A gradual transition from veridical to categorical representations along the visual hierarchy during working memory, but not perception.

Chaipat Chunharas [1], Meike D. Hettwer [2,3], Michael J. Wolff [4], & Rosanne L. Rademaker [4]

[1] Department of Medicine, King Chulalongkorn Memorial Hospital, Chulalongkorn University, Bangkok, Thailand

[2] Max Planck School of Cognition, Max Planck Institute of Human Cognitive and Brain Sciences, Leipzig, Germany

[3] Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Germany

[4] Ernst Strüngmann Institute (ESI) for Neuroscience in Cooperation with the Max Planck Society, Frankfurt, Germany

Abbreviated title: Representational formats of visual perception & working memory

Corresponding authors: chaipat.c@chula.ac.th & rosanne.rademaker@gmail.com

Number of figures: 5 main & 4 supplementary

Conflict of interest: The authors declare no conflict of interest

Acknowledgements: Data acquisition was financed by NEI R01-EY025872 and NIMH R01-MH087214 to John Serences – whose unwavering and continued support has shaped us as scientists. RLR and MDH are funded by the Max Planck Society, and MDH is also funded by the German Federal Ministry of Education and Research (BMBF). We thank the IMP lab at Goethe University & all members of the Rademaker lab for their feedback on the manuscript. We also thank Xue-Xin Wei and Tim Brady for seeding the thoughts that changed our minds, culminating to the modeling work in this paper.

Summary

The ability to stably maintain visual information over brief delays is central to cognitive functioning. One possible way to achieve robust working memory maintenance is by having multiple concurrent mnemonic representations across multiple cortical loci. For example, early visual cortex might contribute to storage by representing information in a “sensory-like” format, while intraparietal sulcus uses a format transformed away from sensory driven responses. As an explicit test of mnemonic code transformations along the visual hierarchy, we quantitatively modeled the progression of veridical-to-categorical orientation representations in human participants. Participants directly viewed, or held in mind, an oriented grating pattern, and the similarity between fMRI activation patterns for different orientations was calculated throughout retinotopic cortex. During direct perception, similarity was clustered around cardinal orientations, while during working memory the obliques were represented more similarly. We modeled these similarity patterns based on the known distribution of orientation information in the natural world: The “veridical” model uses an efficient coding framework to capture hypothesized representations during visual perception. The “categorical” model assumes that different “psychological distances” between orientations result in orientation categorization relative to cardinal axes. During direct perception, the veridical model explained the data well in early visual areas, while the categorical model did worse. During working memory, the veridical model only explained some of the data, while the categorical model gradually gained explanatory power for increasingly anterior retinotopic regions. These findings suggest that directly viewed images are represented veridically, but once visual information is no longer tethered to the sensory world, there is a gradual progression to more categorical mnemonic formats along the visual hierarchy.

Keywords

visual perception; visual working memory; sensory recruitment; visual cortex; parietal cortex; orientation; representational similarity; representational geometry; efficient coding; RSA

Introduction

Holding images in mind over a brief delay is central to cognition, as it allows for the retention and manipulation of information that cannot be viewed directly. Visual working memory (VWM) recruits early visual cortex, including primary visual area V1 – as indexed by response patterns recorded with fMRI (Harrison & Tong, 2009; Serences et al., 2009; Christophel, Hebart, & Haynes, 2012; Riggall & Postle, 2012; Ester, Sprague, & Serences, 2015; Bettencourt & Xu, 2016; Lorenc et al., 2018; Christophel et al., 2018). Being the first cortical processing site of visual inputs, the role of V1 during perception is fundamentally different from its role during visual working memory. This is because in the absence of direct visual input, mnemonic information in V1 and other early visual areas must necessarily be generated internally. Famously, "sensory recruitment theory" posits that higher-order frontal and parietal regions of the brain that are active throughout the working memory delay (Fuster & Alexander, 1971; Funahashi, Bruce, & Goldman-Rakic, 1989; Funahashi, Chafee, & Goldman-Rakic, 1993; Wilson, Scalaidhe, & Goldman-Rakic, 1993; McCarthy et al., 1994; Friedman & Goldman-Rakic, 1994; Goldman-Rakic, 1995; McCarthy et al., 1996; Miller, Erickson, & Desimone, 1996; Chafee & Goldman-Rakic, 1998; Courtney, et al., 1998; Qi et al., 2015), recruit early sensory areas in a top-down manner in order to maintain high fidelity sensory memories (Serences, 2016; Gayet, Paffen, & Van der Stigchel, 2018). Alternatively, recurrent processes in local circuits could sustain information over a memory delay (Compte et al., 2000; Wimmer et al., 2014), but such recurrency is deemed more likely in anterior brain areas (Meijas & Wang, 2022). Irrespective of the exact substrate of working memory maintenance – with inputs from the external sensory world during perception, and from sources within the brain during working memory, it is unlikely that viewed and remembered visual information would be represented in an identical manner in early visual cortex. However, it remains an open question how a cortical area like V1, specialized for processing visual inputs, actually represents visual working memories. Are working memory representations just noisier versions of perceptual representations, or do they differ in a fundamental way? And might there be representational transformations along the visual hierarchy as top-down influences play an increasingly larger role?

Recent work has claimed that early visual cortex (EVC) represents VWM information in a "sensory-like" format that is actually quite similar to representations driven by sensory inputs, while more anterior visual areas like the Intraparietal Sulcus (IPS) were said to represent VWM information in a format that is transformed away from the sensory driven response (Rademaker et al., 2019). This claim of "sensory-likeness" comes from the fact that when participants remember an orientation, response patterns in EVC are similar to response patterns during the direct perception of that orientation. In parietal cortex such cross-generalization from perception to working memory responses fails, despite memories still being decodable when considering only the response patterns during working memory themselves (i.e., without cross-generalization). The idea that visual representations are transformed away from the "sensory-like" into a different, more abstract format is further supported by work similarly using cross-generalization to decode VWM contents (Kwak & Curtis, 2022). In this study, participants remembered one of two visually distinct features – the orientation of a grating, or the direction of a moving dot cloud. During encoding (i.e., perception), the two features evoked distinct response patterns in EVC that did not cross-generalize, likely owing to the distinct retinal inputs evoked by the two features. However, these two features share a spatial component (degrees on a circle), and during the memory delay the EVC response patterns for orientation and direction stimuli did successfully cross-generalize, likely owing to a "line-like" abstraction that is realizable for both features.

However, the possible range of to-be-remembered visual features far outstrips those that can be mapped onto a circle (and into a line). For example, surface features such as contrast or color are not readily abstracted into a "line-like" representation, let alone more complex features such as shapes or objects. This means that a more general principle of abstraction remains to be uncovered. Furthermore, we know that retinotopically organized "sensory-like" representations or abstractions are not generally observed outside of early visual cortex (Rademaker et al., 2019; Kwak & Curtis, 2022; Favila, Kuhl, & Winawer, 2022; Vo et al., 2022). An important step to investigate possible abstractions used for working memory is to consider not only *response patterns*, but also the *representational geometry*. A *response pattern* (also called a "coding scheme", Stokes, 2015) simply refers to the pattern of responses that is measured during an experimental condition. Depending on the measurement technique, this could be a pattern of firing rates across multiple neurons, a pattern of BOLD responses across multiple voxels, or any other kind of response vector measured over a number of units. When it's possible to cross-generalize from one experimental condition to another – for example from perception to memory, or from orientation to direction – we know that the *response pattern* is similar in the two conditions. For example, the specific pattern of brain activity in response to a perceived stimulus with an orientation of 90° would be similar to the pattern measured when that same 90° stimulus is held in working memory. The *representational geometry* captures a lower-dimensional format of a given stimulus set, and can be invariant to changes in the underlying response patterns (Kriegeskorte & Wei, 2021). The pairwise distances between patterns of responses corresponding to a set of stimuli determine the geometry, meaning that even if the underlying response patterns change (e.g., they are inverted, shifted, or undergo some other transformation), the geometry can remain stable. For example, in the case of orientation the geometry may reveal that adjacent orientations (say 90° and 91°) evoke similar underlying response patterns, and that this similarity drops at increasing distances in orientation space. Such representational geometry can be shared between perception and working memory, while a perceived stimulus of 90° may nevertheless evoke a totally different response pattern compared to a 90° stimulus held in working memory. As long as the pattern distances between different orientations are the same during perception and memory, so is the geometry. Indeed, we know that despite dynamic population responses over time, the representational format of a stimulus set can remain remarkably stable (Murray et al., 2016; Spaak, Watanabe, Funahashi, & Stokes, 2017).

By looking at the representational geometry we can investigate the representational formats of perception and working memory in a way that does not depend on response patterns generalizing from one condition to another. Moreover, it allows us to investigate potential systematic differences in the representational geometry between perception and working memory, even when underlying response patterns are still similar enough for successful cross-generalization. To illustrate, what if abstraction happens by compressing part of the stimulus space, making it more categorical? Such compression may warp response patterns for a subset of orientations, without necessarily transforming them to a point where the coding scheme breaks down. In such a case, cross-generalization from perception to working memory could coexist with a change in the representational geometry. Thus, examining the representational geometry allows us more freedom to see if perception and working memory are truly represented similarly, or if the representational formats perhaps differ in fundamental ways.

Here, we introduce a principled approach to investigating sensory-to-memory abstractions across the visual hierarchy. First, we use representational similarity analysis (RSA) to show a clear differentiation between the geometry of perception and working memory representations for orientation in human visual cortex. Second, we model the extent to which the representational geometry is true to a simulated sensory response (the “veridical” model), or abstracted away from the sensory input (the “categorical” model). Our two models are constrained by a single principle, namely, the distribution of orientation information in the natural world (Girshick, Landy, & Simoncelli, 2011; Wei & Stocker, 2015). By applying these two models, we show that sensory inputs are represented in a largely veridical manner in EVC, adapted to the statistics of the natural visual world (i.e., efficient coding). By contrast, working memories are represented more categorically, in a manner predicted from the same landscape of visual input statistics, but using a higher-order metric based on how different any set of orientations may appear to the observer (i.e., their “psychological distance”). Critically, we show that during working memory the representational format becomes increasingly more categorical along the visual hierarchy, uncovering a gradient of abstraction.

Results

To examine neural representations during perception and working memory, we analyze fMRI recordings from six participants who were either directly viewing oriented grating stimuli during a sensory task, or remembering orientations for later recall during a memory task (Figure 1A). We use cross-validated representational similarity analysis (RSA) – a method that projects neural activation patterns into an abstract space that describes the stimulus (here: orientation) (Kriegeskorte, Mur, & Bandettini, 2008a; Kriegeskorte et al., 2008b; Walther et al., 2016; Kriegeskorte & Wei, 2021). Specifically, we created, for each visual cortical Region of Interest (ROI) and each of the two tasks, a matrix capturing the degree of similarity between the neural response patterns to all possible orientations using cross-validated correlations (Figure 1B). To illustrate: When two orientations are represented in a very similar manner, the pattern of voxel responses to the first orientation will correlate strongly with the pattern evoked by the second orientation. Conversely, for orientations with very distinct representations, correlations will be low. Because orientation space is continuous and circular, physically similar orientations (e.g., 10° and 11°) will likely correlate more strongly than physically dissimilar orientations (e.g., 10° and 40°) in areas of the brain that care about orientation.

The representational similarity matrices (RSM’s) constructed for our visual ROI’s (Figure 1C; Supplementary Figure 1) show striking qualitative differences between how orientation is represented during perception (in the sensory task), and working memory (in the memory task). During perception, early visual areas V1–V3 show a strong diagonal component and high degree of similarity around cardinal orientations (180° in particular), with notable transformations away from this representational pattern primarily along the dorsal stream (V3AB and IPS). During working memory, there is a prominent clustering of representational similarity around oblique orientations (45° and 135°). This clustering seems to increase along the visual hierarchy and appears most pronounced in area IPS0.

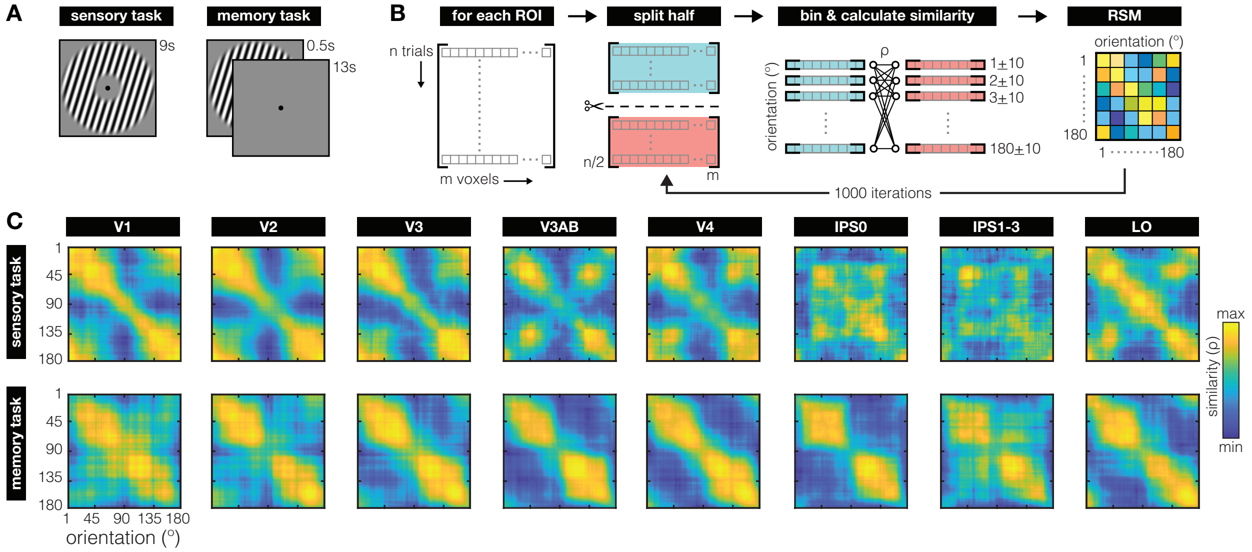


Figure 1: Task and main analysis (A) For the sensory task (left), participants viewed a randomly oriented grating for 9 seconds per trial. Grating contrast was phase-reversed at 5 Hz, with participants reporting brief (200ms) probabilistic instances of contrast dimming. For the working memory task (right), participants remembered a briefly presented (500 ms) randomly orientated grating for 13 seconds, until a 3 second recall epoch (not depicted). On two-thirds of trials a visual distractor (filtered noise or another grating) was presented during the delay (11 seconds, not depicted). Because there was little-to-no quantifiable difference between the different trial types (Rademaker et al., 2019; Supplementary Figure 2), here we analyzed data from all delays combined. For the sensory and memory task, we analyzed average voxel responses from 4.8–9.6 and 5.9–13.9 seconds after stimulus onset, respectively. **(B)** For each Region of Interest (ROI) we employed a split-half randomization procedure to create a Representational Similarity Matrix (RSM) for each participant. On each randomization fold, voxel patterns from all trials (300–340 for sensory, 324 for memory) were randomly split in half. For each half of trials, we averaged the voxel patterns for every degree in orientation space within a ± 10° window. This resulted in 180 vectors with a length equal to the number of voxels for each split of the data. We then calculated the similarity between each vector (or degree) in one half of the data, to all vectors (or degrees) in the second half of the data, using a Spearman correlation coefficient. This resulted in a 180x180 similarity matrix on each fold. This randomization procedure was repeated 1000 times to generate the final RSM for each ROI and each participant. Across all folds, RSM's are near-symmetrical around the diagonal, give-or-take some cross-validation noise. **(C)** Orientation representation (as indexed by RSM's) during sensory perception (top row) and working memory (bottom row), for retinotopically defined ROI's (columns) across all participants. During perception, the clear diagonal pattern in early visual areas V1–V3 indicates that orientations adjacent in orientation space are represented more similarly than orientations further away. Orientations around cardinal (180° in particular) are represented more similarly. During working memory, similarity clusters strongly around oblique orientations (45° and 135°), contrasting starkly with the similarity patterns during perception. Note that the diagonal represents an inherent noise-ceiling, thanks to the cross-validation procedure used. This noise ceiling shows inhomogeneities across orientation space, demonstrating how certain orientations may be encoded with more noise than others. RSM's are scaled to the range of correlations within each subplot to ease visual comparison of representational structure between sensory and memory tasks for all ROI's (for exact ranges, see Supplementary Figure 3). Throughout, we use 0° (and 180°) to denote vertical orientations, and 90° to denote horizontal ones.

To quantify the results in Figure 1C, we contrast two possible models – the “veridical” and the “categorical” model. The veridical model intends to capture how orientations are represented in a manner that faithfully reflects early visual processing of signals from the physical world. The categorical model uses a higher-level concept – the “psychological distance” between orientations – as a basis for abstracting a physically continuous space into discrete categories. Importantly, these two models are jointly constrained by the known distribution of orientation information in the natural world, which has a higher occurrence of cardinal compared to oblique orientations (Figure 2A; Girschick, Landy, & Simoncelli, 2011). According to the “efficient coding” hypothesis, this inhomogeneity in orientation input statistics leads to adaptive changes in the sensory system, with relatively more neural resources dedicated to cardinal compared to oblique orientations (Attneave, 1954; Shannon, 1948; Barlow, 1961; Olshausen & Field, 1997; Simoncelli & Olshausen, 2001). As a result, observers demonstrate a higher resolution (e.g., improved discriminability) around cardinal orientations – a phenomenon paradoxically known as the “oblique effect” (Lennie, 1971; Appelle, 1972; Essock, 1980). Orientation reports also tend to be biased away from cardinal axes (van Bergen et al., 2015; Wei & Stocker, 2015; Henderson & Serences, 2021). Together, this suggests a distinct role for cardinal orientations, both with respect to resolution and bias.

The idea behind the veridical model is that region-wide orientation representations emerge from low-level neural responses to sensory inputs. Specifically, the starting point for this model is a set of idealized tuning functions that tile orientation-space (Figure 2B, top). The amplitude of each orientation tuning function is scaled by the estimated frequency of occurrence for that orientation in the natural world (i.e., by the theoretical “input statistics” function, see Figure 2A and Materials & Methods). Therefore, tuning functions closer to cardinal orientations have relatively higher amplitudes than those closer to obliques. We modulate tuning curve amplitude (and not some other property such as tuning width or density) because of known amplitude differences in the fMRI signal for cardinals compared to obliques (Furmanski & Engel, 2000). For any given stimulus orientation, we can simulate a vector of neural responses by reading out the hypothesized activity from every idealized neural tuning function. Such a simulated response vector can be correlated against simulated responses to all other possible stimulus orientations (analogous to the approach in Figure 1B), to arrive at the veridical model RSM (Figure 2B, bottom). Thus, here we model a veridical early visual representation by accounting for known inhomogeneities of orientation space, based on the principle of efficient coding (Wei & Stocker, 2015). Note that our veridical model is a direct consequence of the choice to use amplitude modulation (instead of e.g., tuning width) as well as correlation (instead of e.g., Euclidian distance), and that there are multiple other possible ways to implement inhomogeneities across orientation space (Wei & Stocker, 2015; Kriegeskorte & Wei, 2021; Harrison, Bays, & Rideaux, 2023).

The idea behind the categorical model is to discretize the physically continuous orientation space into a plausible higher-level psychological abstraction. A principled way to categorize orientation is to again rely on the input statistics function (Figure 2A), and consider how inhomogeneities in orientation space might affect more experiential measures such as perceptual similarity (Schurgin, Wixted, & Brady, 2020). For example, in parts of the orientation space with high resolution (near the cardinals), two physically similar orientations (e.g., 4° apart) can look clearly different from one another, while orientations with the same physical similarity in parts of orientation space with lower resolution (near the obliques) can look indistinguishable. We formalize this "psychological distance" as the distance between any pair of orientations along the theoretical input statistics function (Figure 2C, top). Returning to our example, two near-cardinal orientations (e.g., 88° and 92°) will have a larger psychological distance (i.e., are relatively far apart along this function) compared to two near-oblique orientations (e.g., 43° and 47°) (compare the Figure 2C grating inserts in blue versus red, respectively). The psychological distances between all possible orientations make up the categorical model RSM (Figure 2C, bottom). This RSM shows how orientations in one category (bound by the obliques) are represented similarly to one another, but dissimilarly from orientations in a second category (on the other side of the obliques). Thus, here we model how orientation representations can be categorized based on where an orientation is relative to cardinal – the cardinal axes effectively serving as category boundaries.

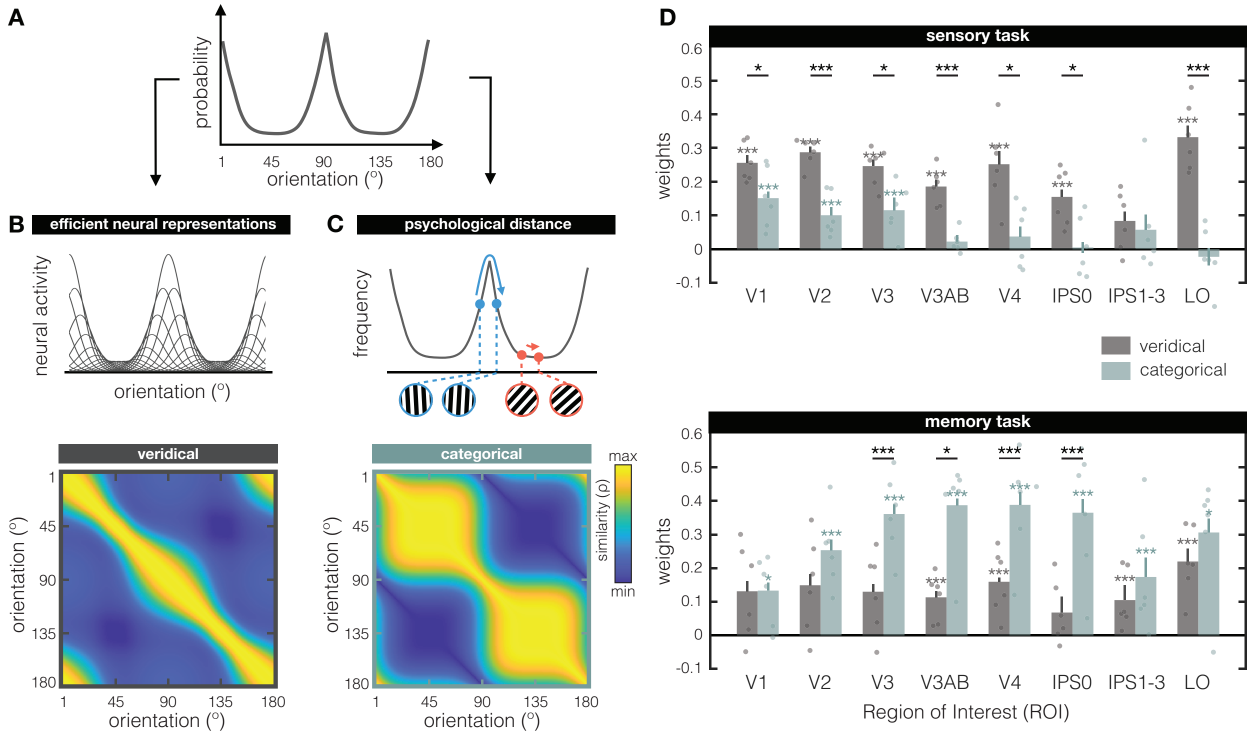


Figure 2: Modeling the representational similarity of perceived and remembered orientations (A) The distribution of visual orientation in the natural world is inhomogeneous, with higher prevalence of orientations closer to cardinal (90° & 180°) compared to oblique (45° & 135°). The function shown here approximates these input statistics, and is used to constrain both the veridical (in **B**) and categorical (in **C**) models. **(B)** The veridical model is based on the principle of efficient coding – the idea that neural resources are adapted to the statistics of the environment. We model this via 180 idealized orientation tuning functions with amplitudes scaled by the theoretical input statistics function (the top panel shows a subset of tuning functions for illustrational purposes). A vector of neural responses is simulated by computing the activity of all 180 orientation-tuned neurons to a given stimulus orientation. Representational similarity is calculated by correlating simulated neural responses to all possible orientations, resulting in the veridical model RSM (bottom panel). Note that while we chose to modulate tuning curve amplitude, there are multiple ways to warp the stimulus space (e.g., by applying non-uniform changes to gain, tuning width, tuning preference, etc. Kriegeskorte &

Wei, 2021; Harrison et al., 2023). **(C)** In the categorical model, categorization is based on people's subjective experience of relative similarity between orientations in different parts of orientation space: If orientations in part of the space appear quite similar, they are lumped together into the same category, while distinctive looking orientations serve as category boundaries. This is quantified via the "psychological distance" – the sum of derivatives along the input statistics function between any pair of orientations (see top panel). The insert shows an example of orientation-pairs near cardinal (in blue) and oblique (in red) that have the same physical distance, but different psychological distances. The psychological distance between each possible pair of orientations yields the categorical model's RSM (bottom panel). **(D)** Fits of the veridical (grey) and categorical (teal) models for the sensory (top) and memory (bottom) tasks. During perception, the veridical model explains a significant amount of the data in all visual ROI's (except IPS1–3), indicating a representational scheme that is largely in line with modeled early sensory responses. The categorical model explains some of the data in areas V1–V3, but to a lesser extent. During working memory, the categorical model gains increasingly more explanatory power over the veridical model along the visual hierarchy, from V1 through IPS0. Weights (on the y-axis) represent the unique contribution of each model after removing the variance explained by the other model. Dots represent individual participants, and error bars represent ± 1 within-participant SEM. Asterisks indicate the significance level of non-parametric two-sided post-hoc comparisons (*$p \leq 0.05$; ***$p \leq 0.001$), with gray and teal asterisks indicating tests against zero, and black asterisks indicating paired-sample tests comparing the two models in each ROI.

How well can the representational geometry during the sensory and memory tasks (Figure 1C) be explained by our veridical (Figure 2B) and categorical (Figure 2C) models? Because our two models are not independent, we evaluated the fit of each model after first removing the variance explained by the other. Specifically, to look at the unique contribution of the veridical model, we fit the veridical model to the residuals left by the categorical model, and vice versa (minimizing the mean squared error; see also Materials & Methods). Model fits are shown in Figure 2D for our sensory (top) and memory (bottom) tasks.

During sensory perception, the veridical model does an overall better job at explaining representational similarity than the categorical model (main effect of model: $F_{(1,5)} = 40.14$, $p = 0.00145$). This advantage is not the same in all ROI's (interaction of model x ROI: $F_{(7,35)} = 3.25$, $p = 0.00917$), with post-hoc tests showing no difference between the two models in IPS1–3 (where neither model explains the data above chance level). The fact that the veridical model outperforms the categorical model during sensory perception helps validate our modeling approach, given that the veridical model is founded on what we know about early sensory processing. This finding partially generalizes to a situation in which oriented gratings are directly viewed but ignored by participants (see Supplementary Figure 4).

During working memory, the categorical model better explains representational similarity than the veridical model (main effect of model: $F_{(1,5)} = 12.43$, $p = 0.0168$). The extent to which the categorical model outperforms the veridical model differs across ROI's (interaction of model x ROI: $F_{(7,35)} = 2.785$, $p = 0.0206$), with the categorical model explaining increasingly more of the representational geometry along the visual hierarchy. A significant difference between the two models emerges in area V3, and persists until area IPS0 – the latter being particularly noteworthy for not showing any sign of a veridical model representation. These results corroborate the qualitative categorization already apparent from the memory task RSM's (Figure 1C), and reveal a posterior-to-anterior gradient in terms of categorization strength.

To verify that our modeling results do not critically depend on the exact shape of the theoretical input statistics function in Figure 2A, we next used behavioral response frequencies from an independent psychophysical experiment to constrain both our models (Figure 3A). A new set of 17 participants each completed 1620 trials of an orientation recall task. For every possible stimulus orientation that was shown (1°–180° in steps of 1°), we calculate the probability of every possible response across all participants. With 27540 total trials in the experiment, this means response probability histograms are based on 153 trials for every possible stimulus orientation. As expected, response probability functions look inhomogeneous along orientation space, with pronounced differences between cardinals and obliques (Figure 3A). We then generated the veridical and categorical models anew from this psychophysical input function (Figure 3B), and again fit both models to our data RSM's to see how well they explained the data (as indexed by the model weights in Figure 3C). We replicated the difference between the sensory and memory tasks, and how their representational geometries are better explained by the veridical and categorical models, respectively. During sensory perception, the veridical model outperformed the categorical model (main effect of model, $F_{(1,5)} = 7.05$, $p = 0.0454$), although this advantage did not differ significantly between ROI's (no interaction of model x ROI $F_{(7,35)} = 2.11$, $p = 0.0683$). During working memory, the categorical model outperformed the veridical model (main effect of model: $F_{(1,5)} = 28.57$, $p = 0.0031$), explaining increasingly more of the representational geometry along the visual hierarchy (interaction of model x ROI: $F_{(7,35)} = 3.63$, $p = 0.0048$).

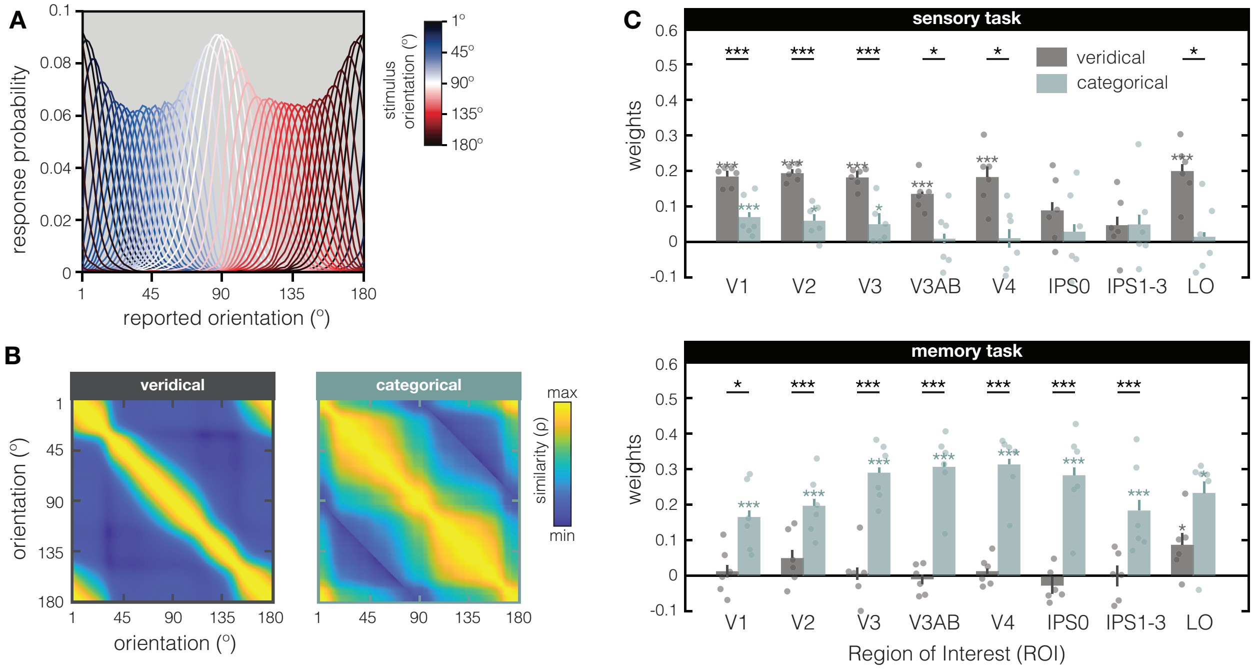


Figure 3: Generating and fitting the veridical and categorical models based on independent behavioral data (A) During an independent psychophysical examination, a new set of participants (N=17) reported the orientation of briefly presented (200ms) and remembered (2s) single gratings by rotating a response dial with a computer mouse (i.e., via method-of-adjustment). For each possible stimulus orientation in the experiment, we calculated the response probability of report. For example, for a grating stimulus of 22° the most probable response might be 22°, but a response of 21° or 23° is also very probable – with response probability tapering off as the distance from the stimulus orientation increases. Here, response probability (y-axis) is plotted against reported orientation for a subset of possible stimulus orientations shown in red-blue colors (see insert). From this psychophysical input function, the veridical and categorical models were generated as described in Figure 2B–C. **(B)** Veridical and categorical models generated from the psychophysical

input function (in **A**). **(C)** Fits of the veridical (in grey) and categorical (in teal) models based on the independent psychophysical data (i.e., the models shown in **A**). During perception (top), the veridical model explains a significant amount of the data in all visual ROI's except for IPS regions. The categorical model explains some of the data in areas V1–V3, but to a lesser extent. During working memory (bottom), the categorical model explains a significant amount of the data in all visual ROI's, while the veridical model fails in all ROI's except Lateral Occipital (LO) cortex. Weights (on the y-axis) represent the unique contribution of each model after removing the variance explained by the other model. Dots represent individual participants, and error bars represent ± 1 within-participant SEM. Asterisks indicate the significance level of non-parametric two-sided post-hoc comparisons (*$p \leq 0.05$; ***$p \leq 0.001$), with gray and teal asterisks indicating tests against zero, and black asterisks indicating paired-sample tests comparing the two models in each ROI.

How might we reconcile the observed *differences* in representational geometry between the sensory and memory tasks, with the *overlap* in coding schemes that is implied by the ability to cross-generalize from the sensory to the memory task in EVC (Rademaker et al., 2019)? To directly relate these two analysis approaches, we applied a minor modification to the typical RSA approach by correlating response patterns from every *perceived* orientation in the sensory task to the response patterns from every *remembered* orientation in the memory task, in what we call "across-task RSA" (Figure 4A). As expected, our across-task RSM of V1 (but not IPS) shows a clear diagonal component that is indicative of overlap in response patterns between perception and working memory, and the ability to cross-generalize. Importantly, this approach also shows how the coding scheme for orientation during working memory is warped with respect to coding scheme during perception – response patters for orientations held in working memory are biased towards what would be patterns associated with obliques during perception (Figure 4B). We validate our "across-task RSA" approach against a conventional multivariate analysis approach known as the inverted encoding model (or "IEM", Brouwer & Heeger, 2009). First, we take into account the predicted gradual drop in representational similarity for orientations at increasing distances from the remembered orientation by creating a "correlation profile" (the sum of correlations between patterns for the remembered and perceived orientations, as shown in grey on top of the panels in Figure 4B, and for all ROI's in Figure 4C). A more "peaked" correlation profile indicates more information about the remembered orientation. Next, we quantify this alternative measure of cross-generalization, the "peakeness" of the correlation profile, with a previously established fidelity metric (as in Rademaker et al., 2019, see also the Materials & Methods), and show that this aligns closely with the same metric applied to results from an IEM using the same data (Figure 4D; Rademaker et al., 2019).

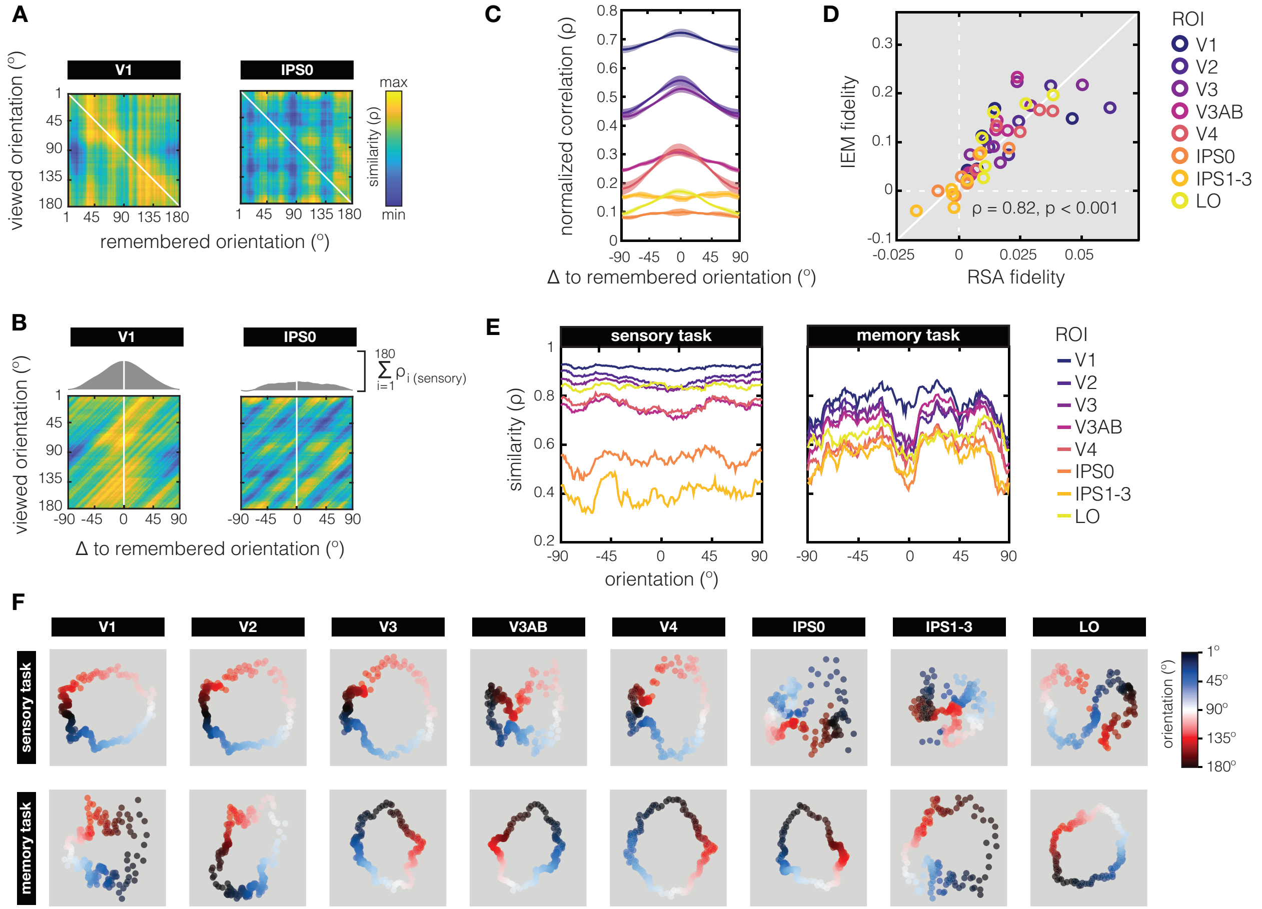


Figure 4: Ability to cross-decode using RSA, and clues on how categorization comes about (A) Using across-task representational similarity analysis, we directly compare orientation response patterns recorded during the sensory task (y-axis), to those measured during the working memory task (x-axis). Here we use V1 (left subplot) and IPS0 (right subplot) as example ROI's. The across-task RSM in V1 shows a clear diagonal component, indicating that response patters for specific orientations in the sensory task are similar (i.e., correlated) to response patterns to those same orientations when they are held in working memory. In IPS0 such pattern similarity for matching orientations in the sensory and memory tasks is less evident. **(B)** We want to quantify the extent to which orientations held in working memory evoke response patterns that overlap most strongly with response patterns from those same orientations when perceived, and the degree to which this similarity drops as a function of distance in orientation space. First, we center our across-task RSM's on the remembered orientation (notice the different x-axis), and then take the sum of correlations relative to the remembered orientation (plotted on top of the across-task RMS's). We refer to this sum of correlations as the "correlation profile" of the remembered orientation. In V1 we see that correlations are highest between response patterns from matching perceived and remembered orientations (indexed by 0° on the x-axis), which explains the ability to cross-decode between sensory and memory tasks as demonstrated in previous work (e.g., Rademaker et al., 2019). By contrast, IPS0 shows a much flatter correlation profile compared to V1. **(C)** Correlation profiles for all retinotopic ROI's in our study, obtained by performing across-task RSA (left panel). Most ROI's show a peaked correlation profile, indicative of shared pattern similarity between the same orientations when perceived and when remembered. The different offsets along the y-axis for different ROI's reflect the overall differences in pattern similarity in different areas of the brain, with pattern similarity being highest in area V1. Shaded areas indicate ± 1 SEM **(D)** To further validate the ability to cross-decode using RSA, we directly compare this new approach (x-axis) to the multivariate analysis performed by Rademaker et al. in 2019 (y-axis). The latter used an inverted encoding model (IEM) that was trained on the sensory task, and tested on the delay period of the memory task. Both the correlation profiles from RSA, and the channel response functions from IEM yield more-or-less peaked functions over orientation space (relative to the remembered orientation) that can be quantified using a fidelity metric (i.e., by convolving with a cosine). Here, we show a high degree of consistency between the fidelity metrics derived with both approaches (dots are near the diagonal), and successful cross-generalization from the sensory to the memory task (as indexed by >0 fidelities in many ROI's). Each color represents a different ROI, and for each ROI we

plot each of the six participants as an individual dot. **(E)** To examine the inhomogeneity or representational similarity throughout orientation space, we plot the diagonals of the RSM's from in Figure 1C. During the memory task, oblique orientations are represented much more similar, and cardinals much less similar, compared the sensory task. This could be a mechanism via which categorization of orientations occurs during working memory. **(F)** We use multidimensional scaling (MDS) to projects high dimensional response patterns into 2 dimensions, in order to better visualize of how orientation space is represented. During perception there is an orderly geometrical progression of orientation space, with the highest similarity between adjacent orientations (and some clustering around cardinal orientations, especially 180º, shown in white and black) in early visual areas V1–V3. There's also a "pinching" of orientation space around the obliques (45º and 135º become very similar) in more anterior visual areas V3AB–IPS and LO. During working memory, the orientation space geometry remains circular in all ROI's, with notable clustering of similarity around the obliques (shown in blue and red).

Might the warping we observe in area V1 during working memory – with biases away from cardinals and towards the obliques (Figure 4B, left panel) – be a possible mechanism through which categorization is realized? If yes, one prediction is that such a compression of parts of the orientation space will make the response patterns around oblique orientations more similar, while patterns around cardinals will become less similar. Indeed, when we look at the representational similarity along orientation space (i.e., the diagonals of our cross-validated within-task RSM's in Figure 1C), we see systematic changes indicating higher similarity for oblique compared to cardinal orientations (Figure 4E). Importantly, this inhomogeneity of pattern similarity across orientation space is much exaggerated during working memory compared to perception, implying a non-linear compression relative to cardinal orientations, that may contribute to the categorization of working memory contents. Multidimensional scaling (Figure 4F) further supports this idea, as it shows stronger clustering around obliques (in red and blue) compared to cardinals (in white and black) during working memory compared to perception.

Thus far, we examined the structure of orientation representations during perception and working memory in individual visual ROI's, and find that orientation representations differ between sensory and memory tasks. We also find that representational geometry *within the same task* is captured by our models to varying extents in different ROI's. To move beyond information in local ROI's (e.g., Figure 1C; Supplementary Figure 1), and more formally assess representational geometries across visual cortex, we use a 2nd level RSA. In this analysis, similarity between different visual cortical areas is calculated by correlating the RSM from every ROI with that of every other ROI. To this end, we use more fine-grained ROI's than in previous analyses, allowing us to look at dorsal versus ventral streams, as well as subregions of IPS and Lateral Occipital (LO) cortex. We evaluate how orientation information is represented across visual cortex in this manner separately for the sensory task and the memory task (Figure 5).

During sensory perception, there is notable shared representational similarity amongst early visual areas (V1–V4) and amongst areas in the intraparietal sulcus (IPS0–3), but low similarity between the two (Figure 5A, top). During the working memory task, orientation geometries across various ROI's show a somewhat different inter-areal organization (Figure 5A, bottom). First, the overall similarity between ROI's appears more pronounced during the working memory task, with higher overall correlations between ROI's compared the sensory perception task (as also already implicated in Figure 4E). This implies that there are fewer transformations of orientation geometry along the visual hierarchy during memory compared to perception. Second, the cluster of early visual ROI's with high representational similarity that was observed during perception (i.e., V1–V4), is shifted "upwards" along the visual hierarchy during memory – with V1 becoming less similar, and IPS becoming more similar to rest of EVC.

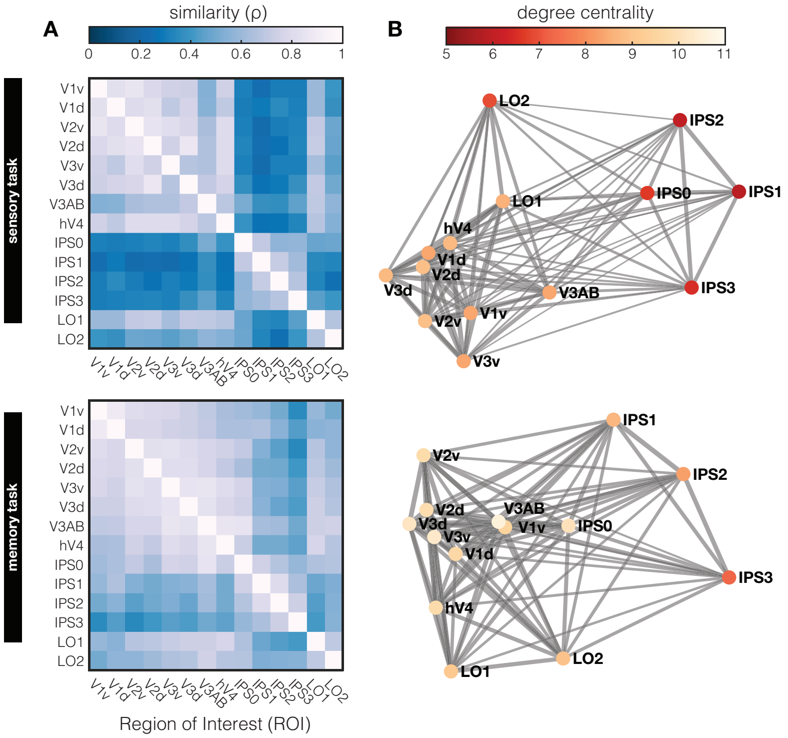


Figure 5: Second level RSA (A) To compare how orientation is represented across different regions of visual cortex, RSM's from fine-grained individual ROI's (Supplementary Figure 1) were correlated in a cross validated 2nd-level similarity analysis. For the sensory task (top panel), we see that representational similarity is high among early visual areas; high among the various IPS regions; and high among LO regions. However, similarity between these three clusters is relatively low. During working memory (bottom panel) there is a slight shift in similarity compared to perception, with V1 becoming less similar, and IPS0 becoming more similar, to areas V2–V4. The distinction between areas is generally less pronounced. **(B)** Representational similarity can also be used as an indicator of connectivity between ROI's based on shared representational geometry: When the geometry is similar, the "connection" is stronger (indicated here by the width of the grey lines connecting different ROI's). The sum of the strength of these connections in a given ROI (i.e., degree centrality) indicates to which degree a local representational geometry resembles that of other ROI's. Degree centrality is highest in early visual cortex and lowest in IPS regions, indicating a higher conservation of geometry across early visual cortical regions.

Finally, we probe the underlying "representational connectivity" structure in individual participants (Kriegeskorte, Mur, & Bandettini, 2008a). Unlike traditional functional connectivity analysis, the representational connectivity approach does not target covarying activation per se, but rather assumes connections on the basis of shared representational geometry. Visualizing these "connections" in a graph (Figure 5B) highlights the dense clustering of early visual cortex, weaker connections to LO, and weakest connections to IPS, both during perception and memory. Based on this graph, we can compute the degree centrality of each ROI (or "node" in this graph) as the sum of connection strengths to other ROI's. A high degree centrality denotes high representational similarity to many –or an especially strong representational similarity with some– other brain regions. The highest degree centrality is observed in early visual cortex, suggesting overlap in representational geometry across these regions. More downstream visual areas (IPS and LO) show the lowest degree centrality, implying a gradual transformation of representational geometry along the visual pathway that results in geometries not present at earlier processing stages. This analysis also shows how *functional* measurements of sensory driven responses to oriented gratings, and even responses during working memory when no stimulus is on the screen, can approximate known *anatomical* structure of ROI's along the visual hierarchy. This demonstrates the power of an approach such as RSA, whereby functionally measured response patterns are transformed into an abstracted representational space.

Discussion

Here we compare how a simple visual stimulus is represented when it is either perceived or temporarily held in working memory, and show fundamental differences in the representational geometry of visual perception and visual working memory throughout human retinotopic cortex. By looking at the similarity of response patterns evoked by grating stimuli of different orientations, combined with a novel modeling approach, we are able to demonstrate relatively veridical representations during perception, and more categorical representations during working memory. We also find that the extent to which working memory representations are categorical increases along the visual hierarchy from posterior-to-anterior visual areas. Importantly, our models make distinct predictions about veridical and categorical representations from a *single* input function based on the statistics of the natural world, which can also be implemented by measuring human behavior with a simple psychophysical task. This makes our modeling approach a potentially powerful tool to apply in other research contexts as well. With clear differences in representational geometry, it seems unlikely that working memory representations are merely noisier versions of perceptual representations. Our data imply a systematic compression of the coding scheme in parts of orientation space as a basis for categorization in working memory. This implies that if noise is involved, it is implemented in a non-linear fashion. Finally, by looking at inter-area representational similarity we recover known anatomical cortical structure, and observe a high degree of similarity for areas within early visual cortex (EVC), intraparietal sulcus (IPS), and lateral occipital cortex (LO) – but relatively low similarity between these respective regions.

Previous work from our group has claimed that visual working memories are represented in a "sensory-like" manner in early visual cortex (Rademaker et al., 2019). This conclusion was drawn from the ability to cross generalize from sensory evoked responses, to responses recorded during the delay of a working memory task, using multivariate decoding techniques. However, there are multiple clues that VWM representations can be abstracted away from sensory evoked responses (Kwak & Curtis, 2022; Favila, Kuhl & Winawer, 2022; Rademaker et al., 2019; Linde-Domingo & Spitzer, 2022; Yan et al., 2023), including the considerable differences between perceptual and working memory geometries unveiled in the present analyses. From a conceptual point of view there may also be good reasons to keep formats distinct, as having identical representations for visual inputs and visual memories might make it difficult to distinguish external reality from internally generated thought (Bettencourt & Xu, 2016; Xu 2017; 2018).

Moreover, some sort of transformation of the information held in mind is often necessary to adequately support behavioral goals and motor output (Henderson, Rademaker, Serences, 2022; van Ede et al., 2019). How can we reconcile the apparent contradiction between successful sensory-to-memory cross-generalization on the one hand, and the mounting evidence favoring abstraction during VWM on the other?

To understand how perception and working memory can evoke overlapping response patterns (i.e., have an overlapping *coding scheme*) while also differing in their representational format (i.e., the *representational geometry*), we will examine the relationship between multivariate decoding and RSA more closely. First, note that within any kind of task, a high degree of similarity along the diagonal of a cross-validated RSM (with lower similarity further away from the diagonal), is a prerequisite for successful multivariate decoding, and vice versa. After all, if a particular stimulus would evoke uncorrelated response patterns every time it is presented (i.e., no similarity along the diagonal), a decoder would not be able to predict such a stimulus from the disparate response patterns that make up its training set (i.e., no decoding). Conversely, in areas of the brain that care about a certain kind of stimulus, you can expect both a clear diagonal component in the RSM, as well as successful within-task decoding. Things are a bit more nuanced for continuously varying stimuli such as orientation. To illustrate: Two identical orientations might evoke similar patterns of responses, give or take some noise, but so will two orientations that are adjacent in orientation space. While two orientations that are further apart are likely to evoke dissimilar responses. Based on the gradual transition in physical similarity between continuously varying stimuli, one would predict an RSM pattern where also off-diagonal similarity can have meaning, albeit with diminishing returns as representational similarity decreases with increasing stimulus distance.

For our data, this means that the clear diagonal component in the RSM's during both perception and working memory (see EVC ROI's in Figure 1C) are indicative of the ability to decode orientation *within* each of these two tasks. However, such apparent overlap in the represenational geometry around the diagonals does not speak to the geometry further away from the diagonals, nor does it speak to the ability to decode *between* the sensory and memory tasks (via cross generalization). With respect to the geometry at larger distances from the diagonal, we know that even relatively subtle transformations (e.g., shifts or warping) of an otherwise fairly stable underlying coding scheme can lead to dynamics in the low-dimensional geometry (Wolff et al., 2020; Kriegeskorte & Wei, 2021). Conversely, the representational geometry can also remain stable in the presence of dynamics in the coding scheme (Murray et al., 2016; Spaak, Watanabe, Funahashi, & Stokes, 2017; Kriegeskorte & Wei, 2021). With respect to cross-generalization, high similarity around the diagonals of both perception and working memory RSM's could stem from totally different response patterns in one task compared to the next as long as the pairwise distances between response pattern are comparable between tasks. We show that in early visual ROI's the underlying coding schemes during the sensory and memory tasks are sufficiently similar to yield a clear diagonal component in an "across-task RSA" (Figure 4). Importantly, we validate our across-task RSA approach against a common implementation of multivariate decoding for continuous stimulus spaces (the so-called inverted encoding model, Brouwer & Heeger, 2009; Rademaker et al., 2019). More interestingly, the across-task RSM also provides some insight into how the coding scheme is warped during working memory compared to perception – we observe strong biases away from cardinal orientations during working memory, with many orientations resulting in patterns that are similar to those of oblique orientations that are directly perceived. To sum up, working memory representations in EVC can be "sensory-like" in that there is considerable overlap in the response patterns during perception and working memory. At the same time, the systematic warping of the coding scheme for orientation during working memory, relative to perception, results in a more categorical representational geometry during VWM with high similarity around obliques.

In addition to using cross-generalization from sensory-to-memory responses to conclude that early visual areas store "sensory-like" working memory representations, our previous work drew upon the *lack* of such cross-generalization (in the presence of high within-task decoding performance) to conclude that IPS stores working memories in a format "transformed away" from sensory-like responses (Rademaker et al., 2019; Iamshchinina et al., 2021). However, our current analyses reveal how the representational geometry during working memory is predominantly categorical *throughout* the visual hierarchy: A significant benefit of the categorical over the veridical model can be seen in V3–IPS0 when using a theoretical input function (Figure 2A, 2D bottom panel), and in *all* retinotopic EVC and IPS regions when using the psychophysical input function (Figure 3A, 3B bottom panel). The reason that cross-generalization from the sensory to the memory task fails in IPS may therefore not be due to a transformation in the representational geometry from earlier-to-later visual areas during working memory.

Instead, IPS might simply process information quite differently during perception than during working memory. Recent recordings from non-human primates reveal that the receptive field of an IPS neuron (in lateral intraparietal "LIP" cortex) that was demarcated by showing the animal visual stimuli on a screen does not necessarily overlap with the receptive field of that same neuron when demarcated by measuring responses during a delayed-match-to-sample task (Krug et al., in prep). This implies a distinct mechanism for representing sensory inputs and working memory contents at the level of single neurons, which plausibly scales up to the level of population recordings as obtained with fMRI. A second reason why cross-generalization from perception-to-memory might be lacking in IPS is because the sensory input is represented rather weakly in IPS in our sensory task. While the full-field grating stimulus was attended, the feature of interest to our analyses (orientation) was not explicitly probed and not directly relevant to the participants' task – which was to monitor and report instances of contrast dimming. It is not yet clear how feature-based attention changes the structure of stimulus representations in IPS. However, given the central role of IPS in attention (Selemon & Goldman-Rakic, 1988; Corbetta & Shulman, 2002; Silver & Kastner, 2009; Bressler & Silver, 2010), and evidence that attention improves decoding of task-relevant stimulus features in EVC (Jehee et al., 2012), it would be interesting to examine how attending different features of the same stimulus might impact stimulus representations, and whether this could explain the relatively noisy RSM's we observed in IPS during our sensory task (Figure 1C, top; Supplementary Figure 1, top).

A big question in the field of VWM concerns the role of primary visual area V1 during memory maintenance. On the one hand, sensory recruitment theory posits that involvement of area V1 is critical to maintaining highly detailed visual representations, as this is the only site thought to have the resolution to do so (Serences, 2016; Adam, Rademaker, & Serences, 2022). In support of this theory, many fMRI studies have shown that VWM contents can be decoded from V1 (Harrison & Tong, 2009; Serences et al., 2009), and correlations between behavioral and decoding performance imply a functional role for V1 (Ester et al., 2013; Iamshchinina et al., 2021). On the other hand, outside of the fMRI literature there is less evidence to support a neural correlate of VWM in area V1, with a general failure to find sustained firing in EVC (Mendoza-Halliday, Torres, & Martinez-Trujillo, 2014; Leavitt, Mendoza-Halliday, & Martinez-Trujillo, 2017, but see also: Bisley et al., 2004; Zaksas & Pasternak, 2006), which has led people to conclude that V1 decoding could be epiphenomenal (Xu, 2017; 2018). Might our findings speak to this discrepancy between fMRI and single cell recording? Receptive fields in V1 are small, so if there is a representational shift from “veridical” during perception, to more “categorical” during VWM (presumably under the influence of top-down feedback), then working memory contents may be coded by a (subtly) different subset of neurons than those that respond to perceptual input. The reasoning that the same neurons may not code for the same stimulus under different task conditions holds true on several levels. For example, multi-unit activity associated with working memory maintenance was restricted to deep and superficial layers in V1, while such activity during perception also included the input layer 4 (van Kerkoerle, Self, & Roelfsema, 2017). Thus, even small shifts in the neural code (from one layer to the next, or from one orientation column to the next) may decrease the chance of finding sustained spiking when a one-to-one mapping between perception and working memory is assumed. Only looking at population wide neural coding, as we do here, can uncover working memory contents that has undergone a shift in representational geometry relative to perception.

A well-known strength of RSA is that it projects response patterns associated with different task conditions (in our case, task conditions are 180 levels of orientation) into an abstracted space where representations can be compared between imaging modalities, species, models, or with behavior (Kriegeskorte et al. 2008a). Thus, RSA is a powerful tool to sidestep the correspondency problem, allowing us to compare the output of systems that differ greatly. For example, one can construct an RSM from behavioral responses and correlate it with an RSM constructed from neural responses of specific brain regions (Kriegeskorte et al. 2008a, 2008b; Bettencourt & Xu, 2016). However, visual perception involves a cascade of processes of increasing complexity, from simple feature-detectors in primary sensory cortex, to more invariant and category-based representations in ventral visual cortex. Behavior is the result of the entire brain working in concert to produce one output (Ungerleider, Courtney, & Haxby, 1998), which means that even for a very simple stimulus such as orientation the representational geometry can differ between areas and tasks. Using a behavior-derived RSM as a model could therefore miss a lot of variability in representational geometry across cortex, or produce misleading conclusions about an area “X” representing orientation while ignoring other areas that match behavioral output less. This is why using behavioral measurements (here: psychophysical input function) in a hypothesis-driven manner as the basis for multiple models (here: veridical and categorical), may allow for deeper insight into which specific areas of the brain are involved in multiple underlying components of a single behavior.

Our models are indeed able to quantify notable differences in representational geometry between orientation perception and working memory tasks, as well as differences between various retinotopic areas within a single task. However, there remain patterns in the data that neither of our current models are designed to capture. For example, data recorded during the sensory task from area V3AB (and arguably also V4, IPS, and LO) show a representational similarity pattern implying that the obliques (45° and 135º) are formatted quite similarly to one another (Figure 1C, top row; Figure 4F, top row). This pattern hints at another possible form of (object-level) abstraction, where all tilted lines, irrespective of their tilt direction, are represented in a similar fashion. This parallels evidence showing that people use the same verbal labeling ("diagonal") for both obliques (Overkott & Souza, 2023). The stronger similarity around vertical (180º) compared to horizonal (90º) orientations is another example where our models do not perfectly capture the data, and might have to do with the predominance of horizontal orientations in natural scenes (Harrison et al., 2023). Given such diversity in RSM patterns, how might we compare the geometry within a given task, while remaining agnostic to the precise patterns in different regions of interest? To evaluate inter-area representational geometry differences in a more hypothesis-free manner, we used a "representational connectivity" analysis, which subsumes all possible patterns by simply quantifying degree of overlap. This allows us to compare how the visual system orchestrates representations across large swaths of cortex during both perception and working memory. One observation that emerges from this approach is that during VWM we see a shift in the interplay between areas, as compared to during perception. Specifically, during working memory the geometry in V1 becomes more differentiable from the geometry in the rest of EVC, and IPS0 becomes more differentiable from the rest of IPS (but more similar to EVC). In other words, we see that the inter-areal structuring of representational geometry differs between perception and memory. Another observation is that compared to perception, geometries during VWM show higher similarity across visual cortical areas. Such homogeneity might be expected if a unitary categorical top-down signal dominates feedback signals to multiple earlier areas. After all, in the absence of visual input, working memory information in V1 must be coming from within the brain itself, either through feedback connections or local recurrent processing.

Using representational similarity, in combination with the novel modeling approach outlined here, can be a powerful tool for studying representational formats during perception and working memory. Once the input statistics are known, either by deriving them from the environment or behavior, these models can be applied to any feature. Thus, in addition to tapping into possible abstractions for spatial features in EVC (Kwak & Curtis, 2022), our approach has potential for surface-based features such as color or contrast, more complex visual objects such as shapes or faces, as well as stimuli in other sensory modalities. Many of the well-known advantages from RSA approaches also apply here, such as the potential to fit models across the entire brain in a manner that is not restricted to early sensory areas for which the receptive field mapping is known. Furthermore, the approach can be used with different measurement techniques that have a higher temporal resolution, allowing additional queries about the temporal progression of representational geometry as stimuli are encoded, remembered, and recalled.

References

Adam, K.C.S., Rademaker, R.L., & Serences, J.T. (2022). Evidence for, and challenges to, sensory recruitment models of visual working memory. *Routledge book on Visual Memory, Chapter 1*

Appelle S. (1972). Perception and discrimination as a function of stimulus orientation: The oblique effect in man and animals. *Psychological Bulletin,* 78, 266–278

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61*(3), 183–193.

Barlow, H.B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01), 217–33

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

Bisley, J. W., Zaksas, D., Droll, J. A. & Pasternak, T. (2004). Activity of neurons in cortical area MT during a memory for motion task. *J. Neurophysiol.,* 91, 286–300

Bressler, D.W. & Silver, M.A. (2010). Spatial attention improves reliability of fMRI retinotopic mapping signals in occipital and parietal cortex. *Neuroimage,* 53, 526–533

Brouwer, G. J. & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *J. Neurosci.,* 29(44), 13992–14003

Chafee, M.V. & Goldman-Rakic, P.S. (1998). Matching patterns of activity in primate prefrontal area 8a and parietal area 7ip neurons during a spatial working memorytask. *Journal of neurophysiology*, 79(6), 2919-40

Christophel, T.B., Hebart, M.N. & Haynes, J.D. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. *J. Neurosci.,* 32, 12983–12989

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

Compte, A., Brunel, N., Goldman-Rakic, P.S., Wang, X-J (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9), 910–23

Corbetta, M., Shulman, G. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci,* 3, 201–215

Courtney, S.M., Petit, L., Maisog, J.M., Ungerleider, L.G. and Haxby, J.V. (1998). An area specialized for spatial working memory in human frontal cortex. *Science*, *279*(5355), 1347-51

Essock, E.A. (1980). The Oblique Effect of Stimulus Identification Considered with Respect to Two Classes of Oblique Effects. *Perception*, 9(1), 37–46

Ester, E.F., Anderson, D.E., Serences, J.T., & Awh, E. (2013). A Neural Measure of Precision in Visual Working Memory. *J Cogn Neurosci*, 25(5), 754–761

Ester, E.F., Sprague, T.C., & Serences, J.T. (2015). Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory, Neuron, 87(4), 893-905

Favila, S.E., Kuhl, B.A. & Winawer, J. (2022). Perception and memory have distinct spatial tuning properties in human visual cortex. *Nat Commun,* 13, 5864

Funahashi, S., Bruce, C.J., & Goldman-Rakic, P.S. (1989). Menomonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.*, 61, 331-49

Funahashi, S., Chafee, M.V., Goldman-Rakic, P.S. (1993). Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. *Nature*, 365, 753–56

Furmanski, C., Engel, S. (2000). An oblique effect in human primary visual cortex. *Nat Neurosci,* 3, 535–36

Fuster, J.M., & Alexander, G.E., (1971). Neuron activity related to short-term memory. *Science,* 173, 652-54

Friedman, H.R. & Goldman-Rakic, P.S. (1994). Coactivation of prefrontal cortex and inferior parietal cortex in working memory tasks revealed by 2DG functional mapping in the rhesus monkey. *Journal of Neuroscience*, 14(5), 2775–88.

Gayet, S., Paffen, C.L.E., Van der Stigchel, S. (2018). Visual Working Memory Storage Recruits Sensory Processing Areas. *TICS*, 22(3), 189-190

Girshick, A.R., Landy, M.S., & Simoncelli, E.P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, 14(7), 926–32

Goldman-Rakic, P.S. (1995) Cellular basis of working memory. *Neuron,* 14, 477–485

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

Harrison, W.J. Bays, P.M., & Rideaux, R. (2023). Neural tuning instantiates prior expectations in the human visual system. *bioRxiv*, https://doi.org/10.1101/2023.01.26.525790

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

Henderson, H., & Serences, J.T. (2021). Biased orientation representations can be explained by experience with nonuniform training set statistics. *Journal of Vision,* 21(8):10

Iamshchinina, P., Christophel, T.B., Gayet, S., & Rademaker, R.L. (2021). Essential considerations for exploring visual working memory storage in the human brain. *Visual Cognition*, 29(7), 425–436

Jehee, J.F.M., Ling, S., Swisher, J.D., van Bergen, R.S., & Tong, F. (2012). Perceptual learning selectively refines orientation representations in early visual cortex. *Journal of Neuroscience,* 32, 16747-16753

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008a). Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.,* (2)

Kriegeskorte, N., Mur, M., Ruff, D., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6), 1126-41

Kriegeskorte, N. & Wei, X-X (2021), Neural tuning and representational geometry. *Nat. Rev. Neurosci*., 22, 703-18

Krug, K. (in prep). Personal communication.

Kwak, Y. & Curtis, C.E. (2022). Unveiling the abstract format of mnemonic representations. *Neuron*, 110(11), 1822-28

Leavitt, M.L., Mendoza-Halliday, D., Martinez-Trujillo, J.C. (2017). Sustained Activity Encoding Working Memories: Not Fully Distributed. *TINS*, 40(6), 328–46

Lennie, P. (1971). Distortions of perceived orientation. *Nature New Biology*, 233(39), 155–156.

Linde-Domingo, J., & Spitzer B. (2022). Geometry of visual working memory information in human gaze patterns. *BioRxiv*, https://doi.org/10.1101/2022.11.17.516917

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

McCarthy, G., Blamire, A.M., Puce, A., Nobre, A.C., Bloch, G., Hyder, F., Goldman-Rakic, P. and Shulman, R.G. (1994). Functional magnetic resonance imaging of human prefrontal cortex activation during a spatial working memory task. *PNAS*, 91(18), 8690–94

McCarthy, G., Puce, A., Constable, T., Krystal, J.H., Gore, J.C. and Goldman-Rakic, P. (1996). Activation of human prefrontal cortex during spatial and nonspatial working memory tasks measured by functional MRI. *Cerebral cortex*, 6(4), 600–11

Mejías, J.F., Wang, X-J (2022). Mechanisms of distributed working memory in a large-scale network of macaque neocortex. *eLife,* 11:e72136

Mendoza-Halliday, D., Torres, S. & Martinez-Trujillo, J. C. (2014). Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. *Nat. Neurosci.,* 17, 1255–1262

Miller, E.K., Erickson, C.A. & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.,* 16, 5154–5167

Murray, J.D., Bernacchia, A., Roy, N.A., Constantinidis, C., Romo, R., & Wang, X-J. (2016). Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *PNAS*, 114(2), 394-399

Olshausen, B.A., & Field, D.J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1?. *Vision research*, *37*(23), 3311-3325

Overkott, C., & Souza, A.S. (2023). The fate of labeled and nonlabelled visual features in working memory. *JEP:HPP*, 49(3), 384–407

Qi, X.-L., Elworthy, A.C., Lambert, B.C. & Constantinidis, C. (2015). Representation of remembered stimuli and task information in the monkey dorsolateral prefrontal and posterior parietal cortex. *J. Neurophysiol.,* 113, 44–57

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

Riggall, A.C. & 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

Schurgin, M.W., Wixted, J.T. & Brady, T.F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nat Hum Behav,* 4, 1156–1172

Selemon, L.D., & Goldman-Rakic, P.S. (1988). Common cortical and subcortical targets of the dorsolateral prefrontal and posterior parietal cortices in the rhesus monkey: evidence for a distributed neural network subserving spatially guided behavior. *J Neurosci,* 8, 4049–68

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

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

Shannon, C.E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423

Silver, M.A. & Kastner, S. (2009). Topographic maps in human frontal and parietal cortex. *Trends Cogn. Sci.,* 13, 488–495

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, *24*(1), 1193-1216

Stokes, M.G. (2015). 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends Cog. Sci.,* 19, 394–405

Spaak, E., Watanabe, K., Funahashi, S. & Stokes, M.G. (2017). Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. *J. Neurosci.*, 37(27), 6503-6516

Ungerleider, L.G., Courtney, S.M. and Haxby, J.V. (1998). A neural system for human visual working memory. *PNAS*, *95*(3), 883–90

van Bergen, R.S., Ji Ma, W.J., Pratte, M.S. & Jehee, J.F.M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nat Neurosci,* 18, 1728–1730

van Ede, F., Chekroud, S.R., Stokes, M.G., & Nobre, A.C. (2019). Concurrent visual and motor selection during visual working memory guided action. *Nat Neurosci,* 22, 477–483

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

Vo, V.A., Sutterer, D.W., Foster, J.J., Sprague, T.C., Awh, E., & Serences, J.T. (2022). Shared Representational Formats for Information Maintained in Working Memory and Information Retrieved from Long-Term Memory. *Cereb Cortex*, 32(5), 1077-92.

Wei X.X., Stocker A.A. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nat Neurosci*., 18(10), 1509-17

Wilson, F.A., Scalaidhe, S.P. and Goldman-Rakic, P.S. (1993). Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science*, *260*(5116), 1955-58

Wimmer K., Nykamp D.Q., Constantinidis C., Compte A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat Neurosci.*, 17(3), 431-9

Wolff M.J., Jochim J., Akyürek E.G., Buschman T.J., Stokes M.G. (2020). Drifting codes within a stable coding scheme for working memory. *PLoS Biol,* 18(3): e3000625

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage*, 137, 188-200

Xu, Y. (2017). Reevaluating the sensory account of visual working memory storage. *TICS*, 21(10), 794–815

Xu, Y. (2018). Sensory cortex is nonessential in working memory storage. *TICS*, 22(3), 192–3

Yan, C., Christophel, T.B., Allefeld, C., & Haynes, J-D. (2023). Categorical working memory codes in human visual cortex. *NeuroImage*, advanced online publication

Zaksas, D. & Paternak, T. (2006). Direction signals in the prefrontal cortex and in area MT during a working memory for visual motion task. *J. Neurosci.,* 26, 11726–1174

Materials and Methods

Stimuli and procedures

We used a publicly available dataset, originally part of a study by Rademaker et al. (2019). This dataset contains both a visual perception task (the "sensory task") and a visual working memory task (the "memory task") performed in the scanner, and presented in the paper as Experiment 1 (6 participants with mean age = 28.67, sd = 3.675, and 5 females). The dataset also contains an independent psychophysical experiment, presented in the paper as Supplementary Figure 9 (21 participants with mean age = 20.12, sd = 2.01, and 14 females. For this psychophysical experiment, only 17 participants were included in the analysis (3 dropped out, and 1 performed at chance level). All participants who contributed to the dataset were neurologically healthy, had normal or corrected-to-normal vision, received monetary reimbursement, and provided their written informed consent. Data were collected at the University of California San Diego.

In the scanner, both the sensory and memory tasks used full-contrast donut-shaped grating stimuli (1.5° and 7° inner and outer radius, respectively) with smoothed edges, a spatial frequency of 2 cycles per degree, random phase, a pseudo-randomized orientation. Stimuli were presented against a uniform grey background, and participants fixated a 0.4° central dot throughout. In the sensory task, donut-shaped gratings were presented in 9 second trials, contrast reversing at 5 Hz. Such donut trials were interleaved with trials showing a circular grating (1.5° radius), and fixation periods (10% of total). Grating contrast was briefly (200ms) and probabilistically reduced to 80% Michelson twice every 9s, and participants' task was to report such contrast changes. Participants completed a total of 300-320 sensory task trials across 3 separate scanning sessions. The sensory task was also used to localize visually responsive voxels (via a donut > circle contrast), and in our current analysis we use this contrast as a mask for all EVC ROI's (but not IPS and LO, where all retinotopically defined voxels are included). In the working memory task, a target grating was shown for 500ms, and recalled 13 seconds later by rotating a white line (spanning 7°) for 3 seconds to match the remembered orientation. Between trials, there could be 3, 5, or 8 second fixation intervals. During the delay of two-thirds of memory task trials, a distractor of 50% Michelson contrast could be presented for 11 seconds during the middle portion of the delay. Distractors could be a grating (1/3rd of trials) or filtered noise (1/3rd of trials), contrast reversing at 4Hz. By ensuring uniform orientations of grating distractors with respect to memory targets, we are able to look at the representations of remembered and distractors orientations independently. Importantly, due to the negligible differences between trials with or without distractors, both in terms of behavior as well as decoding, we collapse the data across all working memory trials for our main analyses. Each participant completed 324 total working memory trials over the course of 3 different scanning sessions.

The independent psychophysical experiment was completed outside of the scanner, and stimuli consisted of gratings presented at 20% Michelson contrast against a uniform grey background. Gratings had a 2° radius, spatial frequency of 2 cycles per degree, random phase, and pseudo-randomized orientation. Each trial started with a 200ms target orientation that was remembered over a 3s delay, and recalled by rotating a dial to match the remembered orientation in an unspeeded manner. Intervals between trials were 800-1000ms. Grating distractors were presented for 200ms during the middle portion of the delay on 90% of trials. As in the scanner, targets and distractors were uncorrelated across trials, allowing for independent analysis of responses to the target. Any biases resulting from distractor presentation were small (Supplementary Figure 9 of Rademaker et al., 2019) and did not exert much influence on responses. Each participant completed 1620 trials over the course of several testing sessions.

For more detailed information on scanning and psychophysical procedures, please reference Rademaker et al. (2019).

Models

First, the function that constrains both the veridical and categorical models, and emulates the frequency distribution of orientations in the natural world, is described by

$$f(x) = \left||\sin x| - 1\right|^2 + b$$

Where $-\pi < x < \pi$, and b is any non-zero baseline (due to z-scoring before fitting this function is scale-free). This function is loosely based on the function defined in Girshick & Simoncelli (2011). Another way to think about this theoretical "input statistics" function, is as the normalized amount of Fisher information at each orientation in orientation-space. We ensured that the results of our model fits were robust to the specific shape of the input function by not only using a theoretical input function based on the statistics of the natural world (Figure 2A), but by also using a psychophysical input function (generated from independent psychophysical measurements, Figure 3A) as the basis for our two models. Irrespective of the input function used (theoretical based on previous literature, or psychophysical based on independent data), model generation, as described next, is identical.

For the veridical model, we assume a set of idealized tuning functions (Figure 2B, top) and use it to simulate neural responses to all possible stimulus orientations. From these simulated responses we calculate the similarity (rho) between all possible pairs of stimulus orientation (as shown in Figure 1B), resulting in a veridical model RSM (Figure 2B, bottom). Each tuning function in the veridical model is defined by a von Mises (circular analogue of a Gaussian distribution),

$$f(x) = a\left(\frac{e^{\kappa\cos(x-\mu)}}{2\pi I_0(\kappa)}\right)$$

with a fixed concentration parameter κ, a center defined by μ, and $-\pi < x < \pi$. $I_0(\kappa)$ is the modified Bessel function of order 0. The amplitude a of each tuning function varies across orientation space as determined by the height of the input statistics function.

For the categorical model, we calculated the psychological distance between all possible pairs of orientation. For any pair of orientations, we sum over the approximated derivatives between these two points along the input statistics function as follows,

$$\sum_{i=start}^{end-1} \left|\frac{y_{i+1} - y_i}{x_{i+1} - x_i}\right|$$

where x is an orientation in orientation space (and wraps around a circle), and y is the amount of normalized y-axis information for that orientation.

The veridical and categorical models described above were converted from radians to degree (spanning the entire orientation space from 1° to 180°, in steps of 1°) to match the dimensions of the data (also in degree). To evaluate if the representational structure in the data RSM's (Figure 1C) is more or less similar to veridical or categorical model RSM's (Figure 2B-C), both data and model RSM's were normalized before fitting. Fitting was done via minimization of the mean squared errors. Because the veridical and categorical models are not independent, each model was fit to the residuals left by the other model. Specifically, to fit model A to the data RSM of a given ROI (RSM_{ROI}), we first fit model B (RSM^{B}) to get its residuals ($\mathcal{E}^{B}$):

$$RSM_{ROI} = w^{B} RSM^{B} + \mathcal{E}^{B}$$

Where w^{B} are the initial weights of model B, and the residuals $\mathcal{E}^{B}$ reflect any pattern in the data unaccounted for by model B. To illustrate, imagine the extreme case where model B explains none of the data RSM_{ROI}, then w^{B} would be 0, and the residuals $\mathcal{E}^{B}$ would be equal to the data RSM itself.

Next, model A is fit to the residuals of model B:

$$\mathcal{E}^{B} \; = \; w^{A} RSM^{A} + \mathcal{E}^{A}$$

Now, the weights of model A (w^{A}) reflect the amount of variance explained by model A independent of model B. To summarize, this procedure ensures that any resemblance (big or small) between the data and model B is first soaked out of the data, after which we obtain the unique contribution made by model A. Thus, the weights of the veridical model are obtained by fitting to the residuals of the categorical model, and the weights of the categorical model by fitting to the residuals of the veridical model (Figure 2D).

Statistics

First, we tested if model fits differed between task (sensory or memory), model (veridical or categorical), or ROI's by running a three-way repeated measures ANOVA using R (version 4.1.1) and RStudio (version 1.4.1717). When using the theoretical input statistics function in Figure 2A as a basis for the veridical and categorical models and their subsequent fits to the data, we found a significant three-way interaction between all factors ($F_{(7,35)} = 2.39$, $p = 0.0415$). We followed this up with two targeted two-way ANOVA's – one for the sensory task, and one for the memory task, as described in the main text. In addition to the theoretical input statistics function based on prior literature (Figure 2A), we also used the behavioral response frequencies from an independent psychophysical experiment to construct both the veridical and categorical models (Figure 3A). After fitting the models generated in this manner, we again evaluated the impact of task, model, and ROI using a three-way ANOVA. This time, the three-way interaction did not quite reach significance ($F_{(7,35)} = 2.177$; $p = 0.0607$), but there were significant two-way interactions between all factors (model x task, $F_{(1,5)} = 35.53$, $p = 0.0019$; model x ROI, $F_{(7,35)} = 4.159$, $p = 0.002$; task x ROI, $F_{(7,35)} = 12.91$, $p < 0.001$). Based on these interactions, and to keep statistical tests consistent with those from the modeling based on the theoretical input function, we followed up with two two-way ANOVAs (for the sensory and memory tasks, as described in the main text).

Significant interactions arising from two-way ANOVA's were further examined via post-hoc tests within each task and ROI, performed using permutation tests over 100.000 iterations using custom written code. First, we compared whether the veridical and categorical model differed significantly in each ROI using two-sided paired-sample t-tests: On each iteration we shuffled the model label assignment (i.e., which weight came from which model) randomly for each subject, after which a t-statistic was calculated across all subjects $t = \frac{\bar{X}_D - \mu_0}{s_D / \sqrt{n}}$ ($\bar{X}_D$ and s_D denote the mean and sd of the pairwise differences, μ_0 is the null hypothesis, and n the number of subjects). The true t-statistic of the intact data was compared to the null distribution of all permuted t-statistics to get the p-value. Second, we tested within each model if the weights differed significantly from zero by using a two-sided t-test: We randomly permuted the sign (– or +) of the model weight for each subject on each iteration and calculated a t-statistic $t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$ ($\bar{x}$ and s denote the mean and standard deviation across subjects, μ_0 is the null hypothesis, and n the subject number). Post-hoc tests were not corrected for multiple comparisons. All post-hoc tests are reported in Supplementary Table 1.

Analyses of 2nd level RSM and representational connectivity

To compare the representational geometry across all retinotopic ROI's during perception and working memory, we employed two approaches first described by Kriegeskorte et al. (2008b): A 2nd level RSA and a 'representational connectivity' analysis. Note that for these analyses we used the smallest possible ROI's that were retinotopically defined, meaning we split early visual areas into their dorsal and ventral parts, and used the individual sub-areas of IPS and LO. For the 2nd level RSA, we calculated the similarity of each across-subject RSM (as shown in Supplementary Figure 1) to every other across-subject RSM using spearman correlation (as RSM's are monotonically, but perhaps not linearly related). This resulted in the 2nd level RSM's in Figure 5A, showing representational similarity between all retinotopic ROI's during perception, and during working memory. For the 'representational connectivity' we similarly computed Spearman correlations between ROI's but at a within-subject level as is the recommended procedure (Kriegeskorte et al., 2008b). Across-subject averages are visualized in a graph (Figure 5B) where each node signifies a ROI, and each edge signifies the correlation coefficient to each other ROI. Thicker and shorter edges indicate higher similarity. We computed degree centrality for each node as the sum of all edges, depicted by the saturation of each node (less saturation indicating higher degree centrality). Thus, higher degree centrality indicates that a ROI shares its representational geometry either strongly with a few, or somewhat strongly with many, other ROI's.

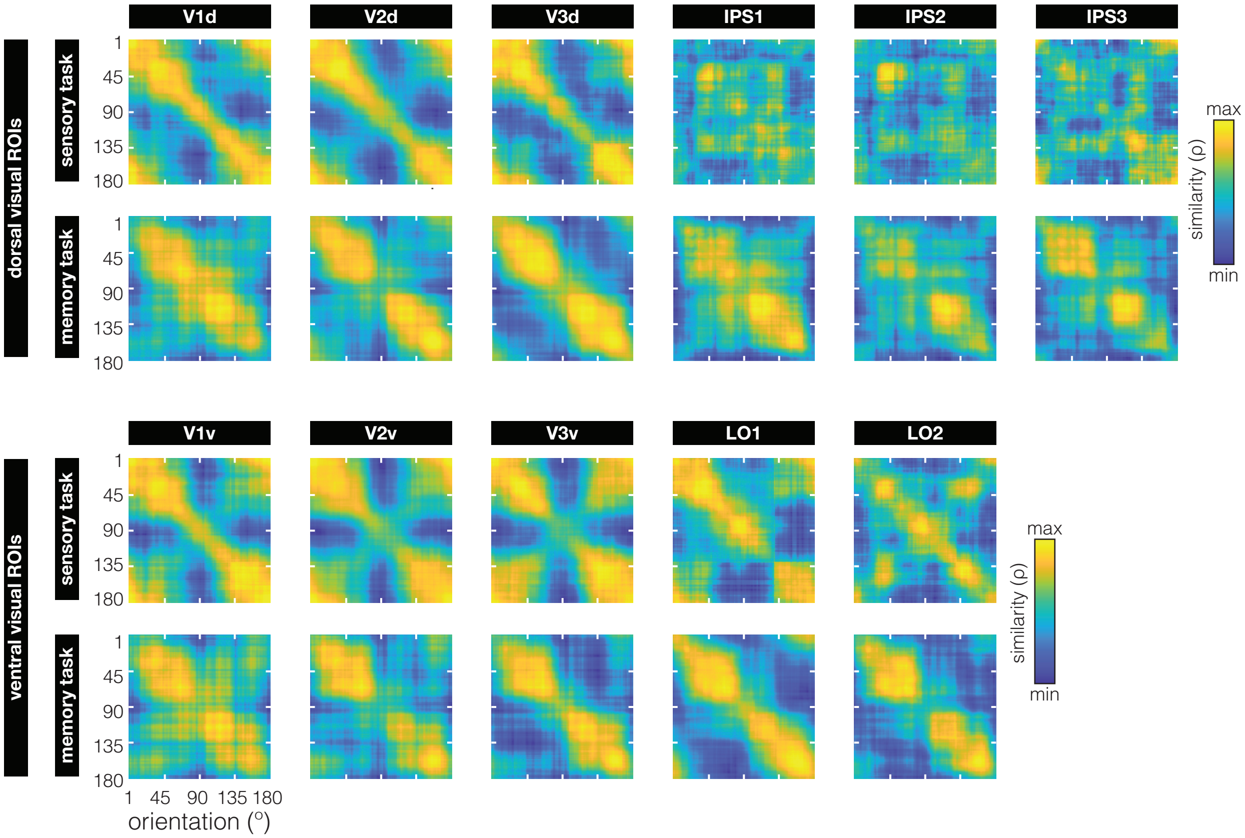
Across-task RSA fidelity

To directly relate cross-generalization from decoding (or more specifically, from the inverted encoding model, or "IEM", as used in Rademaker et al., 2019) to our novel "across-task RSA", we calculate a fidelity metric to quantify how much information there is about the remembered orientation based on the pattern responses to the perceived orientations. Because orientation is a continuous variable, we also take into account the representational similarity to orientations nearby the remembered one, and the expected drop in similarity at increasingly larger distances in orientation space. For each correlation profile (see Figure 4B–D) we calculate fidelity in a manner identical to how this has been calculated for IEM channel reconstructions (in Rademaker et al., 2019", and shown here as the y-axis in Figure 4D). Specifically, we take the correlation value at each degree in orientation space (wrapped onto a 2π circle), and project this vector onto the remembered orientation (centered to zero degrees) via $\cos A_{abs(0º-d)} = \frac{b}{h}$, where *A* is the angle between the remembered orientation (at *0º*) and the degree in orientation space being evaluated (*d*), and *h* is the correlation value at *d* (i.e. the hypotenuse of a right triangle). This procedure was repeated for all 180 degrees in orientation space, and we then calculate the mean of all 180 projected vectors. This fidelity metric captures the amount of information at the remembered orientation, and removes additive offsets.

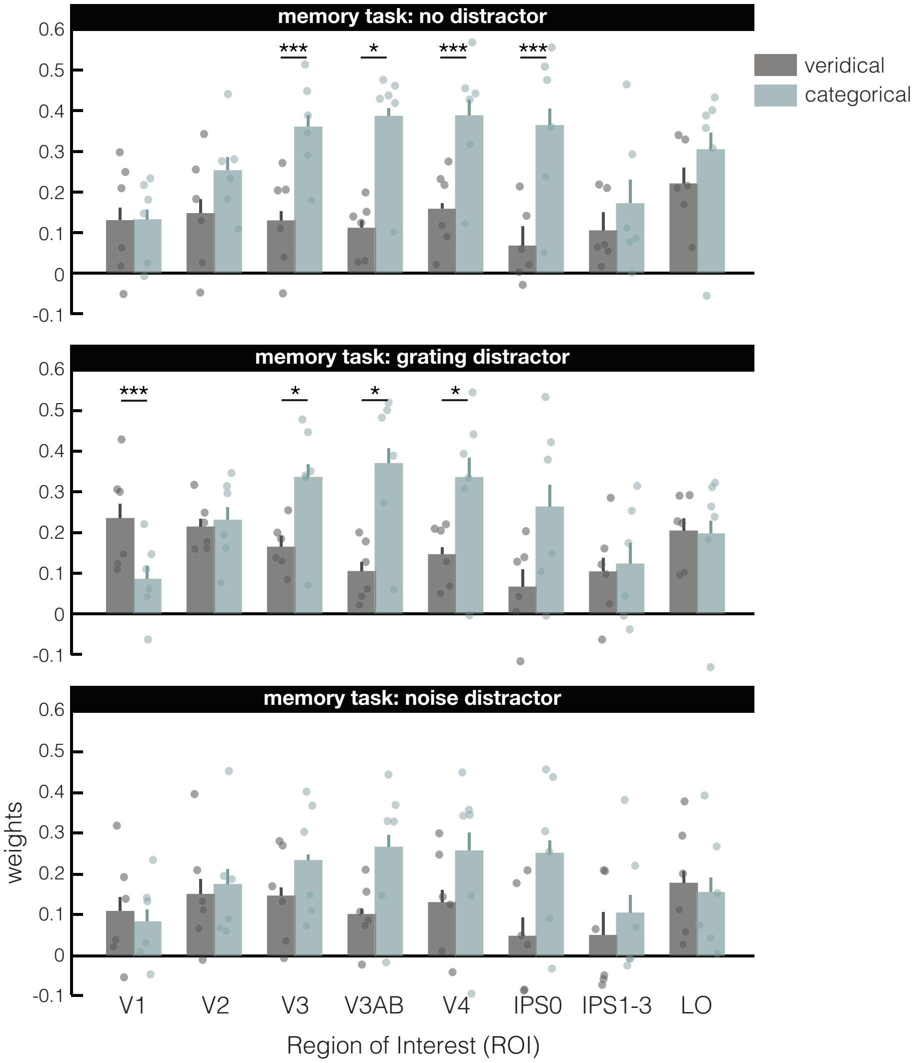
Code accessibility

The data are public and can be accessed via the Open Science Framework (OSF) at https://osf.io/dkx6y which has an accompanying wiki. The code for the analyses in this paper can be found at https://osf.io/placeholder.

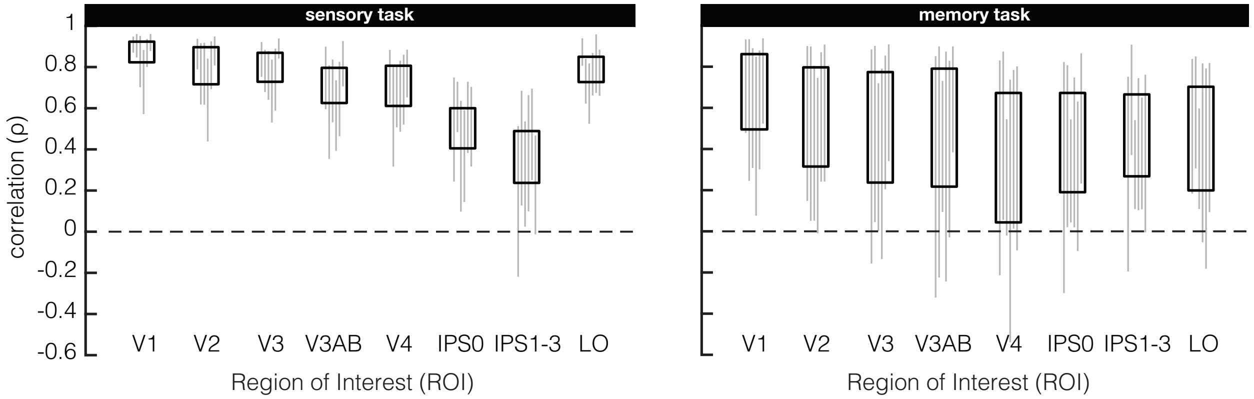
Supplementary Materials



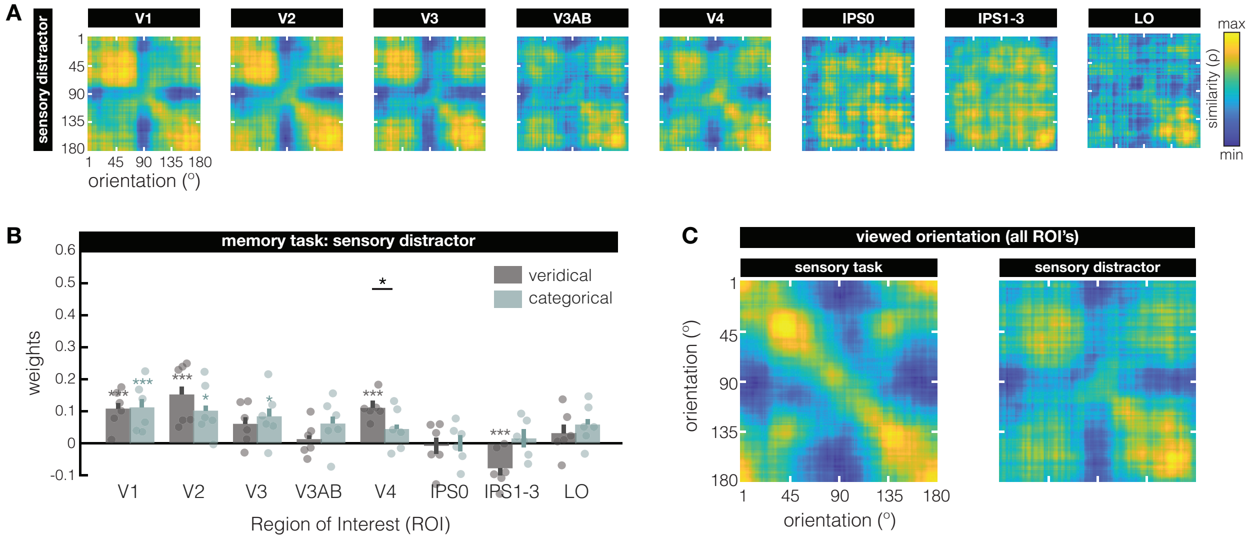
Supplementary Figure 1: Orientation representation (as indexed by RSM's) during sensory perception and working memory for all retinotopically defined ROI's (across all participants) that were not already shown in Figure 1C. ROI's are organized by whether they are located in the dorsal or ventral stream (top and bottom two rows, respectively). Early visual areas V1–V3 were split by their dorsal and ventral portions – used as input to the second-level RSA analyses (Figure 3). Areas IPS1–3 (in the dorsal stream) and LO (in the ventral stream) were split based on their respective sub-portions – and similarly used as input to the second-level RSA analyses. All RSM's are scaled to the range of correlations within each subplot to ease visual comparison of representational structure.



Supplementary Figure 2: Model fits for the 3 different working memory distractor conditions. Overall, the results split by condition are qualitatively similar to the main results across all trails (Figure 2D, bottom panel). The two-way ANOVA's comparing model and ROI showed that the categorical model did a better job at explaining the data without a distractor (main effect model: $F_{(1,5)} = 12.29$; $p = 0.017$), and marginally so with a grating distractor ($F_{(1,5)} = 5.55$; $p = 0.065$) presented during the delay. Both n-distractor and grating-distractor conditions also had significant interactions ($F_{(7,35)} = 2.778$; $p = 0.017$ and $F_{(7,35)} = 4.234$; $p = 0.002$), indicating increasing differences between the veridical and categorical models along the visual hierarchy. For the 108 trials during which a noise distractor was presented during the delay, neither the main effect of model ($F_{(1,5)} = 1.227$; $p = 0.318$), nor the interaction between model and ROI ($F_{(7,35)} = 1.745$; $p = 0.13$) reached statistical significance. Nevertheless, despite using only 1/3rd of the data in each of these sub-plots, the pattern in the data do not change



Supplementary Figure 3: Exact ranges of correlations in the RSM's from Figure 1C. To best show the representational structure for sensory and memory representations across ROI's, and to ease comparison between them, the RSM's in Figure 1C are scaled to the range (min-to-max) of correlations within each subplot. But the minimum and maximum correlations are not identical across subplots, therefore, correlation ranges across all participants (black rectangles) and individual participants (grey lines) are shown here for sensory (left) and memory (right) RSM's.



Supplementary Figure 4: Orientation representations of viewed gratings shown during the working memory delay (A) To test the generalizability of our finding that sensory inputs are represented in a predominantly veridical manner (Figure 2D, top), we constructed RSM's from voxel responses to directly viewed gratings presented as distractors during the delay of our memory task (for retinotopically defined ROI's and across all participants). The grating distractor was shown for 11s during the working memory delay on one-third of trials. Importantly, orientations of the remembered and distractor gratings were random and independent, allowing us to also examine the representational similarity structure of the distractor. In early visual areas, the RSM's show a diagonal pattern (with highest similarity around 180°) combined with some degree of clustering around obliques. Indeed, the modeling results in **(B)** show that both the veridical and categorical models explain the representational pattern to some extent. There is a significant interaction between ROI and model ($F_{(7,35)}$ = 5.21; p = 0.025), indicating that the two models capture the data differently in different ROI's. Only in area V4 did the veridical model outperform the categorical model, but there was no difference between the two models overall (main effect of model, $F_{(1,5)}$ = 1.054; p = 0.308) in V1, V2, and V4 (and an opposite effect in IPS1-3). **(C)** Here we show RSM's of viewed orientations under two different contexts side-by-side (across all visual ROI's). During the sensory task, participants had to attend the grating and detect contrast changes (left). When the sensory distractor was on the screen, participants had to ignore the grating while they were performing a concurrent VWM task (right). In addition to differences in attentional state, the number of trials collected in these different settings also differed (300+ for the sensory task, 108 for the sensory distractor shown during the memory task), likely causing differences in the signal-to-noise ratios. Thus, the differences between these two perceptual contexts may be due to multiple reasons.

		V1	V2	V3	V3AB	V4	IPS0	IPS1–3	LO
theoretical	sensory	t=2.4434 p=0.0319	t=5.1586 p<0.001	t=3.0225 p=0.0312	t=5.9727 p<0.001	t=3.3920 p=0.0307	t=2.6689 p=0.0305	t=0.3260 p=0.7496	t=4.7392 p<0.001
	memory	t=0.0319 p=0.9377	t=2.0768 p=0.0620	t=3.5862 p<0.001	t=4.0215 p=0.0317	t=2.9295 p<0.001	t=3.6337 p<0.001	t=0.7569 p=0.4694	t=0.8297 p=0.4059
behavior	sensory	t=4.3757 p<0.001	t=5.0921 p<0.001	t=4.1982 p<0.001	t=2.4459 p=0.0303	t=2.8264 p=0.0309	t=0.7673 p=0.4686	t=0.0201 p=0.969	t=2.6235 p=0.0318
	memory	t=3.2622 p=0.0315	t=4.1277 p<0.001	t=5.8139 p<0.001	t=6.0233 p<0.001	t=7.3358 p<0.001	t=5.3503 p<0.001	t=2.4268 p<0.001	t=1.6610 p=0.1568

Supplementary Table 1: Post-hoc statistics for two-sided paired t-tests from the theoretical input function based on the statistics in the natural world (in green) and from the psychophysical input function based on independent behavioral measurements (in blue). All significant cells are colored in a lighter shade for the purpose of quick visualization.