Neuroscience

Visual working memories are abstractions of percepts

Ziyi Duan, Clayton E. Curtis

Department of Psychology, New York University, New York, NY 10003, USA • Center for Neural Science, New York University, New York, NY 10003, USA

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint version 1
February 2, 2024 (this version)

Posted to preprint server
December 3, 2023

Sent for peer review
November 24, 2023

Abstract

Pioneering studies demonstrating that the contents of visual working memory (WM) can be decoded from the patterns of multivoxel activity in early visual cortex transformed not only how we study WM, but theories of how memories are stored. For instance, the ability to decode the orientation of memorized gratings is hypothesized to depend on the recruitment of the same neural encoding machinery used for perceiving orientations. However, decoding evidence cannot be used to test the so-called *sensory recruitment hypothesis* without understanding the underlying nature of what is being decoded. Although unknown during WM, during perception decoding the orientation of gratings does not simply depend on activities of orientation tuned neurons. Rather, it depends on complex interactions between the orientation of the grating, the aperture edges, and the topographic structure of the visual map. Here, our goals are to 1) test how these aperture biases described during perception may affect WM decoding, and 2) leverage carefully manipulated visual stimulus properties of gratings to test how sensory-like are WM codes. For memoranda, we used gratings multiplied by radial and angular modulators to generate orthogonal aperture biases despite having identical orientations. Therefore, if WM representations are simply maintained sensory representations, they would have similar aperture biases. If they are abstractions of sensory features, they would be unbiased and the modulator would have no effect on orientation decoding. Results indicated that fMRI patterns of delay period activity while maintaining the orientation of a grating with one modulator (eg, radial) were interchangeable with patterns while maintaining a grating with the other modulator (eg, angular). We found significant cross-classification in visual and parietal cortex, suggesting that WM representations are insensitive to aperture biases during perception. Then, we visualized memory abstractions of stimuli using a population receptive field model of the visual field maps. Regardless of aperture biases, WM representations of both modulated gratings were recoded into a single oriented line. These results provide strong evidence that visual WM representations are abstractions of percepts, immune to perceptual aperture biases, and compel revisions of WM theory.

eLife assessment

This paper provides **valuable** insights into the neural substrates of human working memory. Through clever experimental design and rigorous analyses, the paper provides **compelling** evidence that the working memory representation of stimulus orientation is a reformatted version of the presented stimulus, reflecting the content that is of importance to the task. This work will be of broad interest to cognitive neuroscientists working on the neural bases of visual perception and memory.

Introduction

Following now classic studies demonstrating that fMRI patterns of voxel activity in human early visual cortex can be used to decode the contents of visual working memory (WM) (Harrison and Tong, 2009; Serences et al., 2009), decoding WM content from visual cortex has been a workhorse for neuroimaging studies testing aspects of the sensory recruitment hypothesis of WM. This incredibly influential hypothesis posits that visual WM storage utilizes the encoding machinery in the visual cortex, assuming that memory and perception utilize similar mechanisms (Curtis and D'Esposito, 2003; D'Esposito and Postle, 2015; Postle, 2006; Serences, 2016).

Research has produced evidence for and against this hypothesis. On the one hand, WM representations as early as primary visual cortex (V1) can be used to decode WM representations (Curtis and Sprague, 2021; Harrison and Tong, 2009; Rahmati et al., 2018; Riggall and Postle, 2012; Serences et al., 2009; Sprague et al., 2014). There is even some evidence that classifiers trained on data collected from early visual cortex while participants are simply viewing stimuli (e.g., oriented gratings) can be used to decode the contents of WM (Albers et al., 2013; Harrison and Tong, 2009; Rademaker et al., 2019). The assumption here is that if sensory representations generated via bottom-up processing are interchangeable with WM representations, then the representation itself is perceptual in nature (although see (Lee et al., 2012). Finally, the degree to which WM representations in early visual cortex are epiphenomenal or only support memory under impoverished laboratory conditions remains controversial. Some evidence suggests, however, that the neural circuitry in early visual cortex can simultaneously maintain WM representations while encoding incoming and potentially distracting percepts (Hallenbeck et al., 2021; Lorenc et al., 2018; Rademaker et al., 2019). Moreover, trialwise variations in these decoded WM representations predict key behavioral factors like errors and uncertainty of memory (Li et al., 2021). Distractor induced distortions in WM representations also predict the direction and degree of distractor induced memory errors (Hallenbeck et al., 2021). Together, it appears as if memory guided behaviors depend on a readout of these representations in early visual cortex.

On the other hand, several pieces of evidence are at odds with the sensory recruitment hypothesis of WM. With perhaps the exception of spatial WM (Hallenbeck et al., 2021; Li and Curtis, 2023; Saber et al., 2015; Supèr et al., 2001; van Kerkoerle et al., 2017), persistent activity, the most conclusive neural mechanism of WM, is not characteristic of V1 neurons (Curtis and Sprague, 2021; Leavitt et al., 2017). As mentioned above, fMRI patterns during perception can be used to predict WM content. However, decoding is usually worse compared to when WM data are used to train decoders (Harrison and Tong, 2009; Rademaker et al., 2019), especially in parietal cortex (Albers et al., 2013; Rademaker et al., 2019). WM representations in early visual cortex also appear to change over time from when encoding the memoranda to its maintenance throughout the retention interval. These changes appear to reflect reformatting of the representation from one that is more sensory-like to one during WM that is more connected to the

demands of the memory-guided behavior (Henderson et al., 2022; Kwak and Curtis, 2022; Li and Curtis, 2023) and may explain how WM representations in V1 survive distraction (Hallenbeck et al., 2021; Rademaker et al., 2019).

Most of these studies that provide evidence for and against the sensory recruitment hypothesis of WM relied on decoding the orientation of gratings with fMRI patterns of voxel acitivity. The general linking hypothesis, therefore, assumes that successful orientation decoding depends on the unique patterns of activity originating from inhomogeneous sampling of orientation columns at fine scales across voxels (Boynton, 2005; Haynes and Rees, 2005; Kamitani and Tong, 2005). However, recent research suggests that coarse, not fine, scale biases at the retinotopic map level, such as a global preference for cardinal and radial orientations (Freeman et al., 2013, 2011; Mannion et al., 2010) underlie decoding of orientation during perception. Rather than just a reflection of fine-scale sampling of orientation tuned neurons, orientation decoding also relies on complex interactions between the stimulus's orientation, its bounding aperture, and topographic inhomogeneities across the visual field map (Carlson, 2014; Roth et al., 2018). Despite changes to the hypothesis linking successful decoding of perceived orientation to its underlying causes, it remains unknown how WM decoding might be affected by these coarse-scale biases.

Here, we directly address this gap by testing how aperture biases affect WM decoding, as well as leveraging these carefully manipulated stimulus properties of gratings to test how sensory-like are WM codes. In order to disambiguate the contributions to orientation decoding, we used as memoranda stimuli with aperture biases that were either aligned with or orthogonal to a grating's orientation (Roth et al., 2018). Previewing our results, we found that WM but not perceptual representations in early visual cortex were immune to aperture biases. Using models of V1 (Simoncelli et al., 1992) and techniques to visualize the spatial patterns associated with seeing and remembering oriented gratings (Favila et al., 2022; Kok and de Lange, 2014; Kwak and Curtis, 2022; Yoo et al., 2022; Zhou et al., 2022), WM representations were recoded into line-like patterns across retinotopic cortex. Together, these findings provide strong evidence that visual WM representations are not sensory-like in nature. They are abstractions of percepts and provide evidence that compels revisions to the sensory recruitment hypothesis of WM.

Results

Angular and radial modulators impact orientation decoding during perception but not memory

We measured fMRI blood-oxygen-level-dependent (BOLD) activity in retinotopic visual field maps (**Figure 1A**) in humans when participants performed a delayed orientation WM task using gratings with two types of modulators (**Figure 1B**; WM). Participants also performed a separate perceptual control experiment using the same type of stimuli, but without a WM delay (**Figure 1B**; perception). Stimuli were created by multiplying oriented sinusoidal gratings (the carrier) with an angular or a radial polar grating (the modulator) to generate orthogonal aperture biases despite having the same orientation (Roth et al., 2018). Specifically, the radial modulator evokes a coarse-scale bias aligned with the carrier orientation, while the angular modulator evokes a coarse-scale bias orthogonal to the carrier orientation (**Figure 1C**). We predicted that if the format of the memorized orientation is sensory-like in nature, decoding would conform with the aperture bias.

We first aimed to demonstrate that during a simple perception task without WM the radial and angular modulators induce different aperture biases that impact orientation decoding. As predicted, we replicated (Roth et al., 2018) that classifiers trained to decode the orientation of gratings altered by one type of modulator could only decode the orientation of gratings altered by

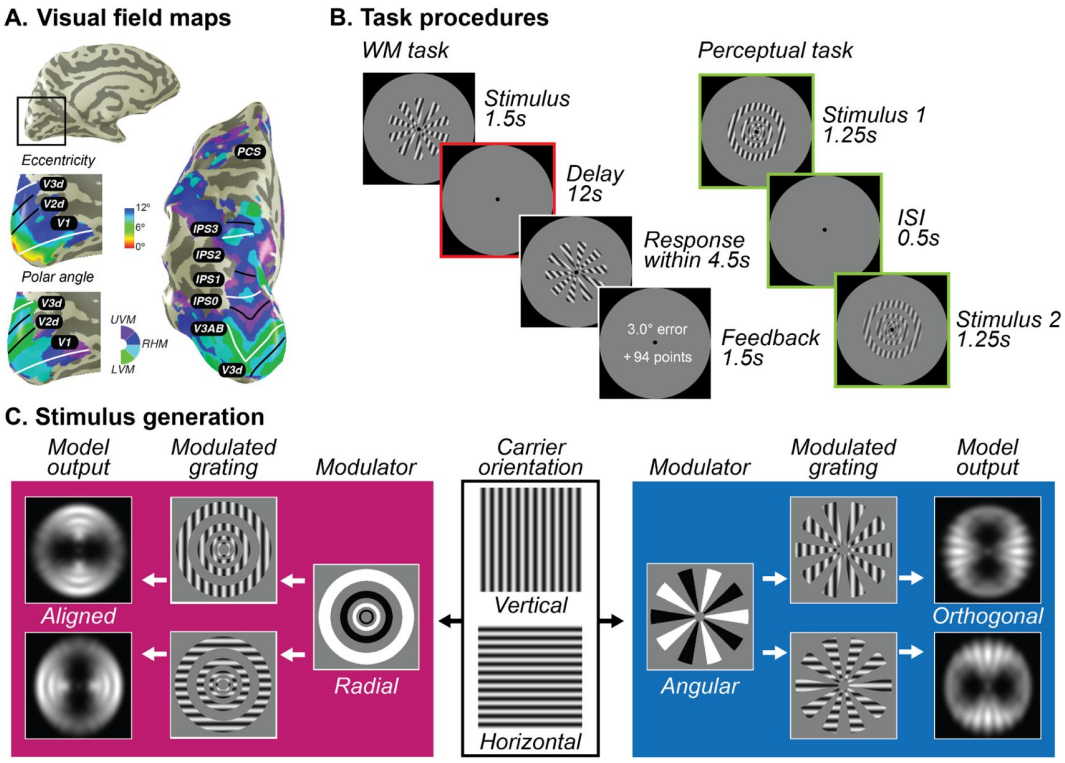

Figure 1.

Population receptive field mapping, trial design, and stimuli generation schema

A. A separate retinotopic mapping session was used to estimate voxel receptive field parameters for defining visual field maps in visual, parietal, and frontal cortices. Example participant's left hemisphere is shown.White lines denote the boundaries at the upper vertical meridian (UVM) and black lines denote the lower vertical meridian (LVM). **B.** For the WM task (left), participants maintained the oriented stimuli over a 12s retention interval and rotated a recall probe to match their memory. More points were awarded for less errors. For the perceptual control task (right), participants viewed the stimuli twice in a row with a short ISI and asked to decide which one has a higher contrast; it places no demand on remembering orientation. Colors denote different epoch of interests, green denotes stimulus presenting epoch while red denotes delay epoch. **C.** Each of the stimuli was created by multiplying a vertical or horizontal grating by a radial or angular modulator. These stimuli were used as input to the model. For radial modulated gratings (left in magenta), the model exhibits a radial preference: larger responses to vertical gratings along the vertical meridian and larger responses to horizontal gratings along the horizontal meridian. However, for angular modulated gratings (right in blue), the orientation preference is tangential: larger responses to vertical gratings along the horizontal meridian and larger responses to horizontal gratings along the vertical meridian.

the same type of modulator (**Figure 2A** & Figure S1; within). Classifiers could not cross-decode orientation gratings altered by the other type of modulator (**Figure 2A** & Figure S1; cross) presumably because classification depends on aperture biases that are orthogonal for radial and angular modulated gratings. Note that these effects were limited to visual field maps in early visual cortex (V1-V3).

Next, we focused on the patterns of late delay period activity during the WM task (**Figure 1B**) when the signals were temporally separated from those evoked during visual stimulation. We used this epoch of data for both training classifiers and testing decoding success. We first validated our methods by replicating successful orientation decoding in visual and parietal cortex separately for each type of modulator (**Figure 2B** & Figure S2; within) (Emrich et al., 2013; Ester et al., 2015; Harrison and Tong, 2009; Kwak and Curtis, 2022; Riggall and Postle, 2012; Sarma et al., 2016; Serences et al., 2009; Yu and Shim, 2017). Turning to the critical test, we asked if a classifier trained on oriented gratings with one type of modulator (e.g., radial) could be used to successfully cross-decode gratings with the other type of modulator (e.g., angular). Indeed, we found that despite the orthogonal aperture biases induced by the two modulators, their patterns during WM maintenance were interchangeable. Within visual field maps in early and mid visual cortex (V1, V2, V3, V3AB), parietal cortex (IPS0/1, IPS2/3), and frontal cortex (sPCS), classifiers trained on different modulators could cross-decode the orientation of gratings (**Figure 2B** & Figure S2; cross). These results indicate that WM representations of orientation are immune to the aperture biases we demonstrated during perception.

To further test if WM representations are similar to perception, we next trained classifiers using data from the perceptual control task and measured the extent to which these classifiers can decode orientation during WM, and what effect the modulators have on decoding. In early visual cortex (V1-V3), we found that classifiers trained during perception can be used to decode orientation information in WM, but only when the aperture bias is aligned with the orientation of the grating (i.e., radial modulator) (**Figure 2C** & Figure S3; within). Similarly, we only observed significant WM decoding across modulator types in early visual cortex when classifiers were trained during perception of the radial (aligned with orientation) but not angular (orthogonal to orientation) modulated grating (**Figure 2C** & Figure S3; cross). These results indicate that WM representations of orientation, which are not biased by aperture, are only similar to perceptual representations when they happen to align with the aperture biases induced during perception. Note that V3AB was a notable exception in that orientation could be decoded regardless of the type of modulator used for training or testing.

WM representations are recoded into abstractions of percepts

The results thus far imply that WM representations in early visual cortex are distinct from perceptual representations. Moreover, WM representations are immune to the aperture biases during perception perhaps because they have been recoded into another format during memory. Next, we aimed to visualize changes in format during perception and WM for oriented gratings with orthogonal aperture biases. We hypothesized that participants recoded in WM the carrier orientation of gratings, regardless of the type of modulator, into line-like images encoded in the spatial distribution of response amplitudes across topographic maps (Kwak and Curtis, 2022; Li et al., 2021). Again using the late delay period activity during the WM task, we constructed the spatial profile of neural activity within visual field maps (Kok and de Lange, 2014; Kwak and Curtis, 2022; Yoo et al., 2022) for both radial and angular modulated orientation gratings (**Figure 3A** & Figure S4). Specifically, for each voxel, we weighted its receptive field (the exponent of a Gaussian distribution) by the delay period amplitude and then summed across all voxels within an ROI (see **Equation 1** in Methods). Then, we rotated the reconstruction map for each orientation such that they were all centered at zero degrees (vertical meridian) and averaged across all orientation conditions. Clearly, the visualization technique confirmed our hypothesis and revealed a line encoded in the amplitudes of voxel activity at the angle matching the target orientation during the WM delay in V1-V3AB and IPS0/1 (**Figure 3A**), but not other ROIs (see

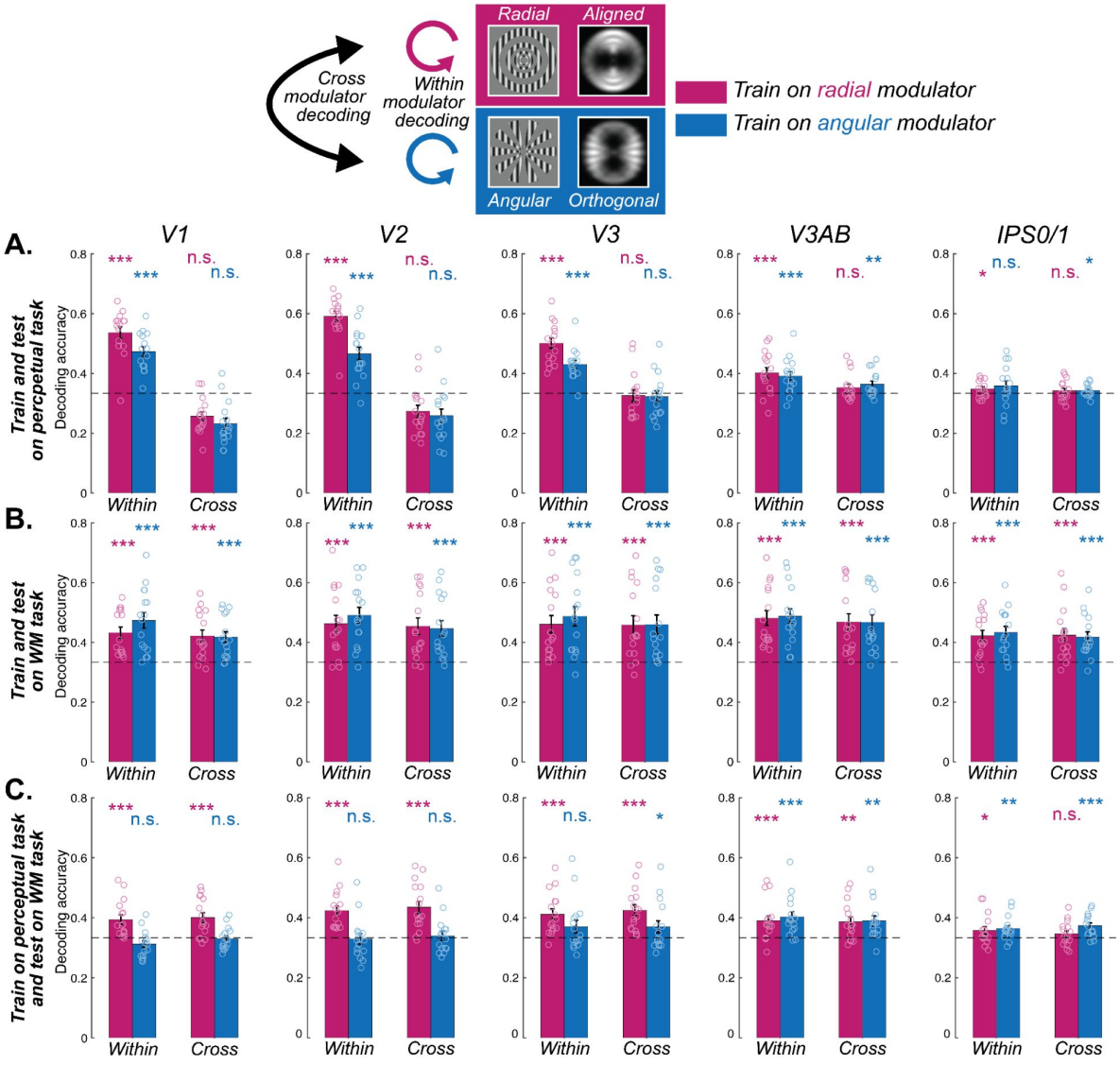

Figure 2.

Decoding orientation during WM and perception

A. Orientations could be decoded only within each kind of modulator, but not across different modulators in visual cortex, indicating the influence of the aperture bias on the stimulus presenting epoch in the perceptual task. **B.** Orientations could be decoded both within and cross modulators in both visual and parietal cortices, suggesting a shared format during the delay epoch in the WM task. **C.** When training the classifier based on the neural pattern of the radial modulator (magenta) in the perceptual task, orientations of both radial (within) and angular (cross) modulators could be decoded during the WM delay epoch in the visual cortex. However, training the classifier based on the angular modulator (blue) could not be generalized, except for V3AB. Results suggest that neural patterns during WM delay are only similar to perceptual representations when their aperture bias aligns with the orientation bias (radial modulator) in early visual cortex (V1-V3). *p<.05, **p<.01, ***p<.001, n.s. Not significant. Error bars represent ±1 SEM. Small circles for each bar represent individual data. Dashed horizontal line denotes theoretical chance level (1/3), but results are based on non-parametric permutation tests. Results for all ROIs can be seen in Supplementary Figure S1-S3.

Figure S4 for details). Critically, these line-like representations were matched to the carrier orientation and not the aperture biases induced by the modulator. We statistically confirmed these effects by quantifying the fidelity of reconstructions of the carrier orientation (**Figure 3B**).

Next, we performed the same analyses using the data from the perception control experiment. The spatial maps in V1 and V2 revealed line-like representations of the gratings, however, they were aligned with the carrier orientation only when it was radial modulated (**Figure 3C**/D, for other ROIs see Figure S5); angular modulated gratings produced line-like representations that were orthogonal carrier orientation reflecting the influence of stimulus vignetting (Roth et al., 2018). We compared those reconstructed spatial maps with the simulated responses from an image-computable model based on the properties of V1 (Roth et al., 2018; Simoncelli et al., 1992). We simulated model outputs of both types of modulated gratings as well as line-like images at angles matching the orientation of the carrier grating (**Figure 4**). The spatial maps based on model responses matched those in the perception control task. They showed a clear orthogonal orientation bias induced by the stimulus aperture, while the results for line-like images matched the spatial profile of neural activity during the WM delay.

Overall, these results provide solid evidence that mnemonic representations are flexibly recoded into a spatial topographic format that is line-like in nature with angles matching the target orientation. WM appears immune to the aperture biases because its format is an abstraction of the perceptual features underlying the biases.

Discussion

In attempts to adjudicate conflicting results between monkey and human studies of the role of the PFC in WM, (Curtis and D'Esposito, 2003) hypothesized that the PFC might be the source of top-down control signals that target neurons in sensory areas where WM representations are stored (see also (Curtis and Sprague, 2021; Postle, 2006)). Although just a speculation at the time, a few years later key evidence emerged. The orientations of memorized gratings could be decoded from the patterns of voxel activity during WM delays in primary visual cortex (Harrison and Tong, 2009; Serences et al., 2009), supporting the prediction that WM representations could be stored in sensory cortex. What became known as the *sensory recruitment hypothesis* of WM emerged shortly after (Curtis and D'Esposito, 2003; D'Esposito and Postle, 2015; Postle, 2006; Serences, 2016), which simply stated that the same neural encoding mechanisms used for perception are also utilized to store WM representations. The findings from the current study have two major and direct implications for this highly influential theory of WM. As we detail next, they provide conclusive evidence that neural representations of percepts are not the same as neural representations of memory, even in early visual cortex. Instead, our evidence indicates that WM representations are reformatted abstractions of percepts.

Orientation decoding during perception and memory depends on distinct mechanisms

First, we situate our results within existing evidence that the neural mechanisms that support perception and WM are shared. The patterns of fMRI voxel activity in early visual cortex during perception of an oriented grating can be used to predict the orientation of a grating stored in WM (Harrison and Tong, 2009; Rademaker et al., 2019). Data such as these have been used to support the idea that the representation of WM features in early visual cortex are sensory-like in nature, presumably because orientation decoding during perception and WM both depend on the activities of neurons with orientation tuning (Ester et al., 2013; Harrison and Tong, 2009; Serences et al., 2009). However, the reason why patterns of voxel activity in early visual cortex can be used to decode orientation in simple perception studies has come under scrutiny. Initially, orientation decoding was thought to reflect random voxel sampling of the fine-scale columnar

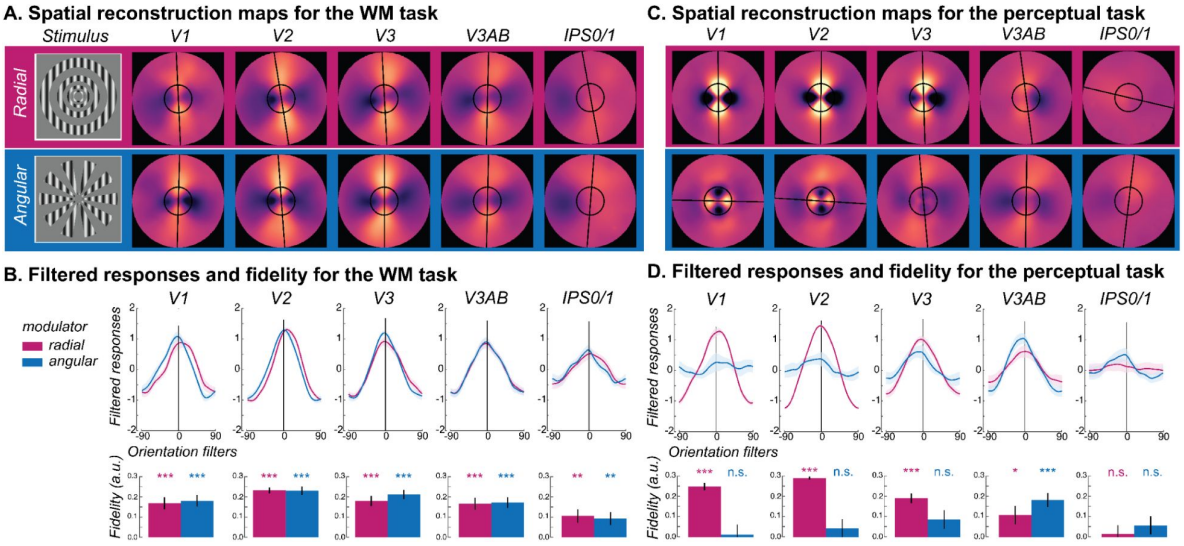

Figure 3.

Visualizing WM and perception of radial and angular modulated oriented gratings

A. Line-like patterns emerged across maps of visual space matching the memorized orientation of carrier gratings regardless of the type of modulator (radial – magenta; angular – blue) during the delay period of the WM task. Spatial maps were rotated such that all orientations were aligned at 0° (top). The warmer colors correspond to increased amplitude of BOLD activity in voxels with receptive fields corresponding to that portion of the visual field. Best fitting lines (black lines) and the size of the stimulus (black circles) are overlaid. **B.** Quantitative analysis confirmed the line-like patterns being aligned with the carrier orientation in the WM task. Filtered responses (top row) represent the sum of pixel values within the area of a line-shaped mask oriented -90° to 90°, where 0° represents the true orientation. Fidelity values (bottom row) are the result of projecting the filtered responses to 0° (see Methods), where higher fidelity values indicate stronger stimulus orientation representations. **C.** Unlike the WM task, during the perception task the angle of the line-like patterns depended on the type of modulator in early (V1 and V2), where the line matched the orientation of the aperture bias, not the carrier. Note how the line is orthogonal to the angular modulated carrier in early visual cortex (V1 and V2) but not in later visual field maps (e.g., V3A/B). **D.** During the perception task, the line-like representations in early visual cortex for radial but not angular modulated orientations result in strong filtered responses and fidelities. *p<.05, **p<.01, ***p<.001. Error bars represent ±1 SEM. Results for all ROIs can be seen in Supplementary Figure S4-S5.

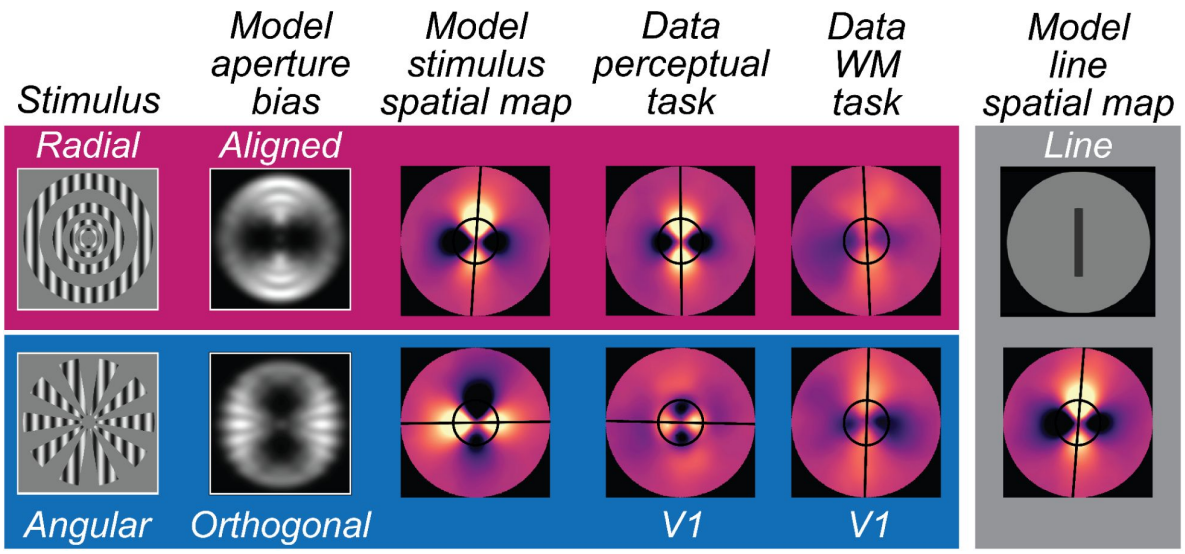

Figure 4.

Modeling and reconstructing spatial maps of perceptual and mnemonic representations in V1

At the left, we illustrate the output of the model of V1 depicting the aperture biases aligned and orthogonal to the carrier orientation for radial and angular modulators, respectively. Using these modeled responses as inputs, we visualized the population code employing the measured pRF parameters from V1 (see Methods). In the modeled stimulus spatial map, line-like representations match the aperture biases, which in turn matches the observed data from V1 during the perception task. Critically, during WM storage, the line-like representations are aligned with the memorized carrier orientation in V1, regardless of modulator type. At the right and using the same model of V1, we visualize a WM representation in V1 assuming that participants are maintaining in WM a simple line that matches the carrier orientation.

distributions of neurons with orientation tuning (Boynton, 2005; Haynes and Rees, 2005; Kamitani and Tong, 2005). Theoretical (Carlson, 2014) and empirical work (Freeman et al., 2013, 2011; Roth et al., 2018) argued that decoding depended instead on coarse-scale factors. Specifically, it appears that orientation decoding relies to some degree on the complex interaction between a grating's orientation, its bounding aperture, and the non-isotropic distribution of orientation tuned neurons across the topographic map of V1. (Roth et al., 2018), using the same modulated orientation gratings we used, demonstrated that fMRI decoding of orientation depends on the coarse-scale aperture biases the modulators evoke. Here, we leveraged the precise control of these aperture biases evoked by the modulators to test if WM representations also depend on these aperture biases. First, we replicated the aperture biases in early visual cortex during perception reported by (Roth et al., 2018) (**Figure 2A**). Second and remarkably, we found that decoding orientation from patterns of activity in early visual cortex during WM delays were immune to the aperture biases noted during perception (**Figure 2B**). Third, when training classifiers based on the perceptual task, we could only decode orientation during WM when the aperture bias was aligned with the orientation of the carrier grating (**Figure 2C**). Together, these results provide strong and direct evidence that the patterns of neural activity during perception of an oriented grating are distinct from the patterns during WM for the same grating.

Seeing is believing: WM representations are abstractions of percepts

Next, we addressed *how* WM representations of orientation are different from those during perception. To do so, we first visualized the spatial pattern of population activity within visual field maps by projecting voxel activity from cortex into spatial maps of activity in the coordinates of the physical screen within which stimuli were presented (Favila et al., 2022; Kok and de Lange, 2014; Kwak and Curtis, 2022; Li and Curtis, 2023; Yoo et al., 2022; Zhou et al., 2022). We found line-like representations of WM across many of the visual field maps in the dorsal stream whose angle matched the orientation of memorized gratings regardless of the modulator (radial, angular) and thus, regardless of the alignment between the carrier orientation and induced aperture bias (**Figure 3A**). During the perception control task, the line-like patterns were also present in the population response, but the angles of these lines matched the axis of the aperture bias rather than the grating's orientation (**Figure 3B**), again confirming differences between perception and memory. Finally, we used a computational model of V1 that simulated the aperture biases induced by the modulators (Roth et al., 2018; Simoncelli et al., 1992). Consistent with our empirical data, we found line-like stripes across retinotopic V1 aligned to the aperture bias and not the carrier orientation, providing a plausible explanation for why WM decoding depends on factors other than orientation. Instead, we propose that WM for oriented gratings, no matter what the aperture is, are reformatted into simple spatial codes, like a line (Kwak and Curtis, 2022; Li and Curtis, 2023). These line-like patterns are remarkably similar to simulations of a physical line input to the model of V1 (**Figure 4**), suggesting people are storing a simplified abstraction of the physical stimulus. These results also may explain why WM representations do not appear to undergo normalization like perceptual representations (Bloem et al., 2018).

Concluding remarks

In summary, we found the WM decoding of orientation is immune to the aperture biases that drive decoding during perceptual studies of orientation. Moreover, WM representations are reformatted into efficient abstractions of percepts such that they most closely support memory guided behavior. Although our previous study also found evidence that oriented gratings were recoded into line-like representations (Kwak and Curtis, 2022), here we demonstrate that those representations are not driven by aperture biases, but instead reflect abstract line-like representations. These results together necessitate revisions to the sensory recruitment hypothesis of WM because the same stimulus is supported by distinct, and not as predicted interchangeable,

patterns of neural activity during perception and memory. If seeing and remembering a stimulus depends on the same encoding mechanisms then one would predict an interchangeable pattern. At the very least, the sensory recruitment hypothesis must be modified to take into account both how WM representations differ from perceptual representations, and how WM representations can morph into different formats that likely depend on the goal of the memory-guided behavior.

Methods

Subjects

Sixteen neurologically healthy volunteers (including the 2 authors; 5 females; 20-54 years old) with normal or corrected-to-normal vision participated in this study. Each participant completed 3 experimental sessions (2 for the WM task and 1 for the control task, ~1h 30 mins each) and 1-2 sessions of retinotopic mapping and anatomical scans (~2hrs). The experiments were conducted with the informed consent of each participant. The experimental protocols were approved by the University Committee on Activities involving Human Subjects at New York University.

Stimuli

Stimuli were created by multiplying two gratings (a carrier and a modulator) (**Figure 1C**) (Roth et al., 2018). The carrier grating consisted of a large, oriented sinusoidal Cartesian grating (contrast = 0.8, spatial frequency = 1 cycle/°) presented within an annulus (inner diameter: 1.2°; outer diameter: 12°). The spatial phase of the carrier grating was either 0 or π, counterbalanced within each run. We generated 180 orientations for the carrier grating to cover the whole orientation space. A gray circular aperture with a diameter of 24.8° (equal to the height of the screen) was presented as the background throughout the experiment.

The modulator grating was polar-transformed and square wave with hard edges, so that when multiplying with the carrier, it alternates the phase of the carrier and creates apertures. On half of the runs, the modulator produced a set of rings starting from the fovea (radial modulator, scaled with eccentricity). While on the other half of the runs, the modulator produced a set of inward-pointing wedges encircling the fovea (angular modulator). Importantly, the image-computable model of V1 (described below) (Roth et al., 2018; Simoncelli et al., 1992) predicted a radial preference for the radial modulated gratings, but a tangential preference for the angular modulated gratings. Therefore, the radial modulator induced a bias that is consistent with the carrier orientation while the angular modulator induced a bias that is orthogonal to the carrier orientation. The modulator grating was either sine phase or cosine phase, counterbalanced within each run. The example of modulators in **Figure 1C** shows one kind of the phase conditions.

Importantly, when changing the orientation of the stimuli for each trial, it only changed the orientation of the carrier grating but not the modulator grating. Therefore, any fMRI activity measured could be attributed to either the orientation of the carrier grating, or an interaction between the orientation of the carrier grating and the static modulator grating.

Apparatus setup

All stimuli were generated by using PsychToolBox in Matlab 2021b and presented by an LCD (VPixx ProPix) projector. The projected image spanned 36.2cm in height and 64.4cm in width. The spatial resolution is 1920 × 1080 for all tasks. The refresh rate is 120 Hz for the two tasks in the current study and 60 Hz for the retinotopic mapping tasks.

fMRI task

Each participant completed 2 sessions for the WM task and 1 session for the control task on separate days. The 2 sessions for the WM task were acquired in two continuous days while having several days intervals between the WM task and the control task to minimize the task confusion. The sequence of the two tasks was randomly assigned (10 subjects did the WM task first). For both tasks, each session consisted of 10 runs, which cost 1.5 to 2 hours. Each run had 12 trials for the WM task and 24 trials for the control task. Thus, participants completed 240 trials in total for both tasks. For each run, the target orientation (i.e., the carrier's orientation) were 15°, 75°, and 135° clockwise from vertical with random jitters (<7°).

The WM task

Participants performed a delayed-estimation WM task where they need to report the remembered orientation for the target stimulus. Each trial began with 0.75s of central fixation (subtended 0.8° diameters) followed by a target stimulus for 1.5s. The stimulus was either radial modulated or angular modulated grating, presented in blocked designs and in interleaved order. After a 12s delay period, participants were asked to rotate a recall probe with a dial to match the remembered orientation within a 4.5s response window. To avoid visual afterimage, we inserted a 0.6s noise mask at the beginning of the delay. The recall probe was the same type as the target stimulus to avoid forcing participants to represent the two stimulus types in an abstract manner. Again, when changing the orientation of the recall probe, only the carrier grating but not the modulator was changing. Participants were provided with feedback on the error they made and the points earned based on the error for each trial (100 points for 0°, no points for ≥ 50°, 2 points for each degree). The feedback was displayed for 1.5s and followed by an inter-trial-interval (ITI) of 6, 9, or 12s.

The perceptual control task

To better compare mnemonic formats with sensory representations, we asked participants to do an additional control task. Instead of asking participants to remember the orientation of the target stimulus, we presented it twice (1.25s for each) with a short inter-stimulus-interval (ISI, 0.5s) and asked participants to discriminate their contrast. The feedback was displayed for 0.5s and followed by an inter-trial-interval (ITI) of 6, 9, or 12s. Thus, the two target stimuli were exactly the same except for their contrast. The contrast for each stimulus was generated from a predefined set of 20 contrasts uniformly distributed between 0.5 and 1.0. We created 20 levels of task difficulty based on the contrast distance between the two stimuli. The task difficulty level was changing based on an adaptive, 1-up-2-down staircase procedure (Levitt, 1971) to maintain performance at approximately 70% correct.

Retinotopic mapping task

Each participant was scanned for a separate retinotopic mapping session (8-12 runs) to identify region-of-interest (ROI) and model each voxel's population receptive field (pRF). Participants ran in either type of attention-demanding tasks: random dot kinematogram (RDK) motion direction discrimination task (2 participants) (Mackey et al., 2017) or an object image rapid serial visual presentation (RSVP) task(14 participants).

In the RDK motion discrimination task, participants maintained fixation at the center of the screen while covertly tracking a bar sweeping slowly but discretely across the screen in four directions (left-to-right, right-to-left, bottom-to-up, up-to-bottom). The bar was divided into three rectangular patches (one central patch and two flanking patches). The dot motion in one of the flanking patches matched the one in the central patch, while the other is the opposite. Participants were asked to discriminate which one is matched. The coherence of dot motions was 100% in the central patch, while the coherence in the flanking patches was staircase by using 2-up-1-down procedure to keep the task difficulty at about 75% accuracy (Levitt, 1971).

In the object image RSVP task, the moving bar that participants need to track consisted of 6 different object images. In each sweep, participants were asked to report whether the target object image existed among the 6 images by pressing a button. The target image was pseudo-randomly chosen for each run and was shown at the start of each run to help participants get familiar with it. The presentation duration of object bars was adjusted based on participants' accuracy in a staircase procedure.

MRI data acquisition

MRI data were acquired on a Siemens Prisma 3T scanner with a 64-channel head/neck coil. For the WM task and the control task, 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, FoV 200 x 200mm, TE/TR of 30/750ms, P → A phase encoding). Intermittently throughout each scanning session, we also acquired distortion mapping scans to measure field inhomogeneities with both forward and reverse phase encoding using a 2D SE-EPI readout and the number of slices matching that of the GE-EPI (TE/TR: 45.6/3537ms, 3 volumes per phase encode direction). BOLD contrast images for the retinotopic mapping task were acquired in a separate session with a higher resolution (MB factor of 4, 56 slices, 2 x 2 x 2mm voxel size, FoV 208 x 208mm, TE/TR: 42/1300ms, P → A phase encoding). Similarly, we collected distortion mapping scans to measure field inhomogeneities with both forward and reverse phase encoding using a 2D SE-EPI readout and the number of slices matching that of the GE-EPI (TE/TR: 71.8/6690ms). Moreover, we also collected 2 or 3 T1 weighted (192 slices, 0.8 x 0.8 x 0.8mm voxel size, FoV 256 x 240mm, TE/TR: 2.24/2400ms) and 1 or 2 T2 weighted (224 slices, 0.8 x 0.8 x 0.8mm voxel size, FoV 256 x 240mm, TE/TR: 564/3200ms) whole-brain anatomical scans using the Siemens product MPRAGE for each participant.

MRI 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 (15 mm FWHM), which was computed as the 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 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. Finally, 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.

Retinotopic mapping and region of interest (ROI) definition

Since the retinotopic mapping scans were acquired with a higher resolution than the experimental scans, we projected the retinotopic time series data onto the surface from its original space (2mm), then from the surface to volume space at the task voxel resolution (2.5mm). This ensured that

variance-explained estimates faithfully reflect the goodness of fit and are not impacted by smoothing incurred from transforming fit parameter values between different voxel grids.

We fit a population receptive field (pRF) model with compressive spatial summation to the averaged time series across all retinotopy runs for each participant after smoothing on the surface with 5mm FWHM Gaussian kernel (Kay et al., 2013; Wandell et al., 2007). Then, 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). We set a threshold to only include voxels with greater than 10% variance explained by the pRF model. We defined bilateral visual ROIs, V1, V2, V3, V3AB, IPS0, IPS1, IPS2, IPS3, iPCS, and sPCS.

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*). We used 'Softmax' and 'cross entropy' as the activation and performance functions, which are suitable for multi-class linear classification problems (Kwak and Curtis, 2022). The scaled conjugated gradient method was used to fit the weights and bias parameters.

WM task decoding analysis

For the main task, we focused on the delay epoch to test the abstract representational format in WM. First, we performed within-modulator decoding from the same modulator type to verify the reliable orientation information during the WM delay epoch. Then, we conducted cross-modulator decoding from different modulator types (e.g., training on the angular modulator and testing on the radial modulator). We were mostly interested in the cross-modulator decoding results to examine whether WM forms an abstract representation across different modulator types.

Decoding analysis was performed on the beta coefficients acquired from running a voxel-wise general linear model (GLM) using AFNI 3dDeconvolve. For each participant, we used GLM to estimate the responses of each voxel to the stimulus encoding, delay, and response epochs. Note that, to better separate data from delay epoch from encoding epoch, we modeled the second half of the whole delay period (late delay). Using the whole delay did not change any of the results we reported here. 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 late delay epoch in performing the 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 epoch of interest while preventing the trial-by-trial variability of other epochs from soaking up a large portion of variance which could potentially be explained by the epoch of interest. Six motion regressors were included to account for movement during the scan. Each voxel's beta coefficients were z-scored within each run independently before the decoding analysis.

We performed a three-way classification to decode the three target orientation conditions, which were 15°, 75°, and 135° clockwise from vertical. For within-modulator decoding, we used leave-one-run-out cross-validation procedure, in which all trials in one run were left out on each iteration to test the performance of the classifier trained on the data from all other runs. For cross-stimulus decoding, the classifier was trained on beta coefficients of all trials in one modulator condition and tested on all trials in the other modulator condition.

Control task decoding analysis

To get better control and verify the existence of the stimulus vignetting effect (Roth et al., 2018), we conducted a purely perceptual task and performed the same analysis on the stimulus epoch data from this task. Based on previous findings, we expected to find reliable above-chance decoding performance for within-modulator decoding, but not for cross-modulator decoding.

Cross-task decoding analysis

We also performed cross-task decoding to test how neural representational formats change for different task goals. For each modulator type (e.g., angular modulator), we trained the classifier based on the stimulus epoch data in the control task and tested it on both the stimulus epoch and the delay epoch data in the WM task for both modulator types (i.e., angular and radial modulators). We were mainly interested in testing the classifier on the late delay epoch data in the WM task. If the WM representations changed to a common format for both modulator types to match the orientation bias, we expected to find a reliable above-chance decoding when training the classifier based on the radial modulator but not the angular modular type. This is because the radial modulator induces a bias that is consistent with the carrier orientation, while the angular modulator induces an orthogonal bias compared to the carrier orientation.

fMRI data analysis: Spatial reconstruction

To visualize the spatial profile of neural activity during the epoch of interest, we projected voxel amplitudes onto the 2D visual field space for each orientation condition and each ROI across all participants. Specifically, we first averaged the beta coefficients (β) from GLM for all trials in each orientation condition. Then, for each voxel, we weighted its receptive field (the exponent of a Gaussian distribution) by the averaged β. Finally, we summed the weighted receptive fields across all voxels within a certain ROI for each orientation condition. To account for the individual differences in the pRF structure, we normalized the spatial profile for each participant and then got the averaged spatial profile across all participants.

For each orientation condition i, the sum S_i of all voxels' weighted receptive fields (assuming the number of voxels in a certain ROI is m) could be computed as Equation 1, where j is the index of each voxel; x_i, y_i and σ_j are the center and width of the voxel's receptive field. x and y are the positions in the reconstruction map at which the receptive fields were evaluated.

$$S_i = \sum_{j=1}^{m} \beta_{j,i} \times e^{-\frac{(x_j-x)^2+(y_j-y)^2}{2\sigma_j^2}} \qquad \text{(Equation 1)}$$

To better visualize the line format, we fit a first-degree polynomial to the reconstructed map in **Figure 3A** and **3C** (black lines). Specifically, we selected pixels with the top 10% image intensity and fit these pixels' coordinates to a first-degree polynomial with a constraint that the fitted polynomial passed through the center. To account for the difference in image intensity between different pixels, we conducted a weighted fit, in which the weight corresponds to the voxel's rank in terms of its image intensity. We were mainly interested in comparing the spatial reconstruction maps between the delay epoch in the WM task and the stimulus epoch in the control task. The visualization provided us with an intuitive understanding of how representational formats changed from perception to WM, and what drove the different decoding results. In addition, we also performed the spatial reconstruction method for the stimulus epoch in the WM task for exploration.

Model simulation: Image-computable model of V1

We used an image-computable model to predict fMRI responses of V1 to different types of stimuli for visual perception (Roth et al., 2018). We first simulated model outputs to different modulator types and then predicted fMRI responses by using pRF sampling analysis. To better visualize the model predictions, we conducted the same spatial reconstruction based on the simulated fMRI responses.

Simulate model outputs

The image-computable model is based on the steerable pyramid model of V1 (Simoncelli et al., 1992), a subband image transform that decomposes an image into orientation and spatial frequency channels. 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. RFs with the same orientation and spatial frequency tuning but different location preferences are channels. In the model, the number of spatial frequency channels, orientation channels, and orientation bandwidth are adjustable. For the model simulation, we used six orientation bands (bandwidth = 360°/6=60°) and a spatial frequency bandwidth of 0.5 octaves as in previous studies (Kwak and Curtis, 2022; Roth et al., 2018). Using four or more bands with correspondingly broader or narrower tuning curves yielded similar results supporting the same conclusions. We used images that were 1920 × 1080 pixels, which resulted in 16 spatial frequency channels for the model. The input images had the same configurations (size of fixation, inner aperture, outer aperture, etc) as the stimuli in both the WM task and the control task. The model outputs were images of the same resolution as the input images, in which each pixel can be thought of as a simulated neuron in the retinotopic map of V1. Importantly, we summed the model responses across all orientation channels, which resulted in a model without any orientation tuning.

For both types of stimuli, we used three target orientations (15°, 75°, and 135° clockwise from vertical), which had two kinds of phases for both the carriers and the modulators. We first generated the model's responses to each target image separately, then averaged the model responses across all phases for each orientation condition. This yielded a set of 3 simulated voxel maps, one for each orientation condition. Within each set, we have an averaged map across all 16 subbands and maps for each subband. We chose the subband with the maximal response differences between the two modulators (level 9) as the final model output for further analysis.

pRF sampling analysis

To simulate an fMRI voxel's response to the stimuli, each participant's pRF Gaussian parameters of V1 were used to weight the model outputs, which resulted in a weighted sum of neural responses corresponding to pRFs. For each orientation condition i, the sampled fMRI BOLD signal (B_i) for voxel j with a pRF centered at x_j, y_j and standard deviation of σ_j, is computed as the dot product between the pRF and the model output (M_i) as in Equation 2. X and y are the positions in the model outputs at which the receptive fields were evaluated.

$$B_{i,j} = \sum_{x,y} M_i \times e^{-\frac{(x_j-x)^2+(y_j-y)^2}{2\sigma_j^2}} \quad \text{(Equation 2)}$$

Finally, we performed the same spatial reconstruction analysis on these simulated BOLD signals after normalizing (z-score) across the three orientation conditions. To account for the individual differences in the pRF structure, we normalized the spatial profile for each participant's simulation and then got the averaged spatial profile across all participants. This was done separately for each of the two modulators.

Eye-tracking setup and analyses

For all imaging sessions, we measured eye position using an EyeLink 1000 Plus infrared video-based eye tracker (SR Research) mounted beneath the screen inside the scanner bore operating at 500 Hz. The camera always tracked the participant's right eye, and we calibrated using either a 9-point (WM task and perceptual control task) or 5-point (retinotopic mapping task) calibration routine at the beginning of the session and as necessary between runs. We monitored gaze data and adjusted pupil/ corneal reflection detection parameters as necessary during and/or between each run.

We preprocessed raw gaze data using fully-automated procedures implemented within iEye_ts (*https://github.com/tommysprague/iEye_ts*). Eye positions were not monitored for S04 during the first and for S16 during both of the two WM task sessions due to technical issues. Overall, 97.72% (radial) and 96.79% (angular) of the total number of eye position sample points during the delay epoch of the WM task across all subjects were within 2° eccentricity from the center (the fixation and the stimulus subtended 0.8° and 12° diameter respectively). The circular correlation between the polar angle of the target orientation and the polar angle of the eye positions is not significant for both the radial (mean=0.030, s.d.=0.103, $t(14)=1.141$, $p=0.273$) and the angular (mean=0.001, s.d.=0.099, $t(14)=0.056$, $p=0.956$) modulator, suggesting that the eye movements could not account for our findings.

Quantification and statistical analysis

All statistical results reported here were based on permutation tests over 1000 iterations. To test whether decoding accuracy was significantly greater than chance level (1/3), we generated permuted null distributions of decoding accuracy values for each participant, ROI, decoding type (within/cross), modulator type (angular/radial), and each time point for the temporal decoding analysis. On each iteration, we shuffled the training data matrix (voxels x trials) for both dimensions so that both voxel information and orientation labels were shuffled. Then, we performed the decoding analysis based on the shuffled data. This procedure was conducted for each of the 16 participants, resulting in 16 null distributions of decoding accuracy. Combining the null decoding accuracy across all participants resulted in one t-statistic per permutation. To test across-participants decoding accuracy against chance level (1/3), we compared the t-statistic calculated from the intact data against the permuted null distribution of t-statistic for each condition and ROI. The p-value is calculated as the proportion of permuted t-statistics that are greater than or equal to the t-statistic using the intact data.

For the spatial reconstruction analysis, we computed reconstruction fidelity to quantify the amount of orientation information in each reconstruction map. Specifically, we first created line filters with orientations evenly spaced between 0° and 180° in steps of 1°. Then, we created masks around these line filters based on two rules. First, coordinates form an acute angle to the oriented line filter (dot product > 0). Second, the projected distance squared is less than a threshold. We chose 1000 here according to the previous paper (Kwak and Curtis, 2022), using different thresholds did not change the results. We chose pixels within these masked areas and summed up the intensities. After z-scoring the summed intensities within each orientation condition, we rotated the response function so that the center is the target orientation. The final tuning curve-like response function was averaged across all three orientation conditions. 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 reconstruction on average points in the correct direction.

The same procedure for statistical analysis was used for the reconstruction fidelity, with the exception that the null hypothesis for the t-statistic was 0. Specifically, the data-derived 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 condition label dimensions, and the shuffled beta coefficients were used to weight the voxels' pRF parameters.

To test whether there were differences in decoding accuracy (and reconstruction fidelity value) between the decoding type and modulator type within each ROI, we used permutation-based two-way repeated-measures analysis of variance (ANOVA). For each permutation, we shuffled the condition labels (decoding type and modulator type) per participant and calculated the null F-statistic. We repeated this procedure 1000 times and got the null distribution of the F-statistic.

We compared the F-statistic derived from the intact data with the null distribution to get the p-value. Significant effects were followed up with post-hoc paired-sample t-tests, and the p-value was calculated by comparing the t-statistic derived by the intact data against a permuted null distribution of t-statistics generated by shuffling condition labels. The p-value was corrected by using a false-discovery rate (FDR) procedure for multiple comparisons.

Data and code availability

The processed fMRI data and raw behavioral data generated in this study have been deposited in the Open Science Framework at *https://osf.io/kws9b/*. Processed fMRI data contains extracted time series from each voxel of each ROI. We also make publicly available all codes that used to analysis the fMRI data, implement the theoretic model of V1, and generate the stimuli.

Acknowledgements

This work was supported by National Institutes of Health Grants R01 EY016407 and EY033925 to CEC. We thank Jonathan Winawer for helpful comments on earlier versions of the manuscript, and NYU's Center for Brain Imaging for support.

Additional information

Funding

Funder	Grant reference number	Author
National Institutes of Health	R01 EY016407	Clayton E Curtis
National Institutes of Health	R01 EY033925	Clayton E Curtis

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Ziyi Duan, Data curation, Formal analysis, Investigation, Visualization, Methodology, Writing–original draft, Writing–review and editing. Clayton E Curtis, Conceptualization, Data curation, Supervision, Funding acquisition, Validation, Investigation, Visualization, Project administration, Writing— review and editing.

Ethics

Human subjects: All subjects gave written informed consent before participating. All procedures were approved by the human subjects Institutional Review Board at New York University.

References

Albers AM, Kok P, Toni I, Dijkerman HC, de Lange FP (2013) **Shared representations for working memory and mental imagery in early visual cortex** *Curr Biol* **23**:1427–1431 https://doi.org/10.1016/j.cub.2013.05.065

Bloem IM, Watanabe YL, Kibbe MM, Ling S (2018) **Visual Memories Bypass Normalization** *Psychol Sci* **29**:845–856 https://doi.org/10.1177/0956797617747091

Boynton GM (2005) **Imaging orientation selectivity: decoding conscious perception in V1** *Nat Neurosci* **8**:541–542 https://doi.org/10.1038/nn0505-541

Carlson TA (2014) **Orientation Decoding in Human Visual Cortex: New Insights from an Unbiased Perspective** *J Neurosci* **34**:8373–8383 https://doi.org/10.1523/jneurosci.0548-14.2014

Curtis CE, D'Esposito M (2003) **Persistent activity in the prefrontal cortex during working memory** *Trends Cogn Sci* **7**:415–423 https://doi.org/10.1016/s1364-6613(03)00197-9

Curtis CE, Sprague TC (2021) **Persistent Activity During Working Memory From Front to Back** *Front Neural Circuits* **15** https://doi.org/10.3389/fncir.2021.696060

D'Esposito M, Postle BR (2015) **The cognitive neuroscience of working memory** *Annu Rev Psychol* **66**:115–142 https://doi.org/10.1146/annurev-psych-010814-015031

Emrich SM, Riggall AC, Larocque JJ, Postle BR (2013) **Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory** *J Neurosci* **33**:6516–6523 https://doi.org/10.1523/JNEUROSCI.5732-12.2013

Ester EF, Anderson DE, Serences JT, Awh E (2013) **A neural measure of precision in visual working memory** *J Cogn Neurosci* **25**:754–761 https://doi.org/10.1162/jocn_a_00357

Ester EF, Sprague TC, Serences JT (2015) **Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory** *Neuron* **87**:893–905 https://doi.org/10.1016/j.neuron.2015.07.013

Favila SE, Kuhl BA, Winawer J (2022) **Perception and memory have distinct spatial tuning properties in human visual cortex** *Nat Commun* **13** https://doi.org/10.1038/s41467-022-33161-8

Freeman J, Brouwer GJ, Heeger DJ, Merriam EP (2011) **Orientation Decoding Depends on Maps, Not Columns** *Journal of Neuroscience* **31**:4792–4804 https://doi.org/10.1523/jneurosci.5160-10.2011

Freeman J, Heeger DJ, Merriam EP (2013) **Coarse-Scale Biases for Spirals and Orientation in Human Visual Cortex** *J Neurosci* **33**:19695–19703 https://doi.org/10.1523/jneurosci.0889-13.2013

Hallenbeck GE, Sprague TC, Rahmati M, Sreenivasan KK, Curtis CE (2021) **Working memory representations in visual cortex mediate distraction effects** *Nature Communications* **12** https://doi.org/10.1038/s41467-021-24973-1

Harrison SA, Tong F (2009) **Decoding reveals the contents of visual working memory in early visual areas** *Nature* **458**:632–635 https://doi.org/10.1038/nature07832

Haynes J-D, Rees G (2005) **Predicting the orientation of invisible stimuli from activity in human primary visual cortex** *Nat Neurosci* **8**:686–691 https://doi.org/10.1038/nn1445

Henderson MM, Rademaker RL, Serences JT (2022) **Flexible utilization of spatial- and motor-based codes for the storage of visuo-spatial information** *Elife* **11** https://doi.org/10.7554/eLife.75688

Kamitani Y, Tong F (2005) **Decoding the visual and subjective contents of the human brain** *Nat Neurosci* **8**:679–685 https://doi.org/10.1038/nn1444

Kay KN, Winawer J, Mezer A, Wandell BA (2013) **Compressive spatial summation in human visual cortex** *J Neurophysiol* **110**:481–494 https://doi.org/10.1152/jn.00105.2013

Kok P, de Lange FP (2014) **Shape perception simultaneously up- and downregulates neural activity in the primary visual cortex** *Curr Biol* **24**:1531–1535 https://doi.org/10.1016/j.cub.2014.05.042

Kwak Y, Curtis CE (2022) **Unveiling the abstract format of mnemonic representations** *Neuron* **110**:1822–1828 https://doi.org/10.1016/j.neuron.2022.03.016

Leavitt ML, Mendoza-Halliday D, Martinez-Trujillo JC (2017) **Sustained Activity Encoding Working Memories: Not Fully Distributed** *Trends in Neurosciences* https://doi.org/10.1016/j.tins.2017.04.004

Lee S-H, Kravitz DJ, Baker CI (2012) **Disentangling visual imagery and perception of real-world objects** *Neuroimage* **59**:4064–4073 https://doi.org/10.1016/j.neuroimage.2011.10.055

Levitt H (1971) **Transformed Up-Down Methods in Psychoacoustics** *J Acoust Soc Am* **49**:467–477 https://doi.org/10.1121/1.1912375

Li H-H, Curtis CE (2023) **Neural population dynamics of human working memory** *Curr Biol* **33**:3775–3784 https://doi.org/10.1016/j.cub.2023.07.067

Li H-H, Sprague TC, Yoo AH, Ma WJ, Curtis CE (2021) **Joint representation of working memory and uncertainty in human cortex** *Neuron* **109**:3699–3712 https://doi.org/10.1016/j.neuron.2021.08.022

Lorenc ES, Sreenivasan KK, Nee DE, Vandenbroucke ARE, D'Esposito M (2018) **Flexible Coding of Visual Working Memory Representations during Distraction** *J Neurosci* **38**:5267–5276 https://doi.org/10.1523/JNEUROSCI.3061-17.2018

Mackey WE, Winawer J, Curtis CE (2017) **Visual field map clusters in human frontoparietal cortex** *Elife* **6** https://doi.org/10.7554/elife.22974

Mannion DJ, McDonald JS, Clifford CWG (2010) **Orientation anisotropies in human visual cortex** *J Neurophysiol* **103**:3465–3471 https://doi.org/10.1152/jn.00190.2010

Postle BR (2006) **Working memory as an emergent property of the mind and brain** *Neuroscience* **139**:23–38 https://doi.org/10.1016/j.neuroscience.2005.06.005

Rademaker RL, Chunharas C, Serences JT (2019) **Coexisting representations of sensory and mnemonic information in human visual cortex** *Nat Neurosci* **22**:1336–1344 https://doi.org/10.1038/s41593-019-0428-x

Rahmati M, Saber GT, Curtis CE (2018) **Population Dynamics of Early Visual Cortex during Working Memory** *J Cogn Neurosci* **30**:219–233 https://doi.org/10.1162/jocn_a_01196

Riggall AC, Postle BR (2012) **The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging** *J Neurosci* **32**:12990–12998 https://doi.org/10.1523/JNEUROSCI.1892-12.2012

Rissman J, Gazzaley A, D'Esposito M (2004) **Measuring functional connectivity during distinct stages of a cognitive task** *Neuroimage* **23**:752–763 https://doi.org/10.1016/j.neuroimage.2004.06.035

Roth ZN, Heeger DJ, Merriam EP (2018) **Stimulus vignetting and orientation selectivity in human visual cortex** *Elife* **7** https://doi.org/10.7554/elife.37241

Saber GT, Pestilli F, Curtis CE (2015) **Saccade planning evokes topographically specific activity in the dorsal and ventral streams** *J Neurosci* **35**:245–252 https://doi.org/10.1523/JNEUROSCI.1687-14.2015

Sarma A, Masse NY, Wang X-J, Freedman DJ (2016) **Task-specific versus generalized mnemonic representations in parietal and prefrontal cortices** *Nat Neurosci* **19**:143–149 https://doi.org/10.1038/nn.4168

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

Serences JT, Ester EF, Vogel EK, Awh E (2009) **Stimulus-Specific Delay Activity in Human Primary Visual Cortex** *Psychological Science* **20**:207–214 https://doi.org/10.1111/j.1467-9280.2009.02276.x

Simoncelli EP, Freeman WT, Adelson EH, Heeger DJ (1992) **Shiftable multiscale transforms** *IEEE Trans Inf Theory* **38**:587–607 https://doi.org/10.1109/18.119725

Sprague TC, Ester EF, Serences JT (2014) **Reconstructions of information in visual spatial working memory degrade with memory load** *Curr Biol* **24**:2174–2180 https://doi.org/10.1016/j.cub.2014.07.066

Supèr H, Spekreijse H, Lamme VA (2001) **A neural correlate of working memory in the monkey primary visual cortex** *Science* **293**:120–124 https://doi.org/10.1126/science.1060496

van Kerkoerle T, Self MW, Roelfsema PR (2017) **Layer-specificity in the effects of attention and working memory on activity in primary visual cortex** *Nat Commun* **8** https://doi.org/10.1038/ncomms13804

Wandell BA, Dumoulin SO, Brewer AA (2007) **Visual field maps in human cortex** *Neuron* **56**:366–383 https://doi.org/10.1016/j.neuron.2007.10.012

Yoo AH, Bolaños A, Hallenbeck GE, Rahmati M, Sprague TC, Curtis CE (2022) **Behavioral Prioritization Enhances Working Memory Precision and Neural Population Gain** *J Cogn Neurosci* **34**:365–379 https://doi.org/10.1162/jocn_a_01804

Yu Q, Shim WM (2017) **Occipital, parietal, and frontal cortices selectively maintain task-relevant features of multi-feature objects in visual working memory** *Neuroimage* **157**:97–107 https://doi.org/10.1016/j.neuroimage.2017.05.055

Zhou Y, Curtis CE, Sreenivasan KK, Fougnie D (2022) **Common Neural Mechanisms Control Attention and Working Memory** *J Neurosci* **42**:7110–7120 https://doi.org/10.1523/JNEUROSCI.0443-22.2022

Article and author information

Ziyi Duan
Department of Psychology, New York University, New York, NY 10003, USA
ORCID iD: 0000-0001-7567-4120

Clayton E. Curtis
Department of Psychology, New York University, New York, NY 10003, USA, Center for Neural Science, New York University, New York, NY 10003, USA
For correspondence: clayton.curtis@nyu.edu
ORCID iD: 0000-0003-0702-1499

Copyright

Editors

Reviewing Editor
Marius Peelen
Radboud University, Nijmegen, Netherlands

Senior Editor
Floris de Lange
Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands

Reviewer #1 (Public Review):

Summary:
The authors aim to test the sensory recruitment theory of visual memory, which assumes that visual sensory areas are recruited for working memory, and that these sensory areas represent visual memories in a similar fashion to how perceptual inputs are represented. To test the overlap between working memory (WM) and perception, the authors use coarse stimulus (aperture) biases that are known to account for (some) orientation decoding in the visual cortex (i.e., stimulus energy is higher for parts of an image where a grating orientation is perpendicular to an aperture edge, and stimulus energy drives decoding). Specifically, the authors show gratings (with a given "carrier" orientation) behind two different apertures: one is a radial modulator (with maximal energy aligned with the carrier orientation) and the other an angular modulator (with maximal energy orthogonal to the carrier orientation).

When the subject detects contrast changes in these stimuli (the perceptual task), orientation decoding only works when training and testing within each modulator, but not across modulators, showing the impact of stimulus energy on decoding performance. Instead, when subjects remember the orientation over a 12s delay, orientation decoding works irrespective of the modulator used. The authors conclude that representations during WM are therefore not "sensory-like", given that they are immune to aperture biases. This invalidates the sensory recruitment hypothesis, or at least the part assuming that when sensory areas are recruited during WM, they are recruited in a manner that resembles how these areas are used during perception.

Strengths:
Duan and Curtis very convincingly show that aperture effects that are present during perception, do not appear to be present during the working memory delay. Especially when the debate about "why can we decode orientations from human visual cortex" was in full swing, many may have quietly assumed this to be true (e.g., "the memory delay has no stimuli, and ergo no stimulus aperture effects"), but it is definitely not self-evident and nobody ever thought to test it directly until now. In addition to the clear absence of aperture effects during the delay, Duan and Curtis also show that when stimulus energy aligns with the carrier orientation, cross-generalization between perception and memory does work (which could explain why perception-to-memory cross-decoding also works). All in all, this is a clever manipulation, and I'm glad someone did it, and did it well.

Weaknesses:
There seems to be a major possible confound that prohibits strong conclusions about "abstractions" into "line-like" representation, which is spatial attention. What if subjects simply attend the endpoints of the carrier grating, or attend to the edge of the screen where the carrier orientation "intersects" in order to do the task? This may also result in reconstructions that have higher bold at areas close to the stimulus/screen edges along the carrier orientation. The question then would be if this is truly an "abstracted representation", or if subjects are merely using spatial attention to do the task.

Alternatively (and this reaches back to the "fine vs coarse" debate), another argument could be that during memory, what we are decoding is indeed fine-scale inhomogenous sampling of orientation preferences across many voxels. This is clearly not the most convincing argument, as the spatial reconstructions (e.g., Figure 3A and C) show higher BOLD for voxels with receptive fields that are aligned to the remembered orientation (which is in itself a form of coarse-scale bias), but could still play a role.

To conclude that the spatial reconstruction from the data indeed comes from a line-like representation, you'd need to generate modeled reconstructions of all possible stimuli and representations. Yes, Figure 4 shows that line results in a modeled spatial map that resembles the WM data, but many other stimuli might too, and some may better match the data. For example, the alternative hypothesis (attention to grating endpoints) may very well lead to a very comparable model output to the one from a line. However testing this would not suffice, as there may be an inherent inverse problem (with multiple stimuli that can lead to the same visual field model).

The main conclusion, and title of the paper, that visual working memories are abstractions of percepts, is therefore not supported. Subjects could be using spatial attention, for example. Furthermore, even if it is true that gratings are abstracted into lines, this form of abstraction would not generalize to any non-spatial feature (e.g., color cannot become a line, contrast cannot become a line, etc.), which means it has limited explanatory power.

Additional context:
The working memory and perception tasks are rather different. In this case, the perception task does not require the subject to process the carrier orientation (which is largely occluded,

and possibly not that obvious without paying attention to it), but attention is paid to contrast. In this scenario, stimulus energy may dominate the signal. In the WM task, subjects have to work out what orientation is shown to do the task. Given that the sensory stimulus in both tasks is brief (1.5s during memory encoding, and 2.5s total in the perceptual task), it would be interesting to look at decoding (and reconstructions) for the WM stimulus epoch. If abstraction (into a line) happens in working memory, then this perceptual part of the task should still be susceptible to aperture biases. It allows the authors to show that it is indeed during memory (and not merely the task or attentional state of the subject) that abstraction occurs.

What's also interesting is what happens in the passive perceptual condition, and the fact that spatial reconstructions for areas beyond V1 and V2 (i.e., V3, V3AB, and IPS0-1) align with (implied) grating endpoints, even when an angular modulator is used (Figure 3C). Are these areas also "abstracting" the stimulus (in a line-like format)?

https://doi.org/10.7554/eLife.94191.1.sa2

Reviewer #2 (Public Review):

Summary:
According to the sensory recruitment model, the contents of working memory (WM) are maintained by activity in the same sensory cortical regions responsible for processing perceptual inputs. A strong version of the sensory recruitment model predicts that stimulus-specific activity patterns measured in sensory brain areas during WM storage should be identical to those measured during perceptual processing. Previous research casts doubt on this hypothesis, but little is known about how stimulus-specific activity patterns during perception and memory differ. Through clever experimental design and rigorous analyses, Duan & Curtis convincingly demonstrate that stimulus-specific representations of remembered items are highly abstracted versions of representations measured during perceptual processing and that these abstracted representations are immune to aperture biases that contribute to fMRI feature decoding. The paper provides converging evidence that neural states responsible for representing information during perception and WM are fundamentally different, and provides a potential explanation for this difference.

Strengths:
1. The generation of stimuli with matching vs. orthogonal orientations and aperture biases is clever and sets up a straightforward test regarding whether and how aperture biases contribute to orientation decoding during perception and WM. The demonstration that orientation decoding during perception is driven primarily by aperture bias while during WM it is driven primarily by orientation is compelling.

2. The paper suggests a reason why orientation decoding during WM might be immune to aperture biases: by weighting multivoxel patterns measured during WM storage by spatial population receptive field estimates from a different task the authors show that remembered - but not actively viewed - orientations form "line-like" patterns in retinotopic cortical space.

Weaknesses:
1. The paper tests a strong version of the sensory recruitment model, where neural states representing information during WM are presumed to be identical to neural states representing the same information during perceptual processing. As the paper acknowledges, there is already ample reason to doubt this prediction (see, e.g., earlier work by Kok & de Lange, Curr Biol 2014; Bloem et al., Psych Sci, 2018; Rademaker et al., Nat Neurosci, 2019; among others). Still, the demonstration that orientation decoding during WM is immune to aperture biases known to drive orientation decoding during perception makes for a compelling demonstration.

2. Earlier work by the same group has reported line-like representations of orientations during memory storage but not during perception (e.g., Kwak & Curtis, Neuron, 2022). It's nice to see that result replicated during explicit perceptual and WM tasks in the current study, but I question whether the findings provide fundamental new insights into the neural bases of WM. That would require a model or explanation describing how stimulus-specific activation patterns measured during perception are transformed into the "line-like" patterns seen during WM, which the authors acknowledge is an important goal for future research.

https://doi.org/10.7554/eLife.94191.1.sa1

Reviewer #3 (Public Review):

Summary:
In this work, Duan and Curtis addressed an important issue related to the nature of working memory representations. This work is motivated by findings illustrating that orientation decoding performance for perceptual representations can be biased by the stimulus aperture (modulator). Here, the authors examined whether the decoding performance for working memory representations is similarly influenced by these aperture biases. The results provide convincing evidence that working memory representations have a different representational structure, as the decoding performance was not influenced by the type of stimulus aperture.

Strengths:
The strength of this work lies in the direct comparison of decoding performance for perceptual representations with working memory representations. The authors take a well-motivated approach and illustrate that perceptual and working memory representations do not share a similar representational structure. The authors test a clear question, with a rigorous approach and provide convincing evidence. First, the presented oriented stimuli are carefully manipulated to create orthogonal biases introduced by the stimulus aperture (radial or angular modulator), regardless of the stimulus carrier orientation. Second, the authors implement advanced methods to decode the orientation information present, in visual and parietal cortical regions, when directly perceiving or holding an oriented stimulus in memory. The data illustrates that working memory decoding is not influenced by the type of aperture, while this is the case in perception. In sum, the main claims are important and shed light on the nature of working memory representations.

Weaknesses:
I have a few minor concerns that, although they don't affect the main conclusion of the paper, should still be addressed.

1. Theoretical framing in the introduction: Recent work has shown that decoding of orientation during perception does reflect orientation selectivity, and it is not only driven by the stimulus aperture (Roth, Kay & Merriam, 2022).

2. Figure 1C illustrates the principle of how the radial and angular modulators bias the contrast energy extracted by the V1 model, which in turn would influence orientation decoding. It would be informative if the carrier orientations used in the experiment were shown in this figure, or at a minimum it would be mentioned in the legend that the experiment used 3 carrier orientations (15{degree sign}, 75{degree sign}, 135{degree sign}) clockwise from vertical. Related, when trying to find more information regarding the carrier orientation, the 'Stimuli' section of the Methods incorrectly mentions that 180 orientations are used as the carrier orientation.

3. The description of the image computable V1 model in the Methods is incomplete, and at times inaccurate. i) The model implements 6 orientation channels, which is inaccurately

referred to as a bandwidth of 60{degree sign} (should be 180/6=30). ii) The steerable pyramid combines information across phase pairs to obtain a measure of contrast energy for a given stimulus.
Here, it is only mentioned that the model contains different orientation and spatial scale channels. I assume there were also 2 phase pairs, and they were combined in some manner (squared and summed to create contrast energy). Currently, it is unclear what the model output represents. iii) The spatial scale channel with the maximal response differences between the 2 modulators was chosen as the final model output. What spatial frequency does this channel refer to, and how does this spatial frequency relate to the stimulus?

4. It is not clear from the Methods how the difficulty in the perceptual control task was controlled. How were the levels of task difficulty created?

https://doi.org/10.7554/eLife.94191.1.sa0