
Research Articles: Behavioral/Cognitive

Flexible coding of visual working memory representations during distraction

Elizabeth S. Lorenc¹, Kartik K. Sreenivasan², Derek E. Nee^{1,3}, Annelinde R. E. Vandenbroucke^{1,4} and Mark D'Esposito^{1,5}

¹Helen Wills Neuroscience Institute, University of California, Berkeley, California, USA

²Division of Science and Mathematics, New York University Abu Dhabi, New York, USA

³Department of Psychology, Florida State University, Tallahassee, USA

⁴Donders Center for Brain and Cognition, Radboud University Nijmegen, The Netherlands

⁵Department of Psychology, University of California, Berkeley, California, USA

DOI: 10.1523/JNEUROSCI.3061-17.2018

Received: 22 October 2017

Revised: 1 March 2018

Accepted: 27 April 2018

Published: 8 May 2018

Author contributions: E.S.L. wrote the first draft of the paper; E.S.L., K.S., A.R.V., D.E.N., and M.D. edited the paper; E.S.L., K.S., A.R.V., D.E.N., and M.D. designed research; E.S.L. performed research; E.S.L. analyzed data; E.S.L. wrote the paper.

Conflict of Interest: The authors declare no competing financial interests.

This research was supported by NIH Grant MH63901 to M. D'Esposito, POF-GA-2013-624380 to A. Vandenbroucke, and pre-doctoral NRSA fellowship NIMH F31MH107157 to E. Lorenc, as well as the National Science Foundation Major Research Instrumentation Program, award number BCS-0821855. In addition, we would like to thank J. Tseng for assistance with participant training and data collection.

Correspondence: Elizabeth S. Lorenc, Helen Wills Neuroscience Institute, University of California, Berkeley, 132 Barker Hall, Berkeley, CA, 94720-3190, USA. elizabeth.lorenc@berkeley.edu

Cite as: J. Neurosci ; 10.1523/JNEUROSCI.3061-17.2018

Alerts: Sign up at www.jneurosci.org/cgi/alerts to receive customized email alerts when the fully formatted version of this article is published.

Accepted manuscripts are peer-reviewed but have not been through the copyediting, formatting, or proofreading process.

Copyright © 2018 the authors

1 **Title: Flexible coding of visual working memory representations during distraction**

2 **Abbreviated title: Distractor effects in visual working memory**

3 **Elizabeth S. Lorenc^{1*}, Kartik K. Sreenivasan², Derek E. Nee^{1,3}, Annelinde R. E.**

4 **Vandenbroucke^{1,4}, & Mark D’Esposito^{1,5}**

5 ¹Helen Wills Neuroscience Institute, University of California, Berkeley, California, USA

6 ²Division of Science and Mathematics, New York University Abu Dhabi, New York, USA

7 ³Department of Psychology, Florida State University, Tallahassee, USA

8 ⁴Donders Center for Brain and Cognition, Radboud University Nijmegen, The Netherlands

9 ⁵Department of Psychology, University of California, Berkeley, California, USA

10 *** Correspondence:** Elizabeth S. Lorenc, Helen Wills Neuroscience Institute, University of

11 California, Berkeley, 132 Barker Hall, Berkeley, CA, 94720-3190, USA.

12 elizabeth.lorenc@berkeley.edu

13 **Number of pages: 31**

14 **Number of figures: 6, Extended Data: 3**

15 **Number of tables: 0**

16 **Number of words in Abstract: 250**

17 **Number of words in Introduction: 599**

18 **Number of words in Discussion: 1,032**

19 **Conflict of Interest:** Nothing to disclose.

20 **Acknowledgements:** This research was supported by NIH Grant MH63901 to M. D’Esposito,

21 PIOF-GA-2013-624380 to A. Vandenbroucke, and pre-doctoral NRSA fellowship NIMH

22 F31MH107157 to E. Lorenc, as well as the National Science Foundation Major Research

Distractor effects in visual working memory

Instrumentation Program, award number BCS-0821855. In addition, we would like to thank J. Tseng for assistance with participant training and data collection.

1. Abstract

Visual working memory (VWM) recruits a broad network of brain regions, including prefrontal, parietal, and visual cortices. Recent evidence supports a “sensory recruitment” model of VWM, whereby precise visual details are maintained in the same stimulus-selective regions responsible for perception. A key question in evaluating the sensory recruitment model is how VWM representations persist through distracting visual input, given that the early visual areas that putatively represent VWM content are susceptible to interference from visual stimulation.

To address this question, we employed an fMRI inverted encoding model approach to quantitatively assess the effect of distractors on VWM representations in early visual cortex and the intraparietal sulcus (IPS), another region previously implicated in the storage of VWM information. This approach allowed us to reconstruct VWM representations for orientation, both before and after visual interference, and to examine whether oriented distractors systematically biased these representations. In our human participants (both male and female), we found that orientation information was maintained simultaneously in early visual areas and IPS in anticipation of possible distraction, and these representations persisted in the absence of distraction. Importantly, early visual representations were susceptible to interference; VWM orientations reconstructed from visual cortex were significantly biased toward distractors, corresponding to a small attractive bias in behavior. In contrast, IPS representations did not show such a bias. These results provide quantitative insight into the effect of interference on VWM representations, and suggest a dynamic tradeoff between visual and parietal regions that allows flexible adaptation to task demands in service of VWM.

Distractor effects in visual working memory

46 **2. Significance Statement**

47 Despite considerable evidence that stimulus-selective visual regions maintain precise visual
48 information in working memory, it remains unclear how these representations persist through
49 subsequent input. Here, we used quantitative model-based fMRI analyses to reconstruct the contents
50 of working memory and examine the effects of distracting input. While representations in the early
51 visual areas were systematically biased by distractors, those in the intraparietal sulcus appeared
52 distractor-resistant. In contrast, early visual representations were most reliable in the absence of
53 distraction. These results demonstrate the dynamic, adaptive nature of visual working memory
54 processes, and provide quantitative insight into the ways in which representations can be affected by
55 interference. Further, they suggest that current models of working memory should be revised to
56 incorporate this flexibility.

Distractor effects in visual working memory

57 **3. Introduction**

58 Visual working memory (VWM) constitutes the brief maintenance and manipulation of visual
 59 information. It allows for a unified representation of the visual world, despite frequent temporal
 60 discontinuities in visual input that occur during eye movements, occlusions, and object motion
 61 (Curtis and D'Esposito, 2003; Serences, 2016). This cognitive ability is supported by a broad
 62 network of brain regions, including lateral prefrontal cortex (Chelazzi et al., 1993; D'Esposito et al.,
 63 1995; 2000; Ester et al., 2015; Riley and Constantinidis, 2015), parietal cortex (Christophel et al.,
 64 2014; Bettencourt and Xu, 2015; Ester et al., 2015; Xu, 2017), and primary sensory cortices
 65 (Harrison and Tong, 2009; Serences et al., 2009). According to a “sensory recruitment” model of
 66 VWM (Pasternak and Greenlee, 2005; Postle, 2006; D'Esposito, 2007), precise visual details are
 67 maintained in the same stimulus-selective regions responsible for primary visual processing during
 68 perception. Indeed, numerous studies have provided evidence that occipital cortex maintains visual
 69 features in working memory (Chelazzi et al., 1993; Miller et al., 1993; Magnussen, 2000; Awh and
 70 Jonides, 2001; Curtis and D'Esposito, 2003; Pasternak and Greenlee, 2005; D'Esposito, 2007;
 71 Harrison and Tong, 2009; Serences et al., 2009; Ester et al., 2013; Pratte and Tong, 2014;
 72 Sreenivasan et al., 2014; D'Esposito and Postle, 2015).

73 However, because early visual cortex is susceptible to interference from subsequent visual input,
 74 it remains unclear how VWM representations in these regions could survive visual interference. In
 75 fact, multiple studies have found that distractors presented during the memory delay period *do*
 76 impair VWM for simple visual features like spatial frequency (Magnussen et al., 1991; Bennett and
 77 Cortese, 1996; Nemes et al., 2011) and motion direction (Magnussen and Greenlee, 1992; Pasternak
 78 and Zaksas, 2003; McKeefry et al., 2007), as well as for more complex stimuli like faces (Yoon et
 79 al., 2006; Sreenivasan and Jha, 2007). This impairment is particularly evident when the distractor

Distractor effects in visual working memory

differs from the item in memory along a task-relevant feature dimension (Magnussen et al., 1991; Lalonde and Chaudhuri, 2002; Nemes et al., 2011), suggesting that visual interference likely acts on VWM representations maintained within segregated feature-specific channels in visual cortex (Sneve et al., 2011). Interestingly, interference from a subsequent distractor can exert an attractive pull on VWM representations of spatial frequency (Huang and Sekuler, 2010; Nemes et al., 2011; Dubé et al., 2014), color (Nemes et al., 2012), or orientation (Rademaker et al., 2015; Wildegger et al., 2015).

Recently, (Bettencourt and Xu, 2015) sought to identify the neural locus of distractor-resistant VWM representations in a human functional magnetic resonance imaging (fMRI) study. They found that remembered orientations could be decoded from the superior intraparietal sulcus (IPS) despite the presentation of irrelevant distractors, and that the discriminability of orientation information in early visual areas was reduced, but not abolished, by distractor presentation. This pattern of results has two equally plausible explanations, with important theoretical consequences for the sensory recruitment model of working memory. First, this pattern of results could suggest that any VWM representations observed in early visual areas are not behaviorally relevant. Alternatively, distraction during VWM could systematically bias early visual representations, as has been observed in behavior (Rademaker et al., 2015; Wildegger et al., 2015), and such biases could account for the observed reduction in orientation discriminability (Ester et al., 2016).

Here, we employed an inverted encoding model orientation reconstruction approach to quantitatively assess the effect of distractors on VWM representations in early visual cortex and the IPS. This approach allowed us not only to identify the regions in which visual information is reliably represented during a working memory delay, but also to investigate whether, and how, these representations are affected by subsequent distractors.

103

104 **4. Materials and Methods**

105 *Participants.* We recruited 21 healthy young adult participants for this study. 9 of these
 106 participants were excluded due to an inability to maintain fixation or an inability to remain alert
 107 or awake during their initial scan session. The final 12 participants (2 male, ages 18-35), which
 108 included two of the authors (E.L. and A.V.), each completed a one-hour training session and four
 109 two-hour MRI scan sessions. Participants were right-handed and had normal or corrected-to-
 110 normal vision, and all procedures were approved by the UC Berkeley Committee for the
 111 Protection of Human Subjects. Participants gave their written informed consent before the study
 112 and were compensated monetarily for their time. In addition, participants were given small
 113 bonuses after each scanning session that scaled with their performance on the cognitive task.

114 *Behavioral training procedure.* Prior to the MRI scan sessions, each participant completed a
 115 one-hour training session to learn and practice the cognitive task, and practice maintaining
 116 central fixation. Participants each completed a minimum of two 12-trial runs of the task, during
 117 which eye position was continuously monitored (Eyelink-1000, SR Research). Feedback about
 118 mean behavioral precision and fixation quality was provided at the end of each practice run. In
 119 addition, participants were familiarized with the retinotopic mapping procedures and practiced
 120 maintaining visual fixation while performing a challenging peripheral visual detection task, in
 121 which they detected the brief appearance of a small gray circle within the polar angle mapping
 122 wedge (see *Region-of-interest creation*).

123 **Experimental Design and Statistical Analyses**

124 *Cognitive task.* The task (Figure 1) was designed to assess the precision of VWM for
 125 orientation, and to capture any memory biases induced by an intervening distractor. Each trial

Distractor effects in visual working memory

126 began with a green 500ms pre-cue at fixation to warn participants that a new trial was about to
 127 start. The pre-cue was immediately followed by a right-lateralized oriented sinusoidal grating,
 128 and participants were instructed to remember the orientation of this grating as precisely as
 129 possible over the remainder of the trial. Because of hemispheric asymmetries in the
 130 representation of visual space in the early visual (Hougaard et al., 2015) and parietal (Sheremata
 131 et al., 2010; Berger et al., 2014; Jeong and Xu, 2016) cortices, we chose to present all stimuli in
 132 the right visual hemifield, rather than counterbalance stimulus hemifield across participants. The
 133 grating was presented for 500 ms, and followed by a 9.5 s blank delay (“Delay 1”). On two
 134 thirds of trials, a distractor grating, which participants were instructed to ignore, was then
 135 presented for 500 ms in the same spatial location as the memory stimulus. On the remaining one
 136 third of trials, the screen remained blank with only a fixation point during this 500 ms interval.
 137 Finally, there was a second 9.5 s blank delay (“Delay 2”) after which participants were presented
 138 with a (randomly oriented) test grating that they had to adjust to the memorized orientation using
 139 an MR-compatible joystick (Current Designs, Inc.) within a 4 s response window. Trials were
 140 separated by a variable inter-trial interval of 8, 10, or 12 s.

141 While distractors were more likely to appear than a blank delay, distractor presentation was
 142 unpredictable on any given trial. The distractor consisted of a sinusoidal grating with an
 143 orientation that was 40 – 50° clockwise (50%) or counterclockwise (50%) of the remembered
 144 orientation. With the exception of two participants (E.L. and A.V.), participants were unaware of
 145 the relationship between the memory and distractor stimulus orientations. In addition, the task
 146 timing was slightly different for these first two participants (see *Functional MRI acquisition and*
 147 *preprocessing*), in that there was a 3 s response window, and the inter-trial interval was 6.67,
 148 8.33, or 10 s.

Distractor effects in visual working memory

Memory orientations were distributed over the entire 180° orientation space; eight equally spaced “base” orientations were chosen (0° to 157.5°, in steps of 22.5°), with an additional positive or negative 1° to 10° of jitter added on each trial. Grouping the orientations into bins allowed us to ensure that similar sets of orientations were shown in each of the three distractor conditions. In order to limit fMRI run lengths, the eight orientation categories were randomly split into two sets for each participant, with one set shown on odd runs and the other on even runs. There were 12 trials per run: one trial from each of four orientation categories for each of the three distractor conditions. Participants completed between 22 and 32 runs (participants always completed an even number of runs for counterbalancing purposes), yielding an average of 108 trials in each distractor condition. Participants were given feedback at the end of each run about the mean precision of their responses.

Stimuli were programmed in MATLAB (v2012b, The MathWorks, Inc., Natick, MA) using the Psychtoolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007) projected onto a screen at the rear of the magnet bore and viewed via a headcoil-mounted mirror. Each circular sinusoidal grating stimulus subtended 10 degrees of visual angle and was centered on a point 7 degrees of visual angle to the right of fixation. The sinusoidal gratings were full contrast, with a spatial frequency of 0.5 cycles/degree, and they alternated in phase at a rate of 12Hz for the duration of their presentation. Participants were required to maintain central fixation throughout each scanning run, and eye position was monitored with an MR-compatible eyetracker (Avotec).

Behavioral analyses. The method-of-adjustment response (see *Cognitive task*) yielded a trial-by-trial measure of memory error in degrees for each distractor condition (no distractor, clockwise distractor, counterclockwise distractor), for each participant. Because we did not have an a priori expectation for any differences between the clockwise and counterclockwise distractor conditions, we flipped the signs of the errors from the counterclockwise distractor

Distractor effects in visual working memory

condition and combined them with those from the clockwise distractor condition, prior to model fitting to increase analysis power. The group data with separate model fits for counterclockwise and clockwise distractor trials is provided as Extended Data, as are the individual subject error histograms and model fits.

First, to examine whether participants' mean response errors reliably differed between the no-distractor and distractor conditions, we calculated the circular distance between the circular means of the no-distractor and distractor error distributions for each participant (Berens, 2009). Then, we calculated circular 95% confidence intervals to determine whether the resulting set of no-distractor / distractor differences significantly differed from zero.

To assess whether the distractors selectively caused an attractive bias, or whether they also affected memory precision or guess rates, we followed this initial analysis with a more detailed mixture model analysis. Using the MemToolbox (Suchow et al., 2013) in MATLAB, each error distribution was fit with a mixture model of a von Mises distribution and a uniform distribution. As in previous studies (Zhang, 2008), three free parameters were estimated: the mean of the von Mises (reflecting any systematic clockwise or counterclockwise biases in participants' responses), the standard deviation of the von Mises (reflecting the average precision of a participant's responses), and the height of the uniform distribution (reflecting the rate of random guesses). This model was fit separately for each condition for each of the twelve participants, using Markov Chain Monte Carlo to compute maximum a posteriori parameter estimates and 95% credible intervals (the Bayesian analog of confidence intervals). We compared the size of the resulting credible intervals across distractor conditions to assess whether the quality of model fit differed by condition. Finally, we performed paired samples t-tests to compare estimated parameters between the no-distractor and distractor conditions.

Distractor effects in visual working memory

196 *Functional MRI acquisition and preprocessing.* MRI data were acquired in the UC Berkeley
 197 Henry H. Wheeler, Jr. Brain Imaging Center with a Siemens TIM/Trio 3T MRI scanner with a
 198 12-channel receive-only head coil. Whole-brain MPRAGE T_1 -weighted scans were acquired for
 199 anatomical localization, normalization, and cortical surface reconstruction. Functional data were
 200 obtained using a one-shot T_2^* -weighted echoplanar imaging (EPI) sequence sensitive to blood
 201 oxygenation level-dependent (BOLD) contrast. The EPI sequence parameters for the first two
 202 participants (TR = 1.6667s, TE = 30ms, flip angle = 55° , field of view = 1110 x 1110, matrix
 203 size = 74 x 74, in-plane resolution = 3 x 3mm, 25 ascending 3mm-thick axial slices separated by
 204 a 0.3mm interslice gap) were slightly different from the sequence used for the remainder of the
 205 participants (TR = 2.0s, TE = 30ms, flip angle = 55° , field of view = 1332 x 1332, matrix size =
 206 74 x 74, in-plane resolution = 3 x 3mm, 30 ascending 3mm-thick axial slices separated by a
 207 0.3mm interslice gap), as minor adjustments were made to improve whole-brain coverage.

208 Functional MRI data were preprocessed with AFNI (Cox, 1996; Cox and Hyde, 1996) and
 209 custom Matlab (v2012b, The MathWorks, Inc., Natick, MA) scripts. Following slice-time
 210 correction, each EPI run was motion-corrected and co-registered to the anatomical scan in a
 211 single resampling step, using a 6-parameter affine registration in `align_epi_anat.py` (Saad et al.,
 212 2009). Finally, each run was linearly de-trended, voxel-wise, in preparation for inverted
 213 encoding model orientation reconstruction.

214 *Regions-of-interest (ROIs).* We chose as our ROIs the left and right early visual areas and the
 215 left and right intraparietal sulcus (IPS). Because the stimuli were always presented in the right
 216 hemifield, the left hemisphere was always contralateral to the memory stimulus, distractor, and
 217 probe, and the right hemisphere was always ipsilateral. We limited our early visual ROIs to V1-
 218 V3, because decoding of orientation information in VWM is of comparable quality for V1

Distractor effects in visual working memory

219 through V3, and somewhat less robust for V3A/V4 (Harrison and Tong, 2009; Pratte and Tong,
220 2014).

221 Gray/white matter boundary segmentation and cortical surface reconstruction was performed
222 with Freesurfer's recon-all tool ((Fischl et al., 2002)
223 <http://surfer.nmr.mgh.harvard.edu/fswiki/recon-all>). Retinotopic mapping was used to delineate
224 early visual areas V1 – V3 (see below), and the anatomical IPS ROIs were defined on the
225 cortical surface using the automated “S_intrapariet_and_P_trans” label from the Destrieux atlas
226 (Destrieux et al., 2010) in Freesurfer. Finally, all surface-based ROIs were transformed back into
227 native volume space for subsequent analyses.

228 **Retinotopic mapping of early visual areas V1 – V3.** During the first scanning session,
229 participants completed six 6.5-minute retinotopic mapping functional runs: two each of
230 clockwise and counterclockwise-rotating polar angle mapping runs, and two eccentricity
231 mapping runs. For the polar angle mapping runs, a 40-degree black and white checkerboard
232 wedge (phase-alternating at 8Hz) rotated around fixation. A complete revolution was
233 completed every 40 seconds, and 10 revolutions were completed per run. For the eccentricity
234 mapping runs, a checkerboard ring stimulus expanded slowly from fixation, with a full cycle
235 completed every 40 seconds, 10 times per run. Participants were instructed to maintain
236 central fixation throughout the run, and to press a button when a small gray circle appeared
237 in a random location within the wedge or ring.

238 The functional retinotopy data was processed with AFNI's @Retino_Proc script
239 (Warnking et al., 2002). Then, left and right visual areas V1, V2, and V3, dorsal and ventral,
240 were individually delineated on the cortical surface in AFNI's SUMA (Saad et al., 2004;
241 Saad and Reynolds, 2012), following standard procedures (Wandell and Winawer, 2011).

Distractor effects in visual working memory

242 *Inverted encoding model analyses.* To reconstruct orientation information from multi-voxel
 243 patterns of BOLD activity, we adopted an inverted encoding model approach, which has also
 244 been termed “forward encoding” (Brouwer and Heeger, 2009; Ester et al., 2013). All of the
 245 following analyses were performed separately within each ROI (left and right V1-V3, left and
 246 right IPS).

247 First, for every trial, we extracted groups of volumes representing stimulus perception (4 to
 248 6s after stimulus onset), the first memory delay (10 to 14s after stimulus onset), and the second
 249 memory delay (20 to 26s after stimulus onset), and averaged each group to yield a single BOLD
 250 intensity pattern for each trial epoch. For the analysis of perception/encoding and first memory
 251 delay representations, the inverted encoding model analysis was completed separately within
 252 each epoch, using a leave-one-run-pair-out cross-validated structure. A *pair* of runs were held
 253 out on each iteration, because only half of the orientation categories were presented on each run;
 254 this ensured that the training and testing datasets contained examples that were evenly
 255 distributed across the entire orientation space. To reconstruct distractor perception and the
 256 second memory delay, the encoding model was trained on the BOLD patterns from the first
 257 memory delay. Training the model on pre-distractor time-points allowed for direct comparison
 258 between reconstructed representations with and without distractors. In addition, this method
 259 avoided any biases that could have been introduced by estimating the model on time-points
 260 containing a mixture of distractor conditions.

261 After averaging the data, we characterized the orientation preferences of each voxel in an
 262 ROI by creating a voxel-by-voxel encoding model. More specifically, we used linear regression
 263 (Equation 1) to calculate the weights (W : 8 channels $\times$ m voxels) that related the BOLD
 264 intensities from each voxel in the training set (B_1 : m voxels $\times$ n trials) to the modeled activation

Distractor effects in visual working memory

of eight hypothetical orientation channels (C_1 : 8 channels $\times$ m voxels), given the orientations that were presented on each trial. These hypothetical orientation basis functions were half-wave rectified sinusoids raised to the 7th power and distributed evenly from 0° to 179°.

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

Next, we inverted the voxel-by-voxel encoding model (Equation 2), and the calculated weights (W) were used to determine the output of each of the hypothetical orientation channels (C_2), given the BOLD activity patterns in the held-out subset of trials (B_2).

$$\text{Equation 2: } C_2 = (W^T W)^{-1} W^T B_2$$

Finally, for every possible orientation (0° to 179°), we calculated the Pearson’s correlation between the reconstructed orientation channel output (C_2) and the eight hypothetical orientation channels (as in Brouwer and Heeger, 2009). The resulting orientation reconstruction function was therefore maximal at the orientation that had most likely been represented on that trial, based on the measured pattern of BOLD activity. To evaluate whether reliable orientation information could be reconstructed across trials, we re-centered each single-trial orientation reconstruction function so that the “correct” orientation on that trial corresponded to 0°, and then averaged across trials to create an average orientation reconstruction function for that ROI in that participant. Each reconstruction function was then averaged across participants within an ROI.

Reconstruction significance was calculated with a permutation procedure, in which the quality of each reconstruction was quantified with a “representational fidelity” metric RF (as in (Sprague et al., 2016)). More specifically, we calculated the vector mean across the entire group-averaged orientation reconstruction (Equation 3), where $r(\theta)$ is the mean Pearson’s correlation at a given polar angle, spanning all -90° to 89° of zero-centered orientation space.

Distractor effects in visual working memory

Equation 3: $RF = \text{mean}(r(\theta) \cos 2*\theta)$

Next, we created a null distribution of representational fidelity values by repeating the above orientation reconstruction analysis for 1000 permutations of the remembered orientation trial labels (within the true run structure of the data) and again applying Equation 3. Finally, we calculated the fraction of samples in the null distribution whose representational fidelity was larger than the observed representational fidelity, which yielded an empirical p-value. Across all reconstructions for which we calculated representational fidelity significance, and comparisons between those reconstructions, we corrected for multiple comparisons by controlling the false discovery rate (FDR) (Benjamini, 2001), $q = 0.05$.

To compare reconstruction quality between ROIs, we simply extended the above procedure to perform a two-tailed test between regions. First, we calculated the absolute value of the difference between the ROIs' representational fidelity values. Then, we performed the same calculation on each of the 1000 sets of permuted results (see above). Finally, the p -value was calculated as the fraction of this null distribution that was larger than the observed difference.

Testing for biases in reconstructed orientation representations. To assess how orientation representations changed post-distractor, we examined the average orientation reconstruction functions during the second memory delay, separately for each participant and ROI.

First, we computed the slope of the best-fit line to the mean reconstruction error sizes from the three distractor conditions (CCW-None-CW). Because this analysis simultaneously incorporates both counterclockwise and clockwise distractor reconstructions, we did not pool the counterclockwise and clockwise distractor data here. The resulting slope would be positive if the distractors exerted an attractive effect on the orientation representations, zero if there were no effect, and negative if the distractors exerted a repulsive effect. Next, we generated a null

Distractor effects in visual working memory

310 distribution by randomly flipping the signs of the individual participant slopes and averaging
 311 across participants, repeating this 1000 times. Finally, we tested for the predicted positive slope
 312 by comparing the observed mean slope across participants to this null distribution.

313 Last, to evaluate whether the quality of the reconstructions were comparable with and
 314 without distractors, irrespective of any bias in the reconstruction center, we performed a
 315 modified version of the above reconstruction fidelity analysis. First, rather than using a zero-
 316 centered cosine function in Equation 3, we used a series of cosine functions centered at each of
 317 the 180 possible orientation centers and took the maximum (hereafter “maximal fidelity”). Then,
 318 for each ROI, we calculated the absolute value of the difference in maximal fidelities between
 319 the no-distractor and distractor reconstructions. Finally, we repeated this analysis for each of the
 320 1000 null reconstructions and calculated the p -value as the fraction of these null maximal fidelity
 321 differences that exceeded the size of the empirical maximal fidelity difference.

322 *Reconstructing orientation representations over time.* Finally, we conducted an exploratory
 323 analysis to examine the temporal evolution of the VWM representations over the course of the
 324 trial, both with and without intervening distractors. This analysis closely resembled the
 325 orientation reconstruction approach described above, in that the encoding model was always
 326 estimated based on the two TRs at the end of the first memory delay (10 to 14 s post-stimulus
 327 onset). However, the resulting voxel weights were then used to create orientation reconstruction
 328 functions within a two-TR sliding window moved one TR at a time over the entire trial.

329

330 5. Results

331 *Behavior.* Prior to each of the following behavioral analyses, we sign-flipped the data from
 332 the counterclockwise distractor trials and combined it with the clockwise distractor trials, to

Distractor effects in visual working memory

yield a single distribution of “distractor” errors to compare against the “no distractor” error distribution within each participant. In this combined distractor data, an attractive bias would then appear as a clockwise shift.

First, to test for an overall difference in the mean errors between trials with and without distractors, we calculated circular 95% confidence intervals around the mean circular distance between the no-distractor and distractor mean error directions. Because the resulting confidence interval did not include zero (mean = 1.042 degrees, circular 95% CI [0.060, 2.025]), this suggests that the mean behavioral responses were shifted toward the distractor orientations.

We followed this coarse analysis with a more specific model-based analysis to examine which parameters of participants’ response distributions were affected by distraction. Using maximum likelihood estimation, we fit each participant’s response error distribution with a mixture of a von Mises and a uniform distribution (Zhang, 2008), separately for trials with and without distractors. Three parameters of these distributions were estimated and then compared across distractor conditions: the mean and standard deviation of the von Mises reflect the mean bias and precision of memory representations, respectively, while the height of the fitted uniform distribution captures the guess rate. Sample no-distractor and distractor response error distributions and resulting model fits for an example participant are depicted in Figure 2A, and the same is shown for all participants in Extended Data Figure 2-1.

A paired-samples t-test revealed that the mean bias (Figure 2B, left) varied by distractor condition ($t(11) = 2.336, p = 0.039$), in that participants showed a significant positive (attractive) bias of about one degree toward presented distractors, but no bias on distractor-free trials. In contrast, there was no effect of distractors on mean guess rate (Figure 2B, center: $t(11) = 0.498, p = 0.628$) or on precision (Figure 2B, right: $t(11) = 0.535, p = 0.603$). The model fits were of

Distractor effects in visual working memory

significantly higher quality for the distractor trials compared to the no-distractor trials (as indexed by the size of the 95% credible intervals around each parameter estimate), because the former included twice the number of trials per participant (repeated measures ANOVA, main effect of distractor condition $F(1) = 31.848, p < 0.001$).

Similar effects were found when data from the counterclockwise and clockwise distractor conditions were fitted separately (Extended Data Figure 2-2), although a reliable bias was only found in the counterclockwise distractor condition.

Orientation reconstruction during perception and pre-distractor memory delay. Using an inverted encoding model, we observed robust reconstruction of perceived orientations (Figure 3, left column) in the contralateral (left hemisphere) early visual areas (mean reconstruction fidelity (RF) = 0.026, $p < 0.001$), as well as in the ipsilateral early visual areas (RF = 0.013, $p = 0.002$). Reconstruction accuracy was superior in contralateral relative to ipsilateral early visual areas (two-tailed $p = 0.003$). In addition, perceived orientations could be reliably reconstructed from activity patterns in the anatomical contralateral IPS ROI (RF = 0.010, $p < 0.001$), with comparable fidelity to the contralateral V1 – V3 (two-tailed $p = 0.348$). While the ipsilateral IPS ROI also yielded accurate orientation reconstructions (RF = 0.004, $p = 0.012$), they were of significantly lower fidelity than in the contralateral V1 – V3 (two-tailed $p = 0.015$).

At the end of the first memory delay, 10 – 14s post-encoding, we found reliable orientation reconstruction in both contralateral and ipsilateral V1 – V3 (RFs = 0.018, both $p < 0.001$), and in the contralateral IPS ROI (RF = 0.013, $p = 0.001$). The average quality of the reconstructions did not differ between the contralateral and ipsilateral early visual areas (two-tailed $p = 0.938$) or between the contralateral early visual areas and IPS (two-tailed $p = 0.386$). In contrast, we did not observe reliable orientation representations in the ipsilateral IPS during this time-point (RF =

Distractor effects in visual working memory

0.001, $p = 0.409$). Therefore, we did not include the ipsilateral IPS in subsequent post-distractor analyses, in which the encoding model was trained on data from this pre-distractor memory delay.

Effect of distractors on subsequent VWM reconstructions. In order to examine whether VWM representations were unaffected, disrupted, or biased by the distractor orientations, we applied a similar inverted encoding model orientation reconstruction approach. However, instead of training and testing within a single time-point, we trained the encoding model on the end of the first memory delay (10 to 14s post-stimulus), and separately reconstructed orientation representations for each of the three distractor conditions, based on an average activity pattern from the three time-points at the end of the second delay (20 to 26s post-stimulus). To increase our power to observe distractor effects, we mirrored the resulting orientation reconstruction functions from the counterclockwise distractor trials and averaged them with the clockwise distractor reconstructions. Therefore, in the “distractor” reconstruction plots, the “correct” orientation is centered at 0 degrees, with the distractor at positive 40 – 50 degrees.

On trials without distractors (Figure 4, left column), remembered orientations could be reliably reconstructed in the contralateral (RF = 0.020, $p = 0.009$) and ipsilateral (RF = 0.014, $p = 0.041$) early visual areas at the end of the second memory delay, although reconstructions in ipsilateral V1-V3 did not exceed the false discovery rate (FDR) corrected threshold of $p = 0.023$. However, there was no difference in reconstruction accuracy between the contra- and ipsilateral early visual areas (two-tailed $p = 0.430$). Finally, orientation information in VWM could not be reliably reconstructed from the contralateral IPS ROI at this time-point on the no-distractor trials (RF = 0.005, $p = 0.264$), although reconstruction quality in the contralateral IPS did not significantly differ from that in the contralateral early visual ROI (two-tailed $p = 0.129$).

Distractor effects in visual working memory

In contrast, after the presentation of a distractor (Figure 4, right column), reconstructions from the contralateral early visual areas were not reliably centered on the correct orientation (RF = 0.006, $p = 0.133$). There were, however, accurate reconstructions in the contralateral IPS (RF = 0.013, $p = 0.023$), and a trend toward accurate reconstructions in the ipsilateral V1 – V3 (RF = 0.008, $p = 0.049$, does not exceed FDR threshold of $p = 0.023$), although neither significantly differed in accuracy from the contralateral early visual area reconstructions (ipsilateral V1-V3: two-tailed $p = 0.685$; contralateral IPS: two-tailed $p = 0.344$). Separate reconstructions for the counterclockwise and clockwise trials are presented in Extended Data Figure 4-1.

Testing for attractive biases in reconstructed orientation representations. The failure to observe accurate reconstructions in early visual areas post-distractor could be due to distractor-related biases in early visual representations. Because the following analyses simultaneously incorporate the data from the counterclockwise and clockwise distractor conditions in looking at the slope across counterclockwise, no-, and clockwise distractor conditions, we did not explicitly pool the distractor data in this analysis.

First, to assess whether reconstructed VWM representations from each ROI were biased toward the previously-presented distractor orientations, as found in behavior, we calculated the orientation at which each participant's orientation reconstruction function was maximal, separately for the counterclockwise, no-, and clockwise distractor conditions, and then calculated the circular mean and standard error of these values across participants for visualization purposes (Figure 5a). We found a linear effect of distractor condition on reconstruction error size in the contralateral early visual ROI (Figure 5a, left), such that reconstructions were biased an average of 15 degrees toward the distractors, but centered near zero when no distractors were presented. We found a similar effect of distractor condition on reconstruction bias in the ipsilateral early

Distractor effects in visual working memory

visual ROI (Figure 5a, center). No effect of distractor condition on reconstruction bias was found in the contralateral IPS ROI (Figure 5a, right).

To quantify these effects, we computed, for each participant and ROI, the slope of the best-fit line to the means of the trial-wise reconstruction errors in the counterclockwise distractor, no distractor, and clockwise distractor conditions. Then, because the measured slopes in the contralateral and ipsilateral early visual areas were correlated across participants (Spearman's $\rho = 0.692$, $p = 0.016$), we averaged the slopes from these ROIs to yield a single “early visual” bias value per participant. Finally, we tested for the predicted positive slope by comparing the observed mean slope across participants to a null distribution generated by randomly permuting the signs of the individual participant slopes (see Methods). We found that there was an attractive biasing effect of the distractors on the VWM representations in the early visual ROIs (Figure 5b, left: $p = 0.041$), but not on the representations in the contralateral IPS (Figure 5b, right: $p = 0.858$; early visual vs. IPS $p = 0.019$).

Finally, to assess whether post-distractor orientation reconstructions were of comparable quality to those on trials without distractors, irrespective of the observed biases, we performed a subsequent reconstruction fidelity analysis that did not require the representation to be centered at zero (see Methods). Representations were of comparable fidelity with and without distractors in all three ROIs (contralateral V1-V3: $p = 0.276$; ipsilateral V1-V3: $p = 0.831$; contralateral IPS: $p = 0.281$, all two-tailed comparisons).

Exploratory time-resolved orientation reconstructions. To examine how the orientation representations evolved over time in each ROI, we repeated the above analysis within a 2-TR sliding window over the entire course of the trial. Because the first two participants had a shorter

Distractor effects in visual working memory

447 TR than the remaining ten (1.667s vs. 2s), their data were not included in this descriptive
448 analysis.

449 On trials without distractors (Figure 6, left column), the orientation in memory could be
450 reliably reconstructed in the contralateral V1 – V3 ROI for the majority of the trial. Although
451 weaker, a similar pattern was observed in the ipsilateral V1 – V3 ROI. In contrast, the
452 contralateral IPS ROI only showed reliable orientation coding during the first half of the trial.

453 On trials with distractors (Figure 6, right column), orientation information could be
454 reconstructed from all three ROIs in the first half of the trial. However, after distractor
455 presentation, representations in both contra- and ipsilateral V1-V3 shifted toward the distractor
456 orientation. Reliable information was observed both before and after the distractor period in the
457 IPS, although an apparent shift in the representation during the time of distractor presentation
458 may correspond to transient processing of the distractor.

459

460 6. Discussion

461 A significant puzzle that remains for models of VWM is how precise working memory
462 representations persist through subsequent perceptual input. To provide insight into this question, we
463 applied an inverted encoding model orientation reconstruction approach to multi-voxel fMRI
464 activity patterns during a delayed estimation task for orientation, with intervening distractors. Given
465 that distractors that differ from a remembered stimulus along a task-relevant feature dimension have
466 been shown to systematically bias VWM reports (Rademaker et al., 2015), we examined early visual
467 cortical and intraparietal sulcus VWM representations during perception and during VWM
468 maintenance, both before and after distraction. The use of an inverted encoding model allowed us to
469 characterize VWM representations at each of these stages of the task, to examine when and where

Distractor effects in visual working memory

470 reliable orientation representations persisted, and to test whether VWM representations in these
471 brain regions exhibited the distractor-related biases that have been observed behaviorally.

472 In line with previous results (Ester et al., 2009; Harrison and Tong, 2009; Serences et al., 2009),
473 we were able to reconstruct reliable orientation information from visual cortex (V1 – V3) during
474 both perception and VWM maintenance. Orientations held in VWM can be decoded from both
475 ipsilateral and contralateral visual cortex (Ester et al., 2009), although decoding can also be specific
476 to the contralateral hemisphere if the task demands that the representation be spatially bound (Pratte
477 and Tong, 2014). Indeed, we found that while perceptual reconstructions showed the expected
478 retinotopic bias, orientation information was represented with comparable quality in both
479 hemispheres of early visual cortex during VWM maintenance. In addition, orientation information
480 could be reconstructed from the IPS contralateral to the remembered stimulus, both at perception
481 and during VWM maintenance (Ester et al., 2015). Finally, the IPS ipsilateral to the remembered
482 stimulus also showed reliable orientation coding during perception, but this representation was
483 significantly less reliable in this region than in the early visual areas and did not persist during
484 VWM maintenance.

485 We next tested whether an oriented grating presented in the same spatial location as the
486 remembered stimulus would a) bias behavior and b) bias or completely disrupt the observed VWM
487 representations. First, at the behavioral level, we found that distractor presentation did not decrease
488 mean VWM precision or increase the frequency of random guesses, but that the reported orientation
489 was biased about one degree toward the orientation of the distractor. Similarly, at the neural level,
490 we found that orientation representations in contralateral V1 – V3 were significantly biased toward
491 distractor orientations. In contrast, the orientation representations in the IPS showed no evidence for
492 a distractor bias, and continued to represent the remembered orientation after visual interference. On
493 trials in which distractors were *not* presented, VWM representations persisted through the end of the

Distractor effects in visual working memory

494 trial in the early visual areas, but not the IPS. This pattern of results is in contrast to previous
 495 findings (Bettencourt and Xu, 2015), which suggested that superior IPS maintains VWM
 496 representations for orientation with and without distraction, and that early visual areas maintain
 497 VWM representations only when distraction is unpredictable. Rather, our data suggests that precise
 498 stimulus-specific VWM representations for orientation are maintained in the early visual areas, and
 499 the involvement of IPS during VWM dynamically adapts to task demands. We found that orientation
 500 information was maintained redundantly in the early visual areas and in the IPS in anticipation of
 501 possible distraction, and it persisted in the early visual areas (and less-so in the IPS), if a distractor
 502 did not appear. However, while these early visual VWM representations were susceptible to
 503 interference, IPS representations were distractor resistant. It could be, therefore, that the difference
 504 we observed between the magnitudes of the neural and behavioral biases is a result of this adaptive
 505 IPS involvement.

506 The finding that content-specific VWM representations can be redundantly maintained in the
 507 visual and parietal cortices, and that those representations are differentially affected by distraction,
 508 underscores the flexibility of the underlying neural mechanisms supporting VWM. The set of
 509 regions primarily involved in VWM maintenance, and the dominant representational scheme (see
 510 (Serences, 2016) for a review), may vary both as a function of task demands and of expectations
 511 regarding the potential for distraction (Bettencourt and Xu, 2015). In our task, participants knew that
 512 distractors were likely, and that they would appear in the same spatial location as the remembered
 513 stimulus. This may have encouraged both a spreading of the representation to ipsilateral early visual
 514 areas and a preparatory engagement of the IPS. Interestingly, it was not the case that early visual
 515 areas switched entirely to coding the distractor orientations. Rather, distractor processing resulted in
 516 biased representations that were intermediate to the remembered and distractor orientations. While
 517 we do not have the cortical layer-specific measurements to fully disentangle early visual area-

Distractor effects in visual working memory

518 intrinsic maintenance from that supported by tonic top-down input from prefrontal (Mendoza-
519 Halliday et al., 2014) or parietal cortices, it is possible that feedback is essential for reducing the
520 influence of task-irrelevant distractors on precise VWM representations.

521 Finally, future studies should investigate whether the observed dynamic trade-off between the
522 early visual areas and the IPS is the result of an explicit strategy shift or whether it is a more
523 automatic process. For instance, it is possible that orientation information is by default redundantly
524 coded in VWM in the early visual areas and the IPS (although perhaps in a more spatial code in the
525 IPS). This IPS representation could then be drawn upon for future behavior if more sensory-based
526 codes have been degraded by visual interference. Further, it could be that the small behavioral biases
527 that have been observed as a result of distraction reflect a weighted averaging of the VWM signals
528 from early visual areas and IPS, informed by a metacognitive readout of VWM representational
529 uncertainty. Indeed, participants show greater behavioral biases in motion perception (Vintch and
530 Gardner, 2014) and memory for spatial frequency (Dubé et al., 2014) under conditions of increased
531 uncertainty. In addition, sensory uncertainty decoded from visual cortex is predictive of a common
532 perceptual bias, in which orientations are perceived as biased away from the cardinal axes (van
533 Bergen et al., 2015). Thus, future studies should examine whether the combination of uncertainty
534 and distractor-related biases decoded from early visual areas and/or IPS during VWM maintenance
535 can predict trial-by-trial behavioral responses.

Distractor effects in visual working memory

536 **7. References**

- 537 Awh E, Jonides J (2001) Overlapping mechanisms of attention and spatial working memory. *Trends*
538 *Cogn Sci (Regul Ed)* 5:119–126.
- 539 Benjamini Y (2001) The control of the false discovery rate in multiple testing under dependency.
540 *Annals of Statistics* 29:1165–1188.
- 541 Bennett PJ, Cortese F (1996) Masking of spatial frequency in visual memory depends on distal, not
542 retinal, frequency. *Vision Research* 36:233–238.
- 543 Berens P (2009) CircStat: a MATLAB toolbox for circular statistics. *J Stat Softw.*
- 544 Berger B, Omer S, Minarik T, Sterr A, Sauseng P (2014) Interacting memory systems-does EEG
545 alpha activity respond to semantic long-term memory access in a working memory task?
546 *Biology (Basel)* 4:1–16.
- 547 Bettencourt KC, Xu Y (2015) Decoding the content of visual short-term memory under distraction
548 in occipital and parietal areas. *Nature Publishing Group* 19:150–157.
- 549 Brainard DH (1997) The Psychophysics Toolbox. *Spat Vis* 10:433–436.
- 550 Brouwer GJ, Heeger DJ (2009) Decoding and reconstructing color from responses in human visual
551 cortex. *The Journal of Neuroscience.*
- 552 Chelazzi L, Miller EK, Duncan J, Desimone R (1993) A neural basis for visual search in inferior
553 temporal cortex. *Nature* 363:345–347.
- 554 Christophel TB, Cichy RM, Hebart MN, Haynes J-D (2014) Parietal and early visual cortices encode
555 working memory content across mental transformations. *NeuroImage.*
- 556 Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance
557 neuroimages. *Comput Biomed Res* 29:162–173.
- 558 Cox RW, Hyde JS (1996) AFNI: Software for analysis and visualization of functional magnetic
559 resonance neuroimages. *Comput-ers and Biomedical Research*, 29, 162-173. Stuttgart:
560 George Thieme Verlag. Thesen.
- 561 Curtis CE, D'Esposito M (2003) Persistent activity in the prefrontal cortex during working memory.
562 *Trends Cogn Sci (Regul Ed)* 7:415–423 Available at:
563 <http://linkinghub.elsevier.com/retrieve/pii/S1364661303001979>.
- 564 D'Esposito M, Detre JA, Alsop DC, Shin RK, Atlas S, Grossman M (1995) The neural basis of the
565 central executive system of working memory. *Nature* 378:279–281.
- 566 D'Esposito M, Postle BR, Rypma B (2000) Prefrontal cortical contributions to working memory:
567 evidence from event-related fMRI studies. *Exp Brain Res* 133:3–11.
- 568 Destrieux C, Fischl B, Dale A, Halgren E (2010) Automatic parcellation of human cortical gyri and
569 sulci using standard anatomical nomenclature. *NeuroImage* 53:1–15.

Distractor effects in visual working memory

- 570 Dubé C, Zhou F, Kahana MJ, Sekuler R (2014) Similarity-based distortion of visual short-term
571 memory is due to perceptual averaging. *Vision Research* 96:8–16 Available at:
572 <http://linkinghub.elsevier.com/retrieve/pii/S0042698913003088>.
- 573 D'Esposito M (2007) From cognitive to neural models of working memory. *Philosophical*
574 *Transactions of the Royal Society of London B: Biological Sciences* 362:761–772 Available at:
575 <http://rstb.royalsocietypublishing.org/content/362/1481/761.full>.
- 576 D'Esposito M, Postle BR (2015) The cognitive neuroscience of working memory. *Annu Rev*
577 *Psychol* 66:115–142.
- 578 Ester EF, Anderson DE, Serences JT, Awh E (2013) A neural measure of precision in visual
579 working memory. *Journal of Cognitive Neuroscience* 25:754–761.
- 580 Ester EF, Rademaker R, Sprague TC (2016) How Do Visual and Parietal Cortex Contribute to
581 Visual Short-Term Memory? *eNeuro* 3.
- 582 Ester EF, Serences JT, Awh E (2009) Spatially Global Representations in Human Primary Visual
583 Cortex during Working Memory Maintenance. *Journal of Neuroscience* 29:15258–15265
584 Available at: <http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4388-09.2009>.
- 585 Ester EF, Sprague TC, Serences JT (2015) Parietal and Frontal Cortex Encode Stimulus-Specific
586 Mnemonic Representations during Visual Working Memory. *Neuron* 87:893–905.
- 587 Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R,
588 Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM (2002) Whole brain
589 segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*
590 33:341–355.
- 591 Harrison SA, Tong F (2009) Decoding reveals the contents of visual working memory in early visual
592 areas. *Nature* 458:632–635.
- 593 Hougaard A, Jensen BH, Amin FM, Rostrup E, Hoffmann MB, Ashina M (2015) Cerebral
594 Asymmetry of fMRI-BOLD Responses to Visual Stimulation. *Stamatakis EA, ed. PLoS ONE*
595 10:e0126477.
- 596 Huang J, Sekuler R (2010) Distortions in recall from visual memory: two classes of attractors at
597 work. *Journal of Vision* 10:24.1–27.
- 598 Jeong SK, Xu Y (2016) The impact of top-down spatial attention on laterality and hemispheric
599 asymmetry in the human parietal cortex. *Journal of Vision* 16:2–2.
- 600 Kleiner M, Brainard D, Pelli D, Ingling A, Murray R (2007) What's new in Psychtoolbox-3.
601 Perception.
- 602 Lalonde J, Chaudhuri A (2002) Task-dependent transfer of perceptual to memory representations
603 during delayed spatial frequency discrimination. *Vision Research* 42:1759–1769.
- 604 Magnussen S (2000) Low-level memory processes in vision. *Trends in Neurosciences* 23:247–251
605 Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0166223600015691>.

Distractor effects in visual working memory

- 606 Magnussen S, Greenlee MW (1992) Retention and disruption of motion information in visual short-
607 term memory. *J Exp Psychol Learn Mem Cogn* 18:151–156.
- 608 Magnussen S, Greenlee MW, Asplund R, Dyrnes S (1991) Stimulus-specific mechanisms of visual
609 short-term memory. *Vision Research* 31:1213–1219.
- 610 McKeefry DJ, Burton MP, Vakrou C (2007) Speed selectivity in visual short term memory for
611 motion. *Vision Research* 47:2418–2425.
- 612 Mendoza-Halliday D, Torres S, Martinez-Trujillo JC (2014) Sharp emergence of feature-selective
613 sustained activity along the dorsal visual pathway. *Nature Neuroscience* 17:1255–1262
614 Available at: <http://www.nature.com/doi/10.1038/nn.3785>.
- 615 Miller EK, Li L, Desimone R (1993) Activity of neurons in anterior inferior temporal cortex during
616 a short-term memory task. *Journal of Neuroscience* 13:1460–1478.
- 617 Nemes VA, Parry NRA, Whitaker D, McKeefry DJ (2012) The retention and disruption of color
618 information in human short-term visual memory. *Journal of Vision* 12:26–26.
- 619 Nemes VA, Whitaker D, Heron J, McKeefry DJ (2011) Multiple spatial frequency channels in
620 human visual perceptual memory. *Vision Research* 51:2331–2339.
- 621 Pasternak T, Greenlee MW (2005) Working memory in primate sensory systems. *Nat Rev Neurosci*
622 6:97–107.
- 623 Pasternak T, Zaksas D (2003) Stimulus Specificity and Temporal Dynamics of Working Memory for
624 Visual Motion. *J Neurophysiol* 90:2757–2762.
- 625 Pelli DG (1997) The VideoToolbox software for visual psychophysics: transforming numbers into
626 movies. *Spat Vis* 10:437–442.
- 627 Postle BR (2006) Working memory as an emergent property of the mind and brain. *Neuroscience*.
- 628 Pratte MS, Tong F (2014) Spatial specificity of working memory representations in the early visual
629 cortex. *Journal of Vision* 14:22–22 Available at:
630 <http://www.journalofvision.org/lookup/doi/10.1167/14.3.22>.
- 631 Rademaker R, Bloem IM, Bloem IM, De Weerd P, De Weerd P, Sack AT (2015) The impact of
632 interference on short-term memory for visual orientation. *Journal of Experimental Psychology:*
633 *Human Perception and Performance* 41:1650–1665 Available at:
634 <http://doi.apa.org/getdoi.cfm?doi=10.1037/xhp0000110>.
- 635 Riley MR, Constantinidis C (2015) Role of Prefrontal Persistent Activity in Working Memory. *Front*
636 *Syst Neurosci* 9:181.
- 637 Saad ZS, Glen DR, Chen G, Beauchamp MS, Desai R, Cox RW (2009) A new method for
638 improving functional-to-structural MRI alignment using local Pearson correlation. *NeuroImage*
639 44:839–848.
- 640 Saad ZS, Reynolds RC (2012) SUMA. *NeuroImage* 62:768–773.

Distractor effects in visual working memory

- 641 Saad ZS, Reynolds RC, Argall B, Japee S, Cox RW (2004) SUMA: An interface for surface-based
642 intra- and inter-subject analysis with AFNI. In, pp 1510–1513. IEEE.
- 643 Serences JT (2016) Neural mechanisms of information storage in visual short-term memory. *Vision*
644 *Research* 128:53–67.
- 645 Serences JT, Ester EF, Vogel EK, Awh E (2009) Stimulus-Specific Delay Activity in Human
646 Primary Visual Cortex. *Psychological Science* 20:207–214 Available at:
647 <http://pss.sagepub.com/lookup/doi/10.1111/j.1467-9280.2009.02276.x>.
- 648 Sheremata SL, Bettencourt KC, Somers DC (2010) Hemispheric asymmetry in visuotopic posterior
649 parietal cortex emerges with visual short-term memory load. *J Neurosci* 30:12581–12588.
- 650 Sneve MH, Alnæs D, Endestad T, Greenlee MW, Magnussen S (2011) Modulation of activity in
651 human visual area V1 during memory masking. Ben-Jacob E, ed. *PLoS ONE* 6:e18651.
- 652 Sprague TC, Ester EF, Serences JT (2016) Restoring Latent Visual Working Memory
653 Representations in Human Cortex. *Neuron* 91:694–707.
- 654 Sreenivasan K, Gratton C, Vytlačil J, D'Esposito M (2014) Evidence for working memory storage
655 operations in perceptual cortex. *Cognitive*.
- 656 Sreenivasan K, Jha AP (2007) Selective attention supports working memory maintenance by
657 modulating perceptual processing of distractors. *Journal of Cognitive Neuroscience* 19:32–41
658 Available at: <http://www.mitpressjournals.org/doi/10.1162/jocn.2007.19.1.32>.
- 659 Suchow JW, Brady TF, Fougner D, Alvarez GA (2013) Modeling visual working memory with the
660 MemToolbox. *Journal of Vision* 13:9–9.
- 661 van Bergen RS, Ji Ma W, Pratte MS, Jehee JFM (2015) Sensory uncertainty decoded from visual
662 cortex predicts behavior. *Nature Publishing Group*.
- 663 Vintch B, Gardner JL (2014) Cortical correlates of human motion perception biases. *Journal of*
664 *Neuroscience*.
- 665 Wandell BA, Winawer J (2011) Imaging retinotopic maps in the human brain. *Vision Research*
666 51:718–737.
- 667 Warnking J, Dojat M, Guérin-Dugué A, Delon-Martin C, Olympieff S, Richard N, Chéhikian A,
668 Segebarth C (2002) fMRI retinotopic mapping--step by step. *NeuroImage* 17:1665–1683.
- 669 Wildegger T, Myers NE, Humphreys G, Nobre AC (2015) Supraliminal But Not Subliminal
670 Distracters Bias Working Memory Recall. *Journal of Experimental Psychology: Human*
671 *Perception and Performance*.
- 672 Xu Y (2017) Reevaluating the Sensory Account of Visual Working Memory Storage. *Trends Cogn*
673 *Sci (Regul Ed)*:1–22.
- 674 Yoon JH, Curtis CE, D'Esposito M (2006) Differential effects of distraction during working
675 memory on delay-period activity in the prefrontal cortex and the visual association cortex.

Distractor effects in visual working memory

676 NeuroImage 29:1117–1126.

677 Zhang W (2008) Discrete fixed-resolution representations in visual working memory. 453:233–235
 678 Available at: <http://www.nature.com/doifinder/10.1038/nature06860>.

679

680 8. Legends

681 **Figure 1.** Delayed-estimation task for orientation, with intervening distractor

682 **Figure 2. A)** Sample no-distractor (left) and distractor (right) response error distributions
 683 and resulting model fits for an example participant. Shaded regions depict the 95%
 684 confidence intervals of the model fits. **B)** Group-level behavioral effect of distractors on
 685 mean response bias (left), guess rate (middle), and precision (right). Error bars represent
 686 between-subject SEM. See Figure 2-1 for data and model fits for all twelve participants
 687 and Figure 2-2 for the results of separately fitting the mixture model to responses from
 688 the counterclockwise and clockwise distractor conditions.

689 **Figure 3.** Orientation reconstruction functions during perception (left column) and at the
 690 end of the first memory delay (right column). Plots display averages across all
 691 participants, with the error bars representing between-subject standard error. Y-axes
 692 depict the Pearson's correlation between the reconstructed orientation channel output
 693 (C2) and the eight hypothetical orientation channels.

694 **Figure 4.** Orientation reconstructions at the end of the second memory delay, split by
 695 distractor condition. Within each participant and ROI, the inverted encoding model was
 696 trained on the first memory delay, and then used to reconstruct orientation functions
 697 separately for trials without distractors (left column), and for trials with distractors
 698 counterclockwise or clockwise to the item in memory. Within each participant,

Distractor effects in visual working memory

reconstructions from trials with counterclockwise distractors were mirrored over zero degrees and averaged with mean clockwise distractor reconstructions, to yield average distractor reconstructions (right column). *P*-values indicate the significance level with which the “correct” orientation [indicated as an orientation of 0° on these plots] was reconstructed within a given ROI, irrespective of the distractor orientation. Plots display averages across all participants, with the error bars representing between-subject standard error. See Figure 4-1 for separate reconstructions for counterclockwise and clockwise distractor conditions.

Figure 5. A) Effect of relative distractor orientation (counterclockwise (CCW) or clockwise (CW)) on mean inverted encoding model reconstruction error for remembered orientations. Error bars represent circular SEM (circular standard deviation divided by the square root of the number of participants). **B)** Effect of relative distractor orientation on subsequent individual subject orientation reconstructions in the early visual areas (left) and contralateral IPS (right). Colored lines depict individual subject best-fit lines (solid lines/filled circles: contralateral, dashed lines/empty circles: ipsilateral), while the black dotted lines illustrate the mean CCW-NO-CW slope across subjects, which was used to assess distractor bias significance.

Figure 6. Orientation reconstructions over time, relative to stimulus onset. Reconstructions from trials with counterclockwise distractors were flipped and averaged with clockwise distractor reconstructions, to yield average reconstructions from all distractor trials, where representations biased toward the distractor orientation appear shifted up from 0 degrees. To aid in visualization, only the positive portions of the reconstruction functions are depicted here. Solid white lines mark the correct orientation, and dashed white lines depict the mean relative distractor orientation. Yellow vertical

Distractor effects in visual working memory

lines indicate distractor onset time.

Figure 2-1. Individual subject error histograms and model fits for the no distractor (top row) and distractor (bottom row) conditions. Shaded regions depict the 95% confidence intervals of the model fits.

Figure 2-2 Group-level behavioral effect of separate counterclockwise (CCW) and clockwise (CW) distractors on mean response bias (left), guess rate (middle), and precision (right). Error bars represent between-subject SEM.

Figure 4-1. Orientation reconstructions at the end of the second memory delay, split by distractor condition. Within each participant and ROI, the inverted encoding model was trained on the first memory delay, and then used to reconstruct orientation functions separately for trials with counterclockwise distractors (left column), no distractors (middle column), and clockwise distractors (right column). *P*-values indicate the significance level with which the “correct” orientation [indicated as an orientation of 0° on these plots] was reconstructed within a given ROI, irrespective of the distractor orientation. Plots display averages across all participants, with the error bars representing between-subject standard error.











