where the fixation visited. Subsequently, two scanpath letter strings were compared to evaluate their similarity. The ScanMatch similarity is the normalized output from a sequence alignment algorithm (Needleman–Wunsch (Needleman & Wunsch, 1970)). This algorithm takes two scanpaths as input and produces a single similarity value representing the degree of similarity between them.

We compared scanpaths for every encoding-delay pair, each producing a single score. That is to say, every trial's encoding period was compared to the delay period of the same trial, as well as the delay periods from every other trial. This generated a 320 x 320 matrix for every participant, with encoding trials as rows and delay trials as columns (**Figure 2C**). We then z-score normalized each similarity matrix to remove across-participant baseline variability. Diagonal values indexed the scanpath similarity from encoding to delay within a trial, which we used to measure the strength of gaze reinstatement during the WM delay (**Figure 2C,** "same image, within trial"). Off-diagonal values for the same image indicated the extent to which image-specific eye-movement patterns generalized across trials (**Figure 2C,** "same image, across trials"). Off-diagonal values for different images served as a baseline comparison denoting whether image-general encoding-delay gaze patterns were shared across trials, regardless of the specific stimulus (**Figure 2C,** "different image, across trials"). Since each unique image was presented multiple times in the experiment, scores were averaged across columns for the "different image" and "across trial" comparisons, to yield three scores for every trial: "within trial", "across trial", and "different image".

Supplementing the string-based analyses (**ScanMatch**), we also examined scanpath similarity along geometric feature dimensions. We used a Python implementation of the original MultiMatch algorithm, multimatch-gaze (Dewhurst et al., 2012; Jarodzka et al., 2010; Wagner et al., 2019). Here, instead of letter sequences, scanpaths were represented as vectors and evaluated along five geometric dimensions: shape, length, position, direction, and duration. We completed the same procedure with both the string-based **ScanMatch** and vector-based **MultiMatch** algorithms. While the former yielded just one similarity score for each comparison, the latter output five scores for the five comparison dimensions.

To evaluate whether scanpath similarity changed in response to visual and semantic task demands, we then tagged each row of the similarity matrix by whether the encoding scanpath (that was shared by all encoding-delay pairs in a row) was from

a visual or semantic trial. For each row, we selected only the scores from columns that had the same trial feature as the base. For example, if the encoding epoch was from a visual trial, then we only paired it with delay epochs (columns) that were also from visual trials (**Figure 3A**). The three scanpath comparisons ("within trial", "across trial", and "different image") were thus partitioned into visual or semantic trial categories, so that we could test the effect of trial feature and whether it emerged within or across trials. We applied the same procedure to the ScanMatch and Multimatch matrices (**Figure 3B** and **3C**).

Lastly, we zeroed in on the "across trial" condition to examine scanpath similarity between the visual and semantic trial types. The idea here was to examine whether generalizable image-specific information in gaze patterns would only be present across trials of the same type (i.e., only within visual trials), or whether one trial type might encourage an encoding strategy that better generalizes between trial types (e.g., visual encoding may share scanpath similarity with a semantic delay, but not the other way around. Therefore, in addition to selecting encoding-delay pairs within the same trial feature (as in the preceding analyses), here we also selected pairs where the encoding and delay came from different trial feature types. This generated four conditions, crossing each encoding and delay trial type: visual-visual, visual-semantic, semantic-visual, or semantic-semantic (**Figure 3D**). As with the other scanpath analyses, similarity values within each condition were averaged to produce one value per condition, per encoding trial.

Representational similarity analysis. In the final analysis, we assessed the timecourse of stimulus representation in gaze patterns through a Representational Similarity Analysis (RSA). We aimed to measure the correlation between a gaze similarity matrix and a stimulus similarity matrix, to determine if the gaze patterns reflected the structure of the stimulus space. To quantify the similarity between naturalistic images in the stimulus set, we used VGG-16, a deep convolutional neural network model, to construct a model-derived similarity matrix (Simonyan & Zisserman, 2015). The stimulus similarity value was determined as the Pearson correlation between the network's response to pairs of images (similar to **Piriyajitakonkij 2024** (Piriyajitakonkij et al., 2024)**)**. Specifically, the modeled similarity for a layer was calculated as

$$RDV_{layer}(base\ image,\ target\ image) = C(F_{layer}(base\ image), F_{layer}(target\ image))$$

Where F_{layer} is the activity in a given layer of the model in response to an image, and C is the Pearson product-moment correlation coefficient. The stimulus similarity

structure consisted of the pairwise correlations between every stimulus image in the sample set (**Figure 4A**).

Next we obtained the gaze similarity matrix for each participant. The gaze data was first epoched to the 5.5s period of interest (i.e. fixation offset to the end of memory delay), then downsampled to 10Hz for efficiency by averaging the data in each 100ms bin. For every participant, we calculated the gaze similarity structure as the Euclidean distance in coordinates between each trial pair, not including to itself (Linde-Domingo & Spitzer, 2024). The Euclidean distance of the gaze coordinates was calculated for every trial pair and time point, then averaged across pairings of the same two images (**Figure 4B**).

This produced a modeled similarity structure between images of size 40 x 40, and a gaze similarity structure of 40 x 40 for each of 55 timepoints. To quantify the stimulus information conveyed in the gaze patterns, we calculated Pearson correlations between the modeled stimulus similarity structure and the gaze similarity structure for each timepoint (**Figure 4B**). Note, the raw values of the modeled similarity ranged from -1 to 1, with 1 signifying maximum similarity (same image). In contrast, in gaze, a smaller difference in Euclidean distance indicated higher similarity. To align the direction of modeled stimulus similarities and gaze pattern, we normalized the model output to the interval [0,1], then subtracted that value from 1, so that 0 represented the strongest similarity (same image). This procedure produced one time series of Pearson correlation for each participant, indexing the degree of stimulus information in their gaze signal.

We first examined the stimulus information in gaze across several layers of the model. We conducted the RSA based on activity in the five max pooling layers and three fully connected layers of VGG-16 (**Figure 4C**;). Since the modeled similarity structure from the last layer (FC3) exhibited the highest correlation with the gaze structure, we used only this last layer similarity for all subsequent analyses.

To address if the stimulus feature that was relevant on each trial shaped how stimulus information was reflected in the gaze signal, we partitioned the trial pairs into three trial feature comparisons: visual-visual, semantic-semantic, and semantic-visual. Note that here, unlike the scanpath analysis, the semantic-visual and visual-semantic conditions are identical. After partitioning the trial pairs, we averaged across any pairings of the same images within each condition, creating three 40 x 40 x 55 matrices. Then we carried out the same RSA correlation separately for each trial feature condition.

The statistical analyses on the RSA time series are similar to previous work (Dong & Kiyonaga, 2024; Zokaei et al., 2019). To test whether the correlations were

significantly above chance, in each iteration, we first randomly flipped the signs of half of the participants' correlations. Then we tested the permuted correlations against zero via a one-sample *t*-test, and this was conducted once for each time point. From the resulting array of p-values, we found the largest cluster that exceeded the significance threshold ($p < 0.05$) and summed the t-values corresponding to this cluster. Then, this procedure was repeated 5000 times to generate a permutation distribution for each condition. Lastly, the same *t*-test was run using the actual correlation scores, and the clusters of significance were identified. The *p*-value of each could be calculated by comparing the actual tsum of the true clusters versus the permutation distribution as $P(T > t_{sum})$. If this value is smaller than 0.05, then such a cluster in the correlation trace would be considered significantly above zero.

Finally, to test if the gaze RSA correlation differed by trial feature, we ran a cluster-based permutation ANOVA, where we first shuffled not the sign, but the labels of the trial feature comparison condition (visual-visual, semantic-semantic, and semantic-visual). Then we applied a one-way ANOVA to each time point, with the shuffled label as the independent variable, and correlation value as the dependent variable. Instead of t-values, here we recorded the sum of f-values of the largest cluster that exceeded significance. This procedure was likewise repeated 5000 times to generate the permutation distribution. We calculated the *F* and *p* values of the real (not permuted) data and identified clusters that exceeded significance. Then the F_{sum} of these clusters was compared to the permutation distribution.

Pupil analysis. For pupil data, after artifact rejection followed the methods described in the *Preprocessing* section, we then applied the same interpolation procedure for sections of missing data. While we used linear of nearest neighbour interpolation to reconstruct the gaze data, we used linear and cubic spline for pupils. The slightly different interpolation methods for the two metrics is meant to better capture their respective physiology (Mathôt & Vilotijević, 2023). Then we also removed any trials that had been excluded based on gaze data quality.

To compare the pupil size under visual and semantic task conditions, we first subtracted each trial from a baseline value. The baseline was taken as the average pupil size during the last 200ms of the fixation period at the beginning of each trial. We then partitioned the data based on the trial feature condition (visual vs. semantic).

We applied cluster-based permutation *t*-tests to identify periods of significant difference in pupil size under different task feature demands (visual and semantic). In each iteration, we randomly shuffled the trial feature labels, and applied independent

t-tests at each timepoint using the shuffled label as the independent variable. The rest of the procedures followed the steps outlined in the previous sections.

References

Bae, G.-Y., & Chen, K.-W. (2024). EEG decoding reveals task-dependent recoding of sensory information in working memory. *NeuroImage*, *297*, 120710. https://doi.org/10.1016/j.neuroimage.2024.120710

Ball, K., Pearson, D. G., & Smith, D. T. (2013). Oculomotor involvement in spatial working memory is task-specific. *Cognition*, *129*(2), 439–446. https://doi.org/10.1016/j.cognition.2013.08.006

Boettcher, S. E. P., Gresch, D., Nobre, A. C., & van Ede, F. (2021). Output planning at the input stage in visual working memory. *Science Advances*, *7*(13), eabe8212. https://doi.org/10.1126/sciadv.abe8212

Brown, L. A., & and Wesley, R. W. (2013). Visual working memory is enhanced by mixed strategy use and semantic coding. *Journal of Cognitive Psychology*, *25*(3), 328–338. https://doi.org/10.1080/20445911.2013.773004

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. https://doi.org/10.1016/j.tics.2016.12.007

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098

Courtney, S. M., Ungerleider, L. G., Keil, K., & Haxby, J. V. (1997). Transient and sustained activity in a distributed neural system for human working memory. *Nature*, *386*(6625), 608–611. https://doi.org/10.1038/386608a0

Coutrot, A., Hsiao, J. H., & Chan, A. B. (2018). Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods*, *50*(1), 362–379. https://doi.org/10.3758/s13428-017-0876-8

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

Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. D. (2010). ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, *42*(3), 692–700. https://doi.org/10.3758/BRM.42.3.692

D'Esposito, M., & Postle, B. R. (2015). The Cognitive Neuroscience of Working Memory. *Annual Review of Psychology*, *66*(1), 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., & Holmqvist, K. (2012). It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods*, *44*(4), 1079–1100. https://doi.org/10.3758/s13428-012-0212-2

Dong, Y., Hung, Y.-C., Xie, C., & Kiyonaga, A. (2025). Pupillary Signatures of Working Memory Content are Flexible to Visual vs. Semantic Task Demands. *Journal of Vision*, *25*(9), 2871. https://doi.org/10.1167/jov.25.9.2871

Dong, Y., & Kiyonaga, A. (2024). Ocular working memory signals are flexible to behavioral priority and subjective imagery strength. *Journal of Neurophysiology*, *132*(1), 162–176. https://doi.org/10.1152/jn.00446.2023

Dotson, N. M., Hoffman, S. J., Goodell, B., & Gray, C. M. (2018). Feature-Based Visual Short-Term Memory Is Widely Distributed and Hierarchically Organized. *Neuron*, *99*(1), 215-226.e4. https://doi.org/10.1016/j.neuron.2018.05.026

Eberhardt, L. V., Strauch, C., Hartmann, T. S., & Huckauf, A. (2022). Increasing pupil size is associated with improved detection performance in the periphery. *Attention, Perception, & Psychophysics*, *84*(1), 138–149. https://doi.org/10.3758/s13414-021-02388-w

Ebitz, R. B., & Moore, T. (2019). Both a Gauge and a Filter: Cognitive Modulations of Pupil Size. *Frontiers in Neurology*, *9*. https://doi.org/10.3389/fneur.2018.01190

Einhäuser, W. (2017). The Pupil as Marker of Cognitive Processes. In Q. Zhao (Ed.), *Computational and Cognitive Neuroscience of Vision* (pp. 141–169). Springer. https://doi.org/10.1007/978-981-10-0213-7_7

Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*(9), 1035–1045. https://doi.org/10.1016/S0042-6989(03)00084-1

Eriksson, J., Vogel, E. K., Lansner, A., Bergström, F., & Nyberg, L. (2015). Neurocognitive Architecture of Working Memory. *Neuron*, *88*(1), 33–46. https://doi.org/10.1016/j.neuron.2015.09.020

Ester, E. F., Anderson, D. E., Serences, J. T., & Awh, E. (2013). A Neural Measure of Precision in Visual Working Memory. *Journal of Cognitive Neuroscience*, *25*(5), 754–761. https://doi.org/10.1162/jocn_a_00357

Fahimi, R., & Bruce, N. D. B. (2021). On metrics for measuring scanpath similarity. *Behavior Research Methods*, *53*(2), 609–628. https://doi.org/10.3758/s13428-020-01441-0

Fukuda, K., & Woodman, G. F. (2017). Visual working memory buffers information retrieved from visual long-term memory. *Proceedings of the National Academy of Sciences*, *114*(20), 5306–5311. https://doi.org/10.1073/pnas.1617874114

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*(2), 129–135.

https://doi.org/10.1016/j.tics.2011.11.014

Hebart, M., & Stoinski, L. (2019). *THINGS object concept and object image database*. https://doi.org/10.17605/OSF.IO/JUM2F

Henderson, M. M., Rademaker, R. L., & Serences, J. T. (2022). Flexible utilization of spatial- and motor-based codes for the storage of visuo-spatial information. *eLife*, *11*, e75688. https://doi.org/10.7554/eLife.75688

Holm, L., & Mäntylä, T. (2007). Memory for scenes: Refixations reflect retrieval. *Memory & Cognition*, *35*(7), 1664–1674. https://doi.org/10.3758/BF03193500

Hustá, C., Dalmaijer, E., Belopolsky, A., & Mathôt, S. (2019). The pupillary light response reflects visual working memory content. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(11), 1522–1528. https://doi.org/10.1037/xhp0000689

Jarodzka, H., Holmqvist, K., & Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 211–218. https://doi.org/10.1145/1743666.1743718

Johansson, R., & Johansson, M. (2014). Look Here, Eye Movements Play a Functional Role in Memory Retrieval. *Psychological Science*, *25*(1), 236–242. https://doi.org/10.1177/0956797613498260

Johansson, R., & Johansson, M. (2020). Gaze position regulates memory accessibility during competitive memory retrieval. *Cognition*, *197*, 104169. https://doi.org/10.1016/j.cognition.2019.104169

Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*, *154*(3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kay, L., Keogh, R., Andrillon, T., & Pearson, J. (2022). The pupillary light response as a physiological index of aphantasia, sensory and phenomenological imagery strength. *eLife*, *11*, e72484. https://doi.org/10.7554/eLife.72484

Keogh, R., & Pearson, J. (2018). The blind mind: No sensory visual imagery in aphantasia. *Cortex*, *105*, 53–60. https://doi.org/10.1016/j.cortex.2017.10.012

Kiyonaga, A., & Egner, T. (2013). Working memory as internal attention: Toward an integrative account of internal and external selection processes. *Psychonomic Bulletin & Review*, *20*(2), 228–242. https://doi.org/10.3758/s13423-012-0359-y

Korda, Ž., Walcher, S., Körner, C., & Benedek, M. (2024). Internal coupling: Eye behavior coupled to visual imagery. *Neuroscience & Biobehavioral Reviews*, *165*, 105855. https://doi.org/10.1016/j.neubiorev.2024.105855

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems*

Neuroscience, *2*. https://doi.org/10.3389/neuro.06.004.2008

Laeng, B., Bloem, I. M., D'Ascenzo, S., & Tommasi, L. (2014). Scrutinizing visual images: The role of gaze in mental imagery and memory. *Cognition*, *131*(2), 263–283. https://doi.org/10.1016/j.cognition.2014.01.003

Leavitt, M. L., Mendoza-Halliday, D., & Martinez-Trujillo, J. C. (2017). Sustained Activity Encoding Working Memories: Not Fully Distributed. *Trends in Neurosciences*, *40*(6), 328–346. https://doi.org/10.1016/j.tins.2017.04.004

Lee, S.-H., Kravitz, D. J., & Baker, C. I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. *Nature Neuroscience*, *16*(8), 997–999. https://doi.org/10.1038/nn.3452

Linde-Domingo, J., & Spitzer, B. (2024). Geometry of visuospatial working memory information in miniature gaze patterns. *Nature Human Behaviour*, *8*(2), 336–348. https://doi.org/10.1038/s41562-023-01737-z

Linde-Domingo, J., Treder, M. S., Kerrén, C., & Wimber, M. (2019). Evidence that neural information flow is reversed between object perception and object reconstruction from memory. *Nature Communications*, *10*(1), 179. https://doi.org/10.1038/s41467-018-08080-2

Liu, B., Alexopoulou, Z.-S., & van Ede, F. (2024). Jointly looking to the past and the future in visual working memory. *eLife*, *12*, RP90874. https://doi.org/10.7554/eLife.90874

Liu, B., Nobre, A. C., & van Ede, F. (2022). Functional but not obligatory link between microsaccades and neural modulation by covert spatial attention. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-31217-3

Loaiza, V. M., & Souza, A. S. (2022). The eyes don't have it: Eye movements are unlikely to reflect refreshing in working memory. *PLOS ONE*, *17*(7), e0271116. https://doi.org/10.1371/journal.pone.0271116

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods*, *50*(1), 94–106. https://doi.org/10.3758/s13428-017-1007-2

Mathôt, S., & Ivanov, Y. (2019). The effect of pupil size and peripheral brightness on detection and discrimination performance. *PeerJ*, *7*, e8220. https://doi.org/10.7717/peerj.8220

Mathôt, S., & Vilotijević, A. (2023). Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis. *Behavior Research Methods*, *55*(6), 3055–3077. https://doi.org/10.3758/s13428-022-01957-7

McAteer, S. M., McGregor, A., & Smith, D. T. (2023). Oculomotor rehearsal in visuospatial working memory. *Attention, Perception, & Psychophysics*, *85*(1), 261–275. https://doi.org/10.3758/s13414-022-02601-4

Miller, J. A., Kiyonaga, A., Ivry, R. B., & D'Esposito, M. (2020). Prioritized verbal working memory content biases ongoing action. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(12), 1443–1457. https://doi.org/10.1037/xhp0000868

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nobre, A. C., & Stokes, M. G. (2019). Premembering Experience: A Hierarchy of Time-Scales for Proactive Attention. *Neuron*, *104*(1), 132–146. https://doi.org/10.1016/j.neuron.2019.08.030

Oberauer, K., & Awh, E. (2022). Is There an Activity-silent Working Memory? *Journal of Cognitive Neuroscience*, *34*(12), 2360–2374. https://doi.org/10.1162/jocn_a_01917

Olivers, C. N. L., & Roelfsema, P. R. (2020). Attention for action in visual working memory. *Cortex*, *131*, 179–194. https://doi.org/10.1016/j.cortex.2020.07.011

Oster, J., Huang, J., White, B. J., Radach, R., Itti, L., Munoz, D. P., & Wang, C.-A. (2022). Pupillary responses to differences in luminance, color and set size. *Experimental Brain Research*, *240*(6), 1873–1885. https://doi.org/10.1007/s00221-022-06367-x

Piriyajitakonkij, M., Itthipuripat, S., Ballard, I., & Pappas, I. (2024). *What Makes a Face Look like a Hat: Decoupling Low-level and High-level Visual Properties with Image Triplets* (arXiv:2409.02241). arXiv. https://doi.org/10.48550/arXiv.2409.02241

Postle, B. R., Idzikowski, C., Sala, S. D., Logie, R. H., & Baddeley, A. D. (2006). The selective disruption of spatial working memory by eye movements. *Quarterly Journal of Experimental Psychology*, *59*(1), 100–120. https://doi.org/10.1080/17470210500151410

Schurgin, M. W. (2018). Visual memory, the long and the short of it: A review of visual working memory and long-term memory. *Attention, Perception, & Psychophysics*, *80*(5), 1035–1056. https://doi.org/10.3758/s13414-018-1522-y

Serences, J. T. (2016). Neural mechanisms of information storage in visual short-term memory. *Vision Research*, *128*, 53–67. https://doi.org/10.1016/j.visres.2016.09.010

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. https://doi.org/10.1111/j.1467-9280.2009.02276.x

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition* (arXiv:1409.1556). arXiv. https://doi.org/10.48550/arXiv.1409.1556

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, & Psychophysics*, *78*(7), 1839–1860. https://doi.org/10.3758/s13414-016-1108-5

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89. https://doi.org/10.1016/j.tics.2013.12.001

Stokes, M. G. (2015). "Activity-silent" working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Trentin, C., Falanga, L., Jeske, J., Olivers, C. N. L., & Slagter, H. A. (2024). Action similarity warps visual feature space in working memory. *Proceedings of the National Academy of Sciences*, *121*(49), e2413433121. https://doi.org/10.1073/pnas.2413433121

Trentin, C., Slagter, H. A., & Olivers, C. N. L. (2023). Visual working memory representations bias attention more when they are the target of an action plan. *Cognition*, *230*, 105274. https://doi.org/10.1016/j.cognition.2022.105274

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, *25*(6), 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

van Ede, F., Chekroud, S. R., Stokes, M. G., & Nobre, A. C. (2019). Concurrent visual and motor selection during visual working memory guided action. *Nature Neuroscience*, *22*(3), 477–483. https://doi.org/10.1038/s41593-018-0335-6

van Ede, F., Deden, J., & Nobre, A. C. (2021). Looking ahead in working memory to guide sequential behaviour. *Current Biology*, *31*(12), R779–R780. https://doi.org/10.1016/j.cub.2021.04.063

Vries, E. de, & Ede, F. van. (2023). *Microsaccades track location-based object rehearsal in visual working memory* (p. 2023.03.21.533618). bioRxiv. https://doi.org/10.1101/2023.03.21.533618

Wagner, A., Halchenko, Y., & Hanke, M. (2019). multimatch-gaze: The MultiMatch algorithm for gaze path comparison in Python. *Journal of Open Source Software*, *4*(40), 1525. https://doi.org/10.21105/joss.01525

Walcher, S., Korda, Ž., Körner, C., & Benedek, M. (2024). How workload and availability of spatial reference shape eye movement coupling in visuospatial working memory. *Cognition*, *249*, 105815. https://doi.org/10.1016/j.cognition.2024.105815

Wynn, J. S., Ryan, J. D., & Buchsbaum, B. R. (2020). Eye movements support behavioral pattern completion. *Proceedings of the National Academy of Sciences*, *117*(11), 6246–6254. https://doi.org/10.1073/pnas.1917586117

Wynn, J. S., Shen, K., & Ryan, J. D. (2019). Eye Movements Actively Reinstate Spatiotemporal Mnemonic Content. *Vision*, *3*(2), Article 2. https://doi.org/10.3390/vision3020021

Xu, Y. (2017). Reevaluating the Sensory Account of Visual Working Memory Storage. *Trends in*

Cognitive Sciences, *21*(10), 794–815. https://doi.org/10.1016/j.tics.2017.06.013

Yang, S., Dong, Y., & Kiyonaga, A. (2025). Response modality modulates spatial and temporal biases in visual working memory. *Journal of Vision*, *25*(9), 2906. https://doi.org/10.1167/jov.25.9.2906

Zokaei, N., Board, A. G., Manohar, S. G., & Nobre, A. C. (2019). Modulation of the pupillary response by the content of visual working memory. *Proceedings of the National Academy of Sciences*, *116*(45), 22802–22810. https://doi.org/10.1073/pnas.1909959116