Neuron

Unveiling the abstract format of mnemonic representations

Highlights

- We revealed the neural nature of abstract WM representations
- Distinct visual stimuli were recoded into a shared abstract memory format
- Memory formats for orientation and motion direction were recoded into a line-like pattern
- Such formats are more efficient and proximal to the behaviors they guide

Authors

Yuna Kwak, Clayton E. Curtis

Correspondence

clayton.curtis@nyu.edu

In brief

Kwak and Curtis demonstrate that the memory formats measured with fMRI for two different visual features, orientation and motion direction, are shared in an abstract line-like format. These results demonstrate that the goal-relevant details of past percepts are recoded into abstract working memory representations in the brain.

Kwak & Curtis, 2022, Neuron *110*, 1–7
June 1, 2022
https://doi.org/10.1016/j.neuron.2022.03.016

CellPress

Neuron

CellPress

Report

Unveiling the abstract format of mnemonic representations

Yuna Kwak[1] and Clayton E. Curtis[1,2,3,*]
[1]Department of Psychology, New York University, New York, NY 10003, USA
[2]Center for Neural Science, New York University, New York, NY 10003, USA
[3]Lead contact
*Correspondence: clayton.curtis@nyu.edu
https://doi.org/10.1016/j.neuron.2022.03.016

SUMMARY

Working memory (WM) enables information storage for future use, bridging the gap between perception and behavior. We hypothesize that WM representations are abstractions of low-level perceptual features. However, the neural nature of these putative abstract representations has thus far remained impenetrable. Here, we demonstrate that distinct visual stimuli (oriented gratings and moving dots) are flexibly recoded into the same WM format in visual and parietal cortices when that representation is useful for memory-guided behavior. Specifically, the behaviorally relevant features of the stimuli (orientation and direction) were extracted and recoded into a shared mnemonic format that takes the form of an abstract line-like pattern. We conclude that mnemonic representations are abstractions of percepts that are more efficient than and proximal to the behaviors they guide.

INTRODUCTION

The precise contents of working memory (WM) can be decoded from the patterns of neural activity in the human visual cortex (Harrison and Tong, 2009; Serences et al., 2009), suggesting that the same encoding mechanisms used for perception also store WM representations (D'Esposito and Postle, 2015). Presumably, representations decoded during memory and perception both reflect activities of neurons selective for encoded stimulus features, and therefore, the representational format of WM is sensory-like in nature (Bettencourt and Xu, 2016; Lorenc et al., 2018; Rademaker et al., 2019). However, patterns during perception of a stimulus are often poor predictors of patterns during WM maintenance. This is especially true in parietal cortex where stimulus-evoked activity fails to predict the contents of WM (Albers et al., 2013; Rademaker et al., 2019). In visual cortex, stimulus-evoked patterns of activity are worse than memory activity at predicting memory content (Harrison and Tong, 2009; Rademaker et al., 2019). Furthermore, under a range of conditions, WM representations in parietal and sometimes even in visual cortices are only slightly impacted by visual distractors (Bettencourt and Xu, 2016; Hallenbeck et al., 2021; Lorenc et al., 2018; Rademaker et al., 2019). Therefore, a reasonable hypothesis is that mnemonic codes are somehow different from perceptual codes, perhaps abstractions of low-level stimulus features. However, the format of these putative abstract representations has thus far remained impenetrable. Here, we demonstrate that different types of visual stimuli can be flexibly recoded into the same WM format when that representation is useful for memory-guided behavior. Specifically, we found that the patterns of activity in visual cortex during WM for gratings and dot motion, two very different retinal inputs, are interchangeable when participants were later tested on the orientation of the grating or the global direction of the motion. Critically, the behaviorally relevant feature of the stimuli was extracted and recoded into a shared mnemonic representation that takes the form of an abstract line-like pattern within spatial topographic maps.

RESULTS

Working memory representations for orientation and motion direction share a common format

We measured fMRI brain activity while participants used their memory to estimate the orientation of a stored grating or the stored direction of a cloud of moving dots after a 12 s retention interval (Figure 1A). Focusing on patterns of delay period activity, we first demonstrate that we could classify both grating orientation and motion direction in several maps (Figure 1B) along the visual hierarchy (Figures 2A and S1A: delay epoch, within-stimulus), consistent with previous investigations (Emrich et al., 2013; Ester et al., 2015; Harrison and Tong, 2009; Riggall and Postle, 2012; Sarma et al., 2016; Serences et al., 2009; Yu and Shim, 2017). Perhaps, neurons with orientation (Hubel and Wiesel, 1962) or directional motion selectivity (Maunsell and Van Essen, 1983) encode and maintain representations of these aspects of the physical stimuli. Alternatively, the format of these mnemonic representations might reflect efficient abstractions of the image-level properties of the stimuli. For instance, memory may take the form of a compressed, low-resolution summary of

Please cite this article in press as: Kwak and Curtis, Unveiling the abstract format of mnemonic representations, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.03.016

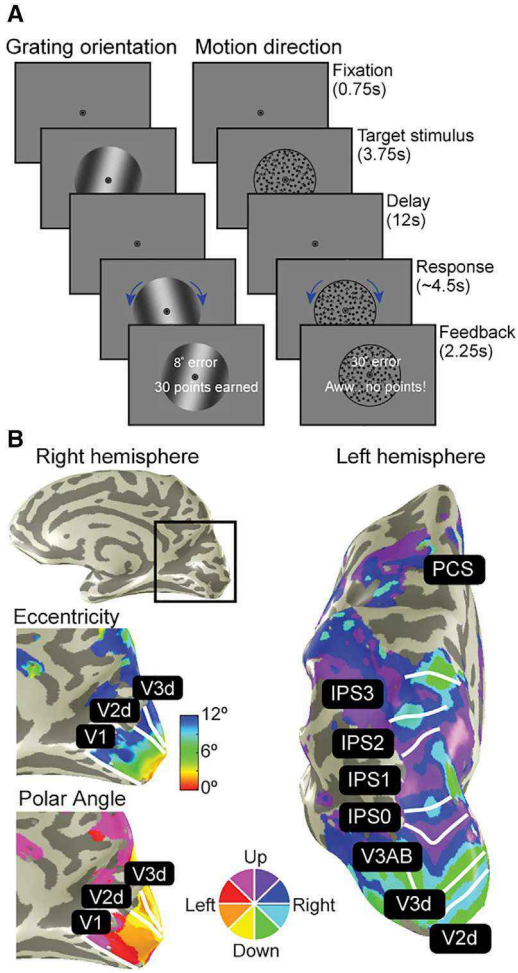

Figure 1. Neuroimaging experiments

(A) Working memory experiments. Participants maintained the orientation of gratings or the direction of dot motion over a 12 s retention interval. After the delay, participants rotated a recall probe to match their memory, and more points were awarded for more accurate memories.

(B) Population receptive field (pRF) mapping. A separate retinotopic mapping session was used to estimate voxel receptive field parameters for defining visual field maps in occipital, parietal, and frontal cortices. Example participant's right and left hemispheres are shown.

the global direction of thousands of dots moving over time akin to a line-like pointer.

From this hypothesis, we predict a similar pattern of delay period activity when abstract WM formats match despite entirely distinct perceptual inputs. In several cortical regions, a classifier trained on one type of stimulus (e.g., orientation) successfully decoded the other type of stimulus (e.g., direction) when angles matched (Figures 2A and S1A: delay epoch, cross-stimulus). Critically, evidence for an abstract WM representation that was shared across stimulus types was limited to the memory delay period. The lack of cross-stimulus decoding during the time epoch corresponding to direct viewing of the stimulus (Figures 2A and S1A: stimulus encoding epoch, cross-stimulus) indicates that the abstract format is specifically mnemonic in nature and not an artifact inherited from some shared perceptual feature during encoding. Neither can it be attributed to gaze instability as we ruled out eye movements as the potential source of significant decoding (see STAR Methods for details). In the temporal generalization matrix using continuous decoding (King and Dehaene, 2014), one can clearly see the emergence and stability of the abstract WM code during the delay period (Figures 2B, 2C, S1B, and S1C).

Format of recoded working memory representations unveiled

Although our evidence supported a WM representation that is abstract in format, we aimed to reveal the latent nature of the WM representation. We hypothesized that participants recoded the sinusoidal gratings and dot motion kinematograms into line-like images at angles matching the orientation and direction, respectively, of the stimuli. We reasoned that the abstract representation might be encoded spatially in the population activity of topographically organized visual field maps. Specifically, we predicted that the spatial distribution of higher response amplitudes across a topographic map forms a line at a given angle, as if the retinal positions constituting a line were actually visually stimulated.

To test this, we reconstructed the spatial profile of neural activity (Kok and de Lange, 2014; Yoo et al., 2022) during WM by projecting the amplitudes of voxel activity during the delay period for each orientation and direction condition into visual field space (Figure 3A) using parameters obtained from models of each map's population receptive field (pRF). Using the following equation, we computed the sum (S) of all voxels' receptive fields (the exponent, which is a Gaussian) weighted by their delay period beta coefficients (β) for each feature condition (θ_i), where i and n are indices of voxels and feature conditions, respectively; x_n, y_n, and σ are the center and width of the pRF; x_0 and y_0 are the positions in the reconstruction map at which the pRFs were evaluated:

$$S_{\theta_i} = \sum_{n=1}^{N} \beta_{i,n} \times e^{-\frac{(x_n - x_0)^2 + (y_n - y_0)^2}{\sigma^2}} \quad \text{(Equation 1)}$$

Remarkably, the visualization technique confirmed our hypothesis and unveiled a stripe encoded in the amplitudes of voxel activity at an angle matching the remembered feature in many of the visual maps (Figures 3B and S2A). This evidence strongly suggests that the neural representation of orientation/direction in memory is similar to that which would be evoked by retinal stimulation of a simple-line stimulus. Interestingly, these line-like representations had greater activation at the end of the line corresponding to the direction of motion, akin to an arrowhead, perhaps allowing for both the storage of the angle and direction of motion (Figure S4B).

To quantify the extent to which the angled stripes in the reconstructed maps aligned with the actual orientation/direction of the stimuli, we computed filtered responses and associated fidelity values (Figures 3C, 3D, S2B, and S2C). The significant fidelity values (Figure 3D) indicate that the mechanism by which orientation and motion direction is stored in WM depends on a spatial recoding into the population's topography. We found the same

Please cite this article in press as: Kwak and Curtis, Unveiling the abstract format of mnemonic representations, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.03.016

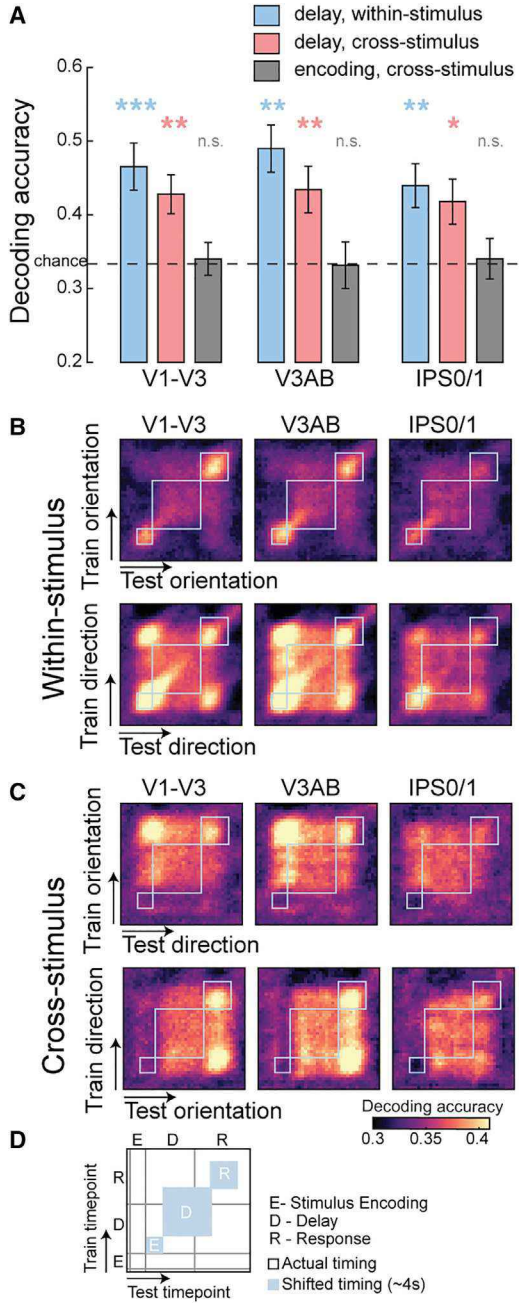

Figure 2. Working memory representations for orientation and motion direction share a common format

(A) Both remembered grating orientation and dot motion direction could be decoded from the pattern of neural activity during the memory delay (blue bars). We successfully decoded not only within but across stimulus types (e.g., training on orientation can decode motion direction; red bars), indicating that the shared patterns do not represent low-level perceptual details of each stimulus type. Moreover, no such cross-stimulus decoding existed during the stimulus encoding epoch (gray bars). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, n.s. not significant, corrected (p values in Table S1). Error bars represent ±1 SEM.

(B and C) Temporal generalization matrix. To evaluate how representations evolve over the time course of a trial, we trained and tested on all possible combinations of time points. Abstract WM codes are stable throughout the delay period. See Figure S1 for other ROIs.

(D) Schematics of the matrix plots. Gray lines denote the actual timing of events, and blue boxes show each of these events shifted by ~4 s assuming hemodynamic lag.

pattern of results when reconstructing the maps for orientation and motion direction trials separately (Figures S2D–S2G), when controlling for potential biases in pRF structure across the visual field, and when reconstructing the neural representation with the weight matrices used to classify orientation and motion direction instead of the magnitude of delay period activity. Therefore, the recoded abstract representations we discovered are robust. In summary, neural representations of gratings and dot motion, despite distinct retinal stimulation with distinct neural encoding mechanisms for perception (Cynader and Chernenko, 1976; He et al., 1998), were recoded in memory into a spatial topographic format that was line-like in nature with angles matching the remembered orientation/direction.

Distinct formats of perceptual and mnemonic representations

Next, we used an image-computable model based on properties of V1 (Roth et al., 2018; Simoncelli et al., 1992) to simulate a potential neural mechanism of the line-like patterns (Figures 4 and S3). The topographic pattern of simulated voxel activity when model inputs were simple-line images (Figure 4, bottom row) closely resembled the line-like patterns that we observed in our experimental data in Figure 3B, further strengthening a putative mechanism involving the maintenance of an imagined line in WM. Critically, using the gratings from our experiment as inputs to the model, the reconstructed maps produced line patterns orthogonal to those observed experimentally (Figure 4, top row). Such orthogonality for grating images is likely due to aperture bias during perception reported previously, where orientation preference of voxels in the early visual cortex is affected by spatial biases induced by stimulus aperture (Freeman et al., 2011; Roth et al., 2018). The dissociation between the simulated reconstructions of gratings and lines demonstrates that the format of the WM representations resembles that of a low-dimensional line rather than a high-resolution pixel-by-pixel image of a grating (Figure S3B). Therefore, based on the V1 model, we demonstrated the feasibility of a line-like WM representation. The model also helped rule out potential confounds specific to visual cortex, confounds that are unlikely to apply to parietal cortex. Furthermore, although the model does not simulate the effects of motion, we found striking evidence from our experimental data that the line-like representations during WM for motion are distinct from those during perception (Wang et al., 2014; Figure S4).

DISCUSSION

Together, these results directly impact key tenets of the sensory recruitment theory of WM, which proposes that the same neural mechanisms that encode stimulus features in sensory cortex during perception are recruited by higher-level control areas such as prefrontal and parietal cortices to support memory (Curtis and D'Esposito, 2003; Postle, 2006; Serences, 2016). First,

Please cite this article in press as: Kwak and Curtis, Unveiling the abstract format of mnemonic representations, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.03.016

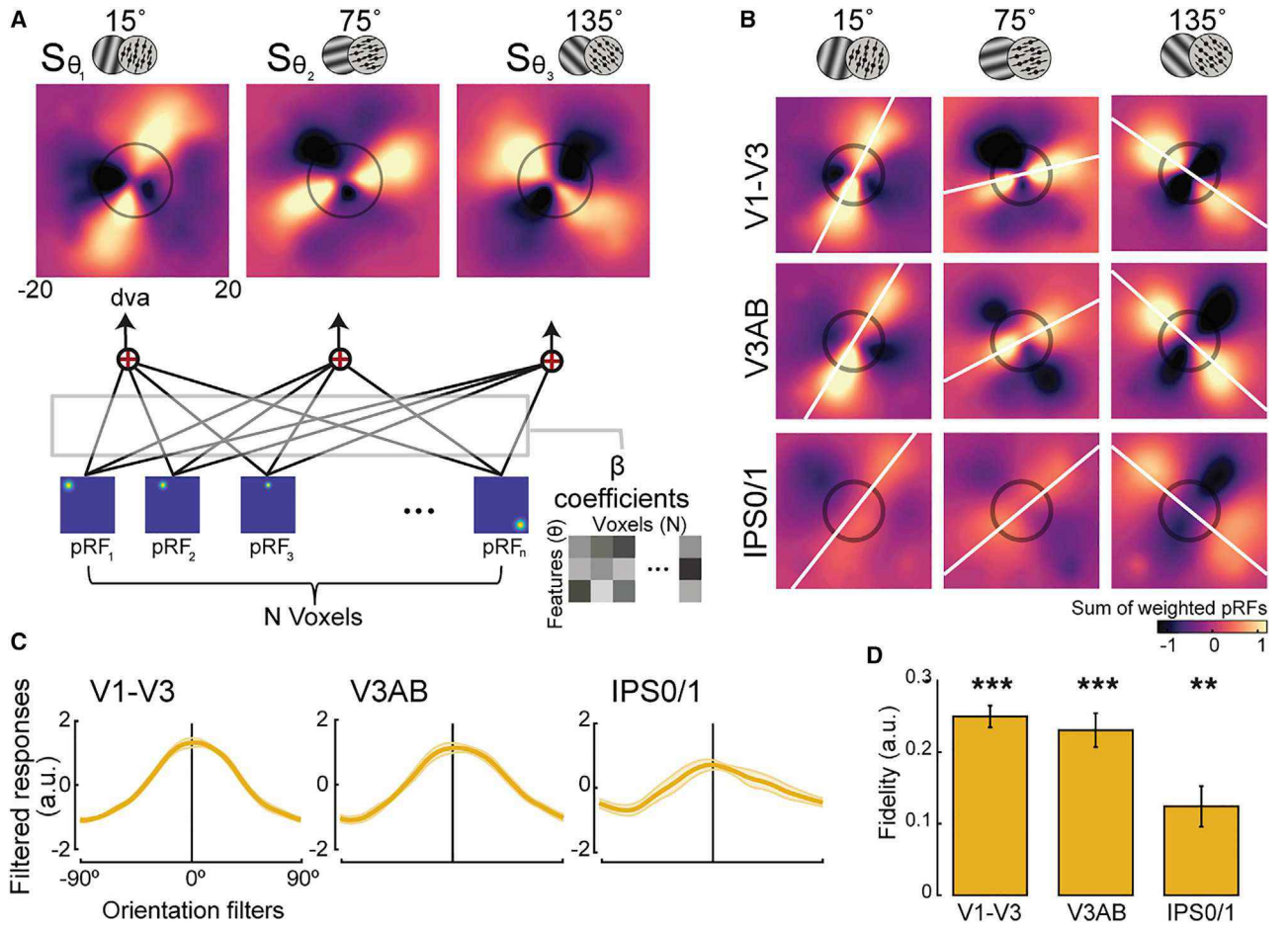

Figure 3. Unveiling the recoded formats of working memory representations for orientation and motion direction

(A) Spatial reconstruction analysis schematics. Voxel activity for each feature condition (orientation/direction; θ) was projected onto visual field space (dva, degrees of visual angle) by computing the sum (S) of voxels' pRFs weighted by their response amplitudes during the memory delay (β) (see Equation 1; STAR Methods for details).

(B) Population reconstruction maps. Lines across visual space matching the remembered angles of the stimuli emerged from the amplitudes of topographic voxel activity. Best fitting lines (white lines) and the size of the stimulus presented during the encoding epoch (black circles) are shown. See Figure S2 for other ROIs.

(C and D) To quantify the amount of remembered information consistent with the true feature, we computed filtered responses and associated fidelity values from the maps in (B). Filtered responses represent the sum of pixel values within the area of a line-shaped mask oriented −90° to 90° (0° represents the true feature), and fidelity values are the result of projecting the filtered responses to 0° (STAR Methods). Higher fidelity values indicate stronger representation. **$p < 0.01$, ***$p < 0.001$, corrected (p values in Table S1). Error bars represent ±1 SEM.

our findings indicate that the mechanisms by which WM and perception are encoded in visual cortex can differ in certain circumstances. The recoding of gratings and dot motion into line-like WM representations explains why previous studies have reported null or weaker effects—brain activation patterns during perception of a stimulus are poor predictors of patterns during WM, especially when compared with training on patterns during WM (Albers et al., 2013; Hallenbeck et al., 2021; Rademaker et al., 2019; Serences et al., 2009; Spaak et al., 2017). Although WM representations may contain some sensory-like information, our results provide robust evidence that perceptual information can be reformatted for memory storage. Second, the spatial nature of the recoded WM representation offers intriguing insights into how the parietal cortex, which lacks neurons with clear orientation or motion directional tuning (Kusunoki et al., 2000), supports WM. Recoding into a spatial format may exploit the prevalent spatial representations in parietal cortex (Heilman et al., 1985; Mackey et al., 2016, 2017) and explain why previous studies have been able to decode features such as orientation from patterns in parietal cortex (Bettencourt and Xu, 2016; Ester et al., 2015; Rademaker et al., 2019; Yu and Shim, 2017). In addition to serving as a mechanism for WM storage, this spatial code in parietal cortex might conceivably reflect the origins of top-down feedback. Indeed, WM representations in visual cortex are thought to depend on feedback signals (van Kerkoerle et al., 2017; Rahmati et al., 2018). This may be why we found the line-like spatial code in the early visual cortex despite the abundance of neurons tuned for orientation and motion (Hubel and Wiesel, 1962; Maunsell and Van Essen, 1983). Perhaps, the retinotopic organization shared between visual

Please cite this article in press as: Kwak and Curtis, Unveiling the abstract format of mnemonic representations, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.03.016

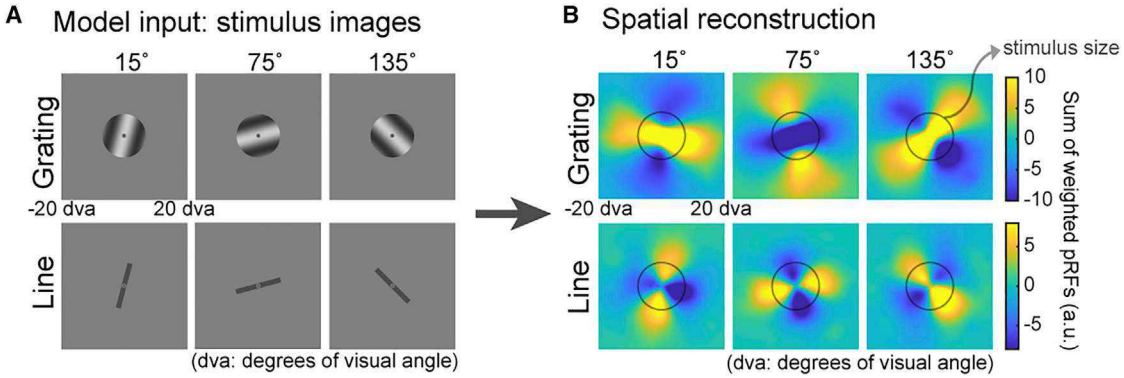

Figure 4. Model simulation for orientation: distinct formats for perceptual and mnemonic representations

(A) Gratings identical to the ones used in the experiment and lines oriented at 15°, 75°, and 135° clockwise from vertical were fed into an image-computable model of V1 to simulate the population response.

(B) We repeated the spatial reconstruction analysis in Figure 3A with simulated V1 voxel amplitudes acquired through sampling the model output neuronal responses (Figure S3A) with pRF parameters (STAR Methods). The simulated topographic patterns when model inputs are lines, but not gratings, closely match our experiment data (Figure 3B), suggesting that mnemonic representations of oriented gratings are similar to perceptual representations of simple line.

and parietal cortices forms an interface where the currency of the feedback is simple retinotopic space. Alternatively, a more complicated feedback mechanism may exist that specifically targets neurons that are both tuned to the memorized feature and contain overlapping spatial receptive fields. Third, distinct representational formats for perception and memory solves the mystery of how visual cortex can simultaneously represent WM content and process incoming visual information without catastrophic loss (Bettencourt and Xu, 2016; Buschman, 2021; Hallenbeck et al., 2021; Rademaker et al., 2019).

These results also impact how we think about the nature of WM representations, especially when we consider the larger constraints and goals of the memory system. A line is an efficient summary representation of orientation and motion direction compared with the complexity of the sensory stimuli from which it is abstracted and may provide a means to overcome the hallmark capacity limits of WM (Miller, 1956). Compare in bits of information the hundreds of dots displaced each frame over several seconds to a single line pointing in the direction of motion. But why a line? The immediate behavioral goal—the "working" part of WM—may largely determine the nature of the code. Consistent with the idea of WM as a goal-directed mechanism, previous studies have found that only the task-relevant feature, but not those irrelevant to the task, is retained in WM (Serences et al., 2009; Yu and Shim, 2017). Therefore, we note that the current experimental task, in which participants were asked to remember orientation and direction among many other visual properties of the stimuli, could have encouraged the recoding into the spatial line representation, a format most proximal to the specific mnemonic demands of the task. In fact, the flexible nature of WM might lead to different representational formats depending on the stimulus being remembered and the task at hand. Written letters or words are converted from a visual into a phonological or sound-based code when stored in WM (Baddeley, 1992). Prospective motor codes are possible when memory-guided responses can be planned or anticipated (Boettcher et al., 2021; Curtis and D'Esposito, 2006; Curtis et al., 2004). When appropriate, WM representations can take the form of the categories abstracted from perceptual exemplars (Lee et al., 2013). On the other hand, it is likely that memory for a stimulus feature in its simplest form, such as visual location that cannot be simplified any more than its spatial coordinates, does not undergo recoding (Hallenbeck et al., 2021; Li et al., 2021).

The idea that memory representations can take a form other than the perceptual features of the stimulus is not new. Our results are striking, however, because they both clearly establish the existence of abstract WM codes in the visual system and more importantly unveil the nature of these WM representations. Visualizing the abstractions of stimuli in the topographic patterns of brain activity is powerful evidence that visual cortex acts as a blackboard for cognitive representations rather than simply a register for incoming visual information (Roelfsema and de Lange, 2016).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Experimental stimuli and task
  - fMRI data acquisition
  - fMRI data preprocessing
  - Regions-of-interest definition
  - fMRI data analysis: Decoding accuracy
  - fMRI data analysis: Spatial reconstruction
  - Model simulation: Image-computable model of V1

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2022.03.016.

ACKNOWLEDGMENTS

We thank David Heeger, Brenden Lake, Masih Rahmati, and Hsin-Hung Li for helpful comments on the project, and Jonathan Winawer for helpful comments on the manuscript. We also thank New York University's Center for Brain Imaging for technical support. This research was supported by the National Eye Institute (NEI) (R01 EY-016407 to C.E.C. and R01 EY-027925 to C.E.C.).

AUTHOR CONTRIBUTIONS

Conceptualization, Y.K. and C.E.C.; methodology, Y.K. and C.E.C.; investigation, Y.K.; visualization, Y.K. and C.E.C.; funding acquisition, C.E.C.; project administration, Y.K. and C.E.C.; supervision, C.E.C.; writing – original draft, Y.K. and C.E.C.; writing – review and editing, Y.K. and C.E.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 14, 2021
Revised: February 3, 2022
Accepted: March 11, 2022
Published: April 7, 2022

REFERENCES

Albers, A.M., Kok, P., Toni, I., Dijkerman, H.C., and de Lange, F.P. (2013). Shared representations for working memory and mental imagery in early visual cortex. Curr. Biol. *23*, 1427–1431.

Baddeley, A. (1992). Working memory. Science *255*, 556–559.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. *57*, 289–300.

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

Boettcher, S.E.P., Gresch, D., Nobre, A.C., and van Ede, F. (2021). Output planning at the input stage in visual working memory. Sci. Adv. *7*, eabe8212.

Brainard, D.H. (1997). The psychophysics toolbox. Spat. Vis. *10*, 433–436.

Buschman, T.J. (2021). Balancing flexibility and interference in working memory. Annu. Rev. Vis. Sci. *7*, 367–388.

Curtis, C.E., and D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. Trends Cogn. Sci. *7*, 415–423.

Curtis, C.E., and D'Esposito, M. (2006). Selection and maintenance of saccade goals in the human frontal eye fields. J. Neurophysiol. *95*, 3923–3927.

Curtis, C.E., Rao, V.Y., and D'Esposito, M. (2004). Maintenance of spatial and motor codes during oculomotor delayed response tasks. J. Neurosci. *24*, 3944–3952.

Cynader, M., and Chernenko, G. (1976). Abolition of direction selectivity in the visual cortex of the cat. Science *193*, 504–505.

D'Esposito, M., and Postle, B.R. (2015). The cognitive neuroscience of working memory. Annu. Rev. Psychol. *66*, 115–142.

Dumoulin, S.O., and Wandell, B.A. (2008). Population receptive field estimates in human visual cortex. Neuroimage *39*, 647–660.

Emrich, S.M., Riggall, A.C., LaRocque, J.J., and Postle, B.R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. J. Neurosci. *33*, 6516–6523.

Ester, E.F., Sprague, T.C., and Serences, J.T. (2015). Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. Neuron *87*, 893–905.

Freeman, J., Brouwer, G.J., Heeger, D.J., and Merriam, E.P. (2011). Orientation decoding depends on maps, not columns. J. Neurosci. *31*, 4792–4804.

Hallenbeck, G.E., Sprague, T.C., Rahmati, M., Sreenivasan, K.K., and Curtis, C.E. (2021). Working memory representations in visual cortex mediate distraction effects. Nat. Commun. *12*, 4714.

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

He, S., Levick, W.R., and Vaney, D.I. (1998). Distinguishing direction selectivity from orientation selectivity in the rabbit retina. Vis. Neurosci. *15*, 439–447.

Heilman, K.M., Bowers, D., Coslett, H.B., Whelan, H., and Watson, R.T. (1985). Directional hypokinesia: prolonged reaction times for leftward movements in patients with right hemisphere lesions and neglect. Neurology *35*, 855–859.

Hubel, D.H., and Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. *160*, 106–154.

King, J.-R., and Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. Trends Cogn. Sci. *18*, 203–210.

Kok, P., and de Lange, F.P. (2014). Shape perception simultaneously up- and downregulates neural activity in the primary visual cortex. Curr. Biol. *24*, 1531–1535.

Kusunoki, M., Gottlieb, J., and Goldberg, M.E. (2000). The lateral intraparietal area as a salience map: the representation of abrupt onset, stimulus motion, and task relevance. Vision Res. *40*, 1459–1468.

Lee, S.-H., Kravitz, D.J., and Baker, C.I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. Nat. Neurosci. *16*, 997–999.

Li, H.-H., Sprague, T.C., Yoo, A.H., Ma, W.J., and Curtis, C.E. (2021). Joint representation of working memory and uncertainty in human cortex. Neuron *109*, 3699–3712.e6.

Lorenc, E.S., Sreenivasan, K.K., Nee, D.E., Vandenbroucke, A.R.E., and D'Esposito, M. (2018). Flexible coding of visual working memory representations during distraction. J. Neurosci. *38*, 5267–5276.

Mackey, W.E., Devinsky, O., Doyle, W.K., Golfinos, J.G., and Curtis, C.E. (2016). Human parietal cortex lesions impact the precision of spatial working memory. J. Neurophysiol. *116*, 1049–1054.

Mackey, W.E., Winawer, J., and Curtis, C.E. (2017). Visual field map clusters in human frontoparietal cortex. Elife *6*, e22974.

Maunsell, J.H., and Van Essen, D.C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. J. Neurophysiol. *49*, 1127–1147.

Miller, G.A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol. Rev. *63*, 81–97.

Postle, B.R. (2006). Working memory as an emergent property of the mind and brain. Neuroscience *139*, 23–38.

Rademaker, R.L., Chunharas, C., and Serences, J.T. (2019). Coexisting representations of sensory and mnemonic information in human visual cortex. Nat. Neurosci. *22*, 1336–1344.

Rahmati, M., Saber, G.T., and Curtis, C.E. (2018). Population dynamics of early visual cortex during working memory. J. Cogn. Neurosci. *30*, 219–233.

Riggall, A.C., and Postle, B.R. (2012). The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. J. Neurosci. *32*, 12990–12998.

Rissman, J., Gazzaley, A., and D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage *23*, 752–763.

Roelfsema, P.R., and de Lange, F.P. (2016). Early visual cortex as a multiscale cognitive blackboard. Annu. Rev. Vis. Sci. *2*, 131–151.

Roth, Z.N., Heeger, D.J., and Merriam, E.P. (2018). Stimulus vignetting and orientation selectivity in human visual cortex. Elife *7*, e37241.

Sarma, A., Masse, N.Y., Wang, X.-J., and Freedman, D.J. (2016). Task-specific versus generalized mnemonic representations in parietal and prefrontal cortices. Nat. Neurosci. *19*, 143–149.

Serences, J.T. (2016). Neural mechanisms of information storage in visual short-term memory. Vision Res. *128*, 53–67.

Serences, J.T., Ester, E.F., Vogel, E.K., and Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. Psychol. Sci. *20*, 207–214.

Simoncelli, E.P., Freeman, W.T., Adelson, E.H., and Heeger, D.J. (1992). Shiftable multiscale transforms. IEEE Trans. Inf. Theor. *38*, 587–607.

Spaak, E., Watanabe, K., Funahashi, S., and Stokes, M.G. (2017). Stable and dynamic coding for working memory in primate prefrontal cortex. J. Neurosci. *37*, 6503–6516.

van Kerkoerle, T., Self, M.W., and Roelfsema, P.R. (2017). Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. Nat. Commun. *8*, 13804.

Wandell, B.A., Dumoulin, S.O., and Brewer, A.A. (2007). Visual field maps in human cortex. Neuron *56*, 366–383.

Wang, H.X., Merriam, E.P., Freeman, J., and Heeger, D.J. (2014). Motion direction biases and decoding in human visual cortex. J. Neurosci. *34*, 12601–12615.

Yoo, A.H., Bolaños, A., Hallenbeck, G.E., Rahmati, M., Sprague, T.C., and Curtis, C.E. (2022). Behavioral prioritization enhances working memory precision and neural population gain. J. Cogn. Neurosci. *34*, 365–379.

Yu, Q., and Shim, W.M. (2017). Occipital, parietal, and frontal cortices selectively maintain task-relevant features of multi-feature objects in visual working memory. Neuroimage *157*, 97–107.

Please cite this article in press as: Kwak and Curtis, Unveiling the abstract format of mnemonic representations, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.03.016

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
fMRI data	This paper	https://osf.io/t6b95/(https://doi.org/10.17605/OSF.IO/T6B95)
Software and algorithms		
MATLAB	MathWorks	https://www.mathworks.com/products/matlab.html
Custom code and algorithm	This paper	https://github.com/yuna.kwak and https://github.com/clayspacelab (https://doi.org/10.5281/zenodo.6342189)
Image-computable model of V1	Roth et al., 2018	https://github.com/elifesciences-publications/stimulusVignetting
Decoding	Princeton MVPA toolbox	https://github.com/princetonuniversity/princeton-mvpa-toolbox

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Clayton Curtis (clayton.curtis@nyu.edu).

Materials availability

This study did not produce new materials.

Data and code availability

The processed fMRI data generated in this study have been deposited in the Open Science Framework https://osf.io/t6b95. Processed fMRI data contains extracted beta coefficients from each voxel of each ROI. The raw fMRI data are available under restricted access to ensure participant privacy; access can be obtained by contacting the corresponding authors. All original code for data analysis is publicly available on GitHub https://github.com/clayspacelab and https://github.com/yunakwak as of the date of publication.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Eleven neurologically healthy volunteers (including the 2 authors; 6 females; 25-50 years old) with normal or corrected-to-normal vision participated in this study. They gave informed consent approved by New York University IRB (protocol IRB-FY2016-852). Each participant completed 2 experimental sessions (~2hrs each) and 1-2 sessions of retinotopic mapping and anatomical scans (~2hrs).

METHOD DETAILS

Experimental stimuli and task

Participants performed a delayed-estimation WM task where they reported the remembered orientation or motion direction. Each trial began with 0.75s of central fixation (subtended 0.7° diameter) followed by the presentation of a target stimulus for 3.75s (presented in donut-shaped circular aperture with 1.5° inner and 15° outer diameter). The stimulus was a drifting grating (contrast = 0.6, spatial frequency = 0.1 cycle/°, phase change = 0.12°/frame) or a random dot kinematogram (RDK; number of dots = 900, size of each dot = 0.12°, speed of each dot = 0.08°/frame, initial coherence = 0.91), presented in blocked designs and in interleaved order. After a 12s delay period, they rotated a recall probe with a dial to match the remembered orientation/direction within a 4.5s response window. The recall probe matched the target stimulus type presented during encoding (e.g., target and recall probe were both gratings for orientation), instead of line bars for example, to prevent from enforcing participants to represent the two stimulus types in an abstract manner. However, there were two exceptions: the recall probe for orientation was a static grating without drifting motion, and the coherence of RDK dots was fixed at 1 throughout the whole response period (see below for adjusting dot coherence level in the stimulus encoding epoch). Participants were provided with feedback on the magnitude of error (°) and the points earned based on the error they made on each trial (40 points for error less than 3°, 30 points for error between 3°-12°, 20 points for error between 12°-21°, 10 points for error between 21°-30°, and no points for error exceeding 30°). The feedback display lasted for 2.25s and was followed by an inter-trial-interval of 7.5s or 9s. All stimuli were presented in a circular aperture spanning the whole visual field.

Please cite this article in press as: Kwak and Curtis, Unveiling the abstract format of mnemonic representations, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.03.016

Data were collected across two sessions for each participant resulting in 22 runs (10 orientation, 12 motion direction) for 8 participants and 21 runs (10 orientation, 11 motion direction) for 3 participants. Each run consisted of 12 trials. For the orientation runs, the target orientations were 15°, 75°, and 135° clockwise from vertical with random jitters (<5°), each repeated 4 times within a run. For motion direction runs, the target directions were 15°, 75°, 135°, 195° (180° apart from 15° and thus opposite in direction from 15°), 255° (180° apart from 75°), and 315° (180° apart from 135°) clockwise from vertical with random jitters (<5°), each repeated 2 times within a run. We collected 1-2 more runs for the motion direction for each participant because we collapsed across the opposite direction trials and labeled them as the same condition for decoding. The orientation/direction of the recall probe was randomized, with the constraint that the direction difference between the target and probe RDKs was less than 90° to match with the orientation trials.

To address the difference in task difficulty across the two stimulus types, dot coherence for the RDK target was staircased. The first trial of each motion direction run began with the coherence of 0.91, and the coherence of the following trials followed a 1-up 2-down staircase method with reference to the performance of the previous orientation run (median of error in trials of the previous orientation run). Although we tried to match the task difficulty across the two stimulus types, performance was significantly lower ($t(10)= 2.525$, $p = 0.023$) for the motion direction trials (mean = 9.26°, s.d. = 3.307°) compared to the orientation trials (mean = 7.72°, s.d. = 2.681°).

fMRI data acquisition

BOLD contrast images were acquired using Multiband (MB) 2D GE-EPI (MB factor of 4, 44 slices, 2.5 x 2.5 x 2.5mm voxel size, TE/TR of 30/750ms). We also acquired distortion mapping scans to measure field inhomogeneities with normal and reversed phase encoding using a 2D SE-EPI readout and number of slices matching that of the GE-EPI (TE/TR of 45.6/3537ms). T1- and T2-weighted images were acquired using the Siemens product MPRAGE (192 slices for T1 and 224 slices for T2, 0.8 x 0.8 x 0.8mm voxel size, TE/TR of 2.24/2400ms for T1 and 564/3200ms for T2, 256 × 240 mm FOV). We collected 2-3 T1 images and 1-2 T2 images per participant.

fMRI data preprocessing

We used intensity-normalized high-resolution anatomical scans as input to Freesurfer's recon-all script (version 6.0) to identify pial and white matter surfaces, which were converted to the SUMA format. This anatomical image processed for each subject was the alignment target for all functional images. For functional preprocessing, we divided each functional session into 2 to 6 sub-sessions consisting of 2 to 5 task runs split by distortion runs (a pair of spin-echo images acquired in opposite phase encoding directions) and applied all preprocessing steps described below to each sub-session independently.

First, we corrected functional images for intensity inhomogeneity induced by the high-density receive coil by dividing all images by a smoothed bias field which was computed by a ratio of signal acquired with the head coil to that of the body coil. Then, to improve co-registration of functional data to the target T1 anatomical image, transformation matrices between functional and anatomical images were computed using distortion-corrected and averaged spin-echo images (distortion scans used to compute distortion fields restricted to the phase-encoding direction). Then we used the distortion-correction procedure in afni_proc.py to undistort and motion-correct functional images. The next step was rendering functional data from native acquisition space into un-warped, motion corrected, and co-registered anatomical space for each participant at the same voxel size as data acquisition (2.5mm iso-tropic voxel). This volume-space data was projected onto the reconstructed cortical surface, which was projected back into the volume space for all analyses.

We linearly detrended activation values from each voxel from each run. These values were then converted to percent signal change by dividing by the mean of the voxel's activation values over each run.

Regions-of-interest definition

For identifying regions of interest (ROIs) and acquiring voxels' population receptive field (pRF) parameters (Figure 1B), we collected data from a separate retinotopic mapping session for each participant (8-12 runs). Participants ran in either one of the two types of attention-demanding tasks: RDK motion direction discrimination task (6 participants) (Mackey et al., 2017) or a rapid serial visual presentation (RSVP) task of object images (5 participants).

In each trial of the RDK motion discrimination task, a bar with pseudo-randomly chosen width (2.5°, 5.0°, and 7.5°) and sweep direction (left-to-right, right-to-left, bottom-to-top, top-to-bottom) swept across 26.4° of the visual field in 12 2.6s steps. Each bar was divided into 3 equal-sized patches (left, center, right). In each sweep, the RDK direction in one of the peripheral patches matched that of the central patch, and participants reported which peripheral patch had the same motion direction with the central patch. The coherence of the RDK in the two peripheral patches was fixed at 50%, and the coherence of the RDK in the central patch was adjusted using a 3-down 1-up staircase procedure to maintain 80% accuracy.

In each trial of the object image RSVP task, a bar consisting of 6 different object images (each image subtended 4.6° x 4.6°) swept across 26.4° of the visual field in one of the 4 sweep directions (left-to-right, right-to-left, bottom-to-top, top-to-bottom) in 12 steps. In each sweep, participants reported whether the target object image existed among the 6 images. The target image was pseudo-randomly chosen in each run and was shown at the beginning of each run at 5 locations in the visual field (center, left, right, up, down). Presentation duration of the object images was set to 400ms on the first trial of each run but was adjusted according to the accuracy on the previous trial. Duration increased with accuracy below 70%, decreased with accuracy above 85%, and stayed the same for accuracy values in between.

Please cite this article in press as: Kwak and Curtis, Unveiling the abstract format of mnemonic representations, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.03.016

BOLD contrast images were acquired using MB 2D GE-EPI (MB factor of 4, 56 slices, 2 x 2 x 2mm voxel size, TE/TR: 42/1300ms). Similar to the main experimental scans (see fMRI data preprocessing), we collected distortion mapping scans to measure field inhomogeneities with normal and reversed phase encoding using a 2D SE-EPI readout and number of slices matching that of the GE-EPI (TE/TR: 71.8/6690ms). The same preprocessing steps used for the main experimental data were applied to the retinotopic mapping data with one exception. Because the experimental and retinotopy scans were acquired with different voxel grid resolution, we projected the retinotopy time series data onto the surface from its native space resolution (2mm), then from the surface to volume space at the task voxel resolution (2.5mm) to compute pRF properties in the same voxel grid as the experimental data.

Using the averaged time series across all retinotopy runs for each participant, we fit a pRF model using vistasoft (github.com/clayspacelab/vistasoft; Dumoulin and Wandell, 2008; Mackey et al., 2017). After estimating the pRF parameters for all the voxels, we projected the best-fit polar angle and eccentricity parameters onto each participant's inflated brain surface map via AFNI and SUMA. ROIs were drawn on the surface based on established criteria for polar angle reversals and foveal representations (Mackey et al., 2017; Wandell et al., 2007), with the variance explained threshold set to ~10%. We defined ROIs V1-V3, V3AB, TO1/2, IPS0/1, IPS2/3, and PCS. We merged ROIs for V1-V3, TO1/2, IPS0/1, IPS2/3, and PCS (sPCS/iPCS) which was justified as the regions being grouped belong to the same cluster defined by overlapping foveal representations (Wandell et al., 2007). For merged ROIs, the voxels were concatenated before multivariate analysis.

fMRI data analysis: Decoding accuracy

All decoding analyses were performed using the multinomial logistic regression with custom code based on the Princeton MVPA toolbox (github.com/princetonuniversity/princeton-mvpa-toolbox) which uses the MATLAB Neural Network Toolbox. 'Softmax' and 'cross entropy' were used as activation and performance functions. These two functions are reasonable choices for multi-class linear classification problems because they normalize the activation outputs to sum to 1. The scaled conjugate gradient method was used to fit the weights and bias parameters.

Our main analysis was performing decoding on the delay epoch representation. Within-stimulus decoding on the delay epoch was performed by training and testing on the same stimulus type (e.g., training and testing on orientation). Cross-stimulus decoding on the delay epoch was conducted by training and testing on the delay representation of different stimulus types (e.g., training on orientation and testing on motion direction). We were mostly interested in the cross-stimulus decoding results on the delay epoch data as we aimed to examine whether a common representational format existed in WM across different stimulus types with similar nature.

Classification was performed on the beta coefficients acquired from running a voxel-wise general linear model (GLM) using AFNI 3dDeconvolve, on all the runs acquired for each participant. We used GLM to estimate the responses of each voxel to the stimulus encoding, delay, response, feedback, and the inter-trial-interval epochs. Each epoch was modeled by the convolution of a canonical model of the hemodynamic impulse response function with a square wave (boxcar regressor) whose duration was equal to the duration of the corresponding epoch. Importantly, we estimated beta coefficients for every trial independently for the epoch that was of main interest (e.g., delay epoch) in performing a particular decoding analysis. Other epochs were estimated using a common regressor for all trials (Rissman et al., 2004). This method was used to capitalize on the trial-by-trial variability of the main epoch of interest (e.g., delay epoch) and prevent the trial-by-trial variability of other epochs soaking up a large portion of variance which could potentially be explained by the epoch of interest. Six motion regressors were also included in the GLM. Each voxel's beta coefficients were z-scored across each run independently for decoding.

We performed a 3-way classification to decode 3 target orientations and 3 motion directions. The orientation conditions were 15°, 75°, and 135° clockwise from vertical, and in the motion direction trials, the two opposite direction conditions sharing the same orientation axis were combined (e.g., 15° and 195° combined into 15°). Therefore, there were 3 target conditions for both stimulus types. For within-stimulus decoding, we implemented the leave-one-run-out cross-validation procedure in which one run was left out on each iteration for testing the performance of the classifier and the rest of the runs were used for training. The decoding accuracies were averaged across all the iterations. For cross-stimulus decoding, the classifier was trained on beta coefficients of all trials in one stimulus type and tested on all other trials in the other stimulus type.

For the temporal generalization decoding analysis, we used z-scored percent signal change values for each TR. A 3TR sliding window was used, meaning that the training and the test data were each voxel's activity values averaged across 3TRs including the TR being trained or tested.

fMRI data analysis: Spatial reconstruction

To visualize neural activity during the delay period, voxel amplitudes for each orientation/direction condition were projected onto the 2D visual field space (Kok and de Lange, 2014). The same GLM beta coefficients (β) extracted for the decoding analysis were averaged across trials for each feature condition and voxel, and were used for weighting the voxels' pRFs in Equation 1. The reconstructed maps reflect aspects of the pRF structure such as increase in size with eccentricity (Dumoulin and Wandell, 2008; Mackey et al., 2017), resulting in a dumbbell shape of the reconstructed line patterns.

Furthermore, to take into account the potential bias from the individual difference in pRF structure, we generated a pRF normalized version of the reconstruction maps. More specifically, the pRF bias map for each participant was computed by summing up all

Please cite this article in press as: Kwak and Curtis, Unveiling the abstract format of mnemonic representations, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.03.016

voxels' pRFs without any weighting. Then, the original reconstruction maps (Equation 1) were divided by the pRF bias map separately for each participant. With this procedure, we aimed to examine whether the underlying pRF structure influences the reconstruction patterns for the different orientation/direction conditions.

We also repeated the same reconstruction analysis with classifier weights, instead of beta coefficients, to weigh the voxel's pRFs. With the same classification algorithm used for the decoding analysis, we estimated the classifier weights of all voxels for each ROI separately for each participant, using GLM beta coefficients. No trials needed to be held out for cross-validation because the main purpose of this analysis was to train the classifier for weight estimation and not to test its performance. Note that no voxels were left out in the weight estimation step but only for computing the weighted sum of pRFs (eccentricity $\leq 20^{\circ}$ of visual angle). As a result, for each participant and ROI, we had a i (number of target orientation/direction conditions) x n (number of voxels in an ROI) weight matrix w. Therefore, Equation 1 was modified as below.

$$S_{\theta_i} = \sum_{n=1}^{N} w_{i,n} \times e^{-\frac{(x_n - x_0)^2 + (y_n - y_0)^2}{\sigma^2}} \quad \text{(Equation 2)}$$

We confirmed that the classifier weights of the voxels had the same spatial structure as the beta coefficients.

For generating all the spatial reconstruction maps, we down-sampled the resolution of the visual field space such that each pixel corresponded to 0.1° of visual angle (10 pixels/$^{\circ}$). Only voxels whose pRF eccentricities were within 20° of visual angle were included in the reconstruction.

To better visualize the line format (white lines in Figure 3B), we fit a first-degree polynomial to the X and Y coordinates of the selected pixels with top 10% image intensity, with the constraint that the fitted polynomial passed through the center. To take into account the difference in magnitude of the image intensity of the selected pixels, we conducted a weighted fit with the weight entry for each pixel corresponding to its rank in terms of image intensity.

Model simulation: Image-computable model of V1

We used the image-computable model to both simulate a putative neural mechanism of maintaining a line in WM, and to rule out the alternative account that the reconstructed lines were caused not by line-like WM representations but by topographic biases in orientation tuning. Previous studies have reported that during perception of oriented gratings, orientation decoding in V1 may depend on a coarse-scale topographic relationship between a voxel's preference for a spatial angle relative to fixation (e.g., 45°, up and to the right) and the matching orientation of the grating (e.g., tilted 45° clockwise from vertical), as well as biases induced by the apertures bounding the stimulus grating (Freeman et al., 2011; Roth et al., 2018). For example, a neuron may exhibit ostensible orientation tuning because of its receptive field overlapping with stimulus edge or change in contrast due to aperture. A neuron may even be orientation-selective, but its orientation selectivity could change depending on stimulus aperture. Therefore, we aimed to compare the model simulation results to gratings used in the present experiment and simple line images (Figure 4).

The image-computable model is based on the steerable pyramid, a subband image transform that decomposes an image into spatial frequency and orientation channels (Roth et al., 2018; Simoncelli et al., 1992). Responses of many linear receptive fields (RFs) are simulated, each of which computes a weighted sum of the stimulus image. The weights determine the spatial frequency and orientation tuning of the linear RFs, which are hypothetical basis sets of spatial frequency and orientation tuning curves of V1. For the model simulation, we used 16 subbands comprised of 4 spatial frequency bands (spatial frequency bandwidth = 0.5 octave) and 4 orientation bands (orientation bandwidth = 90°), which were parameters that could be chosen flexibly. The number of the spatial frequency bands is determined by the size of the stimulus image and the spatial frequency bandwidth parameter. Also, using more than 4 orientation bands (e.g., 6 orientation bands; bandwidth = 60°) corresponding to narrower tuning curves did not change the results.

The inputs to the model were our 1280 x 1024 stimulus images: grating or line images oriented 15°, 75°, 135° clockwise from vertical (Figure 4A; images are shown square for the purpose of illustration to match with the spatial reconstruction maps). Only for the grating images there were 15 phases evenly distributed between 0 and 2π. The input images had the same configuration as the stimuli in the actual experiment (size of fixation, inner aperture, outer aperture, etc), and the outputs of the model were images of the same resolution as the input images. Each pixel of the output image corresponded to the simulated neuron in the retinotopic map of V1. We first measured the model's responses to each stimulus image separately. Then, for the grating images, we averaged across the model responses to 15 different phases for each orientation condition. As a result, we had 3 (number of orientation conditions) neuronal response output images for 16 different subbands. For model outputs to grating images, we present the sum of two subbands whose center spatial frequencies are closest to the spatial frequency of the grating stimulus, one of which is the subband with the maximal response (Figures 4 and S3A). This is a reasonable choice because the grating is a narrow-band stimulus. For model outputs to line images, we present the sum of all subbands instead (Figures 4 and S3A). For completeness, the sum of all subbands for the grating images and the subband with the maximal response for line images are shown in Figures S3C and S3D.

To simulate voxel-level responses based on the model outputs, we conducted a pRF sampling analysis (Roth et al., 2018). For each subband and orientation condition, each participant's pRF gaussian parameters of V1 voxels were used to weigh the model outputs, computing a weighted sum of neuronal responses corresponding to the voxels' pRFs. This sampling procedure resulted in one simulated beta coefficient for each voxel. Therefore, we generated 3 (number of orientation conditions) x N (number of voxel) beta

Please cite this article in press as: Kwak and Curtis, Unveiling the abstract format of mnemonic representations, Neuron (2022), https://doi.org/10.1016/j.neuron.2022.03.016

coefficients for each subband and participant, separately for the grating and line images. We pushed these simulated beta coefficients into the same spatial reconstruction analysis in Figure 3A after z-scoring.

Eye tracking analysis and results

Eye positions were monitored throughout the entire experiment to ensure that participants maintained fixation particularly during the delay epoch. 95.74% of the total number of eye position sample points in the delay epoch across all participants were within 2° eccentricity from the center (the fixation and the stimulus subtended 0.7° and 15° diameter, respectively). Circular correlation between the polar angle of the target orientation/direction and the polar angle of the eye coordinates suggest that the polar angle of the stimulus was not predictive of eye movements (mean = 0.03, s.d. = 0.080, t (10) = 1.24, p = 0.118).

Apparatus

All stimuli were generated via PsychToolBox (Brainard, 1997) in Matlab 2018a and presented via ViewPixx PRoPixx projector (screen resolution: 1280 x1024 for experimental task and RDK retinotopic mapping task, 1080 x 1080 for object image retinotopic mapping task; refresh rate: 60Hz for all tasks). The viewing distance was 63cm, and the projected image spanned 36.1cm height and 45.1cm width. All functional MRI images were acquired at the NYU Center for Brain Imaging 3T Siemens Prisma Scanner with the Siemens 64 channel head/neck coil. Eye tracking data were acquired using an MR-compatible Eye link 1000 infrared eye tracker (SR Research). X, Y coordinates of the eye positions were recorded at 500 or 1000 Hz.

Glitches

For S02, S03, and S07, fMRI data from one of the 22 experimental task runs was not saved due to the computer freezing at the end of the runs. Eye tracking data for one of the experimental task runs was lost for S02 and S08 due to file corruption. Eye positions were not monitored for S07 during one of the two experimental task sessions due to technical issues, and in the session in which eye positions were recorded, eye data for one of the runs was lost due to file corruption.

QUANTIFICATION AND STATISTICAL PROCEDURES

The statistical results reported here are all based on permutation testing over 1000 iterations with shuffled data (see Table S1). *P values* reflect the proportion of a metric (F scores, t scores, reconstruction fidelity) in the permuted null distribution greater than or equal to the metric computed using intact data. Note that the minimum p value achievable with this procedure is 0.001. *P values* were FDR corrected when applicable (Benjamini and Hochberg, 1995).

For statistical testing of fMRI decoding accuracy, we generated permuted null distributions of decoding accuracy values for each epoch (delay/stimulus encoding), decoding type (cross/within), stimulus type (orientation/direction), ROI, and participant. On each iteration, we shuffled the training data matrix (voxels x trials) across both dimensions so that both voxel information and orientation/direction labels could be shuffled. This procedure was conducted for each of the 11 participants resulting in 11 null decoding accuracy estimates per one iteration of permutation. Depending on the statistical tests being performed, we calculated null t scores or F values for each iteration of permutation.

For the spatial reconstruction analysis, we computed reconstruction fidelity to quantify the amount of remembered orientation/direction feature information present in the reconstruction maps (Hallenbeck et al., 2021; Rademaker et al., 2019). Line filters with orientations evenly spaced between -90° and 90° in steps of 1° were used to sum up pixels of the z-scored reconstruction maps within the masked area of the maps. More specifically, a line filter oriented θ° includes pixels with coordinates that form an acute angle to θ° (dot product > 0) and that have a projected distance squared less than 1000. Using different width parameters (projected distance squared) of the orientation filters did not change the results. For each participant, feature condition, and ROI, we generated a tuning curve-like function where x-axis and y-axis represent the orientation of each line filter and the z-scored sum of pixel values masked by the filters, respectively. The filtered responses for each condition were aligned to the true orientation/direction (0°) and then averaged. To compute fidelity, we projected the filtered responses at each orientation filter onto a vector centered on the true orientation (0°) and took the mean of all the projected vectors. Conceptually, this metric measures whether and how strongly the reconstruction on average points in the correct direction. The real fidelity value was compared against the distribution of null fidelity values from shuffled data. To generate the null distribution, the matrix of beta coefficients was shuffled across both the voxel and orientation/direction condition label dimensions, and the shuffled beta coefficients were used to weight the voxels' pRF parameters. The same logic applied to calculating reconstruction fidelity from direction filters, with the exception that the filters were evenly spaced between -90° and 270° in steps of 1° and the filtered responses were aligned so that the true direction peaked at 90° and 180°.

Neuron, Volume *110*

Supplemental information

Unveiling the abstract format of mnemonic representations

Yuna Kwak and Clayton E. Curtis

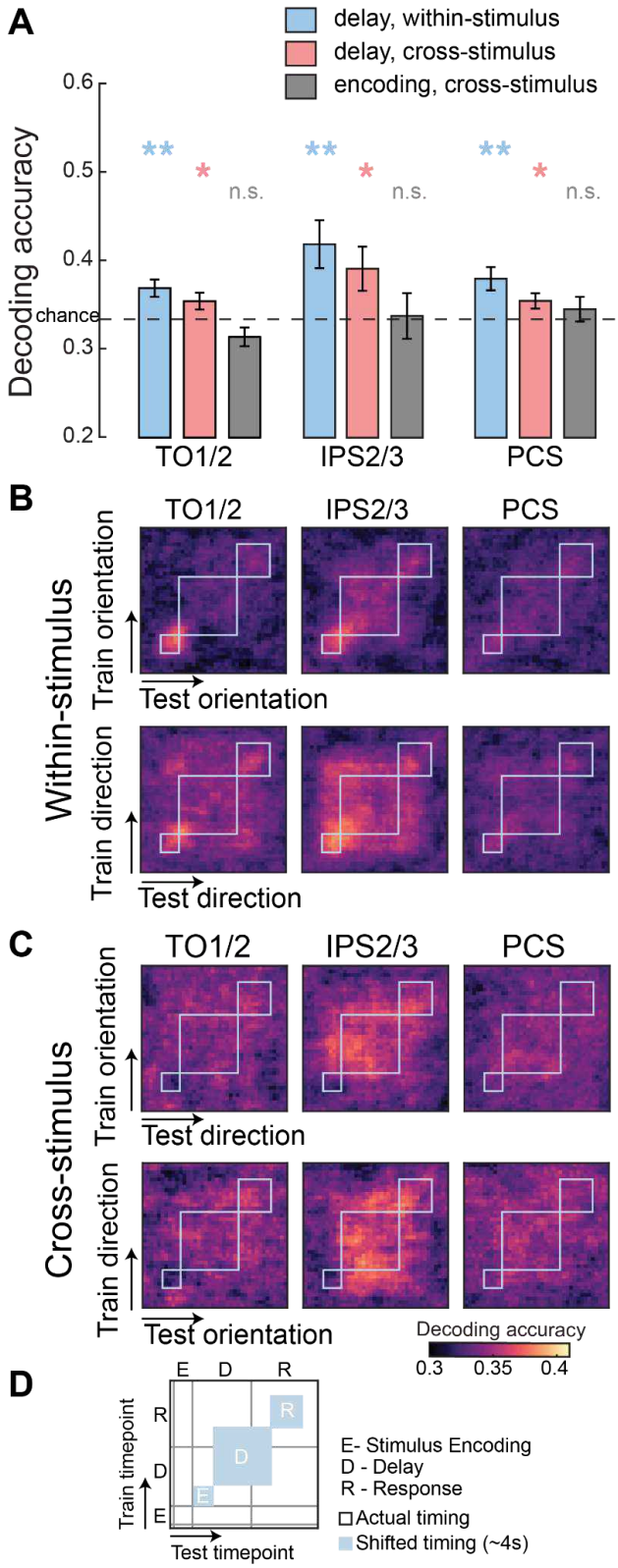

Figure S1. Working memory representations for orientation and motion direction share a common format, related to Figure 2. Decoding analysis was performed for other ROIs. **(A)** In WM, there are robust representations of orientation and motion direction information (blue bars), and these representations share a common format (red bars). The shared representational format is not observed during stimulus encoding (gray bars). **(B-C)** The temporal generalization matrix was generated by training and testing on each time point. **(D)** Schematics of the matrix plots. Gray lines denote the actual timing of events, and blue boxes show each of these events shifted by ~4s assuming hemodynamic lag. *$p < 0.05$, **$p < 0.01$, n.s. not significant, corrected (p values in Table S1). Error bars represent ±1 SEM.

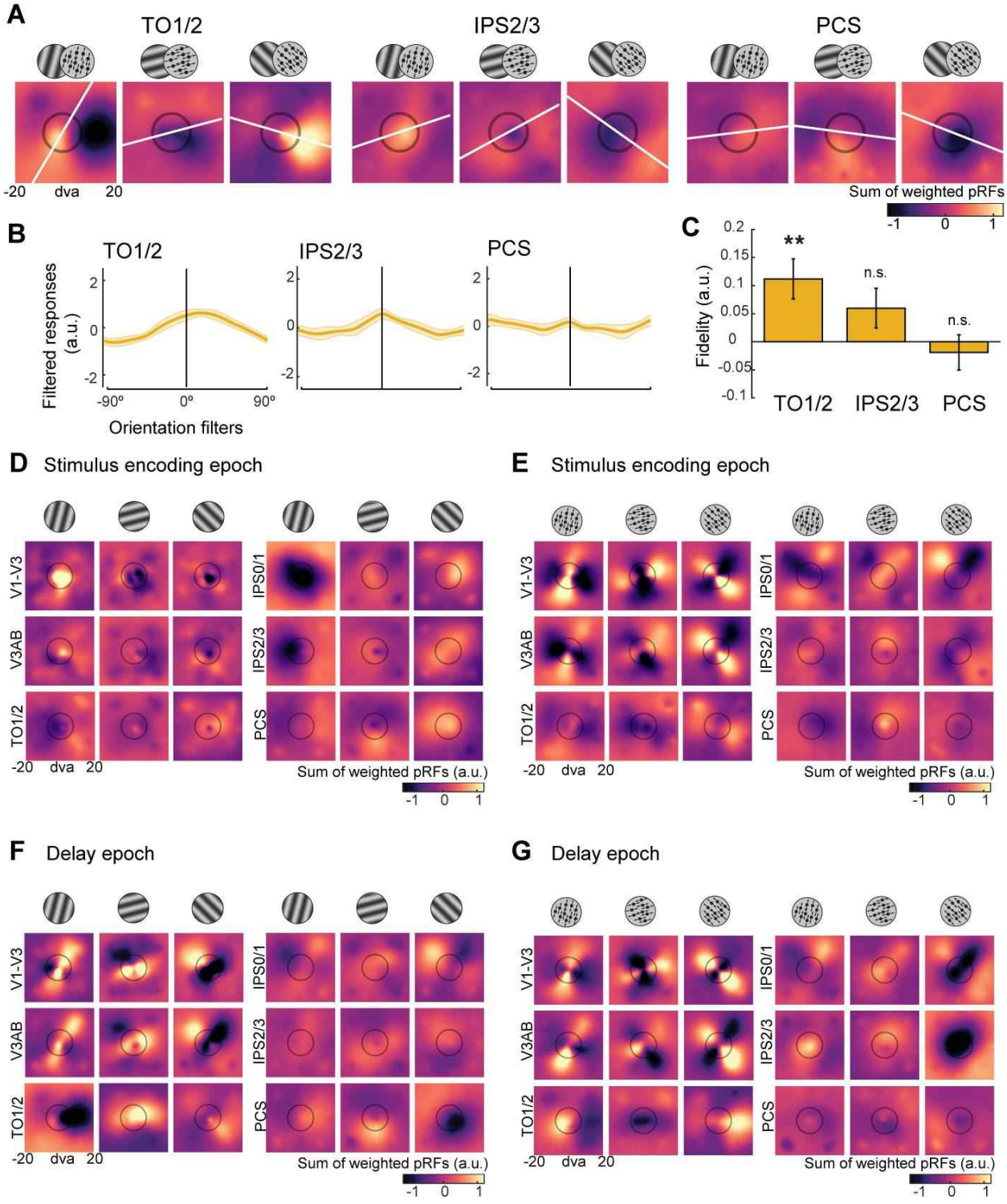

Figure S2. Unveiling the recoded formats of working memory representations for orientation and motion direction, related to Figure 3. The spatial reconstruction analysis was performed for other ROIs.

(A) Population reconstruction maps for orientation and motion direction combined. Best fitting lines of the reconstruction are plotted in white. The size of each map is from -20° to 20° of visual angle (dva), and the stimulus size is shown for reference in black circles.
(B) Filtered responses from reconstruction maps in (A).
(C) Reconstruction fidelity calculated from filtered responses in (B). $^{**}p < 0.01$, n.s. not significant, corrected for multiple comparisons (*p* values in Table S1). Error bars represent ±1 SEM. For more information on filtered responses and fidelity, see STAR Methods.
(D-E) Reconstruction maps during the stimulus encoding epoch, separately for orientation and motion direction. For grating orientation, no line format is observed, likely due to a combination of orthogonal drifting motion and aperture bias during perception, alongside the emergence of orientation representation to be maintained in memory. For motion direction, we do observe a line-like format in some areas, mostly due to the emergence of motion direction representation to be maintained during the delay period.
(F-G) Reconstruction maps during the delay epoch, separately for orientation and motion direction. Both stimulus types were represented as a line format in WM across a wide range of brain regions.
For all figures, the size of each map is from -20° to 20° of visual angle (dva), and the stimulus size is shown for reference in black circles.

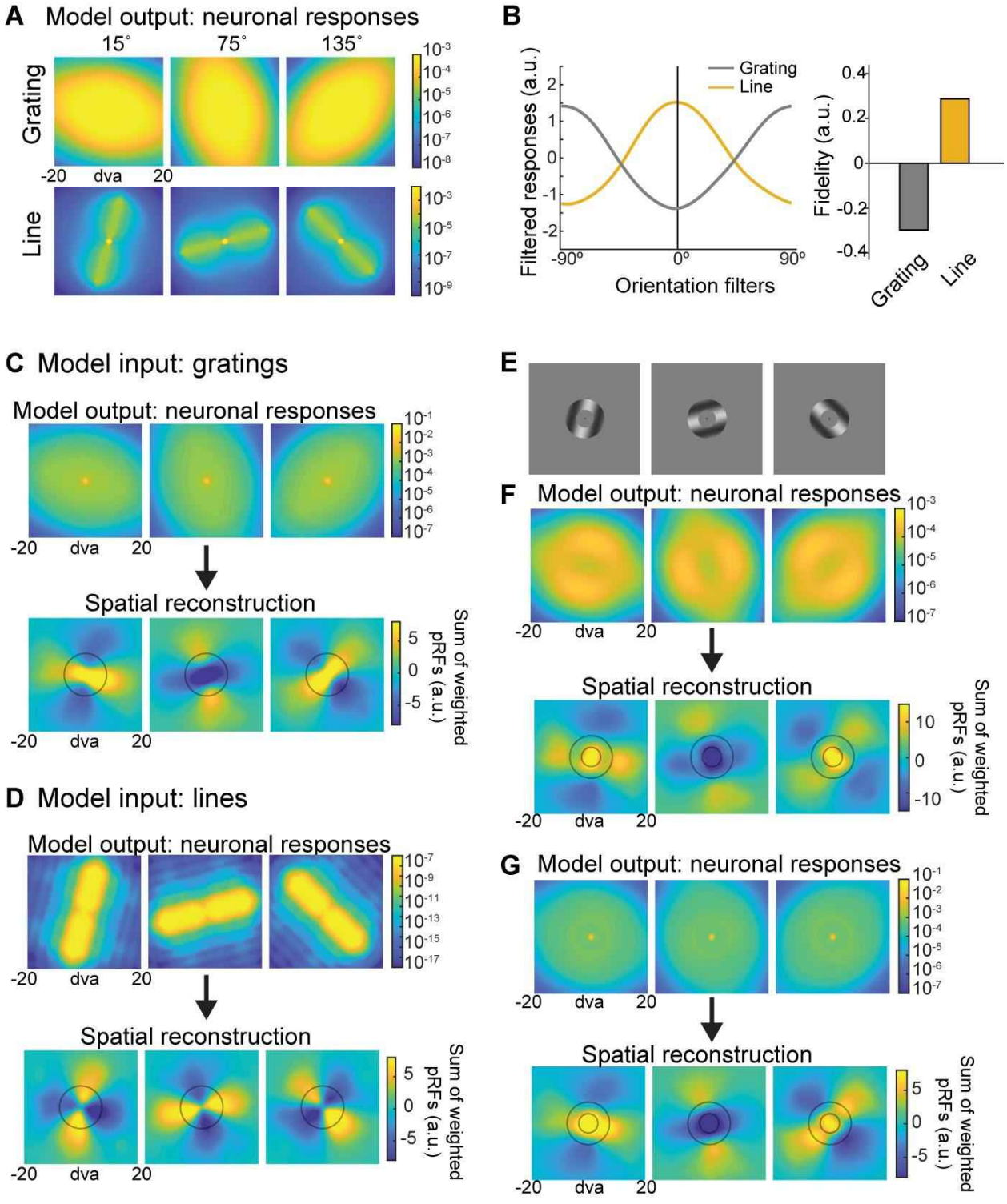

Figure S3. Spatial reconstruction analysis with simulated neuronal responses from the image-computable model of V1, related to Figure 4.

(A) The output neuronal responses from the image-computable model for grating and line images are plotted in log scale. Inputs to the model were grating and line images

oriented 15º, 75º, 135º clockwise from vertical (Figure 4A). For grating images, we present the sum of two subbands of which the center spatial frequencies are the closest to the spatial frequency of the grating stimulus. One of the two chosen subbands corresponds to the subband with the maximal response. For line images, we present the sum of all subbands to be more conservative. For completeness, we also present the sum of all subbands for grating images in (C) and the maximal subband for line images in (D) (See STAR Methods for details on the model). These model output neuronal responses were later used to simulate V1 voxel amplitudes by summing up the neuronal responses within each voxel's pRF. These simulated voxel responses were fed into the spatial reconstruction analysis in Figure 3A which resulted in the reconstruction maps in Figure 4B.

(B) Filtered responses and associated fidelity were computed from the spatial reconstruction maps in Figure 4B for quantification. The filtered responses when input to the model were line images (bottom row of (A) and Figures 4A-B) peak at the true orientation, suggesting that simulated V1 responses are highest along the actual remembered orientation (positive fidelity value). Filtered responses computed from the grating image reconstruction maps (top row of (A) and Figures 4A-B) show an opposite pattern (negative fidelity value). Since the output model responses were identical for all participants and only the pRF parameters differed, statistical analyses were not performed on the model simulations.

(C-D) Neuronal output responses and the reconstruction maps for the sum of all subbands when inputs to the model were grating images in (C) and for the maximal subband when inputs were simple line images in (D). These results resemble the model outputs in (A) and the reconstruction maps in Figure 4B. Therefore, our results are not an artifact of the chosen subband.

(E) For sanity check, we modified our grating stimulus images by increasing the size of the inner aperture (6.75º diameter; corresponding to the inner black circle in the reconstruction maps) to test whether stimulus aperture influences the spatial reconstruction maps. The model outputs and spatial reconstructions maps are shown in (F-G).

(F) The sum of two subbands with the spatial frequency centered on the spatial frequency of the grating stimulus. The spatial reconstruction maps exhibited a slightly different pattern from those in Figure 4B as they were a mixture of patterns parallel and orthogonal to the remembered orientation.

(G) The sum of all subband levels. A larger inner aperture did not change the pattern of reconstruction maps, as it was still orthogonal to the remembered orientation, similar to those from our original stimuli (Figure 4B).

All neuronal output response plots are presented in log scale. For all reconstruction maps, the size of each map is from -20° to 20° of visual angle (dva), and the stimulus size (inner and outer apertures) is shown in black circles for reference.

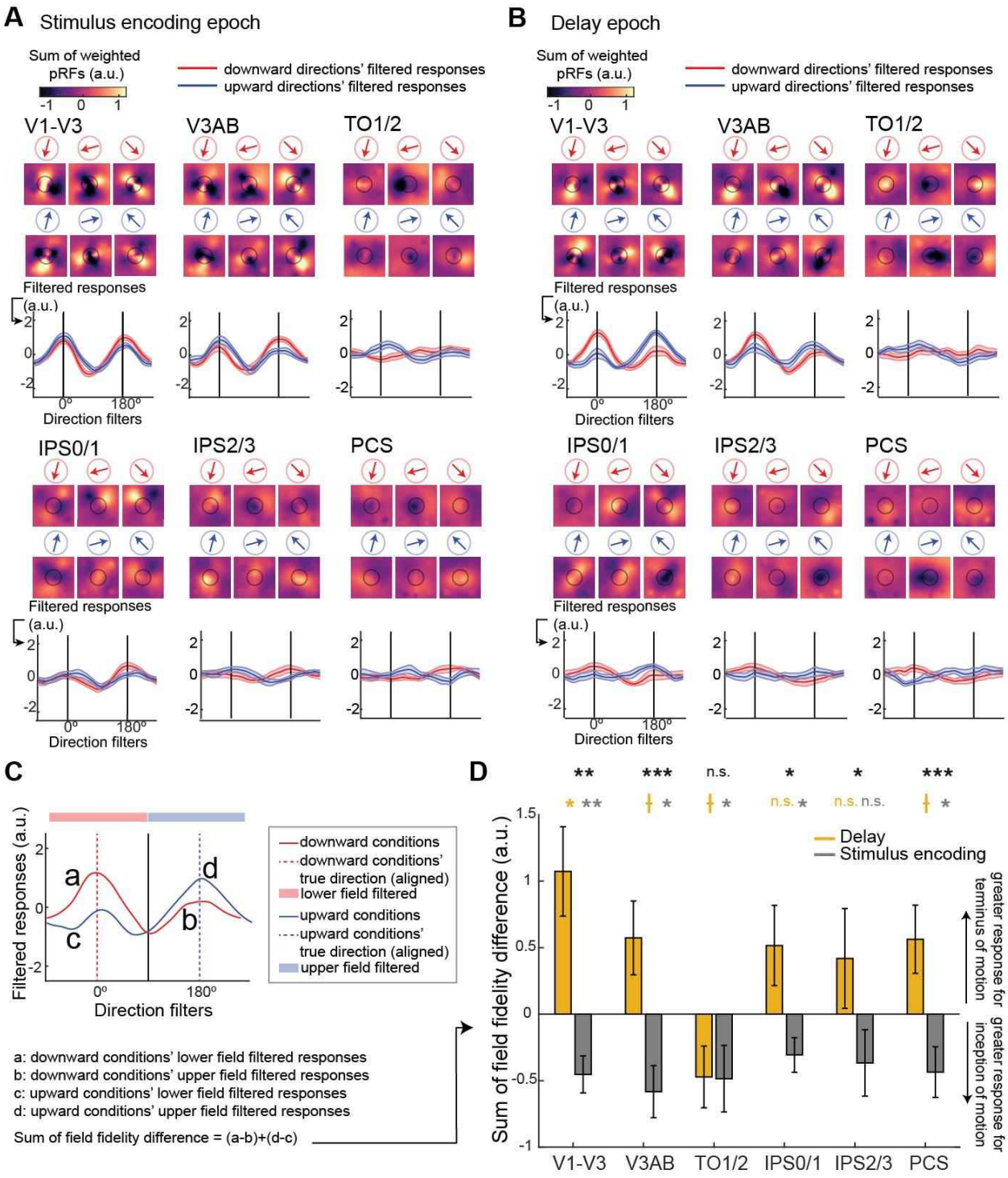

Figure S4. Distinct formats for perceptual and working memory representations of motion direction, related to Figures 3 and 4. Spatial reconstruction was performed on 6 direction conditions separately, as the previously reported aperture bias in perception makes different predictions for opposite motion directions with the same orientation axis (Wang et al., 2014).

(A-B) Reconstruction maps and associated filtered responses for the stimulus encoding in (A) and delay epochs in (B). The direction conditions from which the reconstruction maps are generated are shown in circles (red for downward directions and blue for upward directions). The size of all reconstructed maps is from -20° to 20° of visual angle. Each ROI's filtered response curves were generated by aligning and averaging the individual filtered responses for each of the 6 direction conditions. The first and second bumps at 0º and 180º in the filtered response curves correspond to the information captured by direction filters spanning the lower and upper visual field (360º of polar angle, to quantify whether the representation of remembered direction differs in strength between locations in the topographic map corresponding to the terminus and inception of motion direction). See (C) for a detailed description.
(C) Interpretation of filtered response curves in (A-B). Schematics is shown for the pattern of filtered response curves for most visual maps in the delay epoch in (B). For example, the red curve shows that memory representations for downward direction conditions form a line in topographic space corresponding to the orientation axis of the remembered direction (two bumps), but the portion of the line representation in the lower visual field (terminus of moving dots for downward direction conditions; a) is stronger than that of the upper visual field (inception of the moving dots for downward direction conditions; b). The blue curve shows that memory representations for upward direction conditions show higher response in the upper visual field (terminus of moving dots for upward direction conditions; d), compared to the lower visual field (inception of moving dots for upward direction conditions; c). We additionally computed the sum of field fidelity differences which quantifies the difference in activation between the portion of the topographic maps corresponding to the terminus and the inception of dot motion. Taking into account how the sum of field fidelity differences is computed ((a-b)+(d-c)), larger values indicate greater activation near the terminus of moving dots compared to the inception of moving dots.
(D) As previously reported during dot motion perception (Wang et al., 2014), the line-like patterns we observed early and time-locked to the visible motion stimulus (stimulus encoding epoch) were biased with more activation in the portion of the topographic map near the inception of the moving dots (gray bars and asterisks). During memory (delay epoch) we found the opposite; in most visual maps, there was greater activation near the terminus (yellow bars and asterisks). The difference between these biases during the stimulus encoding and delay epochs was statistically significant in many of the ROIs (black asterisks), demonstrating differences in the representational formats of perceiving and remembering dot motion. The only exception was TO1/2 where the reconstruction patterns observed during stimulus encoding and delay were similar. *p < 0.05, **p < 0.01, ***p < 0.001, n.s. not significant, corrected for multiple comparisons; †p < 0.05 uncorrected (p values in Table S1).

Supplementary Table

Table S1.

Decoding accuracy for the delay epoch[1].						
3-way ANOVA. Decoding type x ROI x Stimulus type						
Decoding type df: (1,10)	***F* = 7.061** ***p* = 0.017** **η_p^2 = 0.414**					
ROI df: (5,50)	***F* = 9.965** ***p* = 0** **η_p^2 = 0.499**					
Stimulus Type df: (1,10)	*F* = 4.001 *p* = 0.063 η_p^2 = 0.286					
Decoding type x ROI df: (5,50)	***F* = 3.039** ***p* = 0.020** **η_p^2 = 0.233**					
Decoding type x Stimulus Type df: (1,10)	*F* = 0.691 *p* = 0.439 η_p^2 = 0.065					
ROI x Stimulus Type df: (5,50)	***F* = 3.576** ***p* = 0.006** **η_p^2 = 0.263**					
Decoding type x ROI x Stimulus Type df: (5,50)	***F* = 2.987** ***p* = 0.023** **η_p^2 = 0.230**					
Post-hoc 2-way ANOVA. Decoding type x Stimulus type for each ROI (FDR thres = 0.005 for decoding type, 0 for stimulus type and decoding type x stimulus type)[2].						
	V1-V3	V3AB	TO1/2	IPS0/1	IPS2/3	PCS
Decoding type df: (1,10)	*F* = 5.505 *p* = 0.040 η_p^2 = 0.351	***F* = 13.948** ***p* = 0.005** **η_p^2 = 0.582**	*F* = 1.508 *p* = 0.232 η_p^2 = 0.131	*F* = 2.015 *p* = 0.194 η_p^2 = 0.168	*F* = 4.254 *p* = 0.065 η_p^2 = 0.298	*F* = 6.470 *p* = 0.022 η_p^2 = 0.393
Stimulus type df: (1,10)	*F* = 6.059 *p* = 0.049 η_p^2 = 0.377	*F* = 6.969 *p* = 0.028 η_p^2 = 0.411	*F* = 0.708 *p* = 0.405 η_p^2 = 0.066	*F* = 1.685 *p* = 0.220 η_p^2 = 0.144	*F* = 0.618 *p* = 0.428 η_p^2 = 0.058	*F* = 0.046 *p* = 0.832 η_p^2 = 0.005
Decoding type x Stimulus type df: (5,50)	*F* = 1.477 *p* = 0.263 η_p^2 = 0.129	*F* = 3.614 *p* = 0.088 η_p^2 = 0.265	*F* = 0.006 *p* = 0.941 η_p^2 = 0	*F* = 0.033 *p* = 0.865 η_p^2 = 0.003	*F* = 1.807 *p* = 0.195 η_p^2 = 0.153	*F* = 0.104 *p* = 0.762 η_p^2 = 0.010
One-sample *t* tests against chance (FDR thres = 0.006 for within-stimulus, 0.029 for cross-stimulus)[3].						

	V1-V3	V3AB	TO1/2	IPS0/1	IPS0/2	PCS
Within-stimulus df: 10	***t* = 4.138** ***p* = 0** ***d* = 1.354**	***t* = 4.880** ***p* = 0.001** ***d* = 1.196**	***t* = 3.537** ***p* = 0.005** ***d* = 1.324**	***t* = 3.591** ***p* = 0.004** ***d* = 1.471**	***t* = 3.117** ***p* = 0.006** ***d* = 1.066**	***t* = 3.432** ***p* = 0.003** ***d* = 1.083**
Cross-stimulus df: 10	***t* = 3.584** ***p* = 0.002** ***d* = 1.081**	***t* = 3.203** ***p* = 0.002** ***d* = 0.966**	***t* = 2.097** ***p* = 0.029** ***d* = 0.632**	***t* = 2.767** ***p* = 0.008** ***d* = 0.834**	***t* = 2.261** ***p* = 0.015** ***d* = 0.682**	***t* = 2.309** ***p* = 0.020** ***d* = 0.696**
3-way ANOVA. Decoding type x ROI x Stimulus type, after matching the training and testing procedures across decoding types.						
Decoding type df: (1,10)	***F* = 7.061** ***p* = 0.027** η_p^2 **= 0.414**					
ROI df: (5,50)	***F* = 9.965** ***p* = 0** η_p^2 **= 0.499**					
Stimulus Type df: (1,10)	*F* = 4.000 *p* = 0.077 η_p^2 = 0.286					
Decoding type x ROI df: (5,50)	***F* = 3.039** ***p* = 0.019** η_p^2 **= 0.233**					
Decoding type x Stimulus Type df: (1,10)	*F* = 0.691 *p* = 0.419 η_p^2 = 0.065					
ROI x Stimulus Type df: (5,50)	***F* = 3.576** ***p* = 0.007** η_p^2 **= 0.263**					
Decoding type x ROI x Stimulus Type df: (5,50)	***F* = 2.987** ***p* = 0.018** η_p^2 **= 0.230**					
Decoding accuracy for the stimulus encoding epoch[4].						
3-way ANOVA. Decoding type x ROI x Stimulus type.						
Decoding type df: (1,10)	***F* = 19.957** ***p* = 0.002** η_p^2 **= 0.666**					
ROI df: (5,50)	***F* = 6.825** ***p* = 0** η_p^2 **= 0.406**					

Stimulus type df: (1,10)	F = 1.126 p = 0.314 η_p^2 = 0.101	
Decoding type x ROI df: (5,50)	***F* = 10.876** ***p* = 0** **η_p^2 = 0.521**	
Decoding type x Stimulus type df: (1,10)	F = 1.503 p = 0.273 η_p^2 = 0.131	

Post-hoc 2-way ANOVA. ROI x Stimulus type for each decoding type (FDR thres = 0 for ROI, Stimulus type, ROI x Stimulus type)[5].

	Within-stimulus	Cross-stimulus				
ROI df: (5,50)	***F* = 15.011** ***p* = 0** **η_p^2 = 0.600**	F = 0.753 p = 0.605 η_p^2 = 0.070				
Stimulus type df: (1,10)	F = 1.319 p = 0.279 η_p^2 = 0.117	F = 0.015 p = 0.908 η_p^2 = 0.002				
ROI x Stimulus type df: (5,50)	***F* = 9.007** ***p* = 0** **η_p^2 = 0.474**	F = 2.192 p = 0.067 η_p^2 = 0.180				

One-sample t tests against chance for cross-stimulus decoding accuracy (FDR thres = 0)[6].

	V1-V3	V3AB	TO1/2	IPS0/1	IPS2/3	PCS
Cross-stimulus df: 10	t = 0.304 p = 0.378 d = 0.092	t = −0.057 p = 0.510 d = −0.017	t = −1.945 p = 0.956 d = −0.586	t = 0.264 p = 0.413 d = 0.079	t = 0.122 p = 0.473 d = 0.037	t = 0.772 p = 0.244 d = 0.233

Reconstruction fidelity for the delay epoch[7].

Orientation and direction trials combined (FDR thres = 0.006).

	V1-V3	V3AB	TO1/2	IPS0/1	IPS2/3	PCS
Fidelity df: 10	***t* = 16.597** ***p* = 0** ***d* = 5.004**	***t* = 9.742** ***p* = 0** ***d* = 2.937**	***t* = 3.133** ***p* = 0.006** ***d* = 0.945**	***t* = 4.379** ***p* = 0.001** ***d* = 1.320**	t = 1.686 p = 0.067 d = 0.508	t = −0.605 p = 0.278 d = −0.182

Reconstruction fidelity on the sum of field reconstruction fidelity differences, for the direction trials[8].

2-way ANOVA. Epoch x ROI.

	Sum of field fidelity differences	
Epoch df: (1,10)	***F* = 33.953** ***p* = 0** **η_p^2 = 0.773**	
ROI df: (5,50)	*F* = 1.990 *p* = 0.088 η_p^2 = 0.166	
Epoch x ROI df: (5,50)	***F* = 3.380** ***p* = 0.007** **η_p^2 = 0.253**	

One-sample *t* tests on sum of field reconstruction fidelity differences, separately for delay epoch and stimulus encoding epoch (FDR thres = 0.002 for delay epoch, 0.036 for stimulus encoding epoch).

	V1-V3	V3AB	TO1/2	IPS0/1	IPS2/3	PCS
Delay epoch df: 10	***t* = 3.190** ***p* = 0.002** ***d* = 0.962**	*t* = 2.068 *p* = 0.048 *d* = 0.623	*t* = −0.232 *p* = 0.030 *d* = −0.613	*t* = 1.710 *p* = 0.055 *d* = 0.516	*t* = 1.117 *p* = 0.147 *d* = 0.337	*t* = 2.194 *p* = 0.028 *d* = 0.662
Stimulus encoding epoch df: 10	***t* = −3.251** ***p* = 0.001** ***d* = −0.980**	***t* = −2.985** ***p* = 0.009** ***d* = −0.900**	***t* = −1.935** ***p* = 0.036** ***d* = −0.583**	***t* = −2.354** ***p* = 0.020** ***d* = −0.710**	*t* = −1.470 *p* = 0.079 *d* = −0.443	***t* = −2.288** ***p* = 0.017** ***d* = −0.690**

Paired-sample *t* tests on sum of field reconstruction fidelity differences: delay epoch vs stimulus encoding epoch (FDR thres = 0.015).

	V1-V3	V3AB	TO1/2	IPS0/1	IPS2/3	PCS
Difference between epochs df: 10	***t* = 3.558** ***p* = 0.002** ***d* = 1.073**	***t* = 5.293** ***p* = 0** ***d* = 1.596**	*t* = 0.069 *p* = 0.474 *d* = 0.021	***t* = 2.779** ***p* = 0.007** ***d* = 0.838**	***t* = 2.395** ***p* = 0.015** ***d* = 0.722**	***t* = 4.421** ***p* = 0** ***d* = 1.333**

Table S1. Non-parametric *p* values.

[1] Significant tests ($p < 0.05$, FDR corrected if applicable) are marked in bold. Factors are decoding type (within-/cross-stimulus), ROI (V1-V3, V3AB, TO1/2, IPS0/1, IPS2/3, PCS), and stimulus type (train on orientation/train on direction). To generate the null distribution, classifiers were trained on data shuffled across both the trial label and voxel dimensions. If applicable, *p* values were FDR corrected across the horizontal dimension of each table.

[2] A post-hoc 2-way ANOVA with decoding type and stimulus type as factors was conducted to examine whether the two stimulus types (train on orientation/train on direction) could be combined for further analyses.

[3] To quantify whether orientation/motion direction information could be decoded from ROIs (within-stimulus) and whether the two features share the same neural representation during WM (cross-stimulus), decoding accuracy averaged across the orientation/direction trials, was compared against the null distribution.

[4] Significant tests ($p < 0.05$, FDR corrected if applicable) are marked in bold. Factors are decoding type (within-/cross-stimulus), ROI (V1-V3, V3AB, TO1/2, IPS0/1, IPS2/3, PCS), and stimulus type (train on orientation/train on direction). To generate the null distribution, classifiers were trained on data shuffled

across both the trial label and voxel dimensions. If applicable, *p* values were FDR corrected across the horizontal dimension of each table.

5 The main purpose of this analysis was to determine whether cross-stimulus decoding accuracies could be combined across the two stimulus types (train on orientation/train on direction). As there was neither the main effect of stimulus type nor an interaction for cross-stimulus decoding, data were combined for further analyses.

6 To examine whether the representational formats for orientation and direction were shared during stimulus encoding, cross-decoding accuracy averaged across the train-on-orientation and train-on-direction conditions was compared with the null distribution.

7 Significant tests ($p < 0.05$, FDR corrected if applicable) are marked in bold. To generate the null distribution, fidelity values were calculated from reconstruction maps computed from shuffled beta coefficients (shuffled across trial label and voxel dimensions). If applicable, *p* values were FDR corrected across the horizontal dimension of each table.

8 Significant tests ($p < 0.05$, FDR corrected if applicable) are marked in bold. For the null distribution, fidelity values were calculated from reconstruction maps generated from shuffled beta coefficients (shuffled across trial label and voxel dimensions). If applicable, *p* values were FDR corrected across the horizontal dimension of each table.