Manipulating attentional priority creates a trade-off between memory and sensory representations in human visual cortex

Rosanne L. Rademaker[1,2] & John T. Serences[2,3]

[1] Ernst Strüngmann Institute for Neuroscience in cooperation with the Max Planck Society, Frankfurt, Germany
[2] Department of Psychology, University of California San Diego, La Jolla, California, USA
[3] Neurosciences Graduate Program, University of California San Diego, La Jolla, California, USA

Running title: attention and memory trade-off
Correspondence: rosanne.rademaker@gmail.com
Conflict of interest: the authors declare no conflict of interest
Main text words count: 7444.
Number of figures: 3 (and 4 Supplementary Figures)

Keywords: visual working memory, distraction, visual attention, fMRI, decoding.

Author Notes: This work was supported by NEI R01-EY025872 (JTS), and by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Grant Agreement No 743941 (RLR).

Abstract

People often remember visual information over brief delays while actively engaging with ongoing inputs from the surrounding visual environment. Depending on the situation, one might prioritize mnemonic contents (i.e., remembering details of a past event), or preferentially attend sensory inputs (i.e., minding traffic while crossing a street). Previous fMRI work has shown that early sensory regions can simultaneously represent both mnemonic and passively viewed sensory information. Here we test the limits of such simultaneity by manipulating attention towards sensory distractors during a working memory task performed by human subjects during fMRI scanning. Participants remembered the orientation of a target grating while a distractor grating was shown during the middle portion of the memory delay. Critically, there were several subtle changes in the contrast and the orientation of the distractor, and participants were cued to either ignore the distractor, detect a change in contrast, or detect a change in orientation. Despite sensory stimulation being matched in all three conditions, the fidelity of memory representations in early visual cortex was highest when the distractor was ignored, intermediate when participants attended distractor contrast, and lowest when participants attended the orientation of the distractor during the delay. In contrast, the fidelity of distractor representations was lowest when ignoring the distractor, intermediate when attending distractor-contrast, and highest when attending distractor-orientation. These data suggest a trade-off in early sensory representations when engaging top-down feedback to attend both seen and remembered features and may partially explain memory failures that occur when subjects are distracted by external events.

Introduction

Interacting with a dynamically changing environment routinely requires the temporary mental storage of relevant information over short periods of time to guide adaptive behaviors – a process that is often referred to as working memory (WM). Depending on task demands, priority can be shifted between maintaining internal representations of recent experiences in memory, and processing new sensory inputs from the outside world. For example, if you're meeting a date for the first time, you may want to hold an image of their face actively in mind as you enter the coffee shop, as not to accidentally miss them. However, it is also important to selectively attend your visual surroundings, so that you don't crash into other coffee shop patrons who are moving about in their singular quest for caffeine. Strategically re-prioritizing either memories of recently viewed information, or directing attention to external visual inputs, is critical to successfully navigating everyday situations.

The abilities to selectively attend to and remember sensory stimuli are thought to be mediated by top-down feedback signals. Specifically, feedback from anterior cortical regions such as the prefrontal cortex (PFC) can modulate neural activity in sensory areas such as early visual cortex (EVC). For instance, selectively attending to a relevant feature, such as the eye color of your new date, leads to enhanced firing rates in sensory neurons that are selectively tuned to that feature (along with modulations in neural variability and population covariance structure, (Martinez-Trujillo and Treue, 2004; David et al., 2008; Jehee et al., 2011; Liu et al., 2011; Rust and Cohen, 2022), and can improve visual sensitivity for the attended feature (Carrasco et al., 2004; Wolfe et al., 2011). In addition to supporting the prioritization of relevant sensory inputs, top-down modulations of EVC have also been implicated as a mechanism that supports WM in the absence of sensory inputs. When a relevant stimulus is no longer present in the environment, top-down signals can engage cortical areas specialized for processing sensory inputs to support detailed memories – an idea termed the *sensory recruitment hypothesis*. For example, auditory cortex can be recruited to remember an exact pitch (Czoschke et al., 2021; Deutsch et al., 2023), holding an object in mind can recruit object-selective inferotemporal (IT) cortex (Miyashita and Chang, 1988; Miller et al., 1991, 1993; Hirabayashi et al., 2013), and the short-term maintenance of simple visual features, such as particular colors or visual orientations, can involve recruitment of early visual cortex (Harrison and Tong, 2009; Serences et al., 2009; Yan et al., 2023). Recruitment of early sensory areas could have functional relevance, as it associated with better memory recall performance (Ester et al., 2013; Iamshchinina et al., 2021).

Although selective attention and working memory are thought to recruit the same general areas of early sensory cortex (Van Kerkoerle et al., 2017), it is unclear if they operate via similar or distinct mechanisms and the extent to which they compete. The degree of overlap and competition is unclear because selective attention and working memory have often been studied in isolation. For example, fMRI studies focused on WM typically use long delay periods that are free from other visual inputs and thus do not mimic real-world scenarios where new sensory inputs are constantly apprehended by the retina and may require your attention (Harrison and Tong, 2009; Serences et al., 2009; Christophel et al.,

2012, 2018; Ester et al., 2015; Sprague et al., 2016; Kwak and Curtis, 2022). Of course, in real life there are many other disruptions such as saccadic eye movements, which often go hand-in-hand with attention. Indeed, it's been argued that using the same cortical resources to concurrently encode new sensory inputs and store mnemonic information could lead to interference and confusion between seen and remembered stimuli (Xu, 2020). Prior fMRI studies found that presenting irrelevant but behaviorally distracting stimuli during the delay period of a WM task interfered with representations of low-level stimulus features in EVC (Bettencourt and Xu, 2016; Rademaker et al., 2019). In addition, performing a concurrent visual search task can disrupt memory representations of high-level objects such as faces in object-selective areas of ventral visual cortex (Kiyonaga et al., 2017). However, subsequent work found that when salient visual distractors failed to impair WM performance that interference at the level of EVC is not obligatory, demonstrating that both behavioral performance and EVC representations can be highly resilient to concurrent visual inputs (Rademaker et al., 2019). Thus, evidence for the degree to which EVC can simultaneously encode new sensory inputs and mnemonic information is mixed. It appears that only when people's memory performance is negatively impacted, implying their attention was drawn away from their memory contents, interference arises at the level of EVC. This suggests that attention and memory may recruit overlapping cortical resources, eliciting a trade-off between these competing top-down demands.

Here we test the hypothesis that interference may depend on how much priority is strategically allocated to sensory versus mnemonic information. We manipulate concurrent processing demands during the memory delay while using fMRI to measure activation patterns in EVC. We show that activation patterns support decoding of both perceived and remembered information, but that there is a trade-off: Ignoring new sensory inputs preserves behavioral performance and the integrity of mnemonic representations in visual cortex, whereas attending to new sensory inputs leads to lower behavioral performance and disrupted memory-related activation patterns. Together, the behavioral and fMRI results suggest that EVC plays a role in maintaining high-fidelity representations in WM and that disrupting these representations via the concurrent engagement of visual attention interferes with the successful retention of information in WM.

Materials and Methods

Participants. Nine volunteers (7 female) between the ages of 21 and 32 years (SD = 3.67) participated in the experiment. Participants had varying amounts of experience with fMRI experiments, ranging from scanner-naïve (S03 and S08) to highly experienced (with > 10 hours in the scanner; S04, S05, and S07). Of our nine volunteers, only the first eight are included for further analyses, as data collection for S09 was stopped at the start of the Covid-19 pandemic. Each of the included volunteers participated in a behavioral training in the lab, and 4–5 testing sessions in the fMRI scanner. The study was conducted at the University of California, San Diego (UCSD), and approved by the local Institutional Review Board. All participants provided written informed consent, had normal or

corrected-to-normal vision, and received monetary reimbursement for their time ($20 an hour).

Stimuli and procedure: Main fMRI task. All stimuli were projected on a 16 x 21.3 cm screen placed inside the scanner bore, viewed from ~40 cm through a tilted mirror. Stimuli were generated using Ubuntu 14.04, Matlab 2017b (Natick, MA), and the Psychophysics toolbox (Brainard, 1997; Kleiner et al., 2007). During the main experiment, memory targets were full contrast donut-shaped sinusoidal grating stimuli (0.75º inner and 10.5º outer radius) with smoothed edges (0.5º kernel with sd = 0.26º), spatial frequency of 2 cycles/º, and mean luminance equal to the grey background. Distractor grating stimuli shown during the delay had the same specifications, with the exception that their contrast was 50% Michelson. Target and distractor grating orientations were pseudo-randomly chosen from one of six orientation bins to ensure a roughly uniform sampling of orientation space (**Supp. Fig. 1**). Importantly, the target and distractor orientations were independent and counterbalanced, such that there was no systematic relationship between them across trials. This counterbalancing was done across 9 consecutive runs of the main task. The recall probe consisted of two black line segments (each 9.25º long and 0.04º wide) such that together the segments spanned the same eccentricity from fixation as the donut-shaped grating stimuli. A central black dot (0.4º) aided fixation throughout.

On every trial we randomly chose a spatial phase for the memory target and the distractor grating (both 0–2π). Each initial grating stimulus was then toggled back and forth between its original and inverted contrast at 4Hz, without blank gaps in between, for as long as the grating was on the screen. Thus, the memory target (500 ms total duration) cycled through 1 contrast-reversal (i.e., 250 ms per contrast), such that afterimages were minimized. Similarly, distractors (11s total duration) contrast-reversed for 22 cycles.

Each trial of the main experiment (**Fig. 1a**) started with a 1.6s change in the color of the central fixation dot, indicating with 100% validity the attention condition during the delay (i.e., ignore the distractor, attend changes in distractor contrast, attend changes in distractor orientation). Cues could be blue, green, or red. The pairing of cue-colors with attention-conditions was randomized across participants. Following the cue, a memory target was shown for 500 ms and participants remembered its orientation over a 15 second delay. A distractor grating was presented for 11 seconds during the middle portion of the delay. Irrespective of the attention condition, there would be 2–4 changes in the contrast of the distractor (lower or higher), and 2–4 changes in the orientation of the distractor (counterclockwise or clockwise) on every trial, with each such change lasting for 250ms. The total number of changes for each feature (contrast and orientation) was counterbalanced across all 3 attention conditions. After the delay, participants used four buttons to rotate a recall probe around fixation, matching the remembered orientation as precisely as possible. The left two buttons rotated the line counterclockwise, while the right two buttons rotated it clockwise. Using the outer- or inner-most buttons would result in faster or slower rotation of the recall probe, respectively. Participants had 3 seconds to respond before being presented with the next memory target 3, 5, or 8 seconds later. Each run of the main experiment consisted of 12 trials, and lasted 5 minutes and 3 seconds. Data for 36 total runs (432 total trials, with 144 trials per condition) were acquired

across 4 separate scanning sessions for all participants (except for S07, for whom all data of the main experiment were collected within the span of 3 longer scanning sessions).

Before starting the fMRI experiment, participants practiced the main task outside the scanner until they were comfortable using the response buttons to recall the target orientation within the temporally restricted response window; their mean absolute response error for the memory target was <10°; and they were >90% accurate for supra-threshold changes on the distractor attention tasks (both contrast and orientation). Practice took between 5 and 28 blocks of trials, performed across 1–4 separate days.

To equate the difficulty of the two distractor attention tasks, we measured each participant's performance threshold for both the contrast and orientation attention tasks. This thresholding task was identical to the main experiment, with two exceptions: (1) The "ignore distractor" attention condition was skipped in the interest of time, and (2) a different Δ contrast or orientation was used for every change of the distractor. Specifically, we used Quest (Watson and Pelli, 1983) to determine the magnitude of change (i.e., contrast increase or decrease; counterclockwise or clockwise orientation deviation) at which participant's discrimination performance was ~75% correct. To get an initial threshold estimate, participants completed 1–4 thresholding runs in the lab, prior to scanning. Subsequently, participants also completed 1–2 additional runs in the scanner prior to each scan session (so while the scanner was still off). This was important for the contrast threshold in particular, since thresholds strongly depend on the specific screen that stimuli are presented on. Once established, these thresholds ensured identical visual inputs irrespective of the attention condition, as the magnitude of contrast and orientation changes was kept constant for each set of 9 consecutive runs of the main experiment. Across participants, the average Δ contrast = 0.116 (SD = 0.075) and the average Δ orientation = 2.074° (SD = 0.547°). As intended, performance on the contrast task (71.85%; SD = 5.848%) and performance on the orientation task (71.56%; SD = 5.718%) did not significantly differ between tasks ($t_{(7)}$ = 0.157; p = 0.858; **Fig. 1b**).

Stimuli and procedure: Localizer tasks. In addition to runs for the main task, we also collected data from an independent sensory localizer task, and an independent memory localizer task. The sensory localizer was used for voxel selection. Both sensory and memory localizer tasks were used for model training purposes (i.e., the analyses shown in **Fig. 3**).

The sensory localizer task consisted of trials showing either a circle-shaped (0.75° radius) or donut-shaped (0.75° inner and 10.5° outer radius) full-contrast grating. On every trial, this grating was shown for 6 seconds (contrast-reversing as in the main experiment), followed by a 3, 5, or 8s inter-trial interval. During grating presentation, a small grey circle was superimposed on the stimulus occasionally (0–3 times per trial) for 250ms. These brief 'blobs' could be centered at any distance from fixation between 0.056° and 13.78°, and at any angle relative to fixation (1–360°). Blobs were scaled for cortical magnification such that all blobs stimulated roughly 1mm of cortex (i.e., blobs had radii between 0.18° to 0.75° of visual angle). No blobs were presented during the first 500ms or the last 500ms of a trial, or within 500ms of each other. The blobs functioned to keep participant's

attention directed at the general spatial location occupied by the grating stimuli. Participants were instructed to indicate detection of a blob via a button press. Grating orientation was randomly chosen from one of 9 orientation bins on each trial, and equally often from each bin within a run. Participants completed 36 trials per run (18 circle-shaped grating trials, and 18 donut-shaped grating trials, randomly interleaved), and each run took 7 minutes and 5s. Participants completed between 12–23 total runs of this sensory localizer. These sensory localizer data were obtained in parallel with data collection for another study (Experiment 2 in Rademaker et al., 2019) and have therefore also been used in this previous work.

The memory localizer task was used to estimate voxel responses while participants were remembering an orientation and *not concurrently viewing* any visual inputs (in the main task, memory is always concurrent with visual input from the distractor). During the memory localizer task, participants remembered a briefly presented grating (500 ms; 11.5^{o} radius; full contrast; contrast-reversing as in the main experiment; pseudo random orientation chosen from 1 of 6 bins) over a 12 second blank delay, after which they recalled the orientation by rotating a dial within a 4 s time window. Each trial was preceded by a 1.4s change in the color of the fixation dot (to alert subjects of the upcoming target and indicate the attention condition), and each trial was followed by a 3, 5, or 8s inter trial interval. For five participants (S01, S04, S05, S06, and S07), these data were collected as part of Experiment 2 in Rademaker et al. (2019), which included trials with and without visual distractors presented during the memory delay. For our current purposes, only the 108 trials with a blank delay period were used (out of the original 324 total trials, collected across 27 total runs). The remaining three participants (S02, S03, and S08) completed 108 total trials of this memory localizer task across a total of 9 runs, and only performed trials without any visual distractors presented during the delay period. Thus, for all participants we have 108 trials with a blank delay period that we use as our memory localizer.

Magnetic resonance imaging. All scans were performed on a General Electric Discovery MR750 3.0T scanner located at the UCSD Keck Center for Functional Magnetic Resonance Imaging (CFMRI). High-resolution (1 mm^3 isotropic) anatomical images were acquired during a retinotopic mapping session, using an Invivo eight-channel head coil. Functional echo-planar imaging (EPI) data for the current experiment were acquired using a Nova Medical 32-channel head coil (NMSC075-32-3GE-MR750) and the Stanford Simultaneous Multi-Slice EPI sequence (MUX EPI), using nine axial slices per band and a multiband factor of eight (total slices = 72; 2 mm^3 isotropic; 0 mm gap; matrix = 104 × 104; field of view = 20.8 cm; repetition time/echo time (TR/TE) = 800/35 ms, flip angle = 52°; inplane acceleration = 1). At sequence onset, the initial 16 TRs served as reference images critical to the transformation from k-space to image space. Un-aliasing and image reconstruction procedures were performed on local servers using CNI-based reconstruction code. Forward and reverse phase-encoding directions were used during the acquisition of two short (17 s) ‘topup’ datasets. From these images, susceptibility-induced off-resonance fields were estimated (Andersson et al., 2003) and

used to correct signal distortion inherent in EPI sequences, using FSL topup (Smith et al., 2004; Jenkinson et al., 2012).

Preprocessing. All imaging data were preprocessed using software tools developed and distributed by FreeSurfer and FSL (free to download at https://surfer.nmr.mgh.harvard.edu and http://www.fmrib.ox.ac.uk/fsl). Cortical surface gray-white matter volumetric segmentation of the high-resolution anatomical image was performed using the 'recon-all' utility in the FreeSurfer analysis suite (Dale et al., 1999). Segmented T1 data were used to generate inflated surfaces on which to draw retinotopic ROIs for use in subsequent analyses. Segmented T1 data were also used for coregistration to the functional data: The first volume of every first functional run of a scanning session was coregistered to the anatomical image. Transformation matrices were generated using FreeSurfer's manual and boundary-based registration tools (Greve and Fischl, 2009). These matrices were then used to transform each four-dimensional functional volume using FSL FLIRT (Jenkinson and Smith, 2001; Jenkinson et al., 2002), such that all across-session data from each single participant was spatially aligned. Next, motion correction was performed using the FSL tool MCFLIRT (Jenkinson et al., 2002) without spatial smoothing, a final sinc interpolation stage, and 12 degrees of freedom. Slow drifts in the data were removed last, using a high pass filer (1/40 Hz cutoff). No additional spatial smoothing was applied to the data apart from the smoothing inherent to resampling and motion correction. Signal amplitude time-series were normalized via Z-scoring on a voxel-by-voxel and run-by-run basis. Z-scored data were used for all further analyses. Because stimulus onsets were jittered with respect to TR, we aligned the onset of every stimulus to the center of the nearest TR (such that stimuli that started at any time poit between -400–400 ms are plotted as time = 0) for every trial (and for every task).

To recover the univariate BOLD time courses for all three attention conditions in the main memory experiment (**Fig. 1c**; **Supp. Fig. 2, top**), and separately for the memory localizer task (**Supp. Fig. 2, bottom**), we estimated the hemodynamic response function for each voxel at each time point of interest (27 TR's from memory target onset). This was done using a finite impulse response function model (Dale, 1999) consisting of a column marking the onset of each event (memory target onset) with a '1', and then a series of temporally shifted versions of that initial regressor in subsequent columns to model the BOLD response at each subsequent time point. Estimated hemodynamic response functions were then averaged across all voxels in each ROI.

To compute average voxel responses for each trial, to be used for subsequent multivariate analyses, we obtained average activity for the delay period of the main experiment by calculating the mean activation from 5.6–15.2s after target onset (i.e., TR's 8–20) for every voxel. For the independent sensory localizer, average responses were calculated over a time window of 3.2–9.6s after grating onset (i.e., TR's 5–13), and for the independent memory localizer we took the average delay period activation from 5.6–12s after memory target onset (TR's 8–16).

Identifying ROI's. Standard retinotopic mapping procedures (Engel et al., 1994; Swisher et al., 2007) were employed to define eight a priori ROI's in early visual (V1–V3, V3AB,

hV4), parietal (IPS0, IPS1–IPS3), and ventral visual (LO1-2) cortex. Retinotopic mapping data were collected during an independent scanning session, using both meridian (i.e., bowtie-shaped checkerboard stimuli shown at either the horizontal or vertical meridian) and polar angle mapping (a slowly rotating wedge-shaped checkerboard stimulus) to identify voxel's visual field preferences (described in more detail in Sprague and Serences, 2013). Analyses to obtain retinotopic maps used a set of custom wrappers around existing FreeSurfer and FSL functionality.

To identify voxels that were visually responsive to the part of the visual field where our donut-shaped grating stimuli were shown, a general linear model was applied to data from the sensory localizer using FSL FEAT (FMRI Expert Analysis Tool, v.6.00). Individual localizer runs were analyzed using the brain extraction tool (Smith, 2002) and data prewhitening using FILM (Woolrich et al., 2001). Predicted BOLD responses were generated for each sensory localizer run by convolving the stimulus sequence (of "donut" and "circle" stimuli) with a canonical gamma hemodynamic response function (phase = 0 s, SD = 3 s, lag = 6 s). The temporal derivative was included as an additional regressor to accommodate slight temporal shifts in the waveform to yield better model fits and to increase explained variance. Individual runs were combined using a standard weighted fixed effects model. Voxels that were significantly more activated by the donut compared to the circle stimulus (P = 0.05; false discovery rate corrected) were defined as visually responsive. Only these visually responsive voxels from each ROI were used for subsequent analyses. Exact voxel counts for each participant in each ROI can be found in **Supplementary Table 1**.

fMRI analyses. For decoding, we used an inverted encoding model (IEM). We chose this approach because it can be used to model continuous feature spaces, in this case orientation, and does not require discrete binning. For our main analyses (**Fig. 2** and **Supp. Fig. 3**) model training and testing was performed using a 4-fold cross-validation procedure. Specifically, each observer participated in 36 runs of the main memory task, with complete counterbalancing of orientation bins and task conditions per set of 9 runs (a single session). Thus, for model training and testing on each fold, 27 runs served as the training set, and 9 runs were held out as the test set. Each set of 9 runs was held out once. Model training was done on the average pattern activity across delay period TR's, and across all 3 attention conditions to ensure we use the same training data throughout. For the results in **Fig. 3** we trained the model on data from independent localizers.

The first step of the IEM estimates an encoding model by modelling the response in each voxel to 9 raised cosine orientation filters, or "channels", to characterize the orientation sensitivity profile for each voxel based on training data. The encoding model for a single voxel has the general form:

$$R_j = \sum_{i}^{9} w_i c_i \qquad \text{Equation (1)}$$

Where R_j is the response R of voxel j, and c_i is the channel magnitude c at the i^{th} channel. A voxel's orientation sensitivity profile is captured by 9 weights w, one for each channel. Channels were modeled as:

$$c(d) = \cos((d - \mu) \cdot \frac{\pi}{180})^8 \qquad \text{Equation (2)}$$

Where d is the distance in degrees from the channel center μ. Channel centers were spaced 20º apart. Equation 1 can be expressed for matrices as:

$$B_1 = WC_1 \qquad \text{Equation (3)}$$

Here, a matrix of observed BOLD responses B_1 (*m* voxels x *n* trials) is related to a matrix of modeled channel responses C_1 (*k* channels x *n* trials) by a weight matrix W (*m* voxels x *k* channels). For each trial, C_1 is the pointwise product of a stimulus mask (i.e., "1" at the true stimulus orientation, "0" at all other orientations) by channels. W quantifies the sensitivity of each voxel at each idealized orientation channel, and can be computed with least-squares linear regression:

$$\hat{W} = B_1 C_1^T (C_1 C_1^T)^{-1} \qquad \text{Equation (4)}$$

Estimating the sensitivity profiles concludes the first encoding step of the IEM. The second step of the IEM inverts the model, using the estimated sensitivity profiles of all voxels $\hat{W}$ (*m* voxels x *k* channels) in combination with a test set of novel BOLD response data B_2 (*m* voxels x *n* trials) to estimate the amount of orientation information at each channel $\hat{C}_2$ (*k* channels x *n* trials):

$$\hat{C}_2 = \left(\hat{W}^T \hat{W}\right)^{-1} \hat{W}^T B_2 \qquad \text{Equation (5)}$$

This step uses the Moore-Penrose pseudoinverse of $\hat{W}$, and uses the sensitivity profiles across all voxels to jointly estimate channel responses $\hat{C}_2$ for each trial of the test set.

Grating orientations could take any integer value between 1º and 180º. To estimate channel responses $\hat{C}_2$ for each degree in orientation space, both the encoding (Equation 4) and inversion (Equation 5) steps of the IEM were repeated 20 times. On each repeat, the centers of the 9 channels (Equation 2) were shifted by 1º, and we estimated the channel responses $\hat{C}_2$ at those 9 centers, until the entire 180º orientation space was estimated in 1º steps. After generating model-based channel-responses for each trial, all single trial channel-responses were re-centered either on the remembered orientation, or on the orientation of the distractor grating.

Model-based channel-responses were quantified using a "vector mean" fidelity metric derived from trigonometry (Rademaker et al., 2019), and applied to the average of 144 single-trial reconstructions per attention condition (for each condition, participant, and ROI). This vector mean fidelity metric was calculated by convolving the channel response with a cosine, which is equivalent to projecting the channel response at each degree in

orientation space onto the center of that space (0°) and taking the mean over all 180 projected vectors. The vector mean reflects if there is modeled information about a remembered or perceived orientation in the channel response profile (it will be zero if not). Note that this metric, by design, gets rid of additive offsets.

Statistical procedures. To compare participant's behavioral recall performance across the 3 attention conditions we used a non-parametric 1-way ANOVA. Significant results were followed up with non-parametric two-sided t-tests to compare recall performance between each pair of conditions. The same t-test was used to compare performance on the contrast and orientation attention tasks. Specifically, non-parametric tests used 1000 permutations. On each permutation we randomly shuffled the condition labels belonging to the mean performance from each participant, and calculated the appropriate test statistic (an F-value in the case of the ANOVA; a t-value in the case of the t-test) across all participants. Across all 1000 permutations we obtained a null distribution of shuffled test-statistics, against which we compared the test statistic calculated from the intact data to determine the significance level.

To compare decoding performance across the 3 attention conditions and across our 8 ROI's, we performed a permutation-based 2-way ANOVA, which works similarly to the 1-way ANOVA described above (condition labels are shuffled, and a test statistic is calculated across 1000 permutations). In the case of a significant 2-way interaction between condition and ROI, we also test for differences between the 3 attention conditions *within* each ROI (again using non-parametric t-tests). Note that while technically there was only one such interaction (for distractor decoding in Fig 2a, right), we also performed these post-hoc tests for memory decoding (Fig 2a, left) for consistency in the figure. To follow up on main effects of attention condition (**Fig. 2a**, **Fig 3a**), we tested for differences between conditions by taking the average vector mean for each participant across all ROI's, and then performing pair-wise post-hoc tests between all 3 conditions using a non-parametric t-test as already described above. Finally, to follow up on main effects of ROI, we checked for significant above-chance decoding in each condition and ROI (**Fig. 3**) by performing one-sided non-parametric t-tests against zero.

To determine if univariate BOLD responses or decoding differed between the 3 attention conditions over time (i.e., TR's within a trial), we used a non-parametric 1-way ANOVA cluster-based permutation test. Specifically, to verify at which timepoints there were significant differences between the 3 attention conditions, we used a threshold of $p = 0.05$ ($F = 2.73$) and calculated the observed cluster mass (i.e., sum of F-values for adjacent timepoint that supersede the threshold) for every cluster in every ROI. Then, we generated a null-distribution of cluster mass sizes: Across 1000 permutations we shuffled the condition labels for the data in each ROI, while keeping the temporal structure of the data intact, taking the largest cluster mass on each permutation. We then compared the cluster mass at each *real* cluster in the data against this null-distribution. This cluster-based permutation test is performed for each ROI separately.

Glitches. We re-measured the screen size and viewing distances in the scanner since the publication of Rademaker et al., 2019, which led to slightly different stimulus sizes reported here for our sensory localizer, compared stimulus sizes reported in reported in Rademaker et al., 2019 (where the same data were used as "second mapping task" in Experiment 2). Partway through data collection, the bulb in the projector burnt out and needed to be replaced. After replacing the bulb, contrast thresholding had to be re-done for affected participants as the new bulb was considerably brighter than the old one. Finally, for S06, we had to abort what was supposed to be the 3rd session after only 3 runs of the main task. This entire session was re-run on another day so that S06 had a complete data set.

Results

First, we wanted to know if participant's performance on the main working memory task was impacted by concurrent attention to a perceived distractor. Indeed, recall performance while subjects were in the scanner differed significantly between the 3 attention conditions ($F_{(2,14)} = 5.889$, $p = 0.002$), with higher absolute recall errors when participants attended orientation changes in the distractor, compared to when they attended contrast changes ($t_{(7)} = 2.623$, $p = 0.018$) or ignored the distractor ($t_{(7)}=2.53$, $p = 0.022$). The absolute recall error did not differ significantly between trials during which participants attended contrast changes compared to when they ignored the distractor ($t_{(7)}=1.497$, $p = 0.182$).

Next, we looked at univariate responses in the main working memory task to see if distractor attention impacts BOLD time course. We show these univariate time courses for several ROI's in **Fig. 1c** (for all ROI time courses see **Supp. Fig. 2, top row**). It is clear that the 11s distractor drives a large and sustained BOLD response, especially in early visual areas (and when compared to the memory localizer, where the delay was without visual input, see **Supp. Fig. 2, bottom row**). Towards the end of the trial in particular there appears to be a larger univariate response when distractor attention is required, compared to when the distractor is ignored. Note that in V1 this difference between conditions only emerges 12.8s after the onset of the distractor, meaning that differences in response preparation (memory recall) could also play a role.

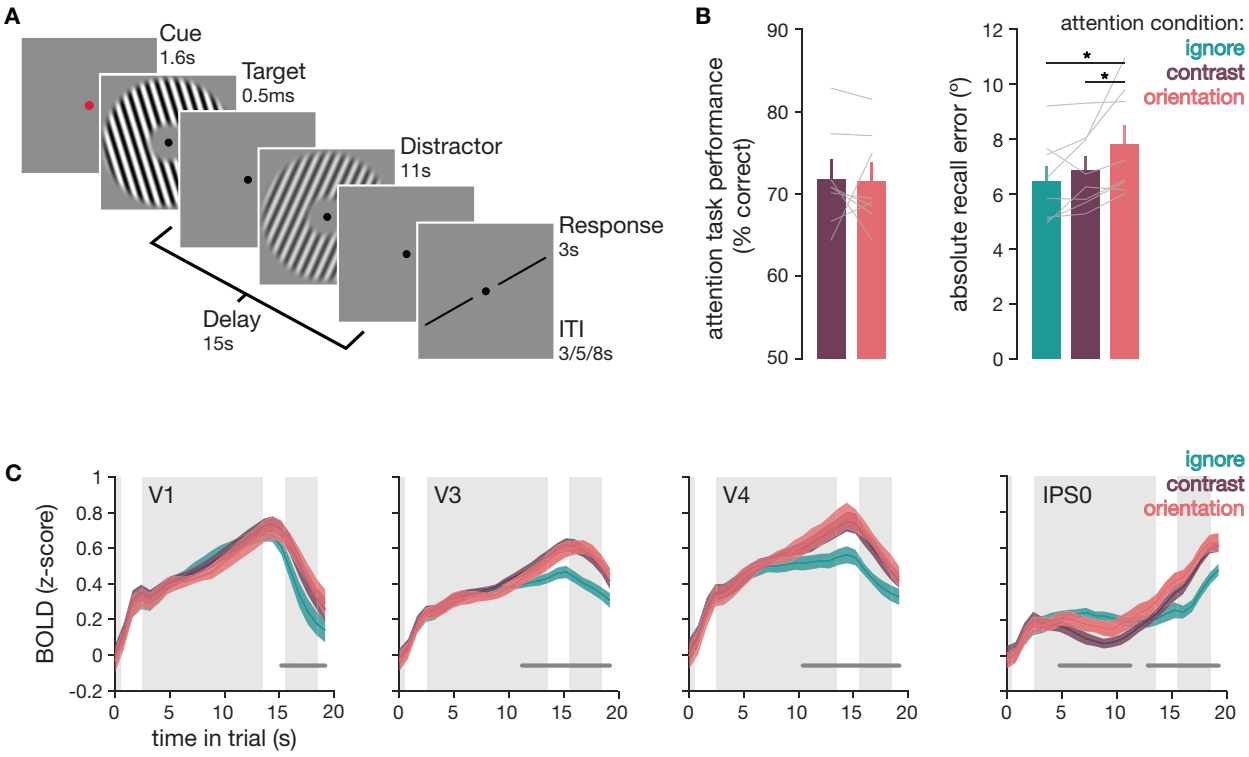


Figure 1. (**a**) Experimental task design. Participants remember the orientation of a brief (0.5s) memory target over a 15s delay, after which they have 3s to rotate a black dial to match the orientation in memory as precisely as possible. On every trial, a distractor grating is shown for 11s during the central-most portion of the delay. And on every trial, this distractor is phase-reversing, and has several small changes to its contrast and orientation. Right before the memory target, participants see a 1.6s cue indicating with 100% validity the upcoming attention condition. Specifically, they need to either (1) ignore the distractor grating, (2) attend and report contrast changes, or (3) attend and report orientation changes. When attending the distractor, participants also report the direction of each change (i.e., whether there is an increase / decrease in contrast, or a clockwise / counterclockwise change in orientation). There is an equal number of trials in each condition, and conditions are randomly interleaved. (**b**) Behavioral data recorded while participants were in the scanner shows that performance on the distractor attention task is well-matched ($t_{(7)} = 0.157$; $p = 0.858$), and participants perform similarly when detecting contrast (71.85% correct) or orientation (71.56% correct) changes (left panel). Recall of the memory target orientation did differ between conditions ($F_{(2,14)} = 5.889$, $p = 0.002$), and was generally worse when participants had to perform a concurrent orientation attention task on the distractor. Bars indicate average performance, and grey lines individual participants. Asterisks indicate significant post-hoc differences between conditions. (**c**) Deconvolved BOLD responses for a few example ROI. Distractors in all 3 attention conditions effectively drove univariate responses in early visual areas, with qualitatively higher responses when attention was deployed towards the distractor (shown in purple and pink for attention to distractor contrast or orientation, respectively). In IPS, responses seem more transient, with the strongest response occurring with attention to distractor orientation. Grey background-panels in each subplot indicate the time during which the memory target (0–0.5s, far left panel), the distractor (2.5–13.5s, middle panel), and the response-dial (15.5–18.5s, right most panel) were on the screen. Darker grey lines just above and parallel to the x-axis indicate clusters of consecutive TR's during which the three attention conditions differ significantly from one another (as calculated with a cluster based permutation test, see Methods).

Next, we tested if we could decode the orientation of the memory target that was held in working memory, while also concurrently decoding the orientation of the distractor grating

that was physically on the screen. Here, we used the average pattern response during the delay to train and test our decoder (an IEM, see Methods). The ability to decode the memory target orientation during the delay differed between attention conditions (main effect: $F_{(2,14)} = 10.814$; $p < 0.001$) which was the case for all ROI's (no condition x ROI interaction: $F_{(14,98)} = 1.024$; $p = 0.439$, and no main effect of ROI: $F_{(7,49)} = 1.048$; $p = 0.417$)(**Fig. 2a, left**). Specifically, the remembered orientation was better decoded when the distractor was ignored, compared to when participants attended changes in distractor contrast ($t_{(7)} = 3.925$; $p < 0.001$) or orientation ($t_{(7)} = 3.586$; $p = 0.012$), without a significant difference between the contrast and orientation attention conditions ($t_{(7)} = 0.212$; $p = 0.814$). These decoding results from visual cortex mirror the differences we see in recall behavior, where performance is also best when the distractor can be ignored.

The ability to decode the orientation of the distractor shown during the delay similarly differed depending on the attention condition (main effect: $F_{(2,14)} = 16.08$; $p < 0.001$), but these differences were not the same in all ROI's (condition x ROI interaction, $F_{(14,98)} = 2.155$; $p = 0.008$, and main effect of ROI: $F_{(7,49)} = 24.781$; $p < 0.001$)(**Fig. 2a, right**). Post-hoc tests in each ROI show that while we observed a difference between attention conditions in every ROI (all $F > 4.5$; all $p < 0.024$), all 3 attention conditions differed from one another in some ROI's but not in others. In general, decoding of the distractor grating was highest when participants had to pay attention to orientation changes – both compared to when participants paid attention to contrast ($t_{(7)} = 3.34$; $p < 0.00$) and when they ignored the distractor ($t_{(7)} = 3.34$; $p < 0.001$). Also attention to contrast changes generally improved decoding compared to when the distractor grating was ignored ($t_{(7)} = 2.961$; $p = 0.024$).

We subsequently evaluated how these differences evolve over time by looking at IEM decoding for each TR of the main working memory trial (**Fig. 2b**; **Supp. Fig. 3**). Despite the increase in noise of single TR data, we nevertheless observed time clusters during which the 3 attention conditions diverged, in line with the results in **Fig 2a**. Orientation decoding of both the memory target and distractor gratings differed depending on the attention condition, and these differences emerged around 6–8 seconds into the trial, likely reflecting the change in cognitive state at the onset of the distractor grating.

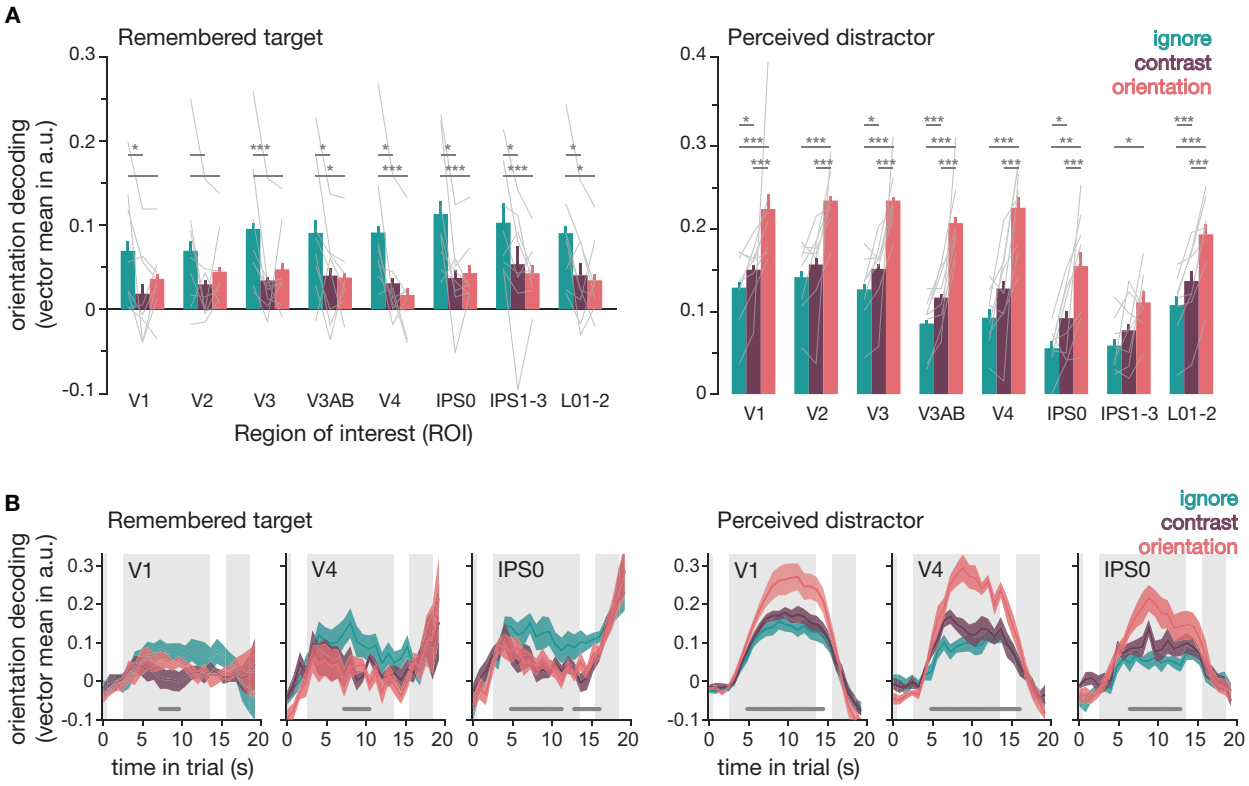


Figure 2. (**a**) Decoding of the target orientation that is held in memory (left) and of the distractor that was physically presented on the screen (right) during the delay period of the main working memory task. The remembered orientation is better decodable when the distractor is ignored (in teal), compared to when it is attended (in purple and pink). The perceived distractor orientation is least decodable when it is ignored (in teal), better decodable when its contrast is attended (in purple), and best decodable when its orientation is attended (in pink). Bars indicate mean orientation decoding averaged across all participants, while light grey lines indicate individual participants. Dark grey lines and asterisks indicate significant post-hoc differences (*p < 0.05; ** p< 0.01; ***p< 0.001) between attention conditions within each ROI (non-parametric t-tests). Note that for decoding of the remembered target, some significance lines are missing an asterisk indicating significance from within-ROI post-hoc t-tests. Due to the lack of an interaction between condition and ROI, these post-hoc t-tests are not technically warranted, and may be prone to type II errors. Rather, these significance lines indicate the post-hoc tests that compare conditions *across* all ROI, which follow the main effect of attention condition. (**b**) Same decoding as in (**a**) for a few example ROI's, but shown TR-by-TR throughout the trial of the main working memory task. For decoding time courses of all ROI's, see **Supp. Fig. 3**. Grey background-panels in each subplot indicate the memory target (0–0.5s, far left panel), distractor (2.5–13.5s, middle panel), and the response (15.5–18.5s, right most panel) periods. Dark grey lines just above and parallel to the x-axis indicate clusters of consecutive TR's during which the three attention conditions differ significantly from one another (as calculated with a cluster based permutation test, see Methods).

In the previous analysis, we trained our decoder across all 3 conditions of the main working memory task to see how well we could retrieve information about remembered and perceived orientations under different attentional conditions. The benefit of this is that we were not biasing our decoder in favor of any one condition. But there are also downsides to such an approach. For example, the signal-to-noise (SNR) ratio may be worse when we ask participants to perform two competing top-down tasks (i.e.,

remembering an orientation while also paying attention to concurrent visual input). By training our decoder on such data, we may not be able to uncover all possible differences between the 3 attention conditions. Another notable downside of training and testing a decoder within the same task is that we remain agnostic to the representational format used to remember or process visual inputs. We just know that activation patterns related to memory can predict the remembered orientation, but we do not know if those patterns are similar to memory representations under conditions without visual input. Similarly, we know that activation patterns related to perception can predict the perceived distractor orientation, but we do not know if those patterns are similar to sensory driven responses evoked under different circumstances. Thus, we next used data from two independent localizer tasks to see if response patterns generalize from other working memory and perception tasks to the patterns we measured in our main working memory task – when participants were concurrently remembering and perceiving two different orientations.

Do working memory representations measured *without* concurrent visual input (i.e., during a blank delay period) generalize to the memory representations in our main task, where orientation was remembered concurrently with visual distraction on every trial? Also when we train on data from the independent memory localizer, decoding of the memory target orientation differed between attention conditions (main effect attention condition: $F_{(2,14)} = 41.248$; $p < 0.001$), which was true for all ROI's (no condition x ROI interaction, $F_{(14,98)} = 1.683$; $p = 0.061$)(**Fig. 3a**). As before, ignoring the distractor resulted in better decoding of the orientation in memory compared to when distractor contrast ($t_{(7)}= 6.66$; $p < 0.001$) or orientation ($t_{(7)}= 6.803$; $p < 0.001$) had to be attended. The remembered orientation was represented marginally better when participants attended contrast as opposed to orientation changes in the distractor ($t_{(7)}= 2.311$; $p = 0.054$). Decoding performance also differed between ROI's (main effect ROI: $F_{(7,49)} = 9.877$; $p < 0.001$), and was generally higher in IPS. While this may arise from less interesting factors such as SNR differences between ROI (e.g., due to differences in the underlying organization of orientation selectivity), it is noteworthy that such differences were *not* observed when training and testing a decoder within the same task (**Fig 2a, left**). Post-hoc tests further showed above-chance memory decoding when distractor attention was required only in parietal cortex (IPS0 and IPS1–3) (**Fig 3a**). Thus, the ability to cross-generalize from memory during a blank delay, to memory during a delay with visual distraction, implies a common format for remembering orientation under these different circumstances. Importantly, this common format appears only when the distractor input is ignored, with the exception of parietal cortex, where we see that this overlap in representational format exists irrespective of whether the visual distractor was ignored or attended.

Second, we wanted to know if response patterns generalize from independently measured sensory driven responses, to sensory distractor representations measured under a concurrent working memory load and with different forms of distractor attention. During the sensory localizer, participants performed a behavioral task that was deliberately orthogonal to all 3 of the attention conditions in the main task. Specifically,

participants had to detect small gray blobs that could be superimposed anywhere on the sensory grating. This ensured that the localizer stimulus was not ignored, and also that neither contrast or orientation were attended. The ability to decode the distractor orientation during the delay did not differ between attention conditions (main effect attention condition: $F_{(2,14)} = 2.199$; $p = 0.133$) (**Fig. 3b**). The most probable reason for this is the complete lack of decoding in parietal areas IPS0 and IPS1–3 (main effect ROI: $F_{(7,49)} = 19.574$; $p < 0.001$). And while there is no significant interaction between attention condition and ROI ($F_{(14,98)} = 743$; $p = 0.741$) differences between conditions can be harder to uncover when decoding is absent in some ROI's (and cannot show any condition differences). It is interesting that, unlike in our main analysis (**Fig 2a**, right), there is no cross-generalization between the sensory localizer and the sensory distractor in IPS, consistent with a change in the representational format of different sensory inputs in IPS.

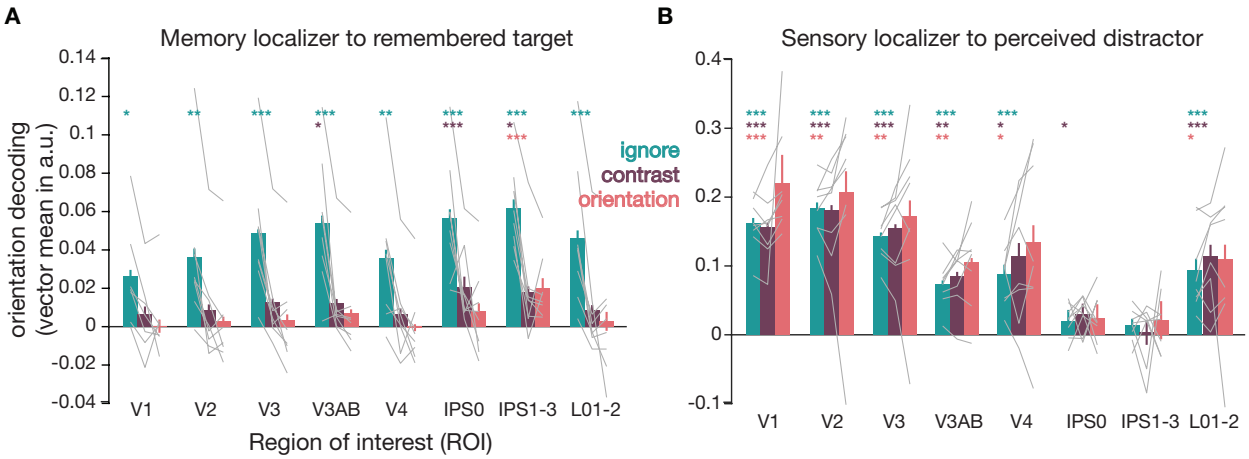


Figure 3. (**a**) Decoding of the target orientation that is held in memory, when training a decoder on an independent memory localizer task. The remembered orientation is better represented when the concurrent visual distractor is ignored (teal) than when its contrast (purple) or orientation (pink) are attended. Only in parietal areas (IPS0 and IPS1–3) is there cross-generalization from memory without visual input (i.e., the memory localizer task) to memory with concurrent visual input that is also attended (i.e., the main memory task when the distractor is also attended). (**b**) Decoding of the sensory distractor in the main memory task, when training a decoder on an independent sensory localizer. Response patterns from the sensory localizer task (which used a blob detection task) generalize to responses to the sensory distractor in the main memory task in most ROI, but not in IPS areas. Possibly because of this, no significant differences between the 3 attention conditions are uncovered. Bars indicate mean orientation decoding averaged across all participants, while light grey lines indicate individual participants. Asterisks in matching condition colors indicate significant post-hoc decoding performance (*$p < 0.05$; ** $p < 0.01$; ***$p < 0.001$) compared to chance.

Discussion

Holding in mind relevant sensory information for future use is a critical cognitive function, and to be useful these memories must be insulated from being overwritten by new sensory inputs. Prior studies in humans using fMRI, and in animal model systems using invasive recordings, suggest that the same general areas of early sensory cortex that

support early perceptual processing are also involved in maintaining sustained memory representations (Harrison and Tong, 2009; Serences et al., 2009; Van Kerkoerle et al., 2017; Yiling et al., 2024). This apparent multiplexing poses a computational challenge as it is unclear how remembered information interacts with new sensory inputs, especially those that are behaviorally relevant and require prioritization via top-down attentional modulations. Here we show that memory recall was relatively spared when an ignored distractor was presented during the delay period, similar to prior reports from our lab and others (Magnussen et al., 1991; Magnussen and Greenlee, 1992; Rademaker et al., 2015, 2019). However, when new inputs were behaviorally relevant and participants engaged in a contrast or orientation detection task, there was a significant drop in both behavioral and neural markers of memory fidelity (**Fig. 1b** & **2a**). Importantly, the decline in the precision of mnemonic representations was accompanied by higher fidelity representations of the distractors, suggesting that attended and remembered information directly competes in areas of early visual cortex (**Fig. 2a** & **3**).

The present results help reconcile prior conflicting reports that presenting distractors during a memory delay interferes with WM representations. For example, in one experiment Bettencourt and Xu (2016) found that activity in visual cortex associated with a remembered orientation was interrupted by the presentation of faces and buildings (gazebos) during the delay period. Rademaker et al. (2019) replicated this result using similar stimuli, but also found that presenting another oriented stimulus during the delay period did not always lead to interference. Based on the present observation that directly manipulating attention reveals graded levels of distractor interference, we speculate that some classes of stimuli – such as faces – may be inherently more interesting and thus more likely to attract attention compared to simple objects comprised of a low-level visual feature like an orientation (Davies and Hoffman, 2002; Langton et al., 2008; Hodsoll et al., 2011; Morrisey et al., 2019)

This framework is also consistent with models of attentional control proposing that attended stimuli automatically gain access to WM and thus have the potential to interact with or disrupt pre-existing WM representations (Bundesen et al., 2005). That said, more recent work in the domain of attentional capture suggests that attended items do not always enter WM: irrelevant distractors can attract spatial attention but be gated out of WM whereas potentially relevant distractors attract spatial attention and are represented in WM (Hakim et al., 2019; Maxwell et al., 2021). For instance, Hakim et al (2021) demonstrated that markers of spatial attention, such as the amplitude of alpha oscillations measured with EEG, track the position of distractors irrespective of their potential relevance. In contrast, markers of encoding into WM, such as the contra-lateral delay activity (CDA) measured with EEG, indicate that only potentially relevant stimuli are gated into WM and thus have the potential to induce interference. These findings are generally consistent with the present results: the “ignored” distractor in our task was presumably spatially attended – at least to some degree – yet it did not interfere with WM because it was completely irrelevant and thus gated out of WM. However, as soon as a distractor

was both attended and task-relevant, we observed interference in both behavioral performance and in the fidelity of cortical representations.

Both selective attention and WM are thought to be sustained by top-down feedback signals from areas of parietal and pre-frontal cortex (PFC) (Desimone and Duncan, 1995; Kastner and Ungerleider, 2000). Our observation of interference between maintaining WM representations and attending to new sensory inputs is consistent with destructive interference occurring between competing top-down modulatory signals, perhaps due to their convergence on common feedback layers in visual cortex (Van Kerkoerle et al., 2017). Moreover, recent work on WM for both simple features and different classes of objects suggests that this feedback may not simply maintain a "sensory-like" representation in visual cortex (Chunharas et al., 2023; Xu, 2023). Instead, the geometry of mnemonic representations in visual cortex morphs over the delay period and along the early visual hierarchy to more closely resemble the geometry of categorical representations in parietal cortex. These geometric rotations towards a more parietal-like representation are consistent with other work in non-human model systems suggesting that dynamic WM codes may serve to insulate remembered information from sensory interference (Libby and Buschman, 2021). While Xu (2023) did not present distractors during the delay period so the functional significance of the observed rotations is not clear, Chunharas et al. (2023) did use data from a task with visual distraction during the delay. It could be beneficial if the brain (e.g., parietal cortex) applied a geometric rotation to remembered information irrespective of the visual input or attentional state dictated by the environment, as a continuous readiness to possible distractions could help ensure robustness of mnemonic contents. In the current data we see that only the representational format in parietal cortex, but not in other visual areas, is shared between memories in the absence of visual input, with ignored visual input, and with attended visual input. However, different continuously visible sensory stimuli – such as the sensory localizer stimuli and the distractor stimuli used in the present study – evoke different response patterns in parietal cortex. Together, this overall pattern of memory-general patterns and sensory-specific patterns is consistent with the idea that parietal cortex employs a stable format to maintain mnemonic information in working memory.

In principle the current paradigm and dataset might be well-positioned to determine if the presence of distractors increases rotational dynamics (Degutis et al., 2024). However, the most common general approach to assessing rotational dynamics in neural codes is to learn a geometry at one timepoint in the delay period and then generalize that geometry to a later timepoint in the delay period. Critically, cross-generalization only yields interpretable results if the SNR is equated at both timepoints in the delay period because if it is not, then a failure to cross-generalize may simply reflect worse signal in one epoch. Similarly, apparent geometrical differences between conditions could also be driven by differences in SNR. In our data set, we observed a precipitous drop-off of memory decoding accuracy – and thus SNR – in the presence of attended distractors, even using models trained on a timepoint-by-timepoint basis that are capable of learning new geometries at each point in the delay period (**Supp Fig. 4**). Thus, this drop in SNR –

which is our key finding of interest – unfortunately also renders cross-generalization analyses difficult to interpret with confidence.

Collectively, our findings provide an explanation to reconcile prior reports of distractor resistance and distractor interference in early visual cortex during WM: distractors do not obligatorily interfere with WM representations, but simultaneously attending to behaviorally relevant new inputs does lead to interference. These results - in particular the joint drop in behavioral performance and in the fidelity of WM representations – further support a role for early visual areas in simultaneously supporting multiple cognitive functions via sustained top-down modulation signals that bias ongoing sensory processing.

Data & Code Availability

We have uploaded all behavioral data and all preprocessed fMRI data, from each participant and ROI, to the Open Science Framework (OSF) at https://osf.io/2wgrv. Code to generate the experimental stimuli used during data collection, and for the analyses used to generate the figures and statistics in our paper can also be found here, with an accompanying wiki to provide an overview of all data and code.

References

Andersson JLR, Skare S, Ashburner J (2003) How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. NeuroImage 20:870–888.

Bettencourt KC, Xu Y (2016) Decoding the content of visual short-term memory under distraction in occipital and parietal areas. Nat Neurosci 19:150–157.

Brainard DH (1997) The Psychophysics Toolbox. Spat Vis 10:433–436.

Bundesen C, Habekost T, Kyllingsbæk S (2005) A Neural Theory of Visual Attention: Bridging Cognition and Neurophysiology. Psychological Review 112:291–328.

Carrasco M, Ling S, Read S (2004) Attention alters appearance. Nat Neurosci 7:308–313.

Christophel TB, Hebart MN, Haynes J-D (2012) Decoding the Contents of Visual Short-Term Memory from Human Visual and Parietal Cortex. J Neurosci 32:12983–12989.

Christophel TB, Iamshchinina P, Yan C, Allefeld C, Haynes J-D (2018) Cortical specialization for attended versus unattended working memory. Nat Neurosci 21:494–496.

Chunharas C, Hettwer MD, Wolff MJ, Rademaker RL (2023) A gradual transition from veridical to categorical representations along the visual hierarchy during working memory, but not perception. Available at: http://biorxiv.org/lookup/doi/10.1101/2023.05.18.541327 [Accessed September 16, 2024].

Czoschke S, Fischer C, Bahador T, Bledowski C, Kaiser J (2021) Decoding Concurrent Representations of Pitch and Location in Auditory Working Memory. J Neurosci 41:4658–4666.

Dale AM (1999) Optimal experimental design for event-related fMRI. Hum Brain Mapp 8:109–114.

Dale AM, Fischl B, Sereno MI (1999) Cortical Surface-Based Analysis. NeuroImage 9:179–194.

David SV, Hayden BY, Mazer JA, Gallant JL (2008) Attention to Stimulus Features Shifts Spectral Tuning of V4 Neurons during Natural Vision. Neuron 59:509–521.

Davies TN, Hoffman DD (2002) Attention to Faces: A Change-Blindness Study. Perception 31:1123–1146.

Degutis JK, Weber S, Soch J, Haynes J-D (2024) Neural dynamics of visual working memory representation during sensory distraction. Available at: https://elifesciences.org/reviewed-preprints/99290v1 [Accessed September 16, 2024].

Desimone R, Duncan J (1995) Neural Mechanisms of Selective Visual Attention. Annu Rev Neurosci 18:193–222.

Deutsch P, Czoschke S, Fischer C, Kaiser J, Bledowski C (2023) Decoding of Working Memory Contents in Auditory Cortex Is Not Distractor-Resistant. J Neurosci 43:3284–3293.

Engel SA, Rumelhart DE, Wandell BA, Lee AT, Glover GH, Chichilnisky E-J, Shadlen MN (1994) fMRI of human visual cortex. Nature 369:525–525.

Ester EF, Anderson DE, Serences JT, Awh E (2013) A Neural Measure of Precision in Visual Working Memory. Journal of Cognitive Neuroscience 25:754–761.

Ester EF, Sprague TC, Serences JT (2015) Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. Neuron 87:893–905.

Greve DN, Fischl B (2009) Accurate and robust brain image alignment using boundary-based registration. NeuroImage 48:63–72.

Hakim N, Adam KCS, Gunseli E, Awh E, Vogel EK (2019) Dissecting the Neural Focus of Attention Reveals Distinct Processes for Spatial Attention and Object-Based Storage in Visual Working Memory. Psychol Sci 30:526–540.

Hakim N, Feldmann-Wüstefeld T, Awh E, Vogel EK (2021) Controlling the Flow of Distracting Information in Working Memory. Cerebral Cortex 31:3323–3337.

Harrison SA, Tong F (2009) Decoding reveals the contents of visual working memory in early visual areas. Nature 458:632–635.

Hirabayashi T, Takeuchi D, Tamura K, Miyashita Y (2013) Functional Microcircuit Recruited during Retrieval of Object Association Memory in Monkey Perirhinal Cortex. Neuron 77:192–203.

Hodsoll S, Viding E, Lavie N (2011) Attentional capture by irrelevant emotional distractor faces. Emotion 11:346–353.

Iamshchinina P, Christophel TB, Gayet S, Rademaker RL (2021) Essential considerations for exploring visual working memory storage in the human brain. Visual Cognition 29:425–436.

Jehee JFM, Brady DK, Tong F (2011) Attention Improves Encoding of Task-Relevant Features in the Human Visual Cortex. J Neurosci 31:8210–8219.

Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. NeuroImage 17:825–841.

Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM (2012) FSL. NeuroImage 62:782–790.

Jenkinson M, Smith S (2001) A global optimisation method for robust affine registration of brain images. Medical Image Analysis 5:143–156.

Kastner S, Ungerleider LG (2000) Mechanisms of Visual Attention in the Human Cortex. Annu Rev Neurosci 23:315–341.

Kiyonaga A, Dowd EW, Egner T (2017) Neural Representation of Working Memory Content Is Modulated by Visual Attentional Demand. Journal of Cognitive Neuroscience 29:2011–2024.

Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C (2007) What's new in psychtoolbox-3. Perception 36:1–16.

Kwak Y, Curtis CE (2022) Unveiling the abstract format of mnemonic representations. Neuron 110:1822-1828.e5.

Langton SRH, Law AS, Burton AM, Schweinberger SR (2008) Attention capture by faces. Cognition 107:330–342.

Libby A, Buschman TJ (2021) Rotational dynamics reduce interference between sensory and memory representations. Nat Neurosci 24:715–726.

Liu T, Hospadaruk L, Zhu DC, Gardner JL (2011) Feature-Specific Attentional Priority Signals in Human Cortex. J Neurosci 31:4484–4495.

Magnussen S, Greenlee MW (1992) Retention and disruption of motion information in visual short-term memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 18:151–156.

Magnussen S, Greenlee MW, Asplund R, Dyrnes S (1991) Stimulus-specific mechanisms of visual short-term memory. Vision Research 31:1213–1219.

Martinez-Trujillo JC, Treue S (2004) Feature-Based Attention Increases the Selectivity of Population Responses in Primate Visual Cortex. Current Biology 14:744–751.

Maxwell JW, Gaspelin N, Ruthruff E (2021) No identification of abrupt onsets that capture attention: evidence against a unified model of spatial attention. Psychological Research 85:2119–2135.

Miller E, Li L, Desimone R (1993) Activity of neurons in anterior inferior temporal cortex during a short-term memory task. J Neurosci 13:1460–1478.

Miller EK, Li L, Desimone R (1991) A Neural Mechanism for Working and Recognition Memory in Inferior Temporal Cortex. Science 254:1377–1379.

Miyashita Y, Chang HS (1988) Neuronal correlate of pictorial short-term memory in the primate temporal cortexYasushi Miyashita. Nature 331:68–70.

Morrisey MN, Hofrichter R, Rutherford MD (2019) Human faces capture attention and attract first saccades without longer fixation. Visual Cognition 27:158–170.

Rademaker RL, Bloem IM, De Weerd P, Sack AT (2015) The impact of interference on short-term memory for visual orientation. Journal of Experimental Psychology: Human Perception and Performance 41:1650–1665.

Rademaker RL, Chunharas C, Serences JT (2019) Coexisting representations of sensory and mnemonic information in human visual cortex. Nat Neurosci 22:1336–1344.

Rust NC, Cohen MR (2022) Priority coding in the visual system. Nat Rev Neurosci 23:376–388.

Serences JT, Ester EF, Vogel EK, Awh E (2009) Stimulus-Specific Delay Activity in Human Primary Visual Cortex. Psychol Sci 20:207–214.

Smith SM (2002) Fast robust automated brain extraction. Human Brain Mapping 17:143–155.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004) Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23:S208–S219.

Sprague TC, Ester EF, Serences JT (2016) Restoring Latent Visual Working Memory Representations in Human Cortex. Neuron 91:694–707.

Sprague TC, Serences JT (2013) Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nat Neurosci 16:1879–1887.

Swisher JD, Halko MA, Merabet LB, McMains SA, Somers DC (2007) Visual Topography of Human Intraparietal Sulcus. J Neurosci 27:5326–5337.

Van Kerkoerle T, Self MW, Roelfsema PR (2017) Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. Nat Commun 8:13804.

Watson AB, Pelli DG (1983) Quest: A Bayesian adaptive psychometric method. Perception & Psychophysics 33:113–120.

Wolfe JM, Võ ML-H, Evans KK, Greene MR (2011) Visual search in scenes involves selective and nonselective pathways. Trends in Cognitive Sciences 15:77–84.

Woolrich MW, Ripley BD, Brady M, Smith SM (2001) Temporal Autocorrelation in Univariate Linear Modeling of FMRI Data. NeuroImage 14:1370–1386.

Xu Y (2020) Revisit once more the sensory storage account of visual working memory. Visual Cognition 28:433–446.

Xu Y (2023) Parietal-driven visual working memory representation in occipito-temporal cortex. Current Biology 33:4516-4523.e5.

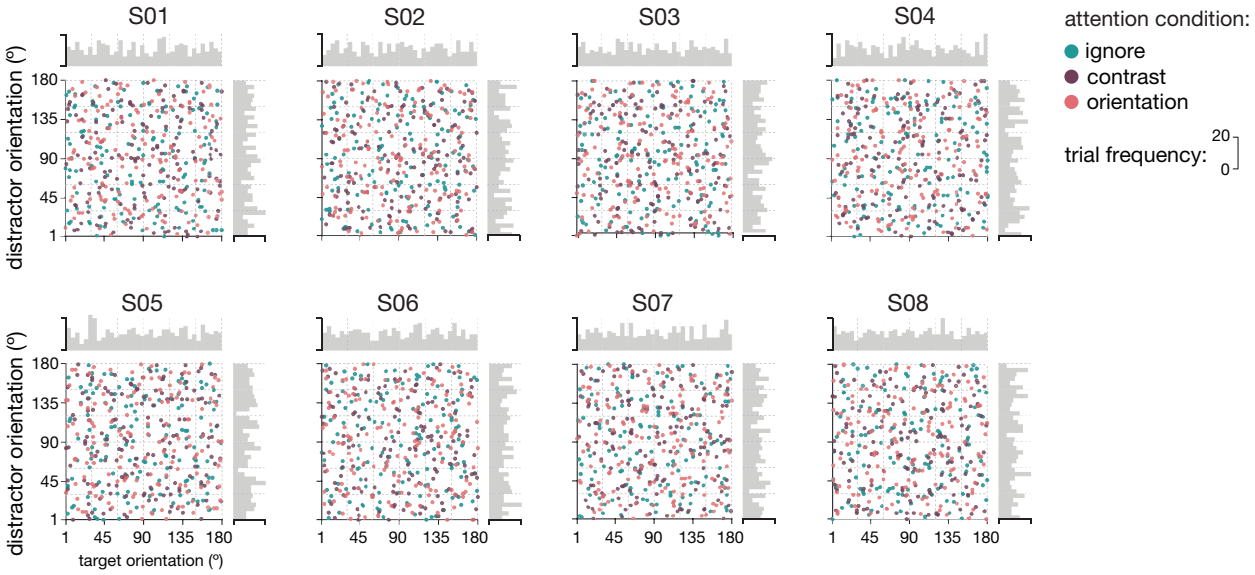
Yan C, Christophel TB, Allefeld C, Haynes J-D (2023) Categorical working memory codes in human visual cortex. NeuroImage 274:120149.

Yiling Y, Klon-Lipok J, Shapcott K, Lazar A, Singer W (2024) Dynamic fading memory and expectancy effects in the monkey primary visual cortex. Proc Natl Acad Sci USA 121:e2314855121.

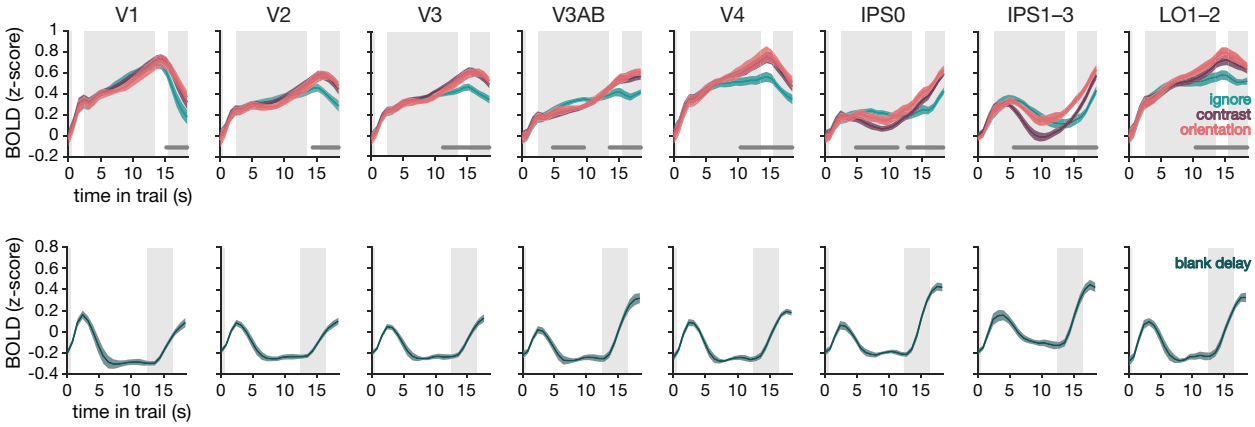
Supplementary Materials

	V1	V2	V3	V3AB	V4	IPS0	IPS1-3	LO1-2
S01	2373 (2032)	2033 (1536)	1936 (1504)	1361 (1001)	461 (327)	691 (372)	1766 (292)	896 (470)
S02	2541 (1912)	2349 (1398)	1610 (997)	1013 (709)	621 (394)	679 (166)	1626 (255)	1381 (551)
S03	2271 (1918)	2377 (1565)	2057 (1296)	1552 (1014)	787 (387)	1183 (499)	1531 (568)	914 (396)
S04	2690 (1898)	2636 (1501)	2219 (1396)	1886 (1066)	958 (414)	1025 (530)	1696 (1044)	1070 (309)
S05	1565 (1184)	1433 (938)	1545 (1142)	1153 (804)	763 (479)	644 (491)	961 (635)	503 (299)
S06	2107 (1685)	2158 (1423)	2531 (1719)	2385 (1421)	737 (313)	1068 (413)	1389 (145)	677 (417)
S07	2620 (1569)	2401 (1436)	1993 (1388)	1592 (910)	882 (402)	1582 (401)	2854 (394)	1323 (410)
S08	2830 (1728)	2598 (1628)	1836 (1359)	840 (686)	474 (388)	397 (327)	1097 (883)	888 (533)

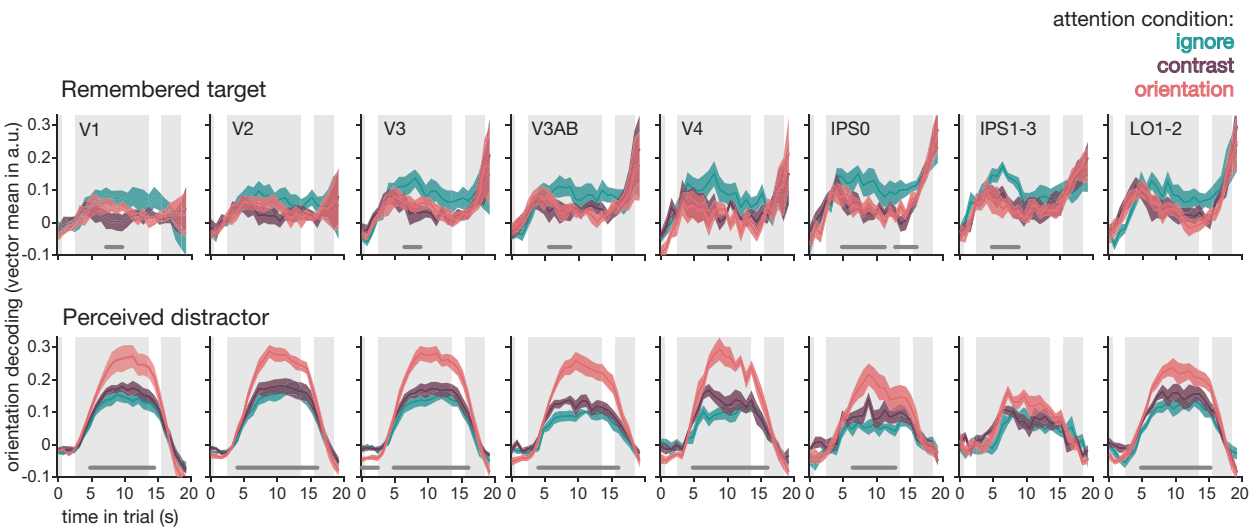
Supplementary Table 1: For each subject, this table holds the number of voxels in each retinotopically defined ROI (numbers shown in bigger font), and the number of voxels within each ROI are visually responsive (numbers shown smaller font in brackets).



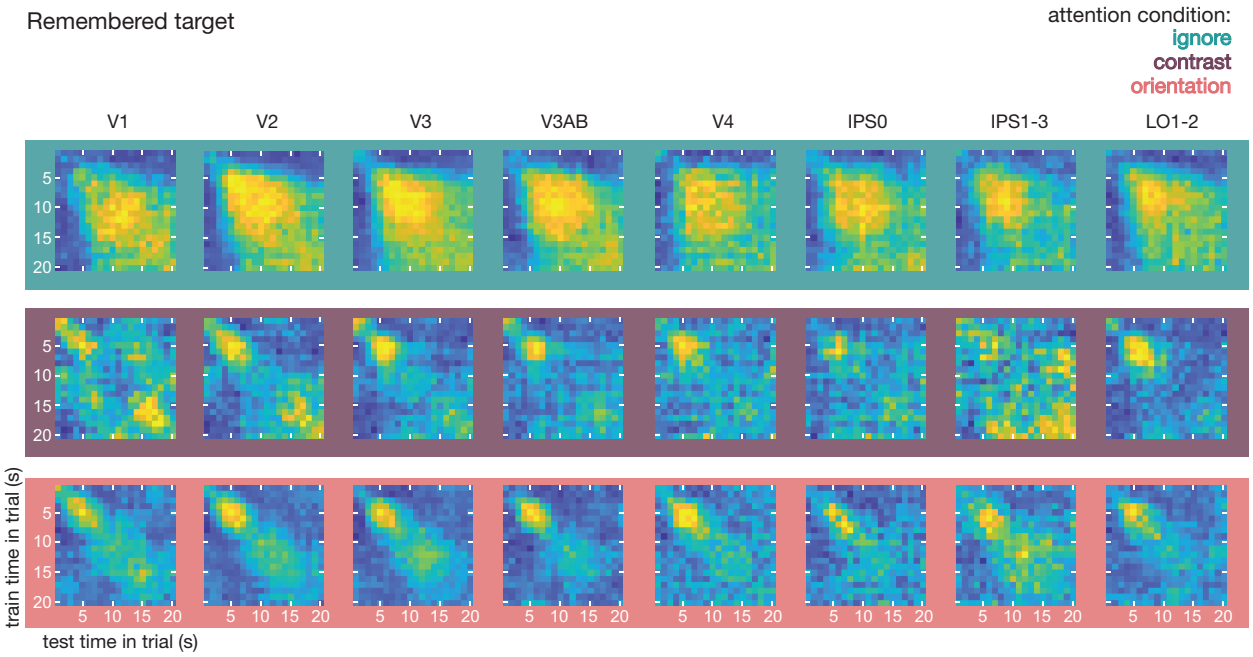
Supplementary Figure 1: There is no systematic relationship between the orientations of the memory targets and the distractors. For each subject, here we plot the distractor orientation (y-axis) against the target orientation (x-axis) on each single trial (dots) in the main working memory experiment. Colors reflect the three different attention conditions – ignore, attend contrast, and attend orientation trials are depicted in teal, purple, and pink, respectively. To ensure relatively uniform sampling of target and distractor orientations across orientation space, both orientations were drawn pseudo-randomly from one of six orientation bins (each bin spanning 30°). The boundaries between these bins are indicated with dashed grey lines. Importantly, orientations were drawn from each bin equally often. Thus, of the 144 total trials in each condition, the target orientation was randomly drawn from the first orientation bin (1°–30°) 24 times, from the second bin (31°–60°) 24 times, and so on. Similarly, distractor orientations were randomly drawn 24 times from each bin. Importantly, draws from target and distractor orientation bins were counterbalanced, such that each bin combination (i.e., each square defined by the dashed-line grid) contains a total of 4 trials. We quantified the relationship between target and distractor orientations via circular correlation (rho), with mean correlations of –0.009 (SEM = 0.008), 0.0008 (SEM = 0.014), and 0.0104 (SEM = 0.013) for the ignore, attend contrast, and attend orientation conditions, respectively. For no subject in no condition was there ever a significant correlation between target and distractor orientations (all p-values $\geq$ 0.44).



Supplementary Figure 2: Deconvolved univariate BOLD time courses measured during the main working memory experiment (top row), and for the working memory localizer (bottom row). For the main working memory task, distractors in all 3 attention conditions effectively drove univariate responses in early visual areas (V1–V4) and LO, with overall qualitatively higher responses when attention was deployed towards the distractor. When participants attended distractor contrast (shown in purple) or distractor orientation (shown in pink), BOLD responses were generally higher then when the distractor was ignored (shown in teal), especially towards the end of the trial. In parietal regions, responses seem more transient, with the strongest response occurring with attention to distractor orientation changes (conversely, attention to contrast changes appears to result in the lowest response). Grey background-panels in each subplot indicate the time during which the memory target (0–0.5s, far left panel), the distractor (2.5–13.5s, middle panel), and the response-dial (15.5–18.5s, right most panel) were on the screen. Darker grey lines just above and parallel to the x-axis indicate clusters of consecutive TR's during which the three attention conditions differ significantly from one another (as calculated with a cluster based permutation test, see Methods). For the memory localizer task there is only a transient response to the memory target, which goes back to baseline in the blank space between target and recall (note the absence of a grey panel during the delay of this task, and an earlier response period from 12.5–16.5s). Lines are group-averaged BOLD responses, with shaded error-bars representing $\pm$ 1 within-subject SEM (N=8 independent subjects that are identical for both tasks).



Supplementary Figure 3: Decoding time courses for remembered and distractor orientations throughout trials of the main working memory task. The orientation held in mind during the working memory delay is better represented in visual cortex when the distractor can be ignored (teal), compared to when the distractor is attended (purple and pink, see also **Fig. 2a**). Here we see that these differences between attention conditions (as indicated by the dark grey lines just above and parallel to the x-axis) become apparent about 7–8 seconds into the trial. Considering the delay in BOLD response, this means differences likely emerge around the time that participants start engaging with the distractor grating (in the contrast and orientation attention conditions). The orientation of the distractor grating is best represented when its orientation is attended, intermediate when its contrast is attended, and worst when the distractor is ignored (see also **Fig. 2a**). Here we see that the onset of these attention condition differences happens relatively early in the trial, around 6s, which likely coincides with the very onset of the distractor grating (taking the BOLD delay into account). Grey background-panels in each subplot indicate the memory target (0–0.5s, far left panel), distractor (2.5–13.5s, middle panel), and the response (15.5–18.5s, right most panel) periods. Significance clusters indicate consecutive TR's during which the three attention conditions differ significantly from one another (as calculated with a cluster-based permutation test, see Methods).



Supplementary Figure 4: Cross-temporal decoding of the remembered orientation during the delay. Here, we train our decoder on every time point during the delay, and test it on every other timepoint. When the distractor is ignored, relatively high decoding (in yellow colors) can be seen between most timepoints during the delay, implying a sustained representation that can cross-generalize from one timepoint during the delay to many others. Under conditions where distractor contrast or orientation are attended, high decoding is more pronounced along the diagonal. However, given that decoding performance is lower in these conditions (see also **Fig 2a** & **Fig. 3**), this could also be a matter of lower SNR when attention is diverted away from the memory stimulus.