UNIVERSITY OF CALIFORNIA, SAN DIEGO

Information content of visual representations depends on attentional priority and working memory load

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Neurosciences with a specialization in Computational Neurosciences

by

Thomas Christopher Sprague

Committee in charge:

Professor John Serences, Chair
Professor Timothy Gentner
Professor Eric Halgren
Professor Donald MacLeod
Professor Tatyana Sharpee

2016

The dissertation of Thomas Christopher Sprague is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2016

DEDICATION

To Cindy & Roy Sprague, Jr., Pat & Tom Nickels, and Pat & Roy Sprague, Sr.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

EEG: electroencephalogram

FDR: false discovery rate

fMRI: functional magnetic resonance imaging

hMT+: human motion processing complex (analog of macaque middle temporal area)

IEM: inverted encoding model

IPS: intraparietal sulcus

MT: middle temporal area

ROI: region of interest

sPCS: superior precentral sulcus

V1: primary visual cortex

WM: working memory

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work would not be possible without my family and friends. Your support and love made the good times better and the hard times easier. My parents – Cindy and Roy – grandparents – Pat and Tom, and Pat and Roy – have always believed I could achieve success in whatever I chose to do. That belief, confidence, and trust is largely responsible for my successes thus far. My friends from college – Tate, Margaret, Chris, Patrick, Alison, Caitlin and Sarah – and from graduate school – Andy, Rachel, Maya, Panid, Erik, Miranda, as well as the entire Neurosciences Graduate Program student body – keep me sane, and insane, in equal and appropriate amounts. My lab members – Anna, Chaipat, Eddie, Javi, Mary, Nuttida, Roseanne, Sameer, Sirawaj, Steph, Tiffany, and Vy – are the most capable, caring, and inviting coworkers one could ask for. Their friendship and scientific expertise played an enormous role in completing this work. My committee – Tim Gentner, Eric Halgren, Don MacLeod, and Tatyana Sharpee – gave invaluable guidance throughout my time at UCSD. Their recommendations and ideas shaped the contents of this dissertation, as well as my thinking as a scientist, and I'm far better for it. And my adviser – John Serences – has been the model of a mentor. Kind, smart, caring, understanding, and considerate only begin to describe him. Most importantly, in addition to incredible scientific mentorship, his ability to maintain a comfortable and happy balance between an academic life and a fulfilling family life make him a wonderful role model. I'm so happy to have made the decision to join John's lab years ago, and can't imagine what graduate school would have been like otherwise.

Finally, my fiancé – Samantha – gives meaning to all I do. She keeps me calm and driven, and always makes sure I don't lose sight of what is really important. I look forward more and more every day to our next set of journeys – first as post-doctoral fellows in New

York, then marriage in Chicago, then wherever life carries us next. It's been a wonderful ride so far, and is getting better every day.

Chapter 1, in full, is a reprint of the material as it appears in a review entitled "Attention mitigates information loss in small- and large-scale neural codes" published in Trends in Cognitive Sciences 2015. Sprague, Thomas C.; Saproo, Sameer; Serences, John T., Cell Press, 2015. The dissertation author was the primary author of the manuscript. Supported by National Institutes of Health (NIH) grant R01-MH092345 and a James McDonnel Foundation Scholar Award to J.T.S., and NIH grant T32-MH20002-15 and a National Science Foundation (NSF) Graduate Research Fellowship to T.C.S. We thank Vy Vo for comments on an earlier version of this manuscript.

Chapter 2, in full, is a reprint of the material as it appears in an article entitled "Attention modulates spatial priority maps in the human visual, parietal, and frontal cortices" published in Nature Neuroscience 2013. Sprague, Thomas C.; Serences, John T., Nature Publishing Group, 2013. The dissertation author was the primary author of the manuscript. We thank Ed Vul and Sirawaj Itthipuripat for assistance with statistical methods and Miranda Scolari and Mary Smith for assistance with parietal cortex mapping protocols. This work was supported by a National Science Foundation Graduate Research Fellowship to T.C.S. and by US National Institutes of Health grant R0I MH-092345 and a James S. McDonnell Scholar Award to J.T.S.

Chapter 3, in full, is a reprint of the material as it appears in a report entitled "Reconstructions of information in visual spatial working memory degrade with memory load" published in Current Biology 2014. Sprague, Thomas C.; Ester, Edward F.; Serences, John T., Cell Press, 2014. The dissertation author was the primary author of the manuscript. We thank Miranda Scolari and Mary Smith for assistance in developing parietal cortex mapping protocols, Anna Byers for assistance with data collection, Sirawaj Itthipuripat, Vy

Vo, and Alexander Heitman for discussion, and Sirawaj Itthipuripat, Vy Vo, and Stephanie Nelli for comments on the manuscript. This work was supported by a NSF Graduate Research Fellowship to T.C.S., NIH T32-MH020002-12 to E.F.E., and NIH R01 MH-092345 to J.T.S.

Chapter 4, in full, is a manuscript entitled “Restoring latent visual working memory representations in human cortex” submitted for publication. Sprague, Thomas C.; Ester, Edward F.; Serences, John T. The dissertation author was the primary investigator and author of the manuscript. We thank Edward Awh, Brad Postle, Sirawaj Itthipuripat, Stephanie Nelli, Samantha Scudder, and Vy Vo for helpful comments on a draft of this manuscript, and Haider Al-Hakeem for helpful discussions and assistance with data collection. Funded by NIH R01-MH092345 and a James S. McDonnell Foundation Scholar Award to J.T.S., NIH T32-MH20002 to T.C.S. and E.F.E., and an NSF Graduate Research Fellowship to T.C.S.

VITA

2010, Bachelor of Arts with Honors, Cognitive Science, Rice University

2016, Doctor of Philosophy, Neurosciences with a specialization in Computational Neurosciences, University of California, San Diego

PUBLICATIONS

Sprague, T.C., Serences, J.T. "Attention modulates spatial priority maps in the human visual, parietal and frontal cortices." 2013. *Nature Neuroscience*: 16(12): 1879-87.

Itthipuripat, S., Garcia J.O., Rungratsameetaweemana N., **Sprague T.C.**, Serences J.T. "Changing the spatial scope of attention alters patterns of neural gain in human cortex." 2014. *The Journal of Neuroscience*: 34(1): 112-23.

Sprague, T.C., Ester, E.F., Serences, J.T. "Reconstructions of information in visual spatial working memory degrade with memory load." 2014. *Current Biology*: 24(18): 2174-80.

Sprague, T.C., Saproo, S., Serences, J.T. "Visual attention mitigates information loss in small- and large-scale neural codes." (Review). 2015. *Trends in Cognitive Sciences*: 19(4): 215-26.

Sprague, T.C., Serences, J.T. "Using human neuroimaging to examine top-down modulations of visual perception" in *An introduction to model-based cognitive neuroscience.* Editors: Forstmann, B., Wagenmakers, E.J. (Review). 2015.

Ester, E.F., **Sprague, T.C.**, Serences, J.T. "Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory." 2015. *Neuron*: 87(4): 893-905.

ABSTRACT OF THE DISSERTATION

Information content of visual representations depends on attentional priority and working memory load

by

Thomas Christopher Sprague

Doctor of Philosophy in Neurosciences, with a specialization in Computational Neurosciences

University of California, San Diego, 2016

Professor John Serences, Chair

Though our experience of the world often appears rich and detailed, less information is immediately accessible than our intuition would suggest. When viewing a complex scene – such as a crowded city street – information about irrelevant features of the environment is lost due to noisy neural processing, while information about relevant features (those selected by visual attention) is spared. Similarly, behavioral experiments demonstrate that when even modest amounts of information must be held briefly in mind (in visual working memory), the amount of available information

about each item is diminished, and this available information decreases with increasing information load. In what manner do visual representations across large-scale neural activity patterns support these behavioral information processing limits? In three studies, we examined the fidelity with which human cortical neural activation patterns measured with functional magnetic resonance imaging represent visual information. To this end, we developed a novel analysis technique whereby we reconstruct images of visual stimuli using neural activation patterns measured over entire brain regions. Using this technique, we established that the neural representation of a relevant visual stimulus is enhanced in its amplitude over a noisy baseline in several visual and parietal cortical regions, suggestive of an increase in the representation's information content. Subsequently, we demonstrated that under conditions where no information is available in a display, the maintenance of a larger number of items in visual working memory is accompanied by a degradation in each item's representation amplitude, indicative of lower population-level information content. Finally, we evaluated the relationship between these two findings by directing participants to attend to one of several items held in visual working memory. Surprisingly, we discovered that degraded representations can recover with visual attention, and the degree of recovery was related to behavioral task performance. Such recovery of degraded information suggests that additional information must be available to the system but invisible to our measurements before attention is allocated. Together, these results demonstrate that behavioral limits on information processing are related to the fidelity with which visual information is represented in large-scale neural codes.

Chapter 1:

Attention mitigates information loss in small- and large-scale neural codes

Visual attention mitigates information loss in small- and large-scale neural codes

Thomas C. Sprague[1], Sameer Saproo[2], and John T. Serences[1,3]

[1] Neurosciences Graduate Program, University of California San Diego, La Jolla, CA 92093-0109, USA
[2] Department of Biomedical Engineering, Columbia University, New York, NY, USA
[3] Department of Psychology, University of California San Diego, La Jolla, CA 92093-0109, USA

The visual system transforms complex inputs into robust and parsimonious neural codes that efficiently guide behavior. Because neural communication is stochastic, the amount of encoded visual information necessarily decreases with each synapse. This constraint requires that sensory signals are processed in a manner that protects information about relevant stimuli from degradation. Such selective processing – or selective attention – is implemented via several mechanisms, including neural gain and changes in tuning properties. However, examining each of these effects in isolation obscures their joint impact on the fidelity of stimulus feature representations by large-scale population codes. Instead, large-scale activity patterns can be used to reconstruct representations of relevant and irrelevant stimuli, thereby providing a holistic understanding about how neuron-level modulations collectively impact stimulus encoding.

Visual attention and information processing in visual cortex

Complex visual scenes contain a massive amount of information. To support fast and accurate processing, behaviorally-relevant information should be prioritized over behaviorally-irrelevant information (Figure 1). For example, when approaching a busy intersection while driving it is crucial to detect changes in your lane's traffic-light rather than one nearby to prevent a dangerous collision. This capacity for selective information processing, or selective visual attention, is supported by enhancing the amount of information that is encoded about relevant visual stimuli relative to the amount of information that is encoded about irrelevant stimuli. Importantly, understanding how relevant visual stimuli are represented with higher fidelity requires considering more than only the impact of attention on the response properties of individual neurons. Instead, examining activity patterns across large neural populations can provide insights into how different unit-level attentional modulations synergistically improve the quality of stimulus representations in visual cortex.

In the scenario above, neurons can undergo several types of modulation in response to the relevant light compared to one that is irrelevant: response amplitudes can increase (response gain), responses can become more

Corresponding authors: Sprague, T.C. (tsprague@ucsd.edu); Serences, J.T. (jserences@ucsd.edu).
Keywords: vision; visual attention; stimulus reconstruction; information theory; neural coding..

1364-6613/
 http://dx.doi.org/10.1016/j.tics.2015.02.005

Glossary

Bit: unit of entropy (base 2).
Decoder: algorithm whereby a feature or features about a stimulus (orientation, spatial position, stimulus identity, etc.) is/are inferred from an observed signal (spike rate, BOLD signal). Typically, the signal is multivariate across many neurons/voxels, but in principle a decoder can use a univariate signal.
Dynamic range: the set of response values a measurement unit can take. An increase in the response gain of a unit will increase the range of possible response values, and this will increase its entropy.
Encoding model: a description of how a neuron (or voxel) responds across a set of stimuli (e.g., a spatial receptive field can be a good encoding model for many visual neurons and voxels, see Box 2).
Entropy: a measure of uncertainty in a random process, such as a coin flip or observation of a neuron's spike count. A variable with a single known value will have 0 entropy, whereas a fair coin would have >0 entropy (1 bit).
Feature space: after reconstruction using the IEM technique, data exist in feature space, with each datapoint being defined by a vector of values corresponding to the activation of a single feature-selective population response (e.g., orientation, spatial position); common across all participants and visual areas.
Inverted encoding model (IEM): when encoding models are estimated across many measurement units, it may be possible to use all encoding models to compute a mapping from signal space into feature space which allows reconstruction of stimulus representations from multivariate patterns of neural activity across the modeled measurement units (Box 2).
Multivariate: when analyses are multivariate, signals from more than one measured unit are analyzed; utilizing information about the pattern of responses across units rather than simplifying the data pattern by taking a statistic over the units (e.g., mean).
Mutual information: the amount of uncertainty about a variable (e.g., state of the environment) that can be reduced by observation of the state of another random variable (e.g., the voxel or the neuron's response).
Noise entropy: variability in one signal that is unrelated to changes in another signal.
Receptive field (RF): region of the visual field which, when visually stimulated, results in a response in a measured neuron or voxel (population RF, or pRF).
Tuning function (TF): the response of a neuron or voxel to each of several values of a feature, such as orientation or motion direction.
Signal entropy: variability in one signal that is related to changes in another signal.
Signal space: data as measured exist in signal space, with a dimension for each measurement unit (fMRI voxel, EEG scalp electrode, electrocorticography subdural surface electrode, animal single cell firing rate, or calcium signal); cannot be directly compared across individual subjects without a potentially suboptimal coregistration transformation.

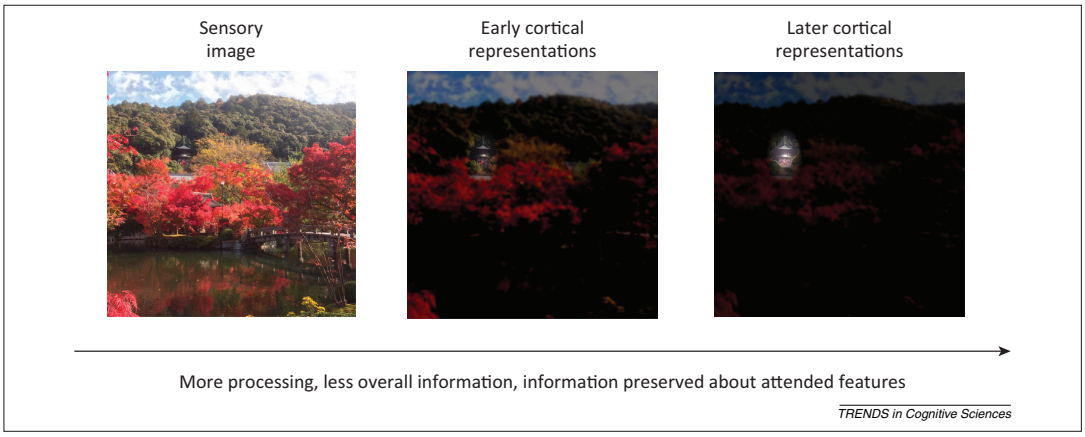


Figure 1. Attention filters behaviorally-relevant information. When viewing a complex natural scene (left), visual processing by a noisy neural system will necessarily result in an overall loss of information. If your eyes were fixated on the center of the image, but you were directing attention to the temple nestled among the trees near the top, information about the attended temple would be selectively preserved from degradation by noisy neural processing – such that, even at successively later stages of computation, information about the attended location and/or features of the image is still maintained, despite substantial loss of information about unattended components of the image (right panel).

reliable, and receptive field properties can shift (e.g., some neurons will shift their spatial receptive field to encompass the attended light). Thus, neural responses associated with attended stimuli generally have a higher signal-to-noise ratio and are more robust compared to responses evoked by unattended stimuli. Accordingly, the behavioral effects associated with visual attention are thought to reflect these relative changes in neural activity: when stimuli are attended, participants exhibit decreased response times, increased discrimination accuracy, and improved spatial acuity ([1–3] for reviews).

This selective prioritization of relevant over irrelevant stimuli follows from two related principles of information theory [4–7] (Box 1). First, the data-processing inequality [7] states that information is inevitably lost when sent via noisy communication channels, and that lost information cannot be recaptured via any amount of further processing. Second, the channel capacity of a communication system is determined by the amount of information that can be transmitted and received, and by the degree to which that information is corrupted during the process of transmission. In the brain, channel capacity is finite because there is a fixed (albeit large) number of neurons and because synaptic connections are stochastic such that information cannot be transmitted with perfect fidelity. Given this framework, different types of attention-related neural modulations can be viewed as a concerted effort to attenuate the unavoidable decay of behaviorally-relevant information as it is passed through subsequent stages of visual processing [8,9]. This framing also highlights the importance of understanding how attention differentially impacts responses across neurons, and, more importantly, how these modulations at the single-unit level interact to support population codes that are more robust to the information-processing limits intrinsic to the architecture of the visual system.

Box 1. Information content of a neural code

Information is related to a reduction in uncertainty [4,5,7]. A code is informative insofar as measurement of one variable (e.g., the firing rate of a single neuron) reduces uncertainty about another variable (e.g., feature of a stimulus). The amount of uncertainty in a random variable (e.g., the outcome of a coin toss or the spiking output of a cell) can be quantified by its entropy, which increases with increasing randomness. Mutual information (MI) is a measure of the reduction in uncertainty of one variable after knowing the state of another variable. MI would be zero for independent variables (e.g., two different coins), whereas MI would be high for two variables that strongly co-vary.

If a neuron noisily responds at the same level to each feature value, then the MI between the state of the stimulus and the state of the neuron's response is low because signal entropy (variability associated with changes in the stimulus) is low and noise entropy (variability unrelated to changes in the stimulus) is high (Figure IA). Instead, if the neuron exhibits a Gaussian-like orientation tuning function (TF; Figure IB,C), then MI is higher because more of the variability in the neuron's response is related directly to changes in the state of the stimulus. In this latter case, if the amplitude of the neuronal TF increases while noise remains approximately constant, then the ratio of signal entropy to noise entropy increases, resulting in greater MI between the neuron's response and the stimulus orientation. However, if the tuning width of the orientation TF changes, this could result in either an increase or decrease in the information about the stimulus, and would be contingent upon several factors such as the original tuning width, noise structure, dimensionality of the stimulus, and the responses of other neurons (Figure IF–H) [37,126–128]. For a widely-tuned neuron, a decrease in tuning width would result in an increase in signal entropy relative to noise entropy, increasing the information content of the neuron about orientation. At the other extreme, for a neuron perfectly tuned for a single stimulus value, with noisy baseline responses to other values, a broadening in tuning would result in greater variability associated with stimulus features, and consequently greater information (Figure IF–H). Thus, an increase in the amplitude of a neural response (under simple noise models) will increase the dynamic range and entropy, whereas a change in tuning width can either increase or decrease the information content of a neural code.

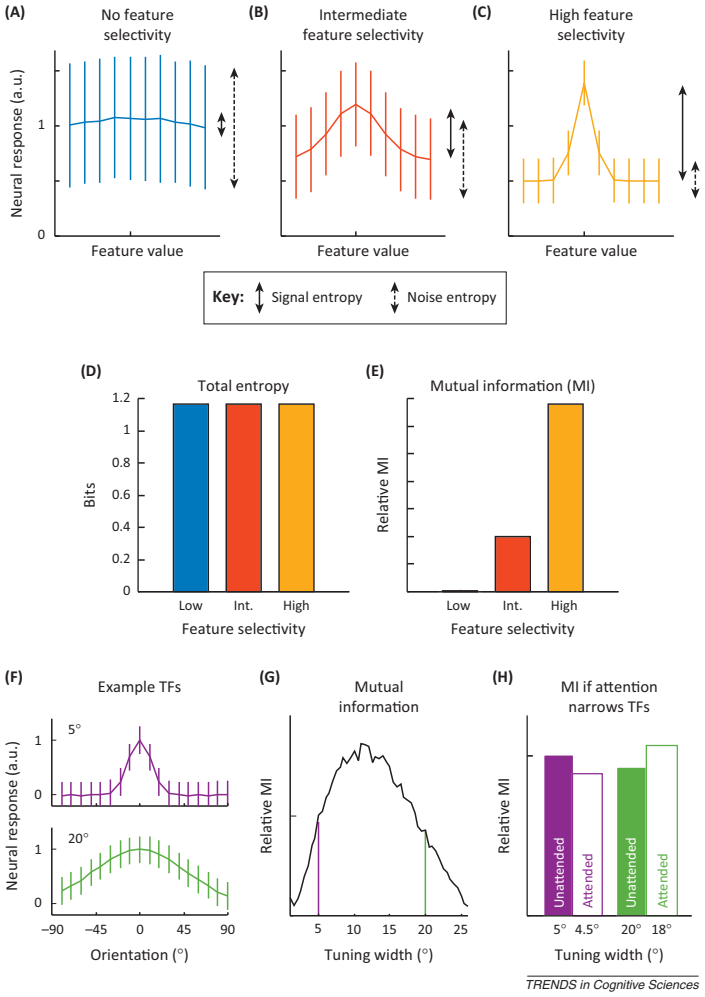


Figure I. Comparison of unit-level information content (mutual information). For a cartoon example of neural tuning functions (TF; measured in arbitrary units, a.u.) with no **(A)**, intermediate **(B)** or high **(C)** feature selectivity, the overall entropy of the neural response (given equal stimulus probability) is the same **(D)** but the mutual information (MI) between the neural response and the feature value **(E)** is much higher when the neuron exhibits selectivity [e.g., (C)] than when it is non-selective [e.g., (A)], with intermediate (Int.) selectivity [e.g., (B)] falling in between. In cases where attention narrows the selectivity of a unit, whether such a change results in improved information content depends on the original selectivity before narrowing [127,128]. **(F)** A cartoon example of neural TFs for stimulus orientation with different tuning widths (top, narrow tuning width/high selectivity; bottom, large tuning width/low selectivity). **(G)** When MI is plotted as a function of tuning width, it is apparent that any changes in tuning width that are associated with attention can either increase or decrease MI depending on the original tuning width. Vertical lines indicate tuning width values of example TFs in (F). **(H)** For a modest attention-related narrowing in tuning (e.g., ~10% reduction in tuning width), a unit with a small tuning width may actually carry less information with attention, whereas a unit with a large tuning width would carry more information. Note: panels (F–H) do not equate total entropy across tuning widths [as in panels (A–C)], accounting for the differences in MI plots between panel (E) and panels (G,H).

With this goal in mind, we first provide a selective overview of recent studies that examine attentional modulations of single measurement units (e.g., single neurons or single fMRI voxels) in visual cortex, with a focus on changes in response amplitude and shifts in spatial sensitivity profiles. We then introduce a framework for evaluating how attention-induced changes in large-scale patterns of activity can shape information processing to counteract the inherent limits of stochastic communication systems. This approach emphasizes reconstructing representations of sensory information based on multivariate patterns of neural signals and relating the properties of these reconstructions to changes in behavioral performance across task demands.

Attention changes the response properties of tuned neurons

Single-neuron firing rates in macaque primary visual cortex (V1) [10–12], extrastriate visual areas V2 and V4

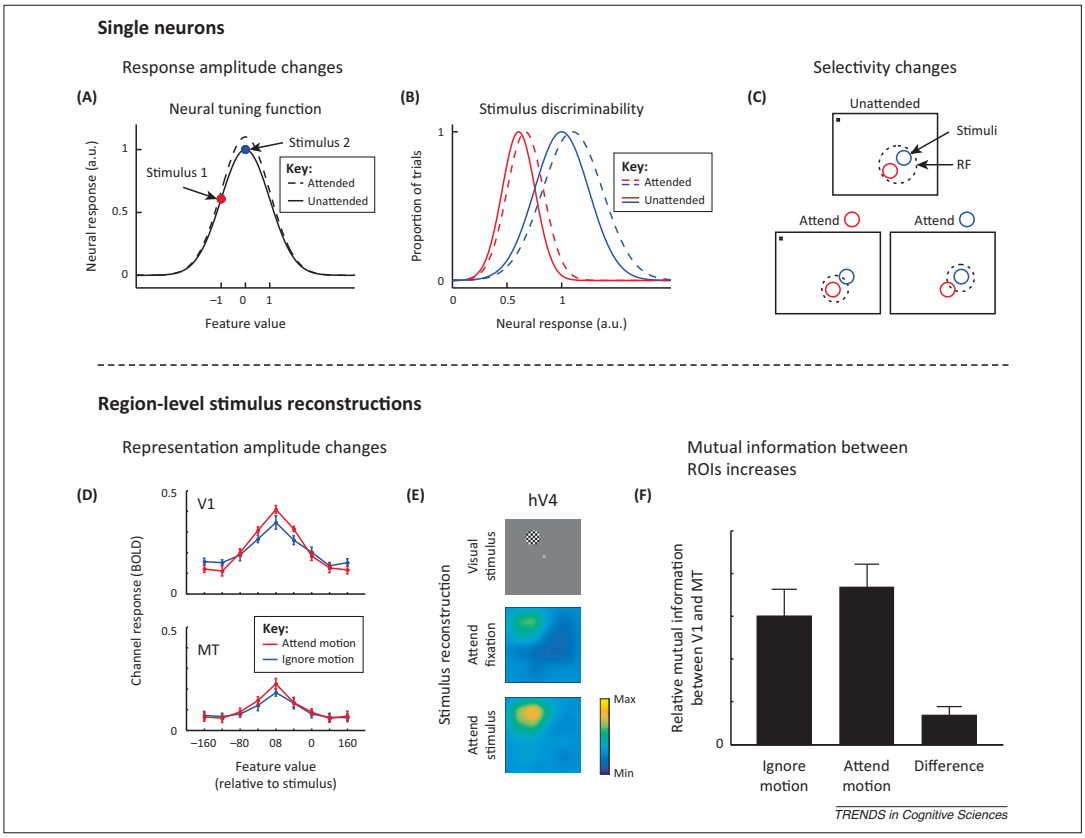


Figure 2. Attention improves the information content of small- and large-scale neural codes. When attention is directed towards a stimulus presented to a feature-selective neuron, several types of responses are commonly observed. **(A)** Response amplitudes often increase, which increases the dynamic range of the response, and accordingly improves the ability of the neuron to discriminate between two stimuli **(B)**. This increased dynamic range enables improved discrimination of multiple stimulus feature values. **(C)** Many neurons show changes in receptive field (RF) properties with attention such that the spatial profile of their response is focused around an attended stimulus placed inside the RF. By contrast, population-level stimulus reconstructions (Box 2) enable assessment of the net impact of all unit-level response changes with attention. **(D)** When participants are instructed to attend to the direction of motion of a moving dot stimulus (as opposed to its contrast), the amplitude of motion direction-selective responses increases in both V1 and middle temporal area (MT). **(E)** When participants attend to a flickering checkerboard disc, the reconstructed stimulus image (Box 2) has a higher amplitude than when they attend to the fixation point, especially in extrastriate visual regions of interest (ROIs) such as human area V4 (hV4). **(F)** When motion is attended [see (D)] the mutual information between V1 and MT is greater than when stimulus contrast is attended, suggesting that attention maximizes the transfer of relevant information between brain regions at a population level. Panels (A–C) are cartoon examples. Panels (D) and (F) were adapted from [9] with permission from the Society for Neuroscience; panel (E) was adapted from [76] with permission from Nature Publishing Group, colormap adjusted.

[12–19], motion-sensitive middle temporal area (MT) [20–22], lateral intraparietal cortex (LIP) [23–27], and frontal eye fields (FEF) [28–31] have all been shown to reliably increase when either a spatial position or a feature of interest is attended (Figure 2A). Heightened neural activity can facilitate the propagation of responses to downstream areas, leading to successively weaker distracter-associated responses compared to target-associated responses. However, even though most studies focus on increases in mean firing rates, many studies also report that a substantial minority of cells show systematic decreases in firing rates with attention (particularly in excitatory cells, e.g., [32]), an important issue when considering population-level neural codes that we revisit below.

In addition to measuring attentional modulations in response to a fixed stimulus set, researchers have also parametrically varied stimuli while an animal maintains a constant focus of attention to measure changes in feature tuning functions (TFs) or spatial receptive fields (RFs; that is, the response profile of a neuron to each member of a set of stimuli, see Glossary). When an animal is cued to attend to a visual feature, such as orientation [15,16], color [33], or motion direction [34], neurons tuned to the cued feature tend to respond more, while those tuned to uncued features tend to respond less [35]. This selective combination of response gain and suppression results in a larger range of possible firing rates, or a larger dynamic range, and thus increases encoding capacity such that different features will evoke a more easily separable neural response (Figure 2B) [36,37]. This increase in encoding capacity with an increase in dynamic range is analogous to switching between a binary and a greyscale image (e.g., a barcode and a black-and-white

photograph): the number of states each pixel can take increases, meaning that more states are discriminable.

When attending to a particular spatial position, the spatial RF of many neurons can also shift to accommodate the attended position in V4 [38], MT [39–42], and LIP [43], and the endpoint of a saccadic eye movement in V4 [44] and FEF [45]. In MT, for example, RFs shrink around the locus of attention when animals are cued to attend to a small region within a neuron's spatial RF (Figure 2C) [13,39,40]. However, when attention is focused immediately outside the penumbra of a neuron's spatial RF, the RF shifts and expands towards the focus of attention [41]. Finally, the tuning of V4 neurons to orientation and spatial frequency (that is, their spectral receptive field) can undergo shifts towards an attended target stimulus when an animal is viewing natural images [46]. These changes in the size and position of spatial RFs – coupled with increases in response amplitude – may lead to a more-robust population code via an increase in the number of cells that respond to relevant features ([2,47,48] for review). For example, an increase in single cell firing rates, coupled with a shift in the selectivity profile of surrounding spatial RFs towards the locus of attention, should generally increase the overall entropy of a population code, and thereby the quality of information encoded about a relevant stimulus (Box 1).

Attentional modulation of large-scale populations

Thus far we have discussed attentional modulations measured from single neurons in behaving monkeys. However, perception and behavior are thought to more directly depend on the quality of large-scale population codes [36,49,50], and it is therefore also necessary to assay how these small-scale modulations jointly impact the information content of larger-scale neural responses. For example, human neuroimaging methods including fMRI and EEG provide a window into the activity of large-scale neural populations [51–53], enabling the assessment of attention-related changes in voxel- or electrode-level signals that reflect the aggregate responses of all constituent neurons.

The firing-rate increases observed in single neurons are echoed by attention-related increases in fMRI blood oxygen level-dependent (BOLD) activation levels [54–59] and amplitude increases in stimulus-evoked EEG signals [60–65]. For example, when attention is directed to one of several stimuli on the screen, the mean BOLD signal measured from visual cortical regions of interest (ROIs) increases [55–59,66,67]. In addition, when fMRI voxels are sorted based on their selectivity for specific features such as orientation ([37,68,69], see also [70]), color [71], face identity [72], or spatial position [73], attention has the largest impact on voxels that are tuned to the attended feature value.

In addition to changes in response amplitude, recently developed techniques can also assess changes in the selectivity of voxel-level tuning functions across different attention conditions. One newly developed method has been used to evaluate how the size of voxel-level population receptive fields (pRFs) changes with attentional demands [74,75]. For example, several studies have measured pRF size as participants view a display consisting of a central fixation point and a peripheral visual stimulus that is used to map the pRF. On different trials, participants either attend to the peripheral mapping stimulus, or they ignore the mapping stimulus and instead attend to the central fixation point. Attending to the peripheral mapping stimulus increases the average size of voxel-level pRFs measured from areas of extrastriate cortex where single-neuron RFs are relatively large. However, no such size modulations are observed in primary visual cortex where single-neuron RFs are smaller [76–78]. At first, this result appears to conflict with neurophysiology studies showing that single-neuron RFs can either shrink or expand depending on the spatial relationship between the neuron's RF and the focus of attention (see above and Figure 2C) [39–42]. However, the response of a voxel reflects the collective response of all single neurons that are contained in that voxel. As a result, when the attended mapping stimulus was anywhere in the general neighborhood of the voxel's spatial RF, many single-neuron RFs within the voxel likely shifted towards the attended stimulus. In turn, this shifting of single-neuron RFs towards attended stimuli in the vicinity of the RF of a voxel should increase the area of visual space over which the voxel would respond (i.e., it would increase the size of the pRF compared to when the fixation point is attended to, and these neuron-level shifts would not occur).

The above studies examined how voxel-level pRFs change when the RF mapping stimulus is attended. A complementary line of work has addressed how attention to a focused region of space alters pRFs measured using an unattended mapping stimulus. In these studies attention was directed either to the left or the right of fixation while a visual mapping stimulus was presented across the full visual field. The center of voxel-level pRFs shift towards the locus of attention [79]; however, because the authors do not report whether pRFs also change in size, it is challenging to fully interpret how shifting the center of a pRF would support enhanced encoding of attended information (Box 1, Figure 3). Furthermore, another study found that increasing the difficulty of a shape-discrimination task at fixation leads to a shift of voxel-level pRFs away from fixation and also to an increase in their size. These modulations may thus result in a lower-fidelity representation of irrelevant stimuli in the visual periphery when a foveated stimulus is challenging to discriminate [80].

Functionally similar examples of information shunting have also been found in other domains: Brouwer and Heeger [71] demonstrated that directing attention to a colored stimulus during a color-categorization task narrows the bandwidth of voxel-level tuning functions, improving the discriminability of voxel responses for distinct colors when the color value is important for the task. Similarly, Çukur *et al.* [81] used a high-dimensional encoding model (which describes the visual stimulus categories for which each voxel is most selective) to show that attending to object categories shifts semantic space towards the attended target to increase the number of voxels responsive to a relevant category (see also [82,83]). Although analogous single-unit neural data are not available for comparison, these results support the notion that shifting the feature selectivity profile of a RF is an important strategy implemented by the visual system to combat

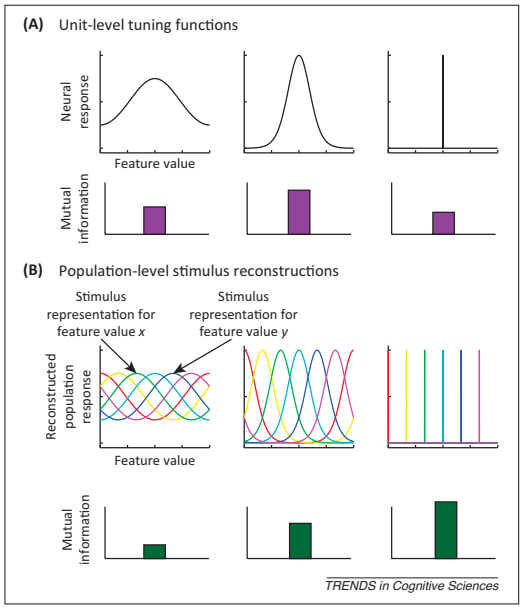


Figure 3. Information content of units and populations. **(A)** As described in Box 1, the mutual information (MI) between the response strength of a unit and the associated stimulus value is a non-monotonic function of its tuning bandwidth. A non-selective unit will have very low MI (as a result of little variability in response associated with variability in the stimulus feature), but a highly selective unit will also have low MI – because it has lower overall entropy. The particular selectivity bandwidth for which a unit has greatest MI about a stimulus feature depends on the shape of the tuning function (TF), noise properties, and the relative frequency of occurrence of different feature values. **(B)** By contrast, for population-level stimulus reconstructions, a narrower reconstructed stimulus representation is more informative because it reflects a greater level of discriminability between different stimulus feature values. Plotted are cartoon reconstructions of different values of a stimulus. Each color corresponds to a different feature value, and each point along each curve corresponds to the reconstructed activation of the corresponding population response (as in Figure 2D,E).

limited channel capacity via increasing the sampling density for relevant information (Figure 1).

With all experiments evaluating responses at the large-scale population level (e.g., the level of a single fMRI voxel), it is important to note that these macroscopic measurements reflect hemodynamic signals related to net changes in the response across hundreds of thousands or more neurons [51]. As a result, it is currently not possible to unambiguously infer whether attentional modulations measured at the scale of single voxels reflect changes in neuron-level feature selectivity or if voxel-level modulations instead reflect non-uniform changes in the response amplitude of neural populations within a voxel that are tuned to different feature values. Despite this limitation, large-scale measurement techniques such as fMRI can provide a unique perspective on the collective impact of small-scale single-neuron modulations on the fidelity of population codes, even though information about the specific pattern(s) of single-neuron modulations may be obscured. In turn, changes in voxel-level selectivity can support some important general inferences about the impact of attention on the encoding capacity of large-scale population responses (Box 1), which are not easily accessible via single-neuron recording methods.

Reconstructing region-level stimulus representations

The techniques used to measure single-neuron and single-voxel response profiles help us understand how changes at the level of single measurement units (whether single neurons or single voxels) can impact encoding capacity to facilitate perception and behavior. However, understanding how individual encoding units behave, either in isolation or at the level of a population average, is only a part of the picture. Indeed, different neurons and voxels are often modulated in different ways even in the context of the same experimental design: some units increase their response amplitude, others decrease [32]; some show (p)RF size increases, whereas others show decreases [39–42,76–78,80]. To understand how these apparently disparate modulations work together to impact the quality of region-level population codes, multivariate methods can be used to directly infer changes in the overall information content of neural response patterns.

An emerging means of evaluating information content of population-codes is via stimulus reconstruction (Box 2). Although population-level reconstruction methods have

Box 2. Inverted encoding models enable evaluation of aggregate effects of multiple unit-level response changes on the quality of a neural code

When firing rates of single units or activation levels of single voxels are measured in response to several different stimulus values (e.g., the orientation of a grating or the position of a stimulus on the screen), it is possible to fit an encoding model to the set of measured responses as a function of feature value. Such an encoding model describes how the neuron or voxel responds to different values of a stimulus, and an accurate encoding model will predict how the neuron or voxel would respond to a novel stimulus value. For example, the best-fit encoding model for an orientation-selective unit would be a circular Gaussian model (Figure IA), whereas the encoding model for a spatially selective unit would be characterized by a 2D Gaussian model (Figure IB). Note that the encoding models need not be visual: measuring firing rates of hippocampal neurons in rodents as they forage for food often reveals a particular region of the environment in which the neuron fires – its place field – which could potentially be described by a 2D Gaussian encoding model for spatial position within the environment.

While the process of estimating encoding models for many single units or single voxels across the brain allows inferences to be made about the manner in which information is measured, computed, and transformed across different stages of processing, the approach remains massively univariate: all encoding models are estimated in isolation, and inferences about neural processing are based on changes in these univariate encoding models in aggregate [76–80]. By contrast, the inverted encoding model approach (IEM) utilizes the pattern of encoding models estimated across an entire brain region (e.g., primary visual cortex) to reconstruct the region-level representation of a stimulus given a measured pattern of activation across the measurement units (Figure IC). These approaches, as implemented presently, rest on assumptions of linearity and are only feasible for simple features of the environment (e.g., orientation, color, spatial position) for which encoding properties are relatively well understood. To date, this general approach has been used to accurately reconstruct feature representations from fMRI voxel activation patterns [9,71,76,88,89,94–101,124], patterns of spike rate in rodent hippocampus [84,119], and human EEG signals measured non-invasively at the scalp [90].

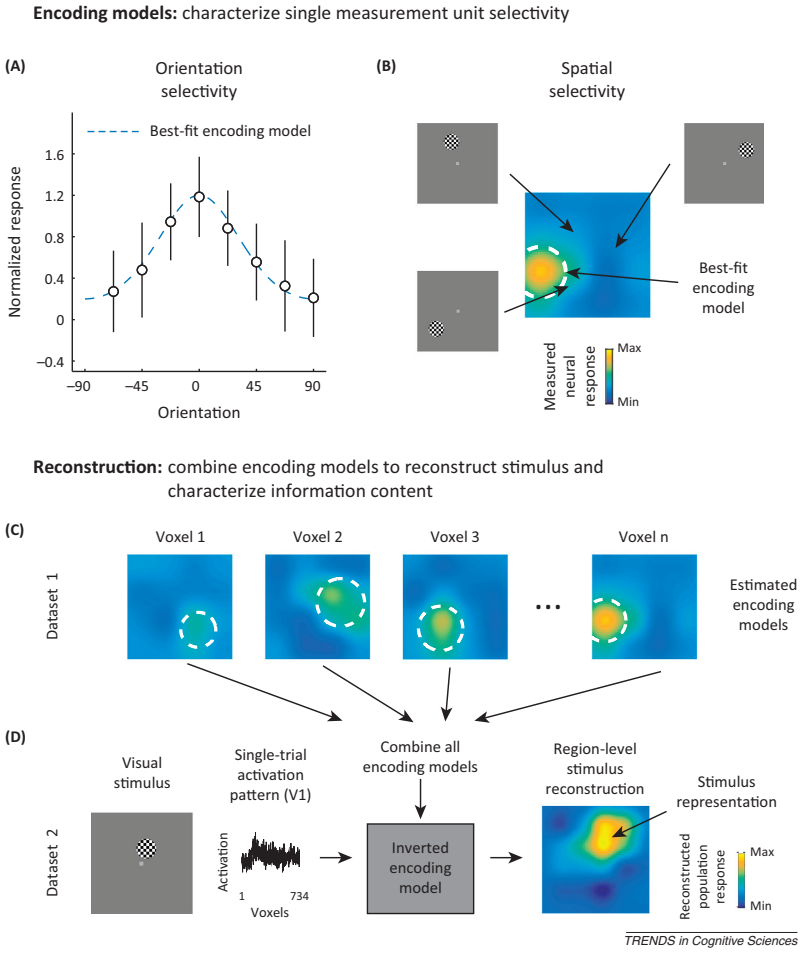


Figure I. Encoding models and stimulus reconstruction. Measurements of neural responses to different stimulus feature values – such as orientation **(A)** or spatial position **(B)** – often reveal selectivity for particular feature values (such as orientation TFs or spatial RFs). Such selectivity can be observed with single-unit firing rates, calcium transients, fMRI BOLD responses, or even scalp EEG. When encoding models are measured for many neurons/voxels/electrodes, it is possible to combine all encoding models to compute an inverted encoding model (IEM) **(C)**. This IEM allows a new pattern of activation measured using a separate dataset to be transformed into a stimulus reconstruction **(D)**, reflecting the population-level representation of a stimulus along the features spanned by the encoding model (here, visual spatial position). This reconstruction (right) reflects data from a single trial, which is inherently noisy. However, when many similar trials are combined (Figure 2E), high-fidelity stimulus representations can be recovered.

existed for decades [84,85], they have recently found widespread application in the field of human neuroimaging [71,75,86–94]. There are many variations on these methods, but all generally involve first estimating an encoding model that describes the selectivity profile (feature TF or spatial RF) of individual measurement units (e.g., single neurons or single voxels). Next, these encoding models are inverted and used to reconstruct the stimulus given a novel pattern of responses across the entire set of units (Box 2). Each computed stimulus reconstruction contains a representation of the stimulus of interest. Thus, for features such as color, orientation, or motion, the results of this procedure reflect a reconstruction of the response across a set of feature-selective populations [9,71,88–90,95–99]; for models based on spatial position, the results reflect reconstructed images of the visual scene viewed or remembered by an observer [75,76,86,100,101]. We call this broad framework – whereby patterns of encoding models are inverted to reconstruct stimulus representations – inverted encoding models (IEM).

The ability to reconstruct an image of visual stimuli based on population-level activation patterns can be used to assess how modulations observed at the level of measurement units are combined to jointly constrain the

amount of information encoded about a stimulus. In contrast to multivariate classification analyses that partition brain states into one of a set of discrete groups [91,102], or Bayesian approaches that generate an estimate of the most likely stimulus feature value [36,49,50,84,103–105], reconstruction enables the quantification of stimulus representations in their native feature space rather than in signal space. In turn, quantifying representations within these reconstructions supports the ability to evaluate attributes such as the amplitude or the precision of the encoded representation. Moreover, because the IEM reconstruction method involves an analog mapping from an idiosyncratic signal space that differs across individuals and ROIs into a common feature space, it is possible to directly compare quantified properties of stimulus reconstructions as a function of attentional demands. This approach thus complements previous efforts to establish the presence of stimulus-specific information by decoding which stimulus of a set was most likely to have caused an observed pattern of activation [75,87,106,107].

Because region-level stimulus reconstructions exploit information contained in the pattern of responses across all measurement units, they may be more closely linked to behavioral measures than to the responses of single neurons or even to mean response changes across a small sample of neurons or brain regions [91,97,108–110]. In addition, representations within these region-level reconstructions can be subjected to similar information theoretic analyses as described in Box 1 [9]. Instead of comparing how the response of a small sample of neurons changes with attention (Figure 3A), it is possible to evaluate how all co-occurring response modulations constrain the ability of a neural population to encode relevant information about a stimulus (Figure 3B).

Although this approach can provide a unique perspective on the quality of large-scale population codes, stimulus-reconstruction methods come at the cost of simplifying assumptions about how information is encoded. For instance, IEMs for simple features (Box 2) will not account for information that is not explicitly modeled. Thus, an IEM for reconstructing spatial representations of simple stimuli [76,101] will not recover any information that was represented about features such as color or orientation, despite the known roles that many visual areas play in encoding these stimulus attributes.

Reconstructions as an assay of population-level information

Attention has been shown to induce a heterogeneous set of modulations at the level of single measurement units such as single cells or voxels. Variability in the magnitude or sign of attention effects is often treated as noise, and the impact of different types of attentional modulation on the quality of stimulus representations is usually not considered (e.g., the joint influence of both gain and bandwidth modulation). IEMs can be used to extend these unit-level results and to evaluate how all types of attentional modulation collectively influence the information content of large-scale population codes.

Similarly to the information content of single-unit responses, when the amplitude of a population-level stimulus reconstruction increases above baseline, then more of the variability in the reconstruction is directly linked with changes in the stimulus (i.e., there is an increase in signal entropy). Importantly, in contrast to single units (Figure 3A, see also Box 1), when a population-level stimulus reconstruction becomes more precise, the population may support more precise inferences about stimulus features by improving the discriminability of responses associated with different stimulus feature values (Figure 3B) (e.g., [71,96,97]). Such a change in stimulus reconstructions could be supported by changes in the selectivity of individual voxels/neurons, non-uniform application of neural gain across the population, or any combination of these response modulations at the unit level.

In one study where participants categorized colors, voxel-level tuning functions for hue narrowed and region-level reconstructions of color response profiles were more clustered in a neural color space compared to when color was irrelevant [71]. This result provides evidence for a neural coding scheme whereby relevant category boundaries for a given task are maximally separated. Because more of the variability in the population response should be associated with changes in the relevant stimulus dimension (greater signal entropy), this modulatory pattern should provide a more-robust population code that can better discriminate different categories in color space (Figure 3B, see also [111,112]).

Similarly, directing attention to the direction of a moving stimulus increased the amplitude of direction-selective representations in both V1 and MT relative to attending stimulus contrast (Figure 2D) [9]. This increase in the dynamic range of responses gives rise to an increase in the information content of the direction-selective representation in both areas (via an increase in signal entropy) when motion was relevant compared to when it was irrelevant. In addition, when attention was directed to motion, the efficacy of feature-selective information transfer between V1 and MT increased relative to when stimulus contrast was attended (Box 1; Figure 2F). This task-dependent increase in the transfer of information between brain regions suggests that attention not only modulates the quality of signals within individual cortical regions but also increases the efficiency with which representations in one region influence representations in another [9]. Although the precise mechanism for such information transfer remains unknown, changes in synaptic efficacy [11] or synchrony of population-level responses such as local field potentials (LFP) and/or spike timing [113–115] likely contribute.

Spatial attention can change the amplitude of single-neuron responses and their spatial selectivity (Figure 2A–C). One study examined how all these changes jointly modulate representations of a visual stimulus within spatial reconstructions of the scene [76]. Participants were asked to perform either a demanding spatial attention task or a demanding fixation task in the scanner. Their fMRI activation patterns were then used to reconstruct stimulus representations from several visual ROIs. Although individual voxel-level pRFs were found to increase in size with attention, no changes were found in the size of stimulus reconstructions with attention. This pattern of results indicates that attention

Box 3. Outstanding questions

- The IEM technique has recently been adapted for use with scalp EEG signals that can provide insights about the relative timing of attentional modulations of stimulus reconstructions with near-millisecond precision [90]. How do these signals measured with scalp electrodes carry information about features like orientation? And what other types of neural signals (such as two-photon *in vivo* calcium imaging [116]) can be used for image reconstruction via IEMs?
- Correlated variability among neurons is an important limiting factor in neural information processing [108,129–133]. How can this correlated variability be incorporated into the IEM approach, and what are the scenarios in which correlated variability helps and hurts the information content of a population code as measured via stimulus reconstructions?
- It is possible to compute region-level feature-specific reconstructions across each of the many visual field maps in cortex [134–136]. However, the role(s) of each of these visual field maps in supporting visual perception and behavior remains largely unknown. By comparing how properties of stimulus representations vary across different visual field maps with measures of behavioral performance, in combination with causal manipulations such as TMS, optogenetic excitation or inhibition of subpopulations of neurons, and electrical microstimulation, the relative contributions of each region's representation to behavioral output can be compared, and accordingly the role(s) of the region in visual behavior may be inferred.
- Application of IEMs for tasks requiring precise maintenance of, or attention to, visual stimulus features such as orientation, color, or motion direction [9,96,97,124] often reveal different results from those requiring attention to or maintenance of spatial positions [76,98,101]. Attending to a feature sharpens or shifts stimulus reconstructions in a manner well-suited for performing the task, whereas attending to a position enhances the amplitude of stimulus reconstructions over baseline. How do the circumstances in which stimulus reconstructions change in their amplitude differ from those in which reconstructions change their precision?

does not sharpen the region-level stimulus representations in this task. However, the study did reveal attention-related increases in the amplitude of stimulus representations (Figure 2E), and this corresponds to more information about the represented stimulus (greater signal entropy) above a noisy, uniform baseline (noise entropy, Box 1; Figure 3B). These results were echoed by a recent report that linked attention-related increases in the size and the gain of spatial pRFs in ventral temporal cortex with improved population-level information about stimulus position [77].

Concluding remarks and future directions

Selective attention induces heterogeneous modulations across single encoding units, and understanding how these modulations interact is necessary to fully characterize their impact on the fidelity of information coding and behavior. At present, analysis techniques that exploit population-level modulations have primarily been implemented with data from large-scale neuroimaging tools such as fMRI and EEG. Applying these analyses to other methods such as two-photon *in vivo* calcium imaging of neurons identified genetically [116,117] or by cortical depth [118], and electrophysiological recordings from large-scale electrode arrays in behaving animals and humans [119,120], will help to bridge gaps in our understanding of how the entire range of neuron-level attentional modulations are related to population-level changes in the quality of stimulus representations (Box 3). Furthermore, the development of improved modeling, decoding, and reconstruction methods as applied to both human neuroimaging and animal physiology and imaging data should enable new inferences about the mechanisms of attention in more complicated naturalistic

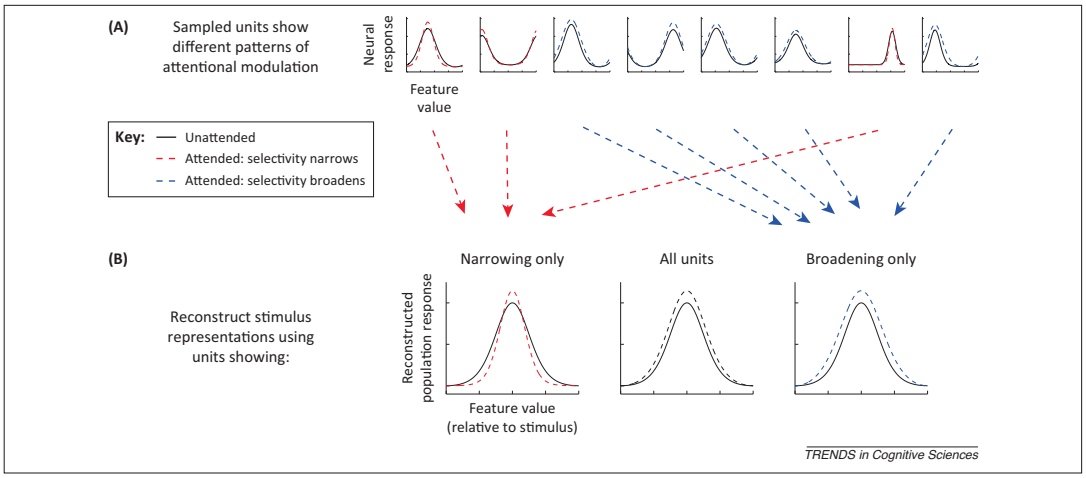


Figure 4. Using stimulus reconstructions to exploit and understand the net impact of heterogeneous response modulations. **(A)** Attentional modulations of different units are often heterogeneous, reflecting combinations of amplitude increases, baseline changes, and selectivity changes. Thus, even though the mean attentional modulations across units often point in the same direction across studies, there is substantial variability within a given sample of neurons (as shown here in simulated cartoon neural TFs), and any information that is encoded by this variability is usually ignored. **(B)** When using the inverted encoding model (IEM) technique, which combines modulations across all constituent units, it is possible to ascertain how different types of unit-level response modulations may contribute to stimulus reconstructions by selecting measurement units *post hoc* that exhibit one type or another of response modulation (e.g., only bandwidth narrowing or broadening) to compute stimulus reconstructions.

settings, potentially even during unrestrained movement [50,91–94,121].

In addition, associating neural modulations with changes in behavioral performance is crucially important as a gold-standard method for evaluating the impact of attention on the quality of perceptual representations. The importance of this brain–behavior link was recently highlighted by a study in which visual attention was correlated with the modulation of single-neuron activity in visual cortex (increased firing rate, among others). However, these modulations in visual cortex were unaffected even after attention-related improvements in behavior were abolished by the transient inactivation of the superior colliculus, an area that is thought to play an important role in attentional control [122,123]. This observation places an important constraint on how we consider different mechanisms of attention: attention results in changes to neural codes in visual cortex that should improve the information content about relevant stimulus features compared to irrelevant features (e.g., firing-rate modulations). However, these attention-related improvements in the quality of local sensory representations will not necessarily be transmitted to downstream areas, and thus may have little or no impact on behavior. Thus, neural responses, both at the neuronal and population levels, need to be systematically evaluated against changes in behavior to establish their overall importance in visual information processing [124].

In future work, one promising approach is to selectively lesion or alter the measured data, post-acquisition, by using only units that show particular encoding or response properties to compute stimulus reconstructions (e.g., measurement units with RFs near or far from the attended stimulus; measurement units that either increase or decrease in response amplitude or RF size [76]; Figure 4). Reconstructions computed using only measurement units with modulations most crucial for improving the fidelity of the neural code for attended stimuli versus unattended stimuli should be associated with increases in mutual information between reconstructions and attended stimuli relative to unattended stimuli, and this can be verified by comparing the information content of reconstructions to measures of behavioral performance across attention conditions [97,98,124]. Using a similar approach, a recent study evaluated the necessity of voxel-level attentional gain on population-level information about spatial position by artificially eliminating attention-related gain from their observed pRFs. They found that pRF gain is not necessary to improve position coding; changes in pRF size and position were sufficient [77]. Finally, the information content of region-level stimulus reconstructions computed in one brain region can be compared to the information content of those measured in other brain regions [125] at different points in time to determine how information is transformed across levels of the visual hierarchy. For example, reconstructions in V1 should primarily reflect information about relevant low-level sensory features, whereas downstream areas in the ventral temporal lobe should encode information about more-holistic object properties such as the features associated with relevant faces or scenes. Comparing successive reconstructions across multiple brain regions may highlight those features of visual scenes undergoing attentional selection, how they are selected, and what happens to features not selected (Figure 1).

By emphasizing the link between the information content of reconstructions across multiple stages of processing and measures of behavioral performance, a more complete picture will emerge about how differently tuned encoding units at each stage – and their associated constellation of attention-induced modulations – can give rise to a stable representation that is more closely linked with the overall perceptual state of the observer.

Acknowledgments

Supported by National Institutes of Health (NIH) grant R01-MH092345 and a James McDonnel Foundation Scholar Award to J.T.S., and NIH grant T32-MH20002-15 and a National Science Foundation (NSF) Graduate Research Fellowship to T.C.S. We thank Vy Vo for comments on an earlier version of this manuscript.

References

1 Carrasco, M. (2011) Visual attention: the past 25 years. *Vision Res.* 51, 1484–1525
2 Anton-Erxleben, K. and Carrasco, M. (2013) Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence. *Nat. Rev. Neurosci.* 14, 188–200
3 Gardner, J.L. (2014) A case for human systems neuroscience. *Neuroscience* Published online 2 July, 2014. http://dx.doi.org/10.1016/j.neuroscience.2014.06.052
4 Shannon, C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423
5 Cover, T. and Thomas, J. (1991) *Elements of Information Theory*, Wiley
6 Quian Quiroga, R. and Panzeri, S. (2009) Extracting information from neuronal populations: information theory and decoding approaches. *Nat. Rev. Neurosci.* 10, 173–185
7 Shannon, C.E. and Weaver, W.B. (1963) *The Mathematical Theory of Communication*, University of Illinois Press
8 Tsotsos, J.K. (1990) Analyzing vision at the complexity level. *Behav. Brain Sci.* 13, 423–445
9 Saproo, S. and Serences, J.T. (2014) Attention Improves transfer of motion information between V1 and MT. *J. Neurosci.* 34, 3586–3596
10 Herrero, J.L. *et al.* (2013) Attention-induced variance and noise correlation reduction in macaque V1 is mediated by NMDA receptors. *Neuron* 78, 729–739
11 Briggs, F. *et al.* (2013) Attention enhances synaptic efficacy and the signal-to-noise ratio in neural circuits. *Nature* 499, 476–480
12 Buffalo, E.A. *et al.* (2010) A backward progression of attentional effects in the ventral stream. *Proc. Natl. Acad. Sci. U.S.A.* 107, 361–365
13 Moran, J. and Desimone, R. (1985) Selective attention gates visual processing in the extrastriate cortex. *Science* 229, 782–784
14 Luck, S.J. *et al.* (1997) Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.* 77, 24–42
15 Motter, B.C. (1993) Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J. Neurophysiol.* 70, 909–919
16 McAdams, C.J. and Maunsell, J.H.R. (1999) Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J. Neurosci.* 19, 431–441
17 McAdams, C.J. and Maunsell, J.H.R. (2000) Attention to both space and feature modulates neuronal responses in macaque area V4. *J. Neurophysiol.* 83, 1751–1755
18 Cook, E.P. and Maunsell, J.H.R. (2002) Attentional modulation of behavioral performance and neuronal responses in middle temporal and ventral intraparietal areas of macaque monkey. *J. Neurosci.* 22, 1994–2004
19 Steinmetz, N.A. and Moore, T. (2014) Eye movement preparation modulates neuronal responses in area V4 when dissociated from attentional demands. *Neuron* 83, 496–506

20 Treue, S. and Maunsell, J.H.R. (1996) Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* 382, 539–541

21 Treue, S. and Maunsell, J.H.R. (1999) Effects of attention on the processing of motion in macaque middle temporal and medial superior temporal visual cortical areas. *J. Neurosci.* 19, 7591–7602

22 Treue, S. and Martinez-Trujillo, J.C. (1999) Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* 399, 575–579

23 Bisley, J.W. and Goldberg, M.E. (2003) Neuronal activity in the lateral intraparietal area and spatial attention. *Science* 299, 81–86

24 Bisley, J.W. and Goldberg, M.E. (2010) Attention, intention, and priority in the parietal lobe. *Annu. Rev. Neurosci.* 33, 1–21

25 Gottlieb, J.P. *et al.* (1998) The representation of visual salience in monkey parietal cortex. *Nature* 391, 481–484

26 Gottlieb, J. (2007) From thought to action: the parietal cortex as a bridge between perception, action, and cognition. *Neuron* 53, 9–16

27 Arcizet, F. *et al.* (2011) A pure salience response in posterior parietal cortex. *Cereb. Cortex* 21, 2498–2506

28 Thompson, K.G. *et al.* (2005) Neuronal basis of covert spatial attention in the frontal eye field. *J. Neurosci.* 25, 9479–9487

29 Armstrong, K.M. *et al.* (2009) Selection and maintenance of spatial information by frontal eye field neurons. *J. Neurosci.* 29, 15621–15629

30 Juan, C-H. *et al.* (2004) Dissociation of spatial attention and saccade preparation. *Proc. Natl. Acad. Sci. U.S.A.* 101, 15541–15544

31 Squire, R.F. *et al.* (2013) Prefrontal contributions to visual selective attention. *Annu. Rev. Neurosci.* 36, 451–466

32 Mitchell, J.F. *et al.* (2007) Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron* 55, 131–141

33 Motter, B.C. (1994) Neural correlates of attentive selection for color or luminance in extrastriate area V4. *J. Neurosci.* 14, 2178–2189

34 Martinez-Trujillo, J.C. and Treue, S. (2004) Feature-based attention increases the selectivity of population responses in primate visual cortex. *Curr. Biol.* 14, 744–751

35 Maunsell, J.H.R. and Treue, S. (2006) Feature-based attention in visual cortex. *Trends Neurosci.* 29, 317–322

36 Butts, D.A. and Goldman, M.S. (2006) Tuning curves, neuronal variability, and sensory coding. *PLoS Biol.* 4, e92

37 Saproo, S. and Serences, J.T. (2010) Spatial attention improves the quality of population codes in human visual cortex. *J. Neurophysiol.* 104, 885–895

38 Connor, C.E. *et al.* (1997) Spatial attention effects in macaque area V4. *J. Neurosci.* 17, 3201–3214

39 Womelsdorf, T. *et al.* (2006) Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nat. Neurosci.* 9, 1156–1160

40 Womelsdorf, T. *et al.* (2008) Receptive field shift and shrinkage in macaque middle temporal area through attentional gain modulation. *J. Neurosci.* 28, 8934–8944

41 Anton-Erxleben, K. *et al.* (2009) Attention reshapes center-surround receptive field structure in macaque cortical area MT. *Cereb. Cortex* 19, 2466–2478

42 Niebergall, R. *et al.* (2011) Expansion of MT neurons excitatory receptive fields during covert attentive tracking. *J. Neurosci.* 31, 15499–15510

43 Ben Hamed, S. *et al.* (2002) Visual receptive field modulation in the lateral intraparietal area during attentive fixation and free gaze. *Cereb. Cortex* 12, 234–245

44 Tolias, A.S. *et al.* (2001) Eye movements modulate visual receptive fields of V4 neurons. *Neuron* 29, 757–767

45 Zirnsak, M. *et al.* (2014) Visual space is compressed in prefrontal cortex before eye movements. *Nature* 507, 504–507

46 David, S.V. *et al.* (2008) Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. *Neuron* 59, 509–521

47 Desimone, R. and Duncan, J. (1995) Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222

48 Zirnsak, M. and Moore, T. (2014) Saccades and shifting receptive fields: anticipating consequences or selecting targets? *Trends Cogn. Sci.* 18, 621–628

49 Jazayeri, M. and Movshon, J.A. (2006) Optimal representation of sensory information by neural populations. *Nat. Neurosci.* 9, 690–696

50 Graf, A.B.A. *et al.* (2011) Decoding the activity of neuronal populations in macaque primary visual cortex. *Nat. Neurosci.* 14, 239–245

51 Logothetis, N.K. (2008) What we can do and what we cannot do with fMRI. *Nature* 453, 869–878

52 Lima, B. *et al.* (2014) Stimulus-related neuroimaging in task-engaged subjects is best predicted by concurrent spiking. *J. Neurosci.* 34, 13878–13891

53 Lopes da Silva, F. (2013) EEG and MEG: relevance to neuroscience. *Neuron* 80, 1112–1128

54 Kastner, S. and Ungerleider, L.G. (2000) Mechanisms of visual attention in the human cortex. *Annu. Rev. Neurosci.* 23, 315–341

55 Buracas, G.T. and Boynton, G.M. (2007) The effect of spatial attention on contrast response functions in human visual cortex. *J. Neurosci.* 27, 93–97

56 Murray, S.O. (2008) The effects of spatial attention in early human visual cortex are stimulus independent. *J. Vis.* 8, 2.1–11

57 Gandhi, S.P. *et al.* (1999) Spatial attention affects brain activity in human primary visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 96, 3314–3319

58 Scolari, M. *et al.* (2014) Functions of the human frontoparietal attention network: Evidence from neuroimaging. *Curr. Opin. Behav. Sci.* 1, 32–39

59 Gouws, A.D. *et al.* (2014) On the role of suppression in spatial attention: evidence from negative BOLD in human subcortical and cortical structures. *J. Neurosci.* 34, 10347–10360

60 Itthipuripat, S. *et al.* (2014) Changing the spatial scope of attention alters patterns of neural gain in human cortex. *J. Neurosci.* 34, 112–123

61 Müller, M.M. *et al.* (1998) The time course of cortical facilitation during cued shifts of spatial attention. *Nat. Neurosci.* 1, 631–634

62 Lauritzen, T.Z. *et al.* (2010) The effects of visuospatial attention measured across visual cortex using source-imaged, steady-state EEG. *J. Vis.* 10, 39

63 Kim, Y.J. *et al.* (2007) Attention induces synchronization-based response gain in steady-state visual evoked potentials. *Nat. Neurosci.* 10, 117–125

64 Störmer, V.S. and Alvarez, G.A. (2014) Feature-based attention elicits surround suppression in feature space. *Curr. Biol.* 24, 1985–1988

65 Itthipuripat, S. *et al.* (2014) Sensory gain outperforms efficient readout mechanisms in predicting attention-related improvements in behavior. *J. Neurosci.* 34, 13384–13398

66 Pestilli, F. *et al.* (2011) Attentional enhancement via selection and pooling of early sensory responses in human visual cortex. *Neuron* 72, 832–846

67 Kastner, S. *et al.* (1999) Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron* 22, 751–761

68 Serences, J.T. *et al.* (2009) Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. *Neuroimage* 44, 223–231

69 Scolari, M. and Serences, J.T. (2010) Basing perceptual decisions on the most informative sensory neurons. *J. Neurophysiol.* 104, 2266–2273

70 Warren, S.G. *et al.* (2014) Featural and temporal attention selectively enhance task-appropriate representations in human primary visual cortex. *Nat. Commun.* 5, 5643

71 Brouwer, G.J. and Heeger, D.J. (2013) Categorical clustering of the neural representation of color. *J. Neurosci.* 33, 15454–15465

72 Gratton, C. *et al.* (2013) Attention selectively modifies the representation of individual faces in the human brain. *J. Neurosci.* 33, 6979–6989

73 Tootell, R.B. *et al.* (1998) The retinotopy of visual spatial attention. *Neuron* 21, 1409–1422

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

75 Kay, K. *et al.* (2008) Identifying natural images from human brain activity. *Nature* 452, 352–355

76 Sprague, T.C. and Serences, J.T. (2013) Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nat. Neurosci.* 16, 1879–1887

77 Kay, K.N. *et al.* (2015) Attention reduces spatial uncertainty in human ventral temporal cortex. *Curr. Biol.* 25, 595–600

78 Sheremata, S.L. and Silver, M.A. (2015) Hemisphere-dependent attentional modulation of human parietal visual field representations. *J. Neurosci.* 35, 508–517
79 Klein, B.P. *et al.* (2014) Attraction of position preference by spatial attention throughout human visual cortex. *Neuron* 84, 227–237
80 De Haas, B. *et al.* (2014) Perceptual load affects spatial tuning of neuronal populations in human early visual cortex. *Curr. Biol.* 24, R66–R67
81 Çukur, T. *et al.* (2013) Attention during natural vision warps semantic representation across the human brain. *Nat. Neurosci.* 16, 763–770
82 Seidl, K.N. *et al.* (2012) Neural evidence for distracter suppression during visual search in real-world scenes. *J. Neurosci.* 32, 11812–11819
83 Peelen, M.V. *et al.* (2009) Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature* 460, 94–97
84 Zhang, K. *et al.* (1998) Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. *J. Neurophysiol.* 79, 1017–1044
85 Stanley, G.B. *et al.* (1999) Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *J. Neurosci.* 19, 8036–8042
86 Thirion, B. *et al.* (2006) Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *Neuroimage* 33, 1104–1116
87 Miyawaki, Y. *et al.* (2008) Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* 60, 915–929
88 Brouwer, G. and Heeger, D. (2009) Decoding and reconstructing color from responses in human visual cortex. *J. Neurosci.* 29, 13992–14003
89 Brouwer, G. and Heeger, D. (2011) Cross-orientation suppression in human visual cortex. *J. Neurophysiol.* 106, 2108–2119
90 Garcia, J. *et al.* (2013) Near-real-time feature-selective modulations in human cortex. *Curr. Biol.* 23, 515–522
91 Serences, J.T. and Saproo, S. (2011) Computational advances towards linking BOLD and behavior. *Neuropsychologia* 50, 435–446
92 Naselaris, T. *et al.* (2011) Encoding and decoding in fMRI. *Neuroimage* 56, 400–410
93 Schoenmakers, S. *et al.* (2013) Linear reconstruction of perceived images from human brain activity. *Neuroimage* 83, 951–961
94 Cowen, A.S. *et al.* (2014) Neural portraits of perception: reconstructing face images from evoked brain activity. *Neuroimage* 94, 12–22
95 Kok, P. *et al.* (2013) Prior expectations bias sensory representations in visual cortex. *J. Neurosci.* 33, 16275–16284
96 Scolari, M. *et al.* (2012) Optimal deployment of attentional gain during fine discriminations. *J. Neurosci.* 32, 1–11
97 Ester, E.F. *et al.* (2013) A neural measure of precision in visual working memory. *J. Cogn. Neurosci.* 25, 754–761
98 Anderson, D.E. *et al.* (2013) Attending multiple items decreases the selectivity of population responses in human primary visual cortex. *J. Neurosci.* 33, 9273–9282
99 Byers, A. and Serences, J.T. (2014) Enhanced attentional gain as a mechanism for generalized perceptual learning in human visual cortex. *J. Neurophysiol.* 112, 1217–1227
100 Kok, P. and de Lange, F.P. (2014) Shape perception simultaneously up- and downregulates neural activity in the primary visual cortex. *Curr. Biol.* 24, 1531–1535
101 Sprague, T.C. *et al.* (2014) Reconstructions of information in visual spatial working memory degrade with memory load. *Curr. Biol.* 24, 2174–2180
102 Tong, F. and Pratte, M.S. (2012) Decoding patterns of human brain activity. *Annu. Rev. Psychol.* 63, 483–509
103 Pouget, A. *et al.* (2003) Inference and computation with population codes. *Annu. Rev. Neurosci.* 26, 381–410
104 Ma, W.J. *et al.* (2006) Bayesian inference with probabilistic population codes. *Nat. Neurosci.* 9, 1432–1438
105 Pouget, A. *et al.* (2000) Information processing with population codes. *Nat. Rev. Neurosci.* 1, 125–132
106 Naselaris, T. *et al.* (2009) Bayesian reconstruction of natural images from human brain activity. *Neuron* 63, 902–915
107 Nishimoto, S. *et al.* (2011) Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* 21, 1641–1646
108 Cohen, M.R. and Maunsell, J.H.R. (2009) Attention improves performance primarily by reducing interneuronal correlations. *Nat. Neurosci.* 12, 1594–1600
109 Cohen, M.R. and Maunsell, J.H.R. (2010) A neuronal population measure of attention predicts behavioral performance on individual trials. *J. Neurosci.* 30, 15241–15253
110 Cohen, M.R. and Maunsell, J.H.R. (2011) Using neuronal populations to study the mechanisms underlying spatial and feature attention. *Neuron* 70, 1192–1204
111 DiCarlo, J.J. *et al.* (2012) How does the brain solve visual object recognition? *Neuron* 73, 415–434
112 Pagan, M. and Rust, N.C. (2014) Dynamic target match signals in perirhinal cortex can be explained by instantaneous computations that act on dynamic input from inferotemporal cortex. *J. Neurosci.* 34, 11067–11084
113 Saalmann, Y.B. *et al.* (2012) The pulvinar regulates information transmission between cortical areas based on attention demands. *Science* 337, 753–756
114 Gregoriou, G.G. *et al.* (2009) Long-range neural coupling through synchronization with attention. *Prog. Brain Res.* 176, 35–45
115 Miller, E.K. and Buschman, T.J. (2013) Cortical circuits for the control of attention. *Curr. Opin. Neurobiol.* 23, 216–222
116 Peters, A.J. *et al.* (2014) Emergence of reproducible spatiotemporal activity during motor learning. *Nature* 510, 263–267
117 Sohya, K. *et al.* (2007) GABAergic neurons are less selective to stimulus orientation than excitatory neurons in layer II/III of visual cortex, as revealed by in vivo functional Ca^{2+} imaging in transgenic mice. *J. Neurosci.* 27, 2145–2149
118 Masamizu, Y. *et al.* (2014) Two distinct layer-specific dynamics of cortical ensembles during learning of a motor task. *Nat. Neurosci.* 17, 987–994
119 Agarwal, G. *et al.* (2014) Spatially distributed local fields in the hippocampus encode rat position. *Science* 344, 626–630
120 Khodagholy, D. *et al.* (2014) NeuroGrid: recording action potentials from the surface of the brain. *Nat. Neurosci.* 18, 310–315
121 Huth, A.G. *et al.* (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76, 1210–1224
122 Zénon, A. and Krauzlis, R.J. (2012) Attention deficits without cortical neuronal deficits. *Nature* 489, 434–437
123 Krauzlis, R.J. *et al.* (2014) Attention as an effect not a cause. *Trends Cogn. Sci.* 18, 457–464
124 Ho, T. *et al.* (2012) The optimality of sensory processing during the speed–accuracy tradeoff. *J. Neurosci.* 32, 7992–8003
125 Haak, K.V. *et al.* (2012) Connective field modeling. *Neuroimage* 66, 376–384
126 Seriès, P. *et al.* (2004) Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nat. Neurosci.* 7, 1129–1135
127 Zhang, K. and Sejnowski, T.J. (1999) Neuronal tuning: to sharpen or broaden? *Neural Comput.* 11, 75–84
128 Pouget, A. *et al.* (1999) Narrow versus wide tuning curves: what's best for a population code? *Neural Comput.* 11, 85–90
129 Ruff, D.A. and Cohen, M.R. (2014) Attention can either increase or decrease spike count correlations in visual cortex. *Nat. Neurosci.* 17, 1591–1597
130 Cohen, M.R. and Kohn, A. (2011) Measuring and interpreting neuronal correlations. *Nat. Neurosci.* 14, 811–819
131 Mitchell, J.F. *et al.* (2009) Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. *Neuron* 63, 879–888
132 Moreno-Bote, R. *et al.* (2014) Information-limiting correlations. *Nat. Neurosci.* 17, 1410–1417
133 Goris, R.L.T. *et al.* (2014) Partitioning neuronal variability. *Nat. Neurosci.* 17, 858–865
134 Silver, M.A. and Kastner, S. (2009) Topographic maps in human frontal and parietal cortex. *Trends Cogn. Sci.* 13, 488–495
135 Wandell, B. *et al.* (2007) Visual field maps in human cortex. *Neuron* 56, 366–383
136 Arcaro, M.J. *et al.* (2009) Retinotopic organization of human ventral visual cortex. *J. Neurosci.* 29, 10638–10652

Chapter 1, in full, is a reprint of the material as it appears in a review entitled "Attention mitigates information loss in small- and large-scale neural codes" published in Trends in Cognitive Sciences 2015. Sprague, Thomas C.; Saproo, Sameer; Serences, John T., Cell Press, 2015. The dissertation author was the primary author of the manuscript. Supported by National Institutes of Health (NIH) grant R01-MH092345 and a James McDonnel Foundation Scholar Award to J.T.S., and NIH grant T32-MH20002-15 and a National Science Foundation (NSF) Graduate Research Fellowship to T.C.S. We thank Vy Vo for comments on an earlier version of this manuscript.

Chapter 2:

Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices

nature neuroscience

Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices

Thomas C Sprague[1] & John T Serences[1,2]

Computational theories propose that attention modulates the topographical landscape of spatial 'priority' maps in regions of the visual cortex so that the location of an important object is associated with higher activation levels. Although studies of single-unit recordings have demonstrated attention-related increases in the gain of neural responses and changes in the size of spatial receptive fields, the net effect of these modulations on the topography of region-level priority maps has not been investigated. Here we used functional magnetic resonance imaging and a multivariate encoding model to reconstruct spatial representations of attended and ignored stimuli using activation patterns across entire visual areas. These reconstructed spatial representations reveal the influence of attention on the amplitude and size of stimulus representations within putative priority maps across the visual hierarchy. Our results suggest that attention increases the amplitude of stimulus representations in these spatial maps, particularly in higher visual areas, but does not substantively change their size.

Prominent computational theories of selective attention posit that basic properties of visual stimuli are encoded in a series of interacting priority maps that are found at each stage of the visual system[1–6]. The maps in different areas are thought to encode different stimulus features (for example, orientation, color or motion) on the basis of the selectivity of component neurons. Two general themes governing the organization of these maps have emerged. First, accurately encoding the spatial location of relevant stimuli is the fundamental goal of these priority maps, as spatial position is necessary to guide saccadic eye movements (and other exploratory and reflexive motor responses). Second, priority maps early in the visual system reflect primarily the physical salience of stimuli in the visual field, whereas priority maps in later areas increasingly index the behavioral relevance of stimuli independent of physical salience[4,5].

Although many studies have investigated the influence of spatial attention on single-unit neural activity over the last several decades[7–17], directly examining the impact of attention on the topographic profile across an entire spatial priority map is a major challenge because single units have access to a limited window of the spatial scene[5]. This is a key limitation, as the relationship between changes in the size and amplitude of individual spatial receptive fields (or voxel-level receptive fields across populations of neurons) and changes in the fidelity of population-level spatial encoding are not related in a straightforward manner (ref. 18 discusses this issue with respect to population codes for orientation). For example, if spatial receptive fields are uniformly shrunk by attention while viewing a stimulus, the population-level spatial representation (or priority map) carried by all of those neurons might shrink or become sharper, but the code may be more vulnerable to uncorrelated noise (as there is less redundant coding of any given spatial position by the population). Alternatively, a uniform increase in spatial receptive field size might blur or increase the size of a spatial representation encoded by a population, but such a representation might be more robust to uncorrelated neural noise because of increased redundancy.

Further complicating matters is the observation that spatial receptive fields have been shown to both increase and decrease in size with attention as a function of where the spatial receptive field is positioned relative to the attended stimulus. Spatial receptive fields tuned near an attended stimulus grow, and spatial receptive fields that fully encompass an attended stimulus shrink[10,19–23]. These receptive field size changes occur in parallel to changes in the amplitude (gain) of neural responses with attention[7–17]. Thus, the net impact of all of these changes on the fidelity of population-level spatial representations is unclear, and addressing this issue requires assessing how attention changes the profile of spatial representations encoded by the joint, region-level pattern of activity.

Here we assessed the modulatory role of attention on the spatial information content of putative priority maps by using an encoding model to reconstruct spatial representations of attended and unattended visual stimuli on the basis of multivariate blood oxygen level–dependent functional magnetic resonance imaging (BOLD fMRI) activation patterns within visually responsive regions of the occipital, parietal and frontal cortices. These reconstructions can be considered to reflect region-level spatial representations, and they allowed us to quantitatively track changes in parameters that characterize the topography of spatial maps within each region of interest (ROI). Notably, this technique exploits the full multivariate pattern of BOLD signal across an entire region to evaluate the manner in which spatial representations are modulated by attention rather than comparing multivariate decoding accuracies or considering the univariate response of each voxel in isolation. This approach can be used to examine mechanisms of attentional modulation that cannot be easily

[1]Neuroscience Graduate Program, University of California San Diego, La Jolla, California, USA. [2]Department of Psychology, University of California San Diego, La Jolla, California, USA. Correspondence should be addressed to T.C.S. (tsprague@ucsd.edu) or J.T.S. (jserences@ucsd.edu).

Received 19 July; accepted 9 October; published online 10 November 2013; doi:10.1038/nn.3574

characterized by measuring changes in either the univariate mean BOLD signal or the decoding accuracy[24–34] (**Fig. 1**).

Our results reveal that spatial attention increases the amplitude of region-level stimulus representations within putative priority maps carried by areas of the occipital, parietal and frontal cortices. However, we found little evidence that attention changes the size of stimulus representations in region-level priority maps, even though we observed increases in spatial filter size at the single-voxel level. In addition, the reconstructed spatial representations that are based on activation patterns in later regions of the occipital, parietal and frontal cortices showed larger attentional modulation than those from early areas, which is consistent with the hypothesis that the representations in later regions increasingly transition to more selectively represent relevant stimuli[4,5]. These changes in the gain of spatial representations should theoretically increase the efficiency with which information about relevant objects in the visual field can be processed and subsequently used to guide perceptual decisions and motor plans[18].

RESULTS

Manipulating attentional demands

To evaluate how task demands influence the topography of spatial representations within different areas of the visual system, we designed a BOLD fMRI experiment that required participants to perform one of three tasks using an identical stimulus display (**Fig. 2a**). In each trial, participants (n = 8) maintained fixation at the center of the screen (Online Methods and **Supplementary Fig. 1**) while a full-contrast flickering checkerboard was presented in 1 of 36 spatial locations that sampled 6 discrete eccentricities (**Fig. 2b**). Participants reported either a faint contrast change at the fixation point (the attend fixation condition) or a faint contrast change of the flickering checkerboard stimulus (the attend stimulus condition) or performed a spatial working memory task in which they compared the location of a probe stimulus, T2, with the remembered location of a target stimulus, T1, presented within the radius of the flickering checkerboard (the spatial working memory condition; **Fig. 2c**). We included the spatial working memory task as an alternate means of inducing focused and sustained spatial attention around the stimulus position[35].

On average, performance in the attend fixation task was slightly, although nonsignificantly, higher than that in the attend stimulus or spatial working memory task (**Fig. 2d**; main effect of condition, $F(2,14) = 0.951$, $P = 0.41$; attend fixation, 87.37 ± 6.46% accuracy (mean ± s.e.m.); attend stimulus, 81.00 ± 6.67%; spatial working memory, 80.00 ± 2.09%). However, we observed a different pattern of response errors across the three task demands: accuracy in the attend fixation condition was lowest in trials in which the flickering checkerboard stimulus was presented near the fixation, whereas accuracy dropped off with increasing stimulus eccentricity in the attend stimulus and spatial working memory tasks (**Fig. 2d**; condition × eccentricity interaction, $F(10,70) = 7.235$, $P < 0.0001$).

Reconstructed spatial representations of visual stimuli

To compare spatial representations carried within different brain regions as a function of task demands, we first functionally identified seven ROIs in each hemisphere of each participant using independent localizer techniques (Online Methods and **Supplementary Table 1**).

Next we used an encoding model[34,36–38] to reconstruct a spatial representation of the stimulus that was presented during each trial using activation patterns from each ROI (**Fig. 3** and **Supplementary Figs. 2** and **3**). This method results in a spatial representation of the entire visual field measured during each trial that is constrained by activation across all voxels within each ROI. As a result, we obtained average spatial representations for each stimulus position for each ROI for each task condition that accurately reflected the stimulus viewed by the observer (**Fig. 4a** and **Supplementary Fig. 4**). This method linearly maps high-dimensional voxel space to a lower-dimensional information space that corresponds to visual field coordinates (Online Methods).

As a point of terminological clarification, we emphasize that we report estimates of the spatial representation of a stimulus display on the basis of the distributed activation pattern across all voxels within a ROI. Throughout the Results section, we therefore refer to our actual measurements as reconstructed spatial representations. However, in the Discussion, we interpret these measurements in the context of putative attentional priority maps that are thought to have a key role in shaping perception and decision making[1–6].

Reconstructed spatial representations based on activation patterns in each ROI exhibited several qualitative differences as a function of stimulus eccentricity, task demands and ROI (which we quantify more formally below). First, representations were very precise in the primary visual cortex (V1) (**Fig. 4a**) and became successively coarser and more diffuse in areas of the extrastriate, parietal and frontal cortices (**Fig. 4b**).

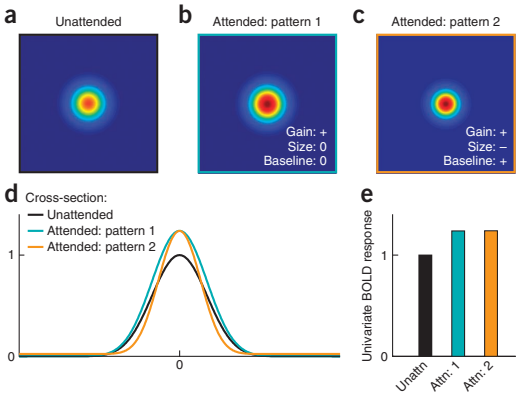


Figure 1 The effects of spatial attention on region-level priority maps. Spatial attention might act through one of several mechanisms to change the spatial representation of a stimulus within a putative priority map. (**a**) The hypothetical spatial representation carried across an entire region in response to an unattended circular stimulus. (**b**) Under one hypothetical scenario, attention might enhance the spatial representation of the same stimulus by amplifying the gain of the spatial representation (i.e., multiplying the representation by a constant greater than 1). (**c**) Alternatively, attention might act through a combination of multiple mechanisms such as increasing the gain, decreasing the size and increasing the baseline activity of the entire region (i.e., adding a constant to the response across all areas of the priority map). (**d**) Cross-sections of **a**–**c**. This is not meant as an exhaustive description of different attentional modulations. (**e**) The different types of attentional modulation can give rise to identical responses when the mean BOLD response is measured across the entire expanse of a priority map. Simple Cartesian representations, such as those in **a**–**c**, may be visualized in early visual areas where retinotopy is well defined at the spatial resolution of the BOLD response. However, later areas might still encode precise spatial representations of a stimulus even when clear retinotopic organization is not evident, so using alternative methods for reconstructing stimulus representations, such as the approach described in **Figure 3**, is necessary to evaluate the fidelity of information encoded in putative attentional priority maps. Unattn, unattended; attn, attended.

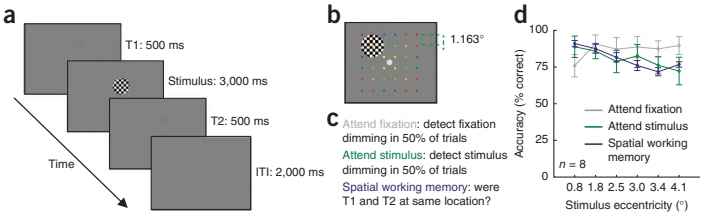


Figure 2 Task design and behavioral results. (**a**) Each trial consisted of a 500-ms target stimulus (T1), a 3000-ms flickering checkerboard (6 Hz, full contrast, 2.33° (or 2.326°) diameter) and a 500-ms probe stimulus (T2). T1 and T2 were at the same location in 50% of trials and were slightly offset in the remaining 50% of trials. During the stimulus presentation period, the stimulus dimmed briefly in 50% of trials, and the fixation point dimmed in 50% of trials (each independently randomly chosen). Participants maintained fixation throughout the experiment, and eye position measured during scanning did not vary as a function of either task demands or stimulus position (**Supplementary Fig. 1**). ITI, intertrial interval. (**b**) In each trial, a single checkerboard stimulus appeared at 1 of 36 overlapping spatial locations with a slight spatial offset between runs (Online Methods). Each spatial location was sampled once per run. This six-by-six grid of stimulus locations probes six unique eccentricities, as indicated by the color code of the dots (which is not present in the actual stimulus display). (**c**) In alternating blocks of trials, participants detected either a dimming of the fixation point (attend fixation) or a dimming of the checkerboard stimulus (attend stimulus) or they indicated if the spatial position of T1 and T2 matched (spatial working memory). Notably, all tasks used a physically identical stimulus display—only the task demands varied. Each participant completed between four and six scanning runs of each of the three tasks. (**d**) For the attend fixation task, performance was better when the stimulus was presented at peripheral locations. In contrast, performance declined with increasing stimulus eccentricity in the attend stimulus and spatial working memory conditions. Error bars, s.e.m.

Similarly, representations of more eccentric stimuli were more diffuse compared to those of more foveal stimuli (for example, when comparing eccentric to foveal representations within each ROI). We also observed higher-fidelity representations of the upper visual field when using only voxels from the ventral aspects of V2 and V3 and higher-fidelity representations of the lower visual field when using only voxels from the dorsal aspects of these regions (**Supplementary Fig. 5a**). These observations, which are consistent with known receptive field properties in nonhuman primates, confirm that our encoding-model method recovered known properties of these visual subregions and these reconstructions were not merely the result of fitting idiosyncratic aspects of our particular data set (i.e., overfitting noise). We further demonstrated this point by using the model to reconstruct representations of completely new stimuli (**Supplementary Fig. 5b**).

Second, the profile of reconstructed spatial representations within many regions also varied with task demands, which is consistent with the notion that these spatial representations reflect spatial maps of attentional priority. Notably, especially in human visual area V4 (hV4), the human middle temporal cortex (hMT+), the intraparietal sulcus (IPS) and the superior precentral sulcus (sPCS), the magnitude of the spatial representations increased when the participant was either attending to the flickering checkerboard stimulus or performing the spatial working memory task compared to when they were performing a task at fixation.

Size of spatial representations across eccentricities and ROIs

Before formally evaluating the effects of attention on the profile of spatial representations, we first sought to quantify changes in the size of these representations due to stimulus eccentricity and ROI for comparison with known properties of the primate visual system. To this end, we fit a smooth surface to the spatial representations associated with each of the three task conditions separately for each of the 36 possible stimulus locations in each ROI (Online Methods and **Supplementary Fig. 2**). These fits generated an estimate of the amplitude, baseline offset and size of the represented stimulus within each reconstructed spatial representation. We averaged the

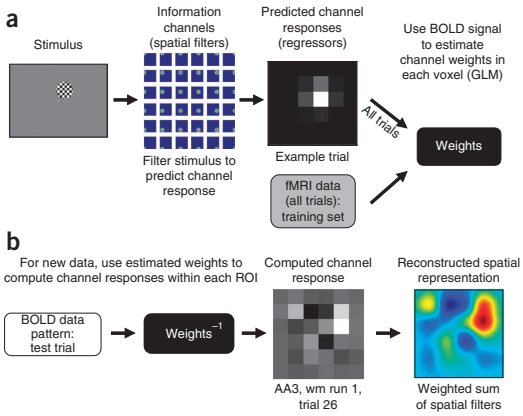


Figure 3 The encoding model that was used to reconstruct spatial representations of visual stimuli. Spatial representations of stimuli in each of the 36 possible positions were estimated separately for each ROI. (**a**) Training the encoding model. Shown is a set of linear spatial filters that forms the basis set, or information channels, that we used to estimate the spatial selectivity of the BOLD responses in each voxel (Online Methods and **Supplementary Figs. 2** and **3**). The shape of these filters determines how each information channel should respond in each trial given the position of the stimulus that was presented (thus forming a set of regressors or predicted channel responses). Then we constructed a design matrix by concatenating the regressors generated for each trial. This design matrix, in combination with the measured BOLD signal amplitude in each trial, was then used to estimate a weight for each channel in each voxel using a standard general linear model (GLM). (**b**) Estimating channel responses. Given the known spatial selectivity (or weight) profile of each voxel as computed in **a**, we then used the pattern of responses across all voxels in each trial in the test set to estimate the magnitude of the response in each of the 36 information channels in that trial. This estimate of the channel responses is thus constrained by the multivariate pattern of responses across all voxels in each trial in the test set and results in a mapping from voxel space (hundreds of dimensions) onto a lower-dimensional channel space (36 dimensions; Online Methods). We then produced a smooth reconstructed spatial representation for every trial by summing the responses of all 36 filters after weighting them by the respective channel responses in each trial. An example of a spatial representation computed from a single trial using data from V1 when the stimulus was presented at the location depicted in **a** is shown on the lower right.

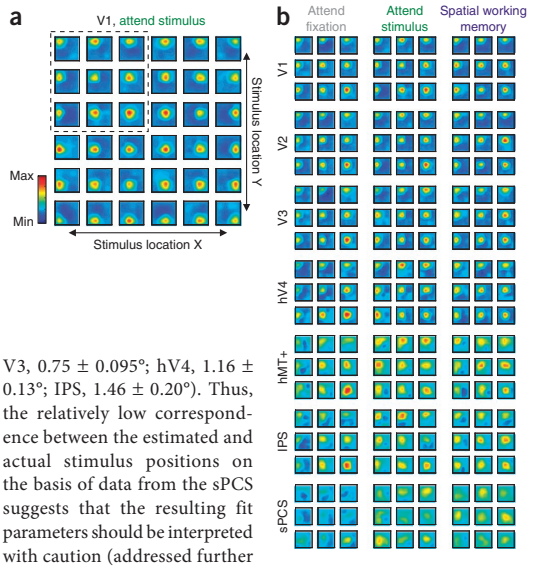


Figure 4 Task demands modulate spatial representations. (**a**) Reconstructed spatial representations of each of 36 flickering checkerboard stimuli presented in a six-by-six grid. All 36 stimulus locations are shown, with each location's representation averaged across participants ($n = 8$) using data from bilateral V1 during attend stimulus runs. One participant was not included in this analysis (AG3; **Supplementary Fig. 4**). Each small image represents the reconstructed spatial representation of the entire visual field, and the position of the image corresponds to the location of the presented stimulus. (**b**) A subset of representations (corresponding to the upper left quadrant of the visual field, represented by the dashed box in **a**) for each ROI and each task condition. The results were similar for the other quadrants (data not shown; **Fig. 5** shows the aggregate quantification of all reconstructions). All reconstructions in **a** and **b** are shown on the same color scale.

fit parameters obtained from each ROI across stimulus locations that were at equivalent eccentricities and then across participants (yielding six sets of fit parameters, with one set for each of the six possible stimulus eccentricities; **Fig. 2b**). We then used these fit parameters to make inferences about how the magnitudes and shapes of the spatial representations of stimuli from each ROI varied across stimulus positions.

First we quantified the accuracy of fits by computing the Euclidean distance between the centroid of the fit function and the actual location of the stimulus across all eccentricities and task conditions. The estimated centroids were generally accurate and closely tracked changes in stimulus location (**Fig. 4a**). However, the distances between the fit centroids and the actual stimulus positions in sPCS were nearly double those of the next least-accurate region, hMT+ (sPCS, 3.01 ± 0.077° (mean ± s.e.m.); hMT+, 1.68 ± 0.17°). The error distances in all other areas were relatively small (V1, 0.67 ± 0.084°; V2, 0.77 ± 0.12°; V3, 0.75 ± 0.095°; hV4, 1.16 ± 0.13°; IPS, 1.46 ± 0.20°). Thus, the relatively low correspondence between the estimated and actual stimulus positions on the basis of data from the sPCS suggests that the resulting fit parameters should be interpreted with caution (addressed further in the Discussion).

In early visual ROIs of V1, V2, V3 and hV4, the size of the reconstructed spatial representations increased with increasing eccentricity regardless of task condition (**Fig. 5** and Online Methods; main effect of eccentricity, two-way analysis of variance (ANOVA) within each ROI, all $P < 0.0004$; unless otherwise specified, all statistical tests on

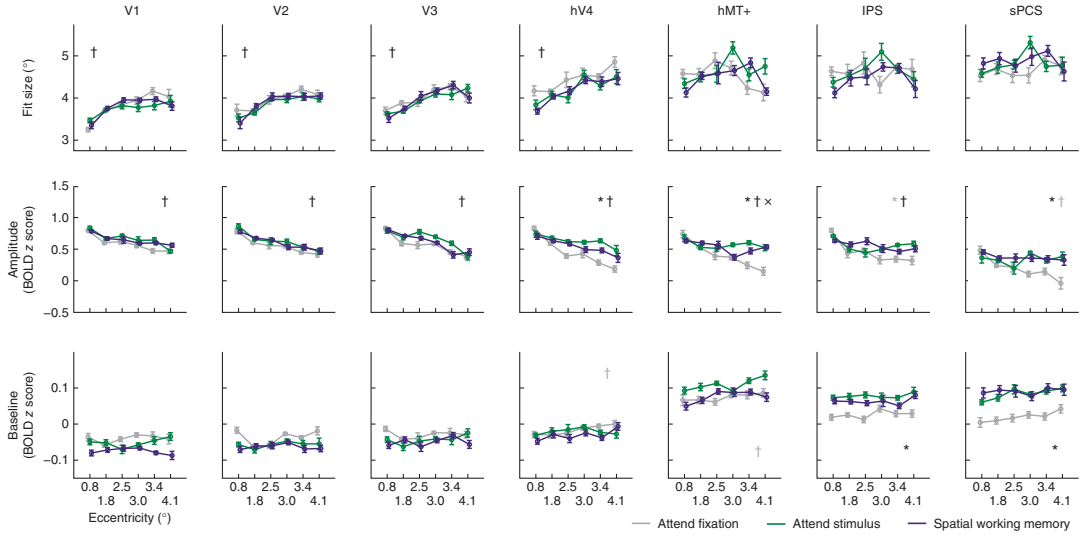
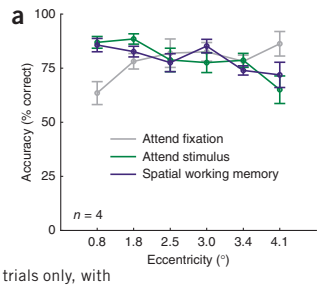


Figure 5 Fit parameters to reconstructed spatial representations averaged across like eccentricities. For each participant, we fit a smooth two-dimensional surface (Online Methods) to the average reconstructed stimulus representation in all 36 locations separately for each task condition and ROI. We allowed the amplitude, baseline, size and center (x, y coordinate) of the fit basis function to vary freely during the fitting. Fit parameters were averaged within each participant across like eccentricities and then averaged across participants. The size of the best-fitting surface varied systematically with stimulus eccentricity and ROI but did not vary as a function of task condition. In contrast, the amplitude of the best fitting surface increased with attention in hV4, hMT+ and sPCS (with a marginal effect in IPS). Shown are the main effect of task condition (*), eccentricity (†) and interaction between task and eccentricity (×) at the $P < 0.05$ level corrected for multiple comparisons (Online Methods). Gray symbols indicate trends at the $P < 0.025$ level uncorrected for multiple comparisons. Error bars, within-participant s.e.m.



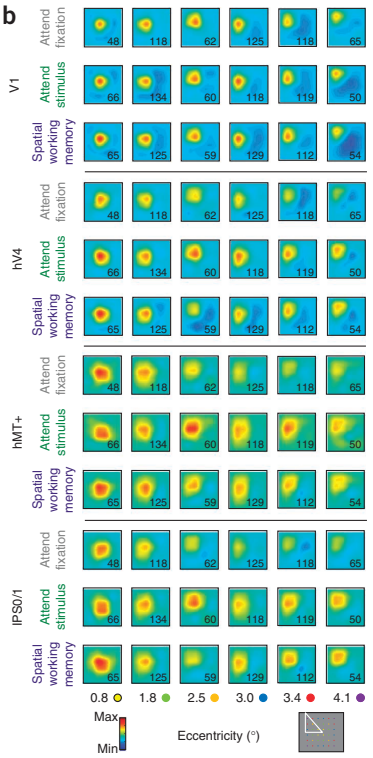


Figure 6 Results are consistent when task difficulty is matched. (**a**) Performance of four participants who were rescanned while carefully matching task difficulty across all three experimental conditions. As in **Figure 2d**, the participants' performance is better in the attend fixation task when the checkerboard is presented in the periphery, and performance in the attend stimulus and spatial working memory tasks is better when the stimulus is presented near the fovea. (**b**) A subset of illustrative reconstructed stimulus representations from V1, hV4, hMT+ and IPS0 and IPS1 (IPS0/1) averaged across like eccentricities (correct trials only, with the number of averaged trials indicated as insets). **Supplementary Figure 7** includes details about IPS subregion identification.

fit parameters to spatial representations employed a nonparametric permutation procedure and were corrected for multiple comparisons). This increase in size with eccentricity was expected, given the use of a constant stimulus size and the well-documented increase in the size of spatial receptive fields in early visual areas with increasing eccentricity[39]. In addition, the size of the reconstructed stimulus representations also increased systematically from V1 to sPCS, which is consistent with the known expansion of mean spatial receptive field sizes in the parietal and frontal cortices[40,41] (three-way ANOVA, significant main effect of ROI on fit size, $P < 0.0001$).

One alternative explanation is that the size of the represented stimulus increases with eccentricity because there is more trial-to-trial variability in the center point of the represented stimulus within reconstructions at more peripheral stimulus locations. In turn, this increase in trial-to-trial variability would 'smear' the spatial representations, leading to larger size estimates. However, our data speak against this possibility, as increased variability in the reconstructed stimulus locations would also result in lower estimated amplitudes, so increases in fit size and decreases in fit amplitude across conditions would always be yoked, and correlating the change in amplitude and the change in size within each eccentricity across each condition pair would reveal a negative correlation (for example, if the size of the spatial representation measured at a given eccentricity increased with attention, then the amplitude would decrease). No combinations of condition pair, eccentricity and ROI revealed a significant correlation between change in amplitude and change in size (all $P > 0.05$, corrected using the false discovery rate (FDR); Online Methods). Furthermore, in a follow-up analysis, we computed the population receptive field (pRF) for each voxel[42], which revealed that voxels tuned to more eccentric visual field positions have a larger pRF size (**Supplementary Table 2**, **Supplementary Results** and Online Methods). This combination of analyses supports the conclusion that increases in fit size with increases in stimulus eccentricity are not due solely to increased variability in reconstructed spatial representations.

Effects of attention on spatial representations

Despite being sensitive to expected changes in representation size on the basis of anatomical properties of the visual system, task demands exerted a negligible influence on the size of the reconstructed spatial representations, with no areas showing a significant effect (hV4 was closest at $P = 0.033$, but this did not survive correction for multiple comparisons, and P values in all other regions were >0.147).

In contrast, the fit amplitudes in hV4, hMT+, IPS and sPCS were significantly modulated by task condition, with a higher amplitude in the attention and working memory conditions than in the fixation condition (**Fig. 5**; three-way ANOVA, main effect of task condition, $P = 0.0003$). For example, in hV4, the amplitude of the best fitting surface to the spatial representations of attended stimuli was higher during the attend stimulus and spatial working memory conditions as compared to the attend fixation condition (two-way ANOVA, main effect of task condition, $P < 0.0001$). We observed similar effects in hMT+ (two-way ANOVA, $P = 0.0007$) and sPCS (two-way ANOVA, $P = 0.0007$). A similar pattern was evident in IPS as well, but it did not survive correction for multiple comparisons (two-way ANOVA, uncorrected $P = 0.011$). Within individual ROIs, there was a significant interaction between task condition and eccentricity in hMT+ ($P = 0.0003$), with larger increases in amplitude observed for more eccentric stimuli. It is notable that this increase in the amplitude of spatial representations with attention corresponds to a focal gain modulation that is restricted to the portion of visual space in the immediate neighborhood of the attended stimulus. Changes in fit amplitude do not result from a uniform, region-wide increase in BOLD signal that equally influences the response across an entire ROI; such a general and widespread modulation would be accounted for by an increase in the baseline fit parameter (**Supplementary Fig. 2**). In addition, the impact of task condition on the amplitude of reconstructed spatial representations was more pronounced in later visual areas (hV4, hMT+, IPS and sPCS) compared to earlier areas (V1, V2 and V3) (three-way interaction between ROI, condition and eccentricity, $P = 0.043$).

In addition to an increase in the fit amplitude of the reconstructed spatial representations, IPS and sPCS also exhibited a spatially global increase in baseline response levels across the entire measured spatial representation in the attend stimulus and spatial working memory conditions compared to the attend fixation condition (**Fig. 5** and

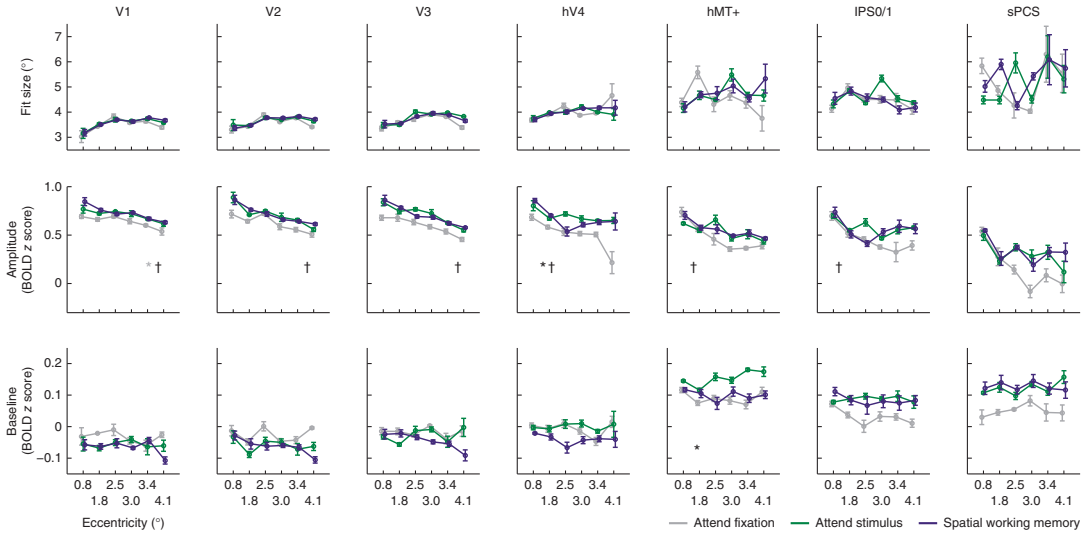


Figure 7 Fit parameters to spatial representations after controlling for task difficulty. As in **Figure 5**, a surface was fit to the averaged, co-registered spatial representations for each participant. However, in this case, task difficulty was carefully matched between conditions, and representations were based solely on trials in which the participant made a correct behavioral response (**Fig. 6b**). The results are similar to those reported in **Figure 5**: attention acts to increase the fit amplitude of spatial representations in hV4 but does not act to decrease size. In hMT+, attention also acted in a nonlocalized manner to increase the baseline parameter. The statistics and symbols are as in **Figure 5**. Error bars, within-participant s.e.m.

Supplementary Fig. 6; two-way ANOVA, main effect of condition, IPS, $P = 0.0014$; sPCS, $P = 0.0012$). The spatially nonselective increases may reflect the fact that spatial receptive fields in these regions are often large enough to encompass the entire stimulus display[40,41], so all stimuli might drive some increase in the response irrespective of spatial position.

Controlling difficulty across task conditions

Slight differences in task difficulty in the first experiment (**Fig. 2d**) might have contributed to the observed changes in spatial representations. To address this possibility, we ran four participants from the original cohort in a second experimental session while carefully equating behavioral performance across all three tasks (**Fig. 6a**). Overall accuracy during this second session did not differ significantly across the three conditions, although we observed a similar interaction between task condition and stimulus eccentricity (**Fig. 6a**; two-way repeated-measures ANOVA, main effect of condition, $F(2,6) = 0.043$, $P = 0.96$; condition × eccentricity interaction, $F(10,30) = 3.28$, $P = 0.005$; attend fixation, 78.8 ± 2.80% (mean ± s.e.m.); attend stimulus, 80.0 ± 2.60%; spatial working memory, 79.8 ± 1.76%). In addition, we also identified IPS visual field maps 0–3 (IPS0–IPS3) using standard procedures so that we could more precisely characterize the effects of attention on stimulus representations in subregions of our larger IPS ROI[31,32,43,44] (Online Methods and **Supplementary Fig. 7**).

To ensure that behavioral performance was not unduly biasing our results, we reconstructed spatial representations using only correct trials (~80% of total trials; **Fig. 6a**). All representations were co-registered on the basis of stimulus eccentricity before averaging (the corresponding eccentricity points are shown in **Fig. 2b**). Even though our sample size was smaller ($n = 4$ as compared to $n = 8$), the influence of attention on the topography of the spatial representations was similar to that in our initial observations (**Fig. 6b**). In addition, mapping out retinotopic subregions of the IPS revealed that the functionally defined IPS ROI (shown in **Fig. 5**) corresponds primarily to IPS0 and IPS1 (**Supplementary Fig. 7a**,**b**).

When examining best-fit surfaces to the spatial representations from this experiment (we computed the fits using co-registered representations and only correct trials for each participant; **Fig. 7** and Online Methods), we found that attention significantly modulated amplitude across all regions (three-way ANOVA, main effect of task condition, $P = 0.0162$). When considered in isolation, only hV4 showed a significant change in amplitude with attention after correction for multiple comparisons (two-way repeated-measures ANOVA, $P = 0.0022$). However, we observed similar trends in V1, V2 and V3 (uncorrected $P = 0.0243$, 0.042 and 0.031, respectively). We found no significant main effect of task condition on the size of the representations (all $P > 0.135$, with the minimum P found for hMT+), and the overall baseline levels significantly increased as a function of task condition in hMT+ only ($P = 0.00197$). Across all ROIs, there was a main effect of eccentricity on fit size (three-way ANOVA, $P = 0.0016$) but no main effect of task condition on fit size (three-way ANOVA, $P = 0.423$).

pRFs expand with attention

For these same four participants, we computed the pRF[42] for each voxel in V1, hV4, hMT+ and IPS0 using data from the behaviorally controlled replication experiment. We computed pRFs by first using the initial step of our encoding-model estimation procedure (**Fig. 3a**) to determine the response of each voxel to each position in the visual field (**Supplementary Figs. 8** and **9** and Online Methods). We then fit each voxel's response profile with the same surface that we used to characterize the spatial representations. By comparing pRFs computed using data from each condition independently, we found that a majority of the pRFs in hV4, hMT+ and IPS0 increased in size during either the attend stimulus or spatial working memory condition as

compared to the attend fixation condition. In contrast, pRF size in V1 was not significantly modulated by attention (**Supplementary Fig. 9** and **Supplementary Results**).

To reconcile the results that voxel-level pRFs expanded with attention yet region-level spatial representations remained at a constant size, we simulated data using estimated pRF parameters from hV4 (a region for which spatial representations increase in amplitude and pRFs increase in size; Online Methods) under different pRF modulation conditions. In the first condition, we generated simulated data using pRFs with sizes centered around two mean values, which resulted in a pRF scaling across all simulated voxels (the average size across voxels increased, but some voxels decreased in size and others increased). Under these conditions, spatial representations increased in size (**Supplementary Fig. 10a**,**b**). In a second pRF modulation scenario, we used the fit pRF values from one participant's hV4 ROI (**Supplementary Fig. 8**) to simulate data. In this case, spatial representations remained the same size but increased in amplitude, which is consistent with our observations using real data (**Figs. 5** and **7** and **Supplementary Fig. 10c**,**d**; this conclusion was also supported when we used pRF data from the other three observers to seed the simulation). Thus, the pattern of pRF modulations across all voxels enhances the amplitude of spatial representations while preserving their size.

DISCUSSION

Spatial attention has previously been shown to alter the gain of single-unit responses that are associated with relevant visual features such as orientation[7–9,12,13,16,17] and motion direction[11,14,15], as well as modulate the size of spatial receptive fields[10,19–23]. Here we show that these local modulations operate jointly to increase the overall amplitude of the region-level spatial representation of an attended stimulus without changing its represented size. Furthermore, these amplitude modulations were especially apparent in later areas of the visual system such as hV4, hMT+ and IPS, which is consistent with predictions made by computational theories of attentional priority maps[4,5].

We were able to reconstruct robust spatial representations across a range of eccentricities and for all three task conditions in all measured ROIs. Notably, even though we used an identical reconstruction procedure in all areas, the size of the reconstructed spatial representations increased from early to later visual areas (**Fig. 5**). Single-unit receptive field sizes across cortical regions are thought to increase in a similar manner[39–41,45,46]. In addition, representations of stimuli presented at higher eccentricities were larger than representations of stimuli presented near the fovea, which also corresponds to known changes in receptive field size with eccentricity[39,42]. Furthermore, simulating data under conditions in which we uniformly scaled the mean size of voxel-level pRFs revealed that such changes are detectable using our analysis method (**Supplementary Fig. 10a**,**b**). Thus, this technique is sensitive to detecting changes in the size of spatial representations of stimuli that are driven by known neural constraints such as relative differences in receptive field size across cortical ROIs and eccentricities, even though these factors are not built in to the spatial encoding model. Together these empirical and modeling results suggest that at the level of region-wide priority maps, the representation of a stimulus does not expand or contract under the attentional conditions tested here, and they underscore the importance of incorporating response changes across all encoding units when evaluating attentional modulations.

The quantification method we implemented for measuring changes in spatial representations across tasks, eccentricities and ROIs involved fitting a surface that was defined by several parameters: center location, amplitude, baseline offset and size (**Supplementary Fig. 2**). Changes in activation that carry no information about stimulus location (such as changes in general arousal or responsiveness to stimuli presented in all locations because of large receptive fields) will influence the baseline parameter, as such changes reflect increased or decreased signal across an entire region. In contrast, a change in the spatial representation that changes the representation of a visual stimulus would result in a change in the amplitude or size parameter (or both). Here we demonstrated that attention operates primarily by selectively increasing the amplitude of stimulus representations in several putative priority maps (**Figs. 5** and **7**) rather than increasing the overall BOLD signal more generally across entire regions.

Notably, spatial reconstructions based on activation patterns from sPCS were relatively inaccurate compared to other ROIs, and this ROI primarily exhibited increases in the fit baseline parameter (**Fig. 5**). This region, which may be a human homolog of the functionally defined macaque frontal eye field[47,48] (FEF), might have showed degraded spatial selectivity in the present study because of the relatively large size of spatial receptive fields observed in many FEF neurons (typically ≥20° diameter[41]) and the small area subtended by our stimulus display (9.31° across horizontally). Consistent with this possibility, previous reports of retinotopic organization in the human frontal cortex used stimuli that were presented at higher eccentricities in order to resolve spatial maps (≥10° (ref. 49) to 25° (ref. 45)).

Attentional priority maps

The extensive literature on spatial salience or priority maps[1–6] postulates the existence of one or several maps of visual space, each carrying information about behaviorally relevant objects within the visual scene. Furthermore, priority maps in early visual areas (for example, V1) are thought to encode primarily low-level stimulus features (for example, contrast), whereas priority maps in later regions are thought to increasingly weight behavioral relevance over low-level stimulus attributes[4]. Although many important insights have stemmed from observing single-unit responses as a function of changes in attentional priority (reviewed in ref. 5), these results provide information about how isolated pixels in a priority map change under different task conditions.

A previous fMRI study used multivariate decoding (classification) analyses to identify several frontal and parietal ROIs that exhibit similar activation patterns during covert attention, spatial working memory and saccade generation tasks[32]. These results provide strong support for the notion that common priority maps support representations of attentional priority across multiple tasks. Here we assessed how the holistic landscape across these priority maps measured using fMRI changed as attention was systematically varied. Our demonstration that spatial representation amplitude was enhanced with attention in later, but not earlier, ROIs supports the hypothesis that priority maps in higher areas are increasingly dominated by attentional factors and suggests that these attentional modulations of priority maps operate by scaling the amplitude of the behaviorally relevant item without changing its represented size.

pRFs

In addition to measuring spatial representations that are carried by the pattern of activation across entire visual regions, we also estimated the voxel-level pRFs[42] for a subset of participants and ROIs by adding constraints to our encoding-model estimation procedure (**Supplementary Figs. 8** and **9** and Online Methods). This alternative tool has been used previously to evaluate the aggregate spatial receptive field profile across all neural populations within voxels that belong to different visual ROIs[42].

Changes in voxel-level pRFs can inform how a region dynamically adjusts the spatial sensitivity of its constituent filters in order to modulate its overall spatial priority map. First, we replicated the typical results that voxel-level pRFs tuned for more eccentric visual field positions are larger in size (**Supplementary Table 2**) and that pRFs for later visual regions tend to be larger than pRFs for earlier visual regions (**Supplementary Fig. 9**). Second, results from this complementary analysis revealed that in regions that showed enhanced spatial representation amplitude with attention (hV4, hMT+ and IPS0), pRF size increased (**Supplementary Figs. 8** and **9**), even though the corresponding region-level spatial representations did not increase in size (**Fig. 7**). This may seem like a disconnect, given that the particular pattern of pRF changes across all voxels within a region jointly shapes how the spatial priority map changes with attention. However, there is not necessarily a monotonic mapping between the size of the constituent filters and the size of population-level spatial representations. Indeed, simulations based on the observed pattern of pRF changes with attention give rise to region-level increases in representation amplitude in the absence of changes in representation size, as we observed in our data (**Supplementary Fig. 10**). This finding, together with our primary results concerning region-level spatial representations, provides evidence that attentional modulation of spatial information encoding is a process that strongly benefits from study at the large-scale population level.

Comparison to previous results

At the level of single-unit recordings, attention has been shown to decrease the size of MT spatial receptive fields when an animal is attending to a stimulus that is encompassed by the recorded neuron's receptive field[19–21] and to increase the size of spatial receptive fields when an animal is attending nearby the recorded neuron's receptive field[20–22]. In V4, spatial receptive fields appear to shift toward the attended region of space in a subset of neurons[10]. With respect to cortical space, these single-unit attentional modulations of spatial receptive fields suggest that unifocal attention may act to increase the cortical surface area that is responsive to a stimulus of constant size. Consistent with this prediction, our measured pRFs for extrastriate regions of hV4, hMT+ and IPS0 increased in size with attention.

In contrast, one previous report suggested that spatial attention instead narrows the activation profile along the cortical surface of the visual cortex in response to a visual stimulus[50]. However, this inference was based on patterns of intertrial correlations between BOLD activation patterns that were associated with dividing attention between four stimuli (one presented in each quadrant). These patterns were suggested to result from a combination of attention-related gain and narrowing of population-level responses[50], that is, a narrower response along the cortical surface with attention.

We did not observe any significant attention-related changes in the size of the reconstructed spatial representations in either the primary visual cortex or other areas in the extrastriate, parietal or frontal cortices. However, the tasks performed by observers and the analysis techniques implemented were very different between these studies. Most notably, observers in the present study and in previous fMRI[24–33] and single-unit studies[10,19–21] were typically required to attend to a single stimulus, whereas population-level activation narrowing was observed when participants simultaneously attended to the precise spatial position of four Gabor stimuli, one presented in each visual quadrant[50]. Furthermore, our observation that pRFs increased in size during the attend stimulus and spatial working memory conditions is compatible with the pattern of spatial receptive field changes in single units[10,19–23], and our data and simulations show that these local changes can result in a region-level representation that changes only in amplitude and not in size (**Supplementary Fig. 10**).

Collectively, it seems probable that the exact task demands (unifocal as compared to multifocal attention) and stimulus properties (single stimulus as compared to multiple stimuli) may have a key role in determining how attention influences the profile of spatial representations. Future work using analysis methods that are sensitive to region-level differences in spatial representations (for example, applying encoding models similar to that described here to data acquired when participants perform different tasks) in conjunction with careful identification of neural receptive field properties across those task-demand conditions (for example, from simultaneous multiunit electrophysiological recordings or *in vivo* two-photon Ca^{2+} imaging in rodents and primates) may provide complementary insights into when and how attention changes the shape and/or amplitude of stimulus representations in spatial priority maps and how those changes are implemented in the neural circuitry.

Notably, although our observations are largely consistent with measured receptive field changes at the single-unit level[10,19–23], we cannot make direct inferences that such single-unit changes are in fact occurring. A number of mechanisms, including one in which only the gain of different populations is modulated by attention, could also account for the pattern of results we saw in both our region-level spatial representations (**Figs. 5** and **7**) and our pRF measurements (**Supplementary Figs. 8** and **9**). We note, however, that some neural mechanisms are highly unlikely given our measured spatial representations and pRFs. For example, we would not observe an increase in pRF size if spatial receptive fields of neurons within those voxels were to exclusively narrow with attention. As a result of these interpretational concerns, we restrict the inferences we draw from our results to the role of attention in modulating region-level spatial priority maps measured with fMRI and make no direct claims about spatial information coding at a neural level.

Information content of attentional priority maps

One consequence of an observed increase in the amplitude of reconstructed priority maps is that the mutual information between the stimulus position and the observed BOLD responses should increase (a more complete discussion is provided in ref. 18). This increase can occur, in theory, because mutual information reflects the ratio of signal entropy (the variability in neural responses that is tied systematically to changes in the stimulus) to noise entropy (the variability in neural responses that is not tied to changes in the stimulus). Thus a multiplicative increase in the gain of the neural responses that are associated with an attended stimulus should increase mutual information because it will increase the variability of responses that are associated with an attended stimulus location, which will in turn increase signal entropy. In contrast, a purely additive shift in all neural responses (reflected by an increase in the fit baseline parameter) will not increase the dynamic range of responses that are associated with an attended stimulus location, causing the mutual information to either remain constant (under a constant additive noise model) or even decrease (under a Poisson noise model, in which noise increases with the mean). Previous fMRI work on spatial attention has not attempted to disentangle these two potential sources of increases in the BOLD signal, highlighting the utility of approaches that can support more precise inferences about how task demands influence region-level neural codes[24–33].

The information content of a neural code is not necessarily monotonically related to the size of the constituent neural filters[18].

Extremely small (pinpoint) or extremely large (flat) spatial filters each individually carry very little information about the spatial arrangement of stimuli within the visual field. Accordingly, the optimal filter size lies somewhere between these two extremes, and thus it is not straightforward to infer whether a change in filter size results in a more or less optimal neural code (in terms of information encoding capacity). By simultaneously estimating changes in filter size across an entire ROI subtending the entire stimulated visual field, we were able to demonstrate that the synergistic pattern of spatial filter (pRF) modulations with attention jointly constrains the region-level spatial representation to maintain a constant represented stimulus size, despite most voxels exhibiting an increase in pRF size (**Supplementary Figs. 8–10**). Together our results demonstrate the importance of incorporating all available information across entire ROIs when evaluating the modulatory role of attention on the information content of spatial priority maps.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank E. Vul and S. Itthipuripat for assistance with statistical methods and M. Scolari and M. Smith for assistance with parietal cortex mapping protocols. This work was supported by a National Science Foundation Graduate Research Fellowship to T.C.S. and by US National Institutes of Health grant R0I MH-092345 and a James S. McDonnell Scholar Award to J.T.S.

AUTHOR CONTRIBUTIONS

T.C.S. and J.T.S. designed the experiments and analysis method and wrote the manuscript. T.C.S. conducted the experiments and implemented the analyses.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Koch, C. & Ullman, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* **4**, 219–227 (1985).
2. Itti, L., Koch, C. & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1254–1259 (1998).
3. Itti, L. & Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2**, 194–203 (2001).
4. Serences, J.T. & Yantis, S. Selective visual attention and perceptual coherence. *Trends Cogn. Sci.* **10**, 38–45 (2006).
5. Fecteau, J.H. & Munoz, D.P. Salience, relevance, and firing: a priority map for target selection. *Trends Cogn. Sci.* **10**, 382–390 (2006).
6. Bichot, N.P. & Schall, J.D. Effects of similarity and history on neural mechanisms of visual selection. *Nat. Neurosci.* **2**, 549–554 (1999).
7. Luck, S.J., Chelazzi, L., Hillyard, S.A. & Desimone, R. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.* **77**, 24–42 (1997).
8. Reynolds, J.H., Chelazzi, L. & Desimone, R. Competitive mechanisms subserve attention in macaque areas V2 and V4. *J. Neurosci.* **19**, 1736–1753 (1999).
9. McAdams, C.J. & Maunsell, J.H.R. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J. Neurosci.* **19**, 431–441 (1999).
10. Connor, C.E., Preddie, D.C., Gallant, J.L. & Van Essen, D.C. Spatial attention effects in macaque area V4. *J. Neurosci.* **17**, 3201–3214 (1997).
11. Treue, S. & Maunsell, J.H.R. Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* **382**, 539–541 (1996).
12. Reynolds, J.H., Pasternak, T. & Desimone, R. Attention increases sensitivity of V4 neurons. *Neuron* **26**, 703–714 (2000).
13. McAdams, C.J. & Maunsell, J.H.R. Attention to both space and feature modulates neuronal responses in macaque area V4. *J. Neurophysiol.* **83**, 1751–1755 (2000).
14. Treue, S. & Maunsell, J.H.R. Effects of attention on the processing of motion in macaque middle temporal and medial superior temporal visual cortical areas. *J. Neurosci.* **19**, 7591–7602 (1999).
15. Seidemann, E. & Newsome, W.T. Effect of spatial attention on the responses of area MT neurons. *J. Neurophysiol.* **81**, 1783–1794 (1999).
16. Motter, B.C. Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J. Neurophysiol.* **70**, 909–919 (1993).
17. Moran, J. & Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784 (1985).
18. Saproo, S. & Serences, J.T. Spatial attention improves the quality of population codes in human visual cortex. *J. Neurophysiol.* **104**, 885–895 (2010).
19. Womelsdorf, T., Anton-Erxleben, K., Pieper, F. & Treue, S. Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nat. Neurosci.* **9**, 1156–1160 (2006).
20. Womelsdorf, T., Anton-Erxleben, K. & Treue, S. Receptive field shift and shrinkage in macaque middle temporal area through attentional gain modulation. *J. Neurosci.* **28**, 8934–8944 (2008).
21. Anton-Erxleben, K., Stephan, V.M. & Treue, S. Attention reshapes center-surround receptive field structure in macaque cortical area MT. *Cereb. Cortex* **19**, 2466–2478 (2009).
22. Niebergall, R., Khayat, P.S., Treue, S. & Martinez-Trujillo, J.C. Expansion of MT neurons excitatory receptive fields during covert attentive tracking. *J. Neurosci.* **31**, 15499–15510 (2011).
23. Anton-Erxleben, K. & Carrasco, M. Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence. *Nat. Rev. Neurosci.* **14**, 188–200 (2013).
24. Liu, T., Pestilli, F. & Carrasco, M. Transient attention enhances perceptual performance and FMRI response in human visual cortex. *Neuron* **45**, 469–477 (2005).
25. Gandhi, S.P., Heeger, D.J. & Boynton, G.M. Spatial attention affects brain activity in human primary visual cortex. *Proc. Natl. Acad. Sci. USA* **96**, 3314–3319 (1999).
26. Kastner, S., Pinsk, M.A., De Weerd, P., Desimone, R. & Ungerleider, L.G. Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron* **22**, 751–761 (1999).
27. Brefczynski, J.A. & DeYoe, E.A. A physiological correlate of the "spotlight" of visual attention. *Nat. Neurosci.* **2**, 370–374 (1999).
28. Silver, M.A., Ress, D. & Heeger, D.J. Neural correlates of sustained spatial attention in human early visual cortex. *J. Neurophysiol.* **97**, 229–237 (2007).
29. Tootell, R.B. *et al*. The retinotopy of visual spatial attention. *Neuron* **21**, 1409–1422 (1998).
30. Murray, S.O. The effects of spatial attention in early human visual cortex are stimulus independent. *J. Vis.* **8**, 2.1–2.11 (2008).
31. Silver, M.A., Ress, D. & Heeger, D.J. Topographic maps of visual spatial attention in human parietal cortex. *J. Neurophysiol.* **94**, 1358–1371 (2005).
32. Jerde, T.A., Merriam, E.P., Riggall, A.C., Hedges, J.H. & Curtis, C.E. Prioritized maps of space in human frontoparietal cortex. *J. Neurosci.* **32**, 17382–17390 (2012).
33. Jehee, J.F.M., Brady, D.K. & Tong, F. Attention improves encoding of task-relevant features in the human visual cortex. *J. Neurosci.* **31**, 8210–8219 (2011).
34. Serences, J.T. & Saproo, S. Computational advances towards linking BOLD and behavior. *Neuropsychologia* **50**, 435–446 (2012).
35. Awh, E. & Jonides, J. Overlapping mechanisms of attention and spatial working memory. *Trends Cogn. Sci.* **5**, 119–126 (2001).
36. Brouwer, G.J. & Heeger, D.J. Decoding and reconstructing color from responses in human visual cortex. *J. Neurosci.* **29**, 13992–14003 (2009).
37. Naselaris, T., Kay, K., Nishimoto, S. & Gallant, J. Encoding and decoding in fMRI. *Neuroimage* **56**, 400–410 (2011).
38. Scolari, M., Byers, A. & Serences, J.T. Optimal deployment of attentional gain during fine discriminations. *J. Neurosci.* **32**, 7723–7733 (2012).
39. Gattass, R. *et al*. Cortical visual areas in monkeys: location, topography, connections, columns, plasticity and cortical dynamics. *Phil. Trans. R. Soc. Lond. B* **360**, 709–731 (2005).
40. Ben Hamed, S., Duhamel, J.R., Bremmer, F. & Graf, W. Visual receptive field modulation in the lateral intraparietal area during attentive fixation and free gaze. *Cereb. Cortex* **12**, 234–245 (2002).
41. Mohler, C.W., Goldberg, M.E. & Wurtz, R.H. Visual receptive fields of frontal eye field neurons. *Brain Res.* **61**, 385–389 (1973).
42. Dumoulin, S.O. & Wandell, B.A. Population receptive field estimates in human visual cortex. *Neuroimage* **39**, 647–660 (2008).
43. Sereno, M.I., Pitzalis, S. & Martinez, A. Mapping of contralateral space in retinotopic coordinates by a parietal cortical area in humans. *Science* **294**, 1350–1354 (2001).
44. Swisher, J.D., Halko, M.A., Merabet, L.B., McMains, S.A. & Somers, D.C. Visual topography of human intraparietal sulcus. *J. Neurosci.* **27**, 5326–5337 (2007).
45. Saygin, A.P. & Sereno, M.I. Retinotopy and attention in human occipital, temporal, parietal, and frontal cortex. *Cereb. Cortex* **18**, 2158–2168 (2008).
46. Kastner, S. *et al*. Modulation of sensory suppression: implications for receptive field sizes in the human visual cortex. *J. Neurophysiol.* **86**, 1398–1411 (2001).
47. Srimal, R. & Curtis, C.E. Persistent neural activity during the maintenance of spatial position in working memory. *Neuroimage* **39**, 455–468 (2008).
48. Paus, T. Location and function of the human frontal eye-field: a selective review. *Neuropsychologia* **34**, 475–483 (1996).
49. Kastner, S. *et al*. Topographic maps in human frontal cortex revealed in memory-guided saccade and spatial working-memory tasks. *J. Neurophysiol.* **97**, 3494–3507 (2007).
50. Fischer, J. & Whitney, D. Attention narrows position tuning of population responses in V1. *Curr. Biol.* **19**, 1356–1361 (2009).

ONLINE METHODS

Participants. Ten neurologically healthy volunteers (five female, 25 ± 2.11 years of age (mean ± s.d.)) with normal or corrected-to-normal vision were recruited from the University of California, San Diego (UCSD). All participants provided written informed consent in accordance with the human participants Institutional Review Board at UCSD and were monetarily compensated for their participation. For the original experiment, participants participated in two to three scanning sessions, each lasting 2 h. Data from two participants (one female) were excluded from the main analysis because of excessive head movement (AJ3) or unusually noisy reconstructions during attend fixation runs (AG3).

In the follow-up experiment in which behavioral performance was carefully controlled and IPS subregions were retinotopically mapped, four participants of our original cohort were scanned for an additional two sessions, each lasting 1.5–2 h.

Stimulus. Stimuli were rear projected on a screen (90-cm width) located 380 cm from the participant's eyes at the foot of the scanner table. The screen was viewed using a mirror attached to the headcoil.

We presented an identical stimulus sequence during all imaging runs while asking observers to perform several different tasks. Each trial began with the presentation of a small red dot (T1) that was presented for 500 ms, followed by a flickering circular checkerboard stimulus at full contrast (stimulus radius (r_{stim}) = 1.163°, 1.47 cycles per degree) that was presented for 3 s and then a probe stimulus (T2) that was identical to T1. A 2-s intertrial interval separated each trial (**Fig. 2a**). T1 was presented between 0.176° and 1.104° from the center of the checkerboard stimulus along a vector of random orientation (in polar coordinates, θ_1 was randomly chosen along the range 0° to 360°, and r_1 was uniformly sampled from the range 0.176° to 1.104°). This ensured that the location of T1 was not precisely predictive of the checkerboard location. In 50% of the trials, T2 was presented in the same location as T1, and in the remaining trials, T2 was presented between 0.176° and 1.104° from the center of the checkerboard along a vector oriented at least 90° from the vector along which T1 was plotted (r_2 was uniformly sampled from the range 0.176° to 1.104°, and θ_2 was randomly chosen by adding between 90° and 270°, uniformly sampled, to θ_1). Polar coordinates used the center of the checkerboard stimulus as the origin. During the working memory condition (see below), participants based their response on whether T1 and T2 were presented in the exact same spatial position.

The location of the checkerboard stimulus was chosen pseudorandomly in each trial from a grid of 36 potential stimulus locations spaced by 1.163°, or r_{stim}. The stimulus location grid was jittered by 0.827° diagonally either up and to the left or down and to the right in each run, allowing for an improved sampling of space. All figures were presented aligned to a common space by removing jitter (see below).

In each run, there were 36 trials (1 trial for each stimulus location) and 9 null trials in which participants passively fixated for the duration of a normal trial (6 s). We scanned participants for between four and six runs of each task, always ensuring each task was repeated an equal number of times.

Tasks. Participants performed one of three tasks during each functional run (**Fig. 2c**). During attend fixation runs, participants responded when they detected a brief contrast dimming of the fixation point (0.33 s), which occurred in 50% of trials. During attend stimulus runs, participants responded when they detected a brief contrast dimming of the flickering checkerboard stimulus (0.33 s), which occurred in 50% of trials. During spatial working memory runs, participants made a button press response to indicate whether T2 was in the same location or a different location as T1. Notably, all three events (T1, checkerboard and T2) occurred during all runs, ensuring that the sensory display remained identical and that we were measuring changes in spatial representations as a function of task demands rather than changes as a result of inconsistent visual stimulation. For the follow-up behavioral control experiment, we dynamically adjusted the difficulty (contrast dimming or T1-T2 separation distance) to achieve a consistent accuracy of ~75% across tasks.

Eye tracking. Participants were instructed to maintain fixation during all runs. Fixation was monitored during scanning for four participants using an ASL LRO-R long-range eye-tracking system (Applied Science Laboratories) with a sampling rate of 240 Hz. We recorded mean gaze as a function of stimulus location and task demands after excluding any samples in which neither the pupil nor the corneal reflection were reliably detected (**Supplementary Fig. 1**).

Imaging. We scanned all participants on a 3T GE MR750 research-dedicated scanner at UCSD. Functional images were collected using a gradient echo planar imaging (EPI) pulse sequence and an eight-channel head coil (19.2 × 19.2 cm FOV, 96 × 96 matrix size, 31 3-mm-thick slices with 0-mm gap, TR = 2,250 ms, TE = 30 ms, flip angle = 90°), which yielded a voxel size of 2 × 2 × 3 mm. We acquired oblique slices with coverage extending from the superior portion of parietal cortex to the ventral occipital cortex.

We also acquired a high-resolution anatomical scan (FSPGR T1-weighted sequence, TR/TE = 11/3.3 ms, TI = 1,100 ms, 172 slices, flip angle = 18°, 1 mm^3 resolution). Functional images were co-registered to this scan. Images were preprocessed using FSL (Oxford, UK) and BrainVoyager 2.3 (BrainInnovations). Preprocessing included unwarping the EPI images using routines provided by FSL, slice-time correction, three-dimensional motion correction (six-parameter affine transform), temporal high-pass filtering (to remove first-, second- and third-order drift), transformation to Talairach space and normalization of signal amplitudes by converting to z scores. We did not perform any spatial smoothing beyond the smoothing introduced by resampling during the co-registration of the functional images, motion correction and transformation to Talairach space. When mapping IPS subregions, we scanned the participants using an identical pulse sequence but instead used a 32-channel Nova Medical headcoil.

Functional localizers. All the ROIs used were identified using independent localizer runs acquired across multiple scanning sessions.

Early visual areas were defined using standard retinotopic procedures[51,52]. We identified the horizontal and vertical meridians using functional data projected onto gray- and white-matter boundary surface reconstructions for each hemisphere. Using these meridians, we defined the areas V1, V2v, V3v, hV4, V2d and V3d. Unless otherwise indicated, data were concatenated across hemispheres and across the dorsal and ventral aspects of each respective visual area. We scanned each participant for between two and four retinotopic mapping runs (n = 3 completed two runs, n = 3 completed three runs, and n = 2 completed four runs).

hMT+ was defined using a functional localizer in which a field of dots either moved with 100% coherence in a pseudorandomly selected direction or were randomly replotted on each frame to produce a visual 'snow' display[53,54]. Dots were each 0.081° in diameter and were presented in an annulus of between 0.63° and 2.26° around the fixation. During coherent dot motion, all dots moved at a constant velocity of 2.71° s^{-1}. Participants attended the dot display for transient changes in velocity (during coherent motion) or replotting frequency (snow). Participants completed between one and three runs of this localizer (n = 2 completed one run, n = 3 completed two runs, and n = 3 completed three runs).

IPS and sPCS ROIs were defined using a functional localizer that required maintenance of a spatial location in working memory, a task that is commonly used to isolate IPS and sPCS, which is the putative human FEF[47,49]. A flickering checkerboard subtending half of the visual field appeared for 12 s, during which time two spatial working memory trials were presented. During the flickering checkerboard presentation, we presented a red target dot for 500 ms, which was followed 2 s later by a green probe dot for 500 ms. After the probe dot appeared, participants indicated whether the probe dot was in the same location or a different location as the red target dot. Here we limited our definition of IPS to the posterior aspect (**Supplementary Table 1**). ROIs were functionally defined with a threshold of FDR-corrected $P < 0.05$ or a more stringent threshold when patches of activation abutted one another. Participants completed between one (n = 2) and two (n = 6) runs of this scan.

We also used data from these IPS and sPCS localizer scans to identify voxels in all other ROIs that were responsive to the portion of the visual field in which stimuli were presented in the main tasks, as the large checkerboard stimuli subtended the same visual area as the stimulus array used in the main task. All ROIs were masked on a participant-by-participant basis such that further analyses only included voxels with significant responses during this localizer task (FDR corrected $P < 0.05$).

Mapping IPS subregions. To determine the likely relative contributions of different IPS subregions to the localized ROI measured for all participants, we scanned the four participants who made up the behaviorally controlled cohort presented

in **Figures 6** and **7** using a polar angle mapping stimulus and an attentionally demanding task.

We used two stimulus types and behavioral tasks to define the borders between IPS subregions[31,32,43,44]. In all runs, we used a wedge stimulus spanning a 72° polar angle and presented between 1.75° and 8.75° eccentricity rotating with a period of 24.75 s. In alternating runs, the wedge was either a 4-Hz flickering checkerboard stimulus (black-white, red-green or blue-yellow) or a field of moving black dots (0.3°, 13 dots per square degree^2, moving at 5° s^{-1}, changing direction every 8 s). During checkerboard runs, participants quickly responded after detecting a brief (250 ms) contrast dimming of a portion of the checkerboard. During moving dots runs, participants quickly responded after detecting a brief (417 ms) increase in dot speed. Targets appeared with 20% probability every 1.5 s. Difficulty was adjusted to achieve approximately 75% correct performance by changing the magnitude of the contrast dimming (checkerboard) or dot speed increment (moving dots) between runs. On average, participants performed with 84.1% accuracy in the contrast dimming task and 75.4% accuracy in the moving dots task. Two participants completed 14 runs (8 clockwise and 6 counterclockwise), and one participant completed 10 runs (AC; 5 clockwise and 5 counterclockwise). One participant was scanned with two different stimulus setups: half of all runs used the parameters described above and half used a wedge that spanned 60° of polar angle and rotated with a period of 36.00 s (AB; 6 runs clockwise, 6 runs counterclockwise).

Preprocessing procedures were identical to those used for the main task. To compute the best visual field angle for each voxel in IPS, we shifted the signals from counterclockwise runs earlier in time by twice the estimated hemodynamic response function (HRF) delay (2 × 6.75 s = 13.5 s), removed the first and last full cycle of data (we removed 22 TRs for all participants except AB, for which we removed 32 TRs) and then reversed the time series so that all runs correspond to the clockwise stimulus presentation. We then averaged these time-inverted counterclockwise runs with the clockwise runs. We computed the power and phase at the stimulus frequency (1/24.75 Hz or 1/36 Hz, participant AB) and subtracted the estimated HRF delay (6.75 s) to align the signal phase in each voxel with the visual stimulus position. We then projected maps onto the reconstructed cortical surfaces for each subject and defined IPS0–IPS3 by identifying the upper and lower vertical meridian responses (**Supplementary Fig. 7a**). Low statistical thresholds were used (computed using normalized power at the stimulus frequency) to identify the borders of IPS subregions. Voxels were selected for further analysis by thresholding their activation during the same independent localizer task that was used to functionally define IPS and sPCS.

Encoding model. To measure changes in spatial representations under different task demands, we implemented an encoding model to reconstruct spatial representations of each stimulus used in the main task[34,36–38]. This technique assumes that the signal measured in each voxel can be modeled as the weighted sum of different discrete neural populations, or information channels, that have different tuning properties[36]. Using an independent set of training data, we estimated weights that approximate the degree to which each underlying neural population contributed to the observed BOLD response in each voxel (**Fig. 3a**). Next, an independent set of test data was used to estimate the activation within these information channels on the basis of the activation pattern across all voxels within an ROI in each test trial using the information channel weights in each voxel that were estimated during the training phase (**Fig. 3b**).

This approach requires specifying an explicit model for how neural populations encode information. Here we assumed a simple model for visual encoding within each ROI that focused exclusively on the spatial selectivity of visually responsive neural populations. To this end, we built a basis set of 36 two-dimensional spatial filters. We modeled these filters as cosine functions raised to a high power: $f(r) = (0.5\cos(r\pi/s) + 0.5)^7$ for $r < s$ and 0 elsewhere (r is the distance from that filter's center; **Supplementary Fig. 2**). This allowed the filters to maintain an approximately Gaussian shape while reaching 0 at a fixed distance from the center (s°), which helped constrain curve-fitting solutions (below). The s (size constant) parameter was fixed at $5r_{stim}$, which is 5.8153°. The 36 identical filters formed a six-by-six grid spanning the visual space subtended by the stimuli. Filters were separated by 2.094°, with the centers tiled uniformly from 5.234° above, below, left and right of the fixation (**Fig. 3a**). The full-width half-maximum (FWHM) of all filters was 2.3103° (**Supplementary Figs. 2** and **3**). This ratio of filter size to spacing was chosen to avoid high correlations between predicted channel responses (caused by too much overlap between channels, which can result in a rank-deficient design matrix) and to accomplish smooth reconstructions (if filters are too small, reconstructed spatial representations are 'patchy'; **Supplementary Fig. 3** shows an illustration of reconstruction smoothness as a function of the filter size to spacing ratio). All filters were assigned identical FWHMs so that known properties of the visual system, such as increasing receptive field size with eccentricity and along the visual stream[39–41], could be recovered without being built in to the analysis.

To avoid circularity in our analysis, we used a cross-validation approach to compute channel responses for every trial. First we used all runs but three (one run for each task condition) to create a training set that had an equal number of trials in each condition. Using this training set, we estimated channel weights within each voxel across all task conditions (i.e., runs one through five of the attend fixation, attend stimulus and spatial working memory tasks were used together to estimate channel weights, which were used to compute channel responses for run six of each task condition). The use of an equal number of trials from each condition in the training set ensures that channel weight estimation is not biased by any changes in BOLD response across task demands. Next the weights estimated across all task demand conditions were used to compute channel response amplitudes for each trial individually. Trials were then sorted according to their task condition and spatial location.

During the training phase, we created a design matrix that contained the predicted channel response for all 36 channels in every trial (**Fig. 3a**). The predicted response for each channel was computed by filtering a mask over the portion of the display subtended by the stimulus on that trial by the channel's basis function. The resulting design matrix was normalized to 1, such that reconstruction amplitudes are in units of BOLD z scores.

To extract relevant portions of the BOLD signal in every trial for computing channel responses, we took an average of the signal over two TRs beginning 6.75 s after trial onset. This range was chosen by examination of BOLD HRFs and was the same across all participants. Qualitatively, results do not change when other reasonable HRF lags are used, such as using two TRs starting 4.5 s after the stimulus.

Using this approach, we modeled voxel BOLD responses as a weighted sum of channel responses comprising each voxel[36,38]. This can be written as a general linear model of the form

$$B_1 = WC_1 \quad (1)$$

where B_1 is the BOLD response in each voxel measured during every trial (m voxels × n trials), W is a matrix that maps channel space to voxel space (m voxels × k channels), and C_1 is a design matrix of predicted channel responses in each trial (k channels × n trials). The weight matrix $\hat{W}$ was estimated by

$$\hat{W} = B_1C_1^T(C_1C_1^T)^{-1} \quad (2)$$

Then, using data from the held out test data set (B_2), the weight matrix estimated above was used to compute channel responses for every trial ($\hat{C}_2$), which were then sorted by task condition and spatial position:

$$\hat{C}_2 = (\hat{W}^T\hat{W})^{-1}\hat{W}^TB_2 \quad (3)$$

Reconstructing spatial representations. To reconstruct the region-wide representation of the visual stimulus viewed in every trial, we computed a weighted sum of the basis set using each channel response as the weight for the corresponding basis function (**Fig. 3b**). Reconstructions were computed out to a 5.234° eccentricity across the horizontal and vertical meridians, although visual stimuli only subtended at a maximum 4.523° eccentricity across the horizontal or vertical meridians. This was done to avoid edge artifacts in the reconstructions. Additionally, at this stage, the reconstructed visual fields were shifted to account for the slight jitter introduced in the presented stimulus locations and to align reconstructions from all trials. Runs in which stimuli were jittered up and to the left were reconstructed by moving the centers of the basis functions down and to the right, and runs in which stimuli were jittered down and to the right were reconstructed by moving the centers of the basis functions up and to the left. These shifts serve to counter the spatial jitter of stimulus presentation for visualization and quantification. By including spatial jitter during stimulus presentation, we are able to attain a more nuanced estimate of channel weights by sampling 72 stimulus locations rather than 36.

We averaged each participant's reconstructions at all 36 spatial locations for each task condition across trials. For **Figure 4**, all $n = 8$ participants' average reconstructions for each task condition were averaged, and reconstructions from all ROIs and task conditions were visualized on a common color scale to illustrate differences in spatial representations across the different task conditions and spatial locations. The three-by-three subset of reconstructions shown in **Figure 4b** was chosen as it is representative and results were similar for all quadrants.

For the follow-up control experiment, we plotted reconstructed spatial representations from only correct trials by co-registering all representations for trials at matching eccentricities and then averaging across all co-registered representations for each participant at each eccentricity. We co-registered representations for like eccentricities to the top left quadrant (**Fig. 6b**). Representations were rotated in 90° steps and flipped across the diagonal (equivalent to a matrix transpose operation on pixel values) as necessary.

Notably, this analysis depends on two necessary conditions. First, individual voxels must respond to certain spatial positions more than others, although the shape of these spatial selectivity profiles is not constrained to follow any particular distribution (for example, it need not resemble a Gaussian distribution). Second, the spatial selectivity profile for each voxel must be stable across time, such that spatial selectivity estimated on the basis of data in the training set can generalize to the held-out test set.

Curve fitting. To quantify the effects of attention on visual field reconstructions, we fit a basis function to all 36 average reconstructions for each participant for each task condition for each ROI using fminsearch as implemented in MATLAB 2012b (which uses the Nelder-Mead simplex search method; Mathworks, Inc).

The error function used for fitting was the sum of squared errors between the reconstructed visual stimulus and the function

$$f(r) = b + a \times (0.5 + 0.5\cos(r\pi / s))^7 \quad \text{for} \quad r < s,\ 0 \text{ elsewhere} \qquad (4)$$

where r is computed as the Euclidean distance from the center of the fit function. We allowed the baseline (b), amplitude (a), location (x, y) and size (s) to vary as free parameters. The size s was restricted so as not to be too large or too small (confined to $0.5815° < s < 26.17°$), and the location was restricted around the region of visual stimulation (x and y lie within stimulus extent borders +1.36° on each side).

Because of the number of free parameters in this function, we performed a two-step stochastic curve-fitting procedure to find the approximate best-fit function for each reconstructed stimulus. First we averaged reconstructions for each spatial location across all three task conditions and performed 50 fits with random starting points. The fit with the smallest sum squared error was used as the starting point around which all other starting points were randomly drawn when fitting to reconstructions from each task condition individually. When fitting individual task condition reconstructions, we performed 150 fits for each condition. We used parameters from the fit with the smallest sum squared error as a quantitative characterization of the reconstructed visual stimulus. Then we averaged the fit parameters across like eccentricities within each task condition, ROI and participant. For the follow-up control experiment, we performed an identical fitting procedure on each of the co-registered representations to directly estimate the best fit parameters at each eccentricity.

Excluded participant. For one participant (AG3), reconstructions from the attend fixation runs were unusually noisy and could not be well approximated by the basis function used for fitting. However, both the attend stimulus and spatial working memory runs for this individual exhibited successful reconstructions (**Supplementary Fig. 4**). As the estimated channel weights used to compute these stimulus reconstructions were identical across the three task conditions, only changes in information coding across task demands could account for this radical shift in reconstruction fidelity. Because this participant's reconstructions could not be accurately quantified for the attend fixation condition, the reconstructions and fit parameters for this individual for all conditions have been left out of the data presented in the Results. However, as noted above, data from this participant are consistent with our main conclusion that attentional demands influence the quality of spatial representations.

Evaluating the relationship between amplitude and size. It may be the case that our observation of increasing spatial representation size with increasing stimulus eccentricity is purely a result of intertrial variability in the reconstructed stimulus position. That is, the same representation could be jittered across trials, and the resulting average representation across trials would appear 'smeared' and would be fit with a larger size and smaller amplitude. If this were true, changes in these parameters would always be negatively correlated with one another—an increase in size across conditions would always occur with a decrease in amplitude.

To evaluate this possibility, for each eccentricity, ROI and condition pair (attend stimulus and attend fixation, spatial working memory and attend stimulus, and spatial working memory and attend fixation), we correlated the change in size with the change in amplitude (each correlation contained eight observations, corresponding to $n = 8$ participants). To evaluate the statistical significance of these correlations, we repeated this procedure 10,000 times, each time shuffling the condition labels separately for size and amplitude, recomputing the difference and then recomputing the correlation between changes in size and changes in amplitude. This resulted in a null distribution of chance correlation values against which we determined the probability of obtaining the true correlation value by chance. After correction for FDR, no correlations were significant (all $P > 0.05$; of note, FDR is more liberal than Bonferroni correction).

Representations from the ventral and dorsal aspects of V2 and V3. For **Supplementary Figure 5a**, we generated reconstructions using a procedure identical to that used for **Figure 4**, except we only used voxels that were assigned to the dorsal or ventral aspects of V2 and V3 instead of combining voxels across the dorsal and ventral aspects, as was done in the main analysis.

Reconstructions of untrained (new) stimuli. For **Supplementary Figure 5b**, we estimated channel weights using all runs of all task conditions from the main task as a training set. We used these weights to estimate channel responses from the BOLD data taken from an entirely new data set, which consisted of responses to a hemi-annulus–shaped radial checkerboard (**Supplementary Fig. 5b**).

This new experiment featured four stimulus conditions: left-in, left-out, right-in and right-out. The inner hemi-annuli subtended at 0.633° to 2.262° eccentricity. The outer hemi-annuli subtended at 2.262° to 4.523° eccentricity. Stimuli were flickered at 6 Hz for 12 s in each trial while the participants performed a spatial working memory task on small probe stimuli presented at different points within the displayed stimulus.

The BOLD signal used for reconstruction was taken as the average of four TRs beginning 4.5 s after stimulus onset. These data were used as the test set. Otherwise, the reconstruction process was identical to that in the main experiment, as were all other scan parameters and preprocessing steps.

pRF estimation. To determine whether the spatial sensitivity of each voxel across all trials and all runs changed across conditions, we implemented a new version of a pRF analysis[42,55]. For this analysis, we estimated the unimodal, isotropic pRF that best accounts for the BOLD responses to each stimulus position within every single voxel. This analysis is complementary to the primary analyses described above.

For four participants (those presented in **Figs. 6** and **7** and **Supplementary Fig. 7**) and four ROIs for each participant (V1, hV4, hMT+ and IPS0, which were chosen because this set includes both ROIs with (hV4, hMT+ and IPS0) and without (V1) attentional modulation), we used data across all runs within each task condition and ridge regression[56] to identify pRFs for each voxel under each task condition. We computed these pRFs using a method similar to that used to compute channel weights in the encoding model analysis (**Fig. 3a** and Online Methods; the univariate step 1 of the encoding model, see equation (1)). We generated predicted responses with the same information channels that were used for the encoding model analysis (**Fig. 3a**), and reconstructed pRFs for each task condition for a given voxel were defined as the corresponding spatial filters weighted by the computed weight for each channel (**Supplementary Fig. 8a**).

In the main analysis in which we computed spatial reconstructions on the basis of activation patterns across an entire ROI (**Figs. 4** and **6c**), any spatial information encoded by a voxel's response could be exploited; this is true even if the voxel's response to different locations was not unimodal (it need not follow any set distribution, as long as it responds consistently). However, univariate pRFs

computed on a voxel-by-voxel basis cannot be well characterized by an isotropic function if they are not unimodal[57]. Thus, to ensure that most of the pRFs were sufficiently unimodal to fit an isotropic function, we used ridge regression[56,57] when computing spatial filter weights for the pRF analysis. The regression equation for computing channel weights then becomes

$$\hat{W}^T = (C_1 C_1^T + \lambda I)^{-1} C_1 B_1^T \quad (5)$$

where I is an identity matrix ($k \times k$). To identify an optimal ridge parameter (λ), we computed the Bayes information criterion[58] value across a range of λ values (0 to 500) for each voxel using data concatenated across all three task conditions. This allowed for an unbiased selection of λ with respect to task condition. The λ with the minimum mean Bayes information criterion value across all voxels within a ROI was selected, and this λ was used to compute channel weights for each of the three task conditions separately. An increasing λ value results in greater sparseness of the best-fit channel weights for each voxel, and a λ value of 0 corresponds to ordinary least-squares regression.

After computing pRFs for each task condition, we fit each pRF with the same function that was used to fit the spatial representations (equation (4)) using a similar optimization procedure. We restricted the fit size (FWHM) to be at most 8.08°, which corresponds to nearly the full diagonal distance across the stimulated visual field. This boundary was typically encountered only for hMT+ and IPS0 and served to discourage the optimization procedure from fitting large, flat surfaces. Then we computed an R^2 value for each fit and used only voxels for which the minimum R^2 across conditions was greater than or equal to the median of the minimum R^2 across conditions from all voxels in that participant's ROI (**Supplementary Fig. 8a**,**b**).

Because we only have a single parameter estimate for each condition for each voxel, we evaluated whether fit size is more likely to increase or decrease between each pair of task conditions (attend stimulus compared to attend fixation, spatial working memory compared to attend stimulus and spatial working memory compared to attend fixation) for each region for each participant by determining the percentage of voxels that lie above the unity line in a plot of one condition against another (**Supplementary Fig. 8d**).

Simulating data with different pRF properties. In order to assess whether our region-level multivariate spatial representation analysis would be sensitive to changes in voxel-level univariate pRFs, we generated simulated data using two different pRF modulation models.

For the first model (**Supplementary Fig. 10a**,**b**), we randomly generated 500 pRF functions so as to uniformly sample the visual field for each of two conditions (condition A, smaller pRFs; condition B, larger pRFs). Across the two conditions, each simulated voxel's pRF maintained its preferred position while its amplitude and baseline were each randomly and independently sampled across conditions from the same normal distribution (amplitude: $\mu = 0.8513$, $\sigma = 0.25$; baseline: $\mu = -0.1952$, $\sigma = 0.25$; these values were taken from the average-fit pRF parameters across all participants for hV4 in the attend fixation and attend stimulus conditions; **Supplementary Fig. 9a**). pRF size (FWHM) was sampled from a normal distribution with $\sigma = 0.5$ and a mean of $\mu = 4.405°$ for condition A (mean of pRF size for hV4, attend fixation) and $\mu = 4.89°$ for condition B (mean of pRF size for hV4, attend stimulus; an increase of 11%). In our simulation, this resulted in 79% of simulated voxels showing larger pRF sizes in condition B compared to condition A. For the second model (**Supplementary Fig. 10c**,**d**), we used the upper median split of fit pRFs for the single participant shown in **Supplementary Figure 8c**, hV4 ROI, to generate the simulated BOLD data. This allowed us to simulate region-level BOLD data for each attention condition tested in our experiment and enabled us to determine whether the changes in univariate voxel-level pRF size we observed (**Supplementary Fig. 9**) are consistent with the multivariate region-level spatial representations presented in the main text (**Figs. 5** and **7**).

After generating voxel-level pRFs using each of the two models described above, we added noise to the simulated weights (Gaussian noise added independently to each channel weight, $\sigma = 0.1$) and presented model voxels with six runs of all 36 spatial positions for each condition. We simulated each voxel's BOLD response as the predicted channel response (response of corresponding spatial filter; **Fig. 3a**) to each stimulus weighted by the corresponding channel weights. We added Gaussian noise to the resulting BOLD data for each simulated voxel independently ($\sigma = 0.1$). Then all analyses of multivariate spatial representations proceeded identically to those described above. We computed spatial representations using estimated channel weights computed across all conditions within a model (i.e., condition A and condition B (**Supplementary Fig. 10a**,**b**) or attend fixation, attend stimulus and spatial working memory (**Supplementary Fig. 10c**,**d**)) and then fit the average spatial representations with a smooth surface (Online Methods) to determine the amplitude and size of each spatial representation. We then averaged these parameters across all 36 positions.

Statistical procedures. All behavioral analyses on accuracy data were performed using two-way repeated-measures ANOVA, with task condition and stimulus eccentricity modeled as fixed effects (three levels and six levels, respectively; **Figs. 2d** and **6a**).

To assess whether fit parameters to reconstructed spatial representations reliably changed as a function of task demands, we performed a multistage permutation testing procedure. This nonparametric procedure was adopted because the spatial filters (basis functions) used to estimate the spatial selectivity of each voxel during the training phase (**Fig. 3a**) overlapped and were not independent (violating a key assumption of standard statistical tests).

For each parameter (the rows in **Figs. 5** and **7**), we first found ROI-parameter combinations that showed an omnibus main effect in a repeated-measures ANOVA (1 factor, 18 levels) corrected using a FDR algorithm[59] across all ROIs. Then we computed F scores for a two-way repeated measures design with eccentricity and condition as factors (six levels and three levels, respectively) for ROIs with significant omnibus main effects.

For all tests, because we had a relatively small n ($n = 8$ for **Fig. 5**, and $n = 4$ for **Fig. 7** and **Supplementary Fig. 7**) and the range of parameters was in some cases restricted to be positive (size), we computed an F distribution for the null hypothesis that there is no main effect of the omnibus factor (omnibus test) or that there is no main effect of condition, eccentricity or their interaction (for a follow-up two-way test) by shuffling trial labels within each participant 100,000 times. For each data permutation, we computed a new F score for the omnibus test, and for ROI-parameter combinations with a significant omnibus effect, we computed a main effect of condition, eccentricity and their interaction. P values were estimated as the probability that the F score computed based on the shuffled data was equal to or greater than the F scores computed using the actual data. These additional tests were corrected for multiple comparisons using Bonferroni's method within each parameter. We also occasionally highlight trends in the data by reporting P values that did not reach significance under correction for multiple comparisons at this sample size as marginal effects, and such P values are always reported as being uncorrected in the text. For display purposes, marginally significant tests are shown in **Figures 5** and **7** at uncorrected $P < 0.025$.

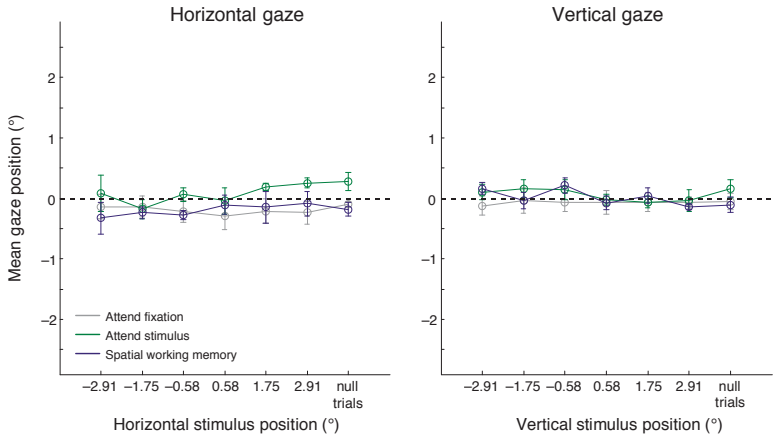
In addition, we performed a three-factor repeated-measures ANOVA with ROI, task condition and eccentricity modeled as fixed effects to determine whether the fit parameters changed across ROIs ($n = 8$ for **Fig. 5**, and $n = 4$ for **Fig. 7**). We implemented the same permutation procedure described above to compute P values (10,000 iterations).

To determine whether pRF size increases at higher eccentricities, we computed a linear fit to a plot of each voxel's pRF size compared to its pRF eccentricity for each ROI for each condition for each participant (**Supplementary Fig. 8c**). To determine whether the slope of the fit line was reliably positive for a given ROI, participant and condition, we computed confidence intervals around the best-fit slopes using bootstrapping (resampled all voxels with replacement 10,000 times), and the related P value was defined as the as the probability that the slope was ≤0. We used a Bonferroni-corrected significance threshold for 48 planned comparisons (4 ROIs × 4 participants × 3 conditions) of $\alpha = 0.001$ (**Supplementary Results** and **Supplementary Table 2**).

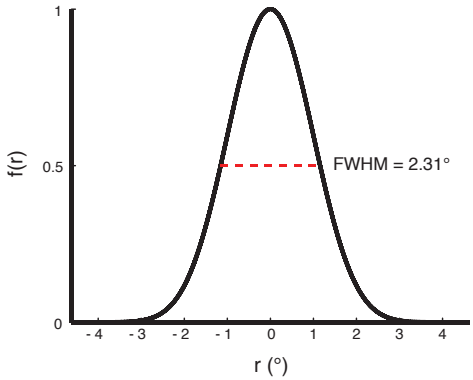
To evaluate the statistical significance of the pRF size increase (**Supplementary Fig. 9**), we first performed a two-way repeated-measures ANOVA with ROI and condition modeled as fixed effects and participant modeled as a random effect in which we shuffled ROI and condition labels for each participant and recomputed the percentage of voxels that increased in size across each condition pair. We repeated this shuffling procedure 10,000 times and compared F scores computed using the real labels to the distribution generated using the shuffled labels, as described above. Then we compared whether each condition pairing resulted in a significant change in pRF size for each ROI by computing a T score testing against the null hypothesis that 50% of voxels show an increase in pRF size. As described above, we generated a null T distribution by shuffling condition labels

within each participant 10,000 times. For this analysis, we used a Bonferroni-corrected significance threshold for 12 planned comparisons (4 ROIs × 3 conditions) of $\alpha = 0.0042$.

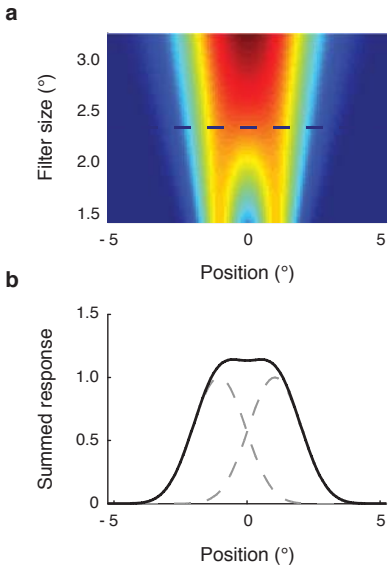
51. Engel, S.A. *et al.* fMRI of human visual cortex. *Nature* **369**, 525 (1994).
52. Sereno, M.I. *et al.* Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* **268**, 889–893 (1995).
53. Tootell, R.B. *et al.* Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. *J. Neurosci.* **15**, 3215–3230 (1995).
54. Serences, J.T. & Boynton, G.M. The representation of behavioral choice for motion in human visual cortex. *J. Neurosci.* **27**, 12893–12899 (2007).
55. Kay, K.N., Naselaris, T., Prenger, R.J. & Gallant, J.L. Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).
56. Hoerl, A.E. & Kennard, R.W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
57. Lee, S., Papanikolaou, A., Logothetis, N.K., Smirnakis, S.M. & Keliris, G.A. A new method for estimating population receptive field topography in visual cortex. *Neuroimage* **81**, 144–157 (2013).
58. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
59. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).



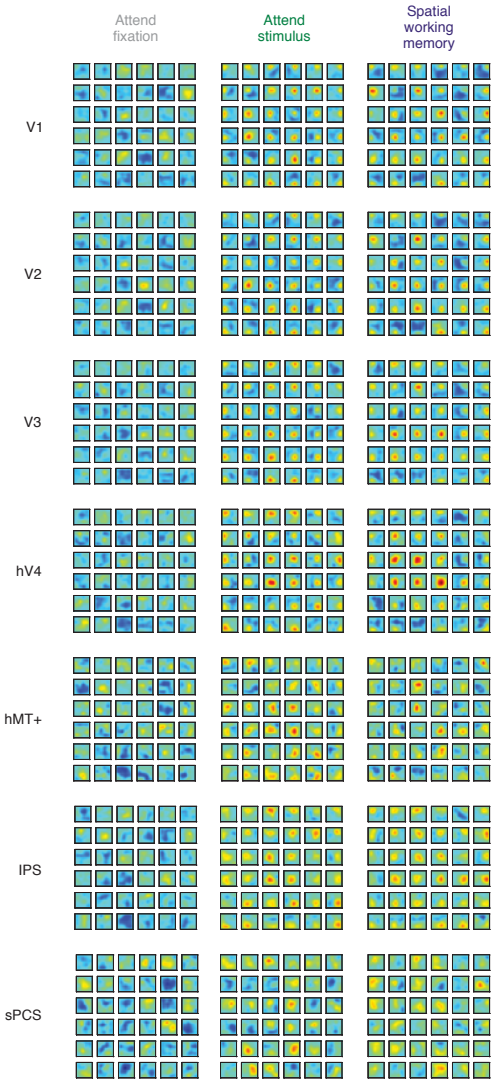
Supplementary Figure 1 Participants maintained fixation in the scanner during all three task conditions, related to Figure 2. Average horizontal and vertical gaze position across each 3 s trial in each task condition. Neither horizontal nor vertical gaze varied as a function of either stimulus position or task demands. 2-way ANOVA for each gaze direction, with task condition and stimulus position (grouped into 6 bins corresponding to *x* or *y* coordinate for horizontal and vertical gaze plots, respectively) as factors: minimum *p* for main effects/interactions = 0.2725, which was for main effect of vertical stimulus position on vertical gaze. Note that data from null trials were not entered into the ANOVA, but subjects maintained steady fixation on these trials as well. Eyetracking data gathered in the scanner for 4 of the 8 participants. Error bars ±1 S.E.M. across subjects.



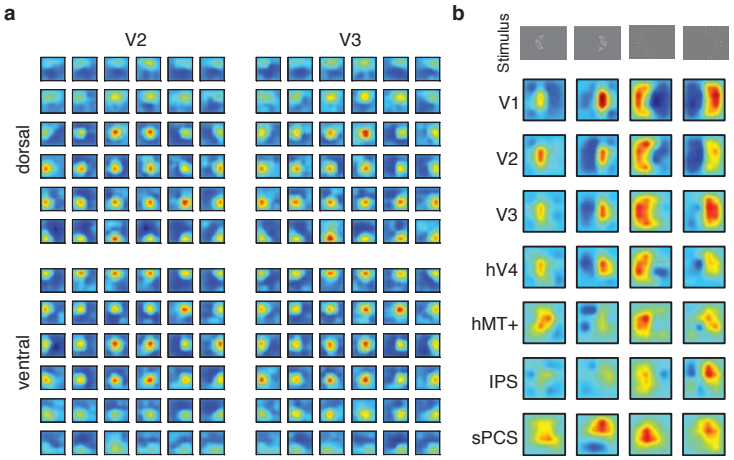
Supplementary Figure 2 One-dimensional cross-section of 2D basis function, related to Figure 3. Cross-section through the center of a single basis function (**Figure 3a**). FWHM is the full-width at half-maximum. The size constant, *s,* was set to $5r_{stim}$ (see Online Methods: *Encoding model*, **Supplementary Fig. 3**), where r_{stim} is 1.17°. This corresponds to the distance from the center at which the filter amplitude reaches 0.



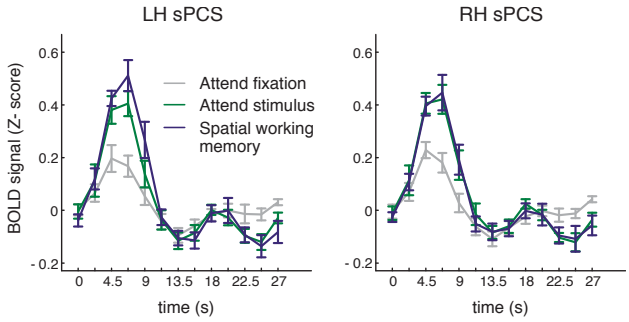
Supplementary Figure 3 The relationship between basis function size and spacing changes the smoothness of reconstructions, related to Figures 3 and 4. (**a**) For a constant spatial filter separation distance of 2.09° (which matches that used in the main analysis), we varied the size parameter (**Supplementary Fig. 2**) of 2 neighboring spatial filters, then plotted their sum as a function of position in space and filter size (which was continuously varied). Summed response is indicated by the image colorscale. (**b**) A slice from (**a**) at the FWHM of the filters used in the main analysis (dashed line in panel **a**). This value resulted in smooth reconstructions to which we could accurately fit surfaces to quantify the spatial representations (see Online Methods: *Curvefitting*), but also resulted in sufficient filter separation so that adjacent filters did not excessively overlap (see below). Smaller FWHM values would result in speckled reconstructed spatial representations which would be poorly fit using a single surface (this would be seen as a dipped black solid line in panel **b**; see panel **a** at small filter size values), and larger FWHM values would result in poorly discriminable predicted channel responses because neighboring filters would account for much of the same variance in the signal due to a high degree of overlap (see **a**, high FWHM values). At high enough FWHM values, the model cannot be estimated because overly high correlations between adjacent filters result in a rank deficient design matrix (Equation 1 in Online Methods).



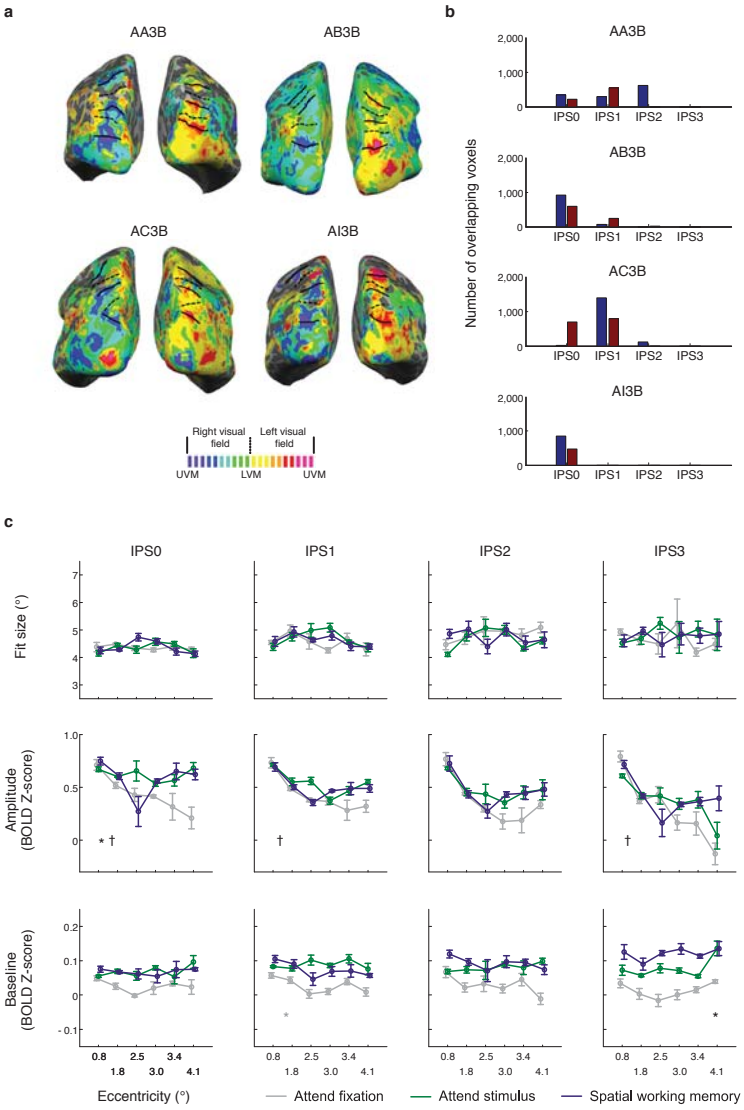
Supplementary Figure 4 Poor reconstructions during attend fixation condition for participant AG3, related to Figure 4. Plotted as in **Figure 4**. All images on same color scale. Poor reconstructed spatial representations were measured during attend fixation runs across all ROIs, but more typical looking reconstructed spatial representations were observed for both of the other task conditions. Behavioral performance for this participant indicated they were awake and vigilantly performing the fixation task. This was the only participant with this issue, and their data were not included in **Figures 4** or **5** (see Online Methods: *Excluded participant*). Note that the same estimated channel weight matrix was used here as was used to reconstruct spatial representations during the attend stimulus and spatial working memory tasks. Furthermore, note that these data support our conclusion of higher amplitude spatial priority maps with attention and they were excluded solely because of the noisy fits in the fixation condition. All of our reported effects would be more pronounced if this participant was included (see data included in the html version of this report).



Supplementary Figure 5 Encoding model does not overfit data and generalizes to novel stimuli, related to Figure 4. (**a**) Reconstructions from all 36 stimulus locations under the attend stimulus condition across all 8 observers using voxels from only the ventral and dorsal aspects of V2 and V3. Color scale is identical to that used in Figure 4. Note that spatial reconstructions in the dorsal & ventral aspects of V2 and V3 are more robust in the lower and upper visual field, respectively. This pattern matches the known selectivity of dorsal and ventral areas V2 and V3. (**b**) Encoding model can be generalized to reconstruct novel stimuli that were not part of the *training set*. An encoding model trained using all attend fixation, attend stimulus & spatial working memory runs was able to accurately reconstruct a novel, untrained stimulus set acquired during a different scanning session on 7 of 8 participants presented in **Figures 4-5** (novel test data was not available for this 8th participant, AA3). This novel stimulus display consisted of four half-circle stimuli presented at one of two eccentricities (see top row), and the model was able to reconstruct these four stimuli with a high degree of precision (see Online Methods: *Stimulus reconstructions – novel stimuli* for more details).

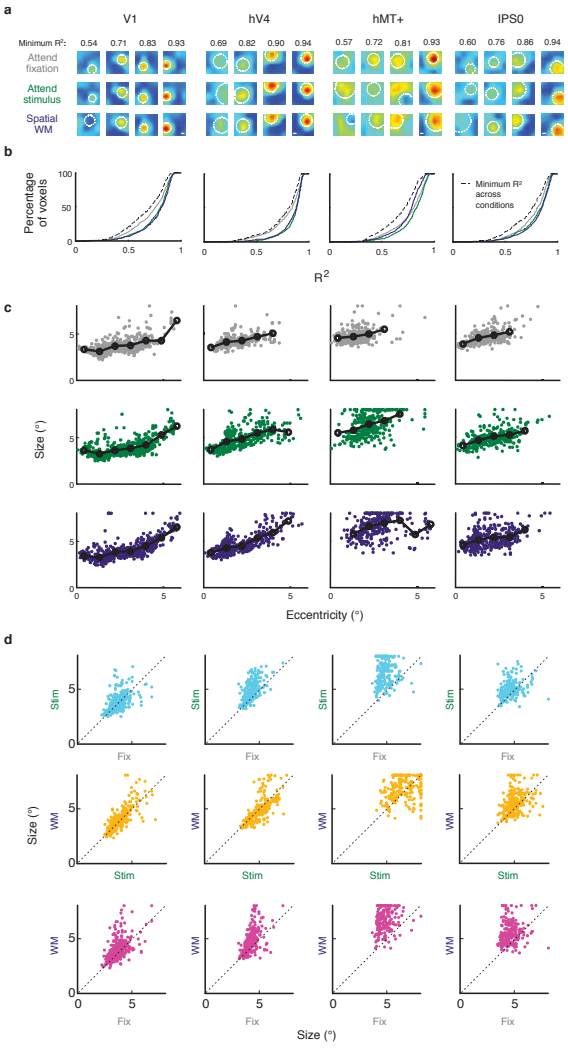


Supplementary Figure 6 sPCS exhibits larger responses, averaged across all voxels within the sPCS, in the attend stimulus and spatial working memory conditions, related to Figures 4 and 5. Both left and right sPCS exhibit strong hemodynamic responses to stimuli, with increased averaged (i.e. univariate) responses during attend stimulus and spatial working memory task conditions compared to the attend fixation condition. Additionally, this mean signal increase under conditions of attention to the stimulus or spatial working memory likely accounts for much of the significant increase in the baseline offset in the reconstructed stimulus representations reported in **Figure 5**. Error bars ±1 SEM across subjects.



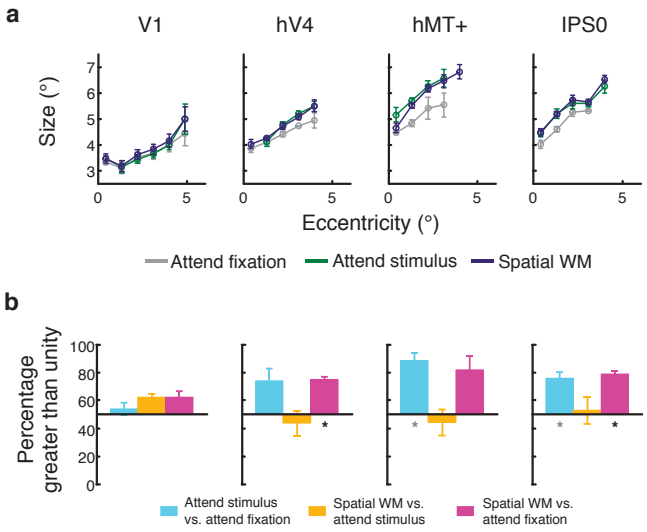
Supplementary Figure 7 IPS ROI primarily corresponds to IPS 0/1, related to Figures 6 and 7. (**a**) Polar angle preferences for each voxel plotted on the inflated surface of 4 participants' cortical sheets. Maps are liberally thresholded to show any voxel with normalized power at the stimulus frequency > 0.005. Smooth polar angle transitions were used to delineate four retinotopic regions of IPS (termed IPS 0-3) in each of these 8 hemispheres. Dashed lines: lower vertical meridian (LVM); solid lines: upper vertical meridian (UVM). (**b**) For each participant and each hemisphere we compared the number of overlapping voxels between our original localizer-defined IPS ROI (see Online Methods: *Mapping IPS subregions*) and each of these 4 retinotopically mapped IPS subregions . The original IPS ROI primarily overlaps with areas IPS 0 and 1, and is therefore

labeled as such in **Figure 7**. Blue: left hemisphere, Red: right hemisphere. (**c**) Fit parameters to reconstructed spatial representations estimated from activation patterns across each IPS subregion for these 4 participants (analysis identical to that implemented for **Fig. 7**). Critically, fit parameters in all regions are similar to those observed for the original IPS subregion (**Fig. 7**). Spatial representations of presented stimuli do not narrow in size when stimuli are attended or a target is remembered, but amplitude increases for representations in IPS0 ($p = 0.012$), and baseline increases in IPS3 ($p = 0.002$; statistics as in **Figs. 5** & **7**; error bars within-participant S.E.M.)

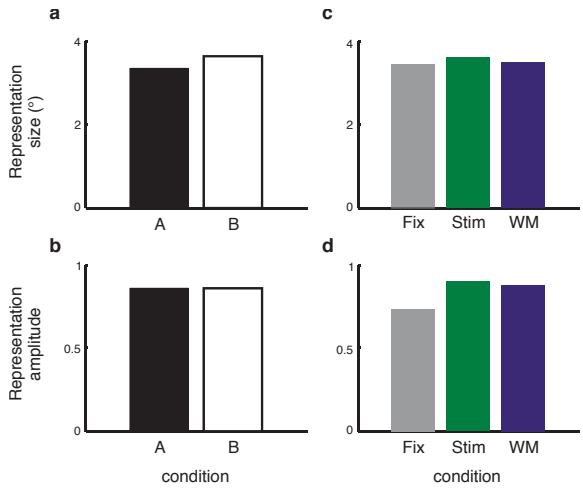


Supplementary Figure 8 Population receptive field analyses: example participant AA3B, related to Figure 7. (**a**) Reconstructed pRFs and best-fit isotropic function for voxels at each interquartile boundary. White dashed circles are plotted at half-maximum of fit function. Quartiles were split by minimum R^2 across all task conditions (see Online Methods: *Population receptive field estimation*). Above each column of pRFs is the minimum R^2 value across the 3 conditions shown below. The right 3 columns (top 50%) are voxels that were included in subsequent analyses. White horizontal scale bars correspond to 1° visual angle. (**b**) Distribution of R^2 (colored lines) and minimum R^2 across conditions (black lines) for each voxel, plotted as a cumulative distribution. (**c**) Size vs. eccentricity for each condition for each ROI. Each data point corresponds to a single voxel. Black circles/lines are the mean size at each eccentricity bin which contains ≥ 5 voxels (these are the points which are included in **Supplementary Fig. 9a**). All slopes for this example participant are significantly >

0 after Bonferroni correction across all 48 tests (4 participants × 4 ROIs × 3 conditions, corrected $\alpha = 0.001$), except hMT+, spatial working memory condition ($p = 0.006$, see Online Methods: *Statistical Procedures*). (**d**) Distribution of pRF size for each voxel across condition pairs. The percentage of voxels which lie above the unity line (that is, the percentage of voxels for which the size increases) within a ROI and condition pair is used to evaluate whether task demands significantly change pRF size (see **Supplementary Fig. 9b**, Supplementary Results, Online Methods).



Supplementary Figure 9 Population receptive fields increase size with attention, related to Figure 7. (**a**) Summary of pRF size as a function of eccentricity across $n = 4$ participants. Each data point is plotted if ≥ 3 participants each had ≥ 5 voxels within that eccentricity bin. Error bars S.E.M. across included participants. (**b**) Summary of pRF size changes across each condition pair. For each ROI for each participant, we computed the percentage of voxels in which the pRF size was greater for the first condition than the second (e.g., cyan bars indicate the percentage of voxels in which pRF size was greater for the attend stimulus condition than for the attend fixation condition; this corresponds to the percentage of voxels which lie above unity when plotted as in **Supplementary Fig. 8d**). Black asterisks indicate significant size changes across a condition pair for an ROI, Bonferroni-corrected (two-tailed t-test, see Online Methods: *Statistical procedures*). Gray asterisks indicate a significant size change using a one-tailed t-test. Error bars indicate S.E.M. across participants ($n = 4$).



Supplementary Figure 10 Simulations demonstrate that uniform changes in voxel-level pRFs are reflected in changes in region-level spatial representations, related to Figure 7. (**a-b**) For 500 simulated voxels, we generated data for 2 conditions in which we only manipulated the simulated pRF size (condition B uses pRFs that were on average 11% larger than pRFs in condition A, which corresponds to the measured increase in pRF size between "attend stimulus" and "attend fixation" conditions in hV4 across all 4 participants). Under these conditions, the size of the multivariate spatial representations scaled with pRF size (**a**, smaller *spatial representation* sizes in condition A than in condition B). However, note that in this scenario, there is no change in fit amplitude (**b**). This demonstrates that (1) multivariate spatial representations are sensitive to changes in pRF size, given that the changes occur uniformly across a region, and (2) that our analysis technique can detect size changes mediated by uniform changes in pRF size in the absence of amplitude changes, were they occurring. This rules out an important possibility that representation size changes might be occurring in our dataset, but they could be too small to measure (see Results: *Size of spatial representations across eccentricity and ROI*). (**c-d)** In panels **a-b** we demonstrate that multivariate region-level spatial representations can increase in size, reflecting uniform changes in the underlying univariate voxel-level pRFs. Here, we used the fit pRF parameters for 1 example participant (AA3B, shown in **Supplementary Fig. 8**) and 1 ROI (hV4), which undergo non-uniform size changes across conditions, to simulate data for all 3 task conditions in the main experiment. Even with pRF size increases observed across conditions (**Supplementary Fig. 8d**), multivariate spatial representations are shown to maintain a constant size (**c**), but increase in amplitude (**d**, mirroring our data in **Figs. 5, 7**). This pattern of results was also found in the other three participants (not shown). This demonstrates a decoupling of pRF size/amplitude and the size/amplitude of multivariate region-level spatial representations, and underscores the importance of exploiting all of the information available in a region to estimate the fidelity of spatial encoding.

	X	Y	Z	volume (mL)
RH-IPS	27.78 ± 3.37	-71.21 ± 4.23	29.33 ± 3.38	0.961 ± 0.37
LH-IPS	-26.80 ± 2.38	-71.32 ± 5.44	27.35 ± 4.30	1.301 ± 0.57
RH-sPCS	30.77 ± 6.36	-5.54 ± 5.00	49.67 ± 3.74	1.411 ± 0.37
LH-sPCS	-28.76 ± 4.22	-8.14 ± 4.44	46.95 ± 1.72	1.17 ± 0.28
RH-hMT+	39.54 ± 3.08	-66.15 ± 5.95	3.03 ± 4.06	1.01 ± 0.12
LH-hMT+	-43.79 ± 6.55	-70.31 ± 5.60	3.32 ± 4.97	0.894 ± 0.20

Supplementary Table 1 Mean ROI sizes and locations, related to Figures 4-5. Mean ± 1 standard deviation locations of ROI centers in Talairach coordinates and volumes for hMT+, IPS and sPCS ROIs.

Mean slope ± standard error across participants (significant participants/4)	**V1**	**hV4**	**hMT+**	**IPS0**
Attend fixation	0.332 ± 0.028 (4)	0.330 ± 0.038 (3)	0.484 ± 0.171 (3)	0.554 ± 0.032 (4)
Attend stimulus	0.381 ± 0.030 (4)	0.536 ± 0.039 (4)	0.621 ± 0.190 (3)	0.479 ± 0.031 (4)
Spatial WM	0.414 ± 0.055 (4)	0.495 ± 0.082 (4)	0.647 ± 0.148 (3)	0.451 ± 0.057 (4)

Supplementary Table 2 pRF size vs. eccentricity slope, related to Figure 7. Each cell contains mean slope in units of pRF size (°)/eccentricity (°), as well as the number of participants with significantly nonzero size vs. eccentricity slopes. All participants, regardless of significance, are included in the mean and standard error. Number of significant participants is evaluated using a Bonferroni-corrected alpha value for 48 comparisons of $\alpha = 0.001$.

Supplementary Results

Population receptive fields (pRFs)

As a complementary analysis, we identified pRFs for each voxel for each condition for a subset of participants and ROIs (see Online Methods: *Population receptive field estimation*). We restricted our analysis to the half of all voxels in each region for which reconstructed pRFs were well-fit with a unifocal isotropic function (**Supplementary Fig. 8a**). In our implementation of the pRF analysis, we reconstruct a map of the portion(s) of the visual field which best drive the BOLD response in each voxel (see **Supplementary Fig. 8a** for example pRFs in each ROI for 1 participant). Then, we fit a smooth 2d surface to each of these reconstructed pRFs, and use the best fit position and size to characterize pRF properties across different attention conditions (best fits shown by white circles in **Supplementary Fig. 8a**). Note that, while most fits appeared accurate, some "best" fits do not accurately capture the positive region of the pRF. To choose "good" fits without bias, we computed an R^2 value for each condition. Then for every voxel we used the minimum R^2 across conditions to determine a median R^2 for every ROI for every participant (**Supplementary Fig. 8b**). Voxels with minimum R^2 greater than or equal to the corresponding median value were included in subsequent analyses.

Because our stimulus set and analysis method was not designed to evaluate pRFs at the resolution that dedicated pRF mapping protocols are (e.g., ref 42), we found generally larger pRFs than have been observed previously. However, we replicated the key pRF result that pRF size increases with eccentricity (**Supplementary Fig. 8c**, **Supplementary Fig. 9a**). At least 3 of the 4 participants had significantly positive slopes for each condition/ROI pairing (see **Supplementary Table 2** for mean slopes and number of significant participants), and all significantly non-zero slopes were positive.

Next, we compared whether the best-fit pRF size increased or decreased in more voxels between each pair of task conditions (attend stimulus vs. attend fixation, spatial WM vs. attend stimulus and spatial WM vs. attend fixation). **Supplementary Fig. 8d** shows this analysis for an example participant. At the group level, the percentage of voxels which lie above the unity line (**Supplementary Fig. 9b**) changed as a function of condition pair (significant main effect of condition pair, $p = 0.007$), and we also observed a condition × ROI interaction ($p = 0.047$). In 2 of 3 regions in which we observed significant increases in the amplitude of spatial representations with attention (**Fig. 5**, hV4 and IPS), we also observed significant size increases in at least one condition pair (hV4: spatial WM vs. attend fixation, $p < 0.001$; IPS0: spatial WM vs. attend fixation, $p = 0.003$; all others n.s. after Bonferroni correction for 12 comparisons, $\alpha = 0.0042$).

Simulated spatial representations

When pRF size is uniformly modulated across conditions (**Supplementary Fig. 10a-b**), we observed changes in the size of spatial representations. Spatial representation size reflects changes in pRF size. When pRF size was increased by 11%, spatial representation size increased by 9.28%. Our analysis method is thus sensitive to particular types of pRF modulation with attention, and so if pRF size is uniformly decreasing or increasing, the measured size of multivariate spatial representations would shrink or expand as a result.

When we used best-fit pRFs to data from each condition from a single participant's hV4 ROI, in which pRF size non-uniformly increases, we observed stable spatial representation size across conditions, as well as an increase in amplitude of spatial representations across attention conditions (**Supplementary Fig. 10c-d**), mirroring our results from the main text (**Figs. 5**, **7**). This demonstrates that, while our analysis method is sensitive to changes in representation size as a result of uniform pRF modulation with attention (**Supplementary Fig. 10a**), such a modulation is not observed when we account for the more nuanced pattern of pRF modulations that is present in our dataset (**Supplementary Fig. 10c**). Additionally, these results demonstrate that while voxel-level univariate pRFs might be subject to particular types of modulation with attention, the size of multivariate region-level spatial representations can remain stable and either increase or decrease in amplitude. This result also underscores the importance of constraining estimates of modulations in spatial information content of a region by the modulatory pattern across all component units (voxel-level pRFs) within the region, as there is not a one-

to-one mapping of changes in pRF size/amplitude to changes in the size/amplitude of region-wide spatial representations.

Chapter 2, in full, is a reprint of the material as it appears in an article entitled "Attention modulates spatial priority maps in the human visual, parietal, and frontal cortices" published in Nature Neuroscience 2013. Sprague, Thomas C.; Serences, John T., Nature Publishing Group, 2013. The dissertation author was the primary author of the manuscript. We thank Ed Vul and Sirawaj Itthipuripat for assistance with statistical methods and Miranda Scolari and Mary Smith for assistance with parietal cortex mapping protocols. This work was supported by a National Science Foundation Graduate Research Fellowship to T.C.S. and by US National Institutes of Health grant R0I MH-092345 and a James S. McDonnell Scholar Award to J.T.S.

Chapter 3:

Reconstructions of information in visual spatial working memory degrade with memory load

Current Biology 24, 2174–2180, September 22, 2014 ©2014 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2014.07.066

Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load

Report

Thomas C. Sprague,[1,*] Edward F. Ester,[2] and John T. Serences[1,2,*]
[1]Neurosciences Graduate Program, University of California, San Diego, La Jolla, CA 92093-0109, USA
[2]Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109, USA

Summary

Working memory (WM) enables the maintenance and manipulation of information relevant to behavioral goals. Variability in WM ability is strongly correlated with IQ [1], and WM function is impaired in many neurological and psychiatric disorders [2, 3], suggesting that this system is a core component of higher cognition. WM storage is thought to be mediated by patterns of activity in neural populations selective for specific properties (e.g., color, orientation, location, and motion direction) of memoranda [4–13]. Accordingly, many models propose that differences in the amplitude of these population responses should be related to differences in memory performance [14, 15]. Here, we used functional magnetic resonance imaging and an image reconstruction technique based on a spatial encoding model [16] to visualize and quantify population-level memory representations supported by multivoxel patterns of activation within regions of occipital, parietal and frontal cortex while participants precisely remembered the location(s) of zero, one, or two small stimuli. We successfully reconstructed images containing representations of the remembered—but not forgotten—locations within regions of occipital, parietal, and frontal cortex using delay-period activation patterns. Critically, the amplitude of representations of remembered locations and behavioral performance both decreased with increasing memory load. These results suggest that differences in visual WM performance between memory load conditions are mediated by changes in the fidelity of large-scale population response profiles distributed across multiple areas of human cortex.

Results

To assess the functional role that population codes in different visually responsive occipital, parietal, and frontal regions of interest (ROIs) play in spatial working memory (WM), we presented participants (n = 4, four scanning sessions each) with two target stimuli (Figure 1A) followed by a postcue instructing them to remember the location(s) of zero (R0), one (R1), or two (R2) stimuli. In behavioral testing sessions performed outside of the scanner, participants used a mouse click to indicate the exact position of the remembered target. During scanning, participants performed a two-alternative forced-choice (2AFC) discrimination task in which they compared the position of a probe stimulus to that of the corresponding remembered target stimulus (Figure 1A). We chose to test precise memory for spatial positions using either a recall task (outside the scanner) or a "same/different" task (during scanning) so that participants were required to encode exact spatial positions rather than use a verbal code or only encode a single dimension (e.g., "8 o'clock," "far to the left").

Behavioral performance on the analog recall task performed outside the scanner revealed lower mnemonic precision when two target locations were remembered compared to when a single target location was remembered (Figure 1C; $p < 0.001$, resampling test). During scanning, response accuracy did not significantly differ across set size conditions, although three out of four participants performed slightly worse with increasing set size (Figure 1D, $p = 0.174$, resampling test; see the Experimental Procedures). However, response times (RTs) were significantly longer when two stimuli were remembered compared to when a single stimulus was remembered (Figure 1E; $p < 0.001$, resampling test). Increased RTs during scanning suggest that memory representations in the R2 condition were degraded and were thus less accessible during behavioral report, consistent with previous observations of increased RTs after manipulations that impair spatial WM (e.g., [17]). Together, the behavioral data recorded inside and outside of the scanner are consistent with a degraded representation of each remembered location in the R2 condition compared to the R1 condition.

To characterize neural responses associated with WM maintenance, we first compared averaged blood-oxygenation-level-dependent (BOLD) functional magnetic resonance imaging (fMRI) responses in a set of functionally defined occipital (V1–hV4 and V3A), parietal (IPS0–IPS3), and frontal (sPCS; thought to be the human homolog of macaque frontal eye fields [18, 19]) ROIs as a function of memory load. We replicated previous reports that BOLD responses in frontal and parietal ROIs were larger on R2 trials compared to R1 trials [6, 20, 21] (Figure S1 available online). Interestingly, in early visual areas (V2–V3A and hV4) we observed a *larger* mean BOLD amplitude on R0 trials compared to R1 or R2 trials (Figure S1B, $p < 0.001$, resampling test). We also observed similar results using a complementary exploratory analysis in which we searched for any voxels with increased activation for larger memory loads (Figure S1C).

Next, we used a multivariate image reconstruction technique based on a spatial encoding model [16] to reconstruct remembered locations in spatial WM based on the *pattern* of activation across all voxels within each ROI (Figure 2). In contrast to analyses that focus solely on mean signal intensity (Figure S1), neural firing rates, or multivariate classification accuracy, this analysis uses an independently estimated model of the spatial sensitivity profile across all voxels in each ROI to transform BOLD activation patterns into an image of the remembered stimulus position(s) carried by those patterns (Figure 2; Experimental Procedures). Importantly, this analysis provides additional information compared to some other methods such as univariate population receptive field (pRF) [22] estimation or multivariate linear classification [9]: by yielding a reconstructed image of the remembered stimulus location(s), covert information held in WM can be directly visualized, quantified, and

*Correspondence: tsprague@ucsd.edu (T.C.S.), jserences@ucsd.edu (J.T.S.)

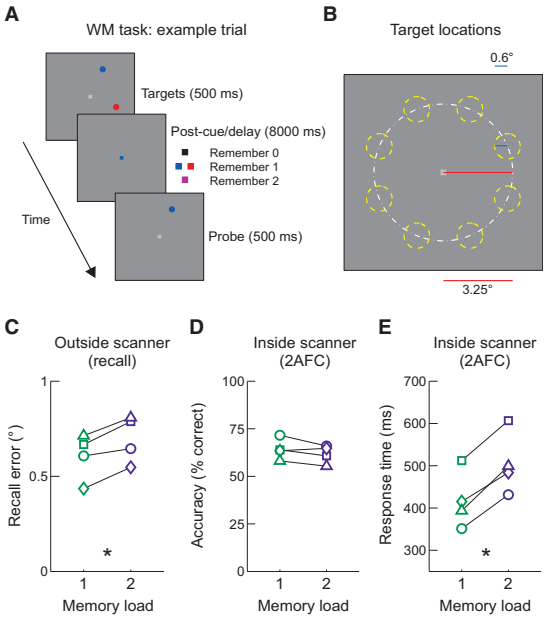


Figure 1. Visual Spatial WM Task and Behavioral Performance

(A) Participants (n = 4) viewed two target stimuli and were postcued to passively fixate for the remainder of the trial (remember zero), remember the precise position of a single target stimulus (remember one), or remember the precise position of both target stimuli (remember two). After a 8 s delay, participants either determined whether a probe stimulus was in exactly the same or a slightly different position as the corresponding target (during fMRI scanning sessions, 0.1–1.5° offset) or precisely recalled the remembered position using a computer mouse (during behavioral sessions).

(B) So that implementation of a "digital" encoding strategy could be discouraged, each target was presented within one of eight discs with uniform jitter equally spaced around fixation and offset from horizontal and vertical meridians.

(C) During behavioral testing sessions outside of the scanner, spatial positions were remembered less precisely with larger memory load as indicated by increased behavioral recall error distance ($p < 0.001$), and this is qualitatively observed for each participant. Each symbol is a single participant, and symbols match those presented in (D) and (E) and Figures S1 and S2.

(D) During scanning, behavioral accuracy was approximately equal across set sizes ($p = 0.174$).

(E) Response times inside the scanner were significantly longer for larger memory load trials ($p < 0.001$).

Throughout all figures, unfilled symbols refer to single-participant data; filled symbols refer to across-participant means. Asterisks reflect significant across-participant resampling tests; see the Experimental Procedures. See also Figure S1.

related to behavior [16]. These *reconstructions* can be thought of as an image of the spatial WM contents in visual field coordinates (rather than coordinates relative to the cortical surface), and we interpret the focal bright spots found at target positions as *target representations*.

Spatial WM reconstructions computed based on patterns of delay-period activation from occipital (V1–hV4v/V3A), parietal (IPS0–IPS3), and frontal (sPCS) cortex revealed highly robust representations of remembered target positions on R1 trials, but not on R0 trials (Figure 3; see Figure S2A for data from individual participants), suggesting that these images reflect memory-related activation changes rather than lingering sensory signals. Furthermore, reconstructed images contain representations of *both* remembered target locations on R2 trials that were robust in many occipital and posterior parietal ROIs (Figure 3C; V1–V3A, hV4, IPS0, and IPS1) but became less separable in anterior parietal and frontal ROIs (IPS2–IPS3 and sPCS). The relative decline in separability of R2 target representations in these anterior parietal and frontal ROIs may reflect the rather small screen size that we used relative to the large size of spatial RFs typical of these ROIs [23, 24]. Finally, we examined the temporal structure of WM reconstructions from all ROIs over the course of the entire trial. We could readily reconstruct images of both remembered locations during target presentation when the positions were encoded into WM, but we could only reconstruct images of locations held in WM during the delay interval (Movie S1).

Next, we sought to quantify how spatial WM reconstructions differ across ROIs and under different memory loads. To do so, we rotated and shifted the reconstruction on each trial to a common reference location such that the target positions were in alignment and averaged all coregistered reconstructions together (Figure 2E; see Movie S2 for coregistered reconstructions through time). Then, because the target position across all trials was now aligned, we quantified attributes of the averaged target representation by fitting a 2D surface (Figures 4A and 4B) characterized by several independent parameters (see Figures S4A–S4D for a demonstration that these parameters reflect dissociable properties of target representations). The *size* parameter reflects the spread (full-width half-maximum, FWHM) of the delay-period target representation: an increased fit size would reflect a less spatially precise representation of the remembered target location (note that here and elsewhere, we use "spatial" with reference to visual field space, not cortical space). The *amplitude* parameter reflects the height of the target representation over baseline: increased fit amplitude would correspond to a more prominent representation of the target over baseline activation not related to the target location. The baseline parameter reflects the non-spatially-selective response amplitude (i.e., a constant offset across the entire reconstructed visual field): a change in baseline reflects a change in mean signal amplitude across an entire ROI that does not carry spatial information and thus does not directly change the spatial information content of the reconstruction.

Increasing memory load did not change the size of the best-fit surfaces to the target representations within WM reconstructions that were based on activation patterns in occipital and posterior parietal ROIs (Figure 4D; V1–IPS0; all statistics were computed via nonparametric resampling methods and Bonferroni corrected for multiple comparisons; Table S1; see the Experimental Procedures). However, fit surface size did increase with memory load in anterior parietal (IPS2–IPS3) and frontal (sPCS) ROIs. Note that in these ROIs, we did not observe strongly disjoint target representations during R2 trials (Figure 3C), so these size increases may partially reflect an inability to separately quantify the representation of each location. It is likely that a larger display and more stimulus separation would enable a more accurate reconstruction and quantification of each remembered target representation in these anterior parietal and frontal areas (like in the early visual and posterior IPS ROIs). We evaluated the possibility that observed size increases may be partially an artifact of coregistering reconstructions and averaging over target positions on R2 trials, even if the "true" target representations are constant in size, by simulating reconstructions under the null assumption that target representations were equal in size

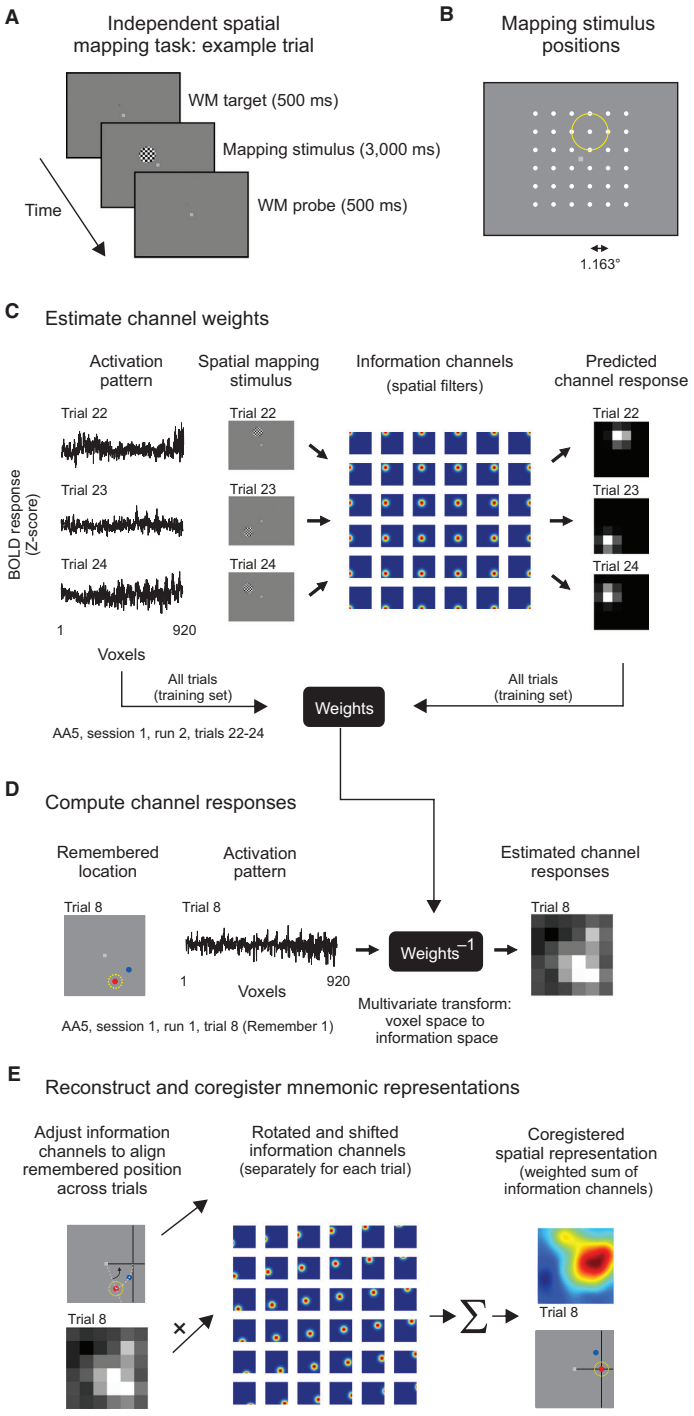


Figure 2. Inverted Spatial Encoding Model for Reconstructing the Contents of Spatial WM

(A) Each participant was scanned for three to four independent spatial mapping runs for encoding model estimation per session (see the Supplemental Experimental Procedures). Participants performed a challenging spatial WM task in which they determined whether a probe stimulus (500 ms) was in the exact same position or a slightly different position from a remembered target position (500 ms; 2AFC; see [16]). During the brief delay period (3,000 ms), a flickering checkerboard stimulus was presented near the remembered target position. This stimulus was irrelevant to the task performed by the participant but was used to drive large sensory responses to estimate a voxel-level encoding model used for computing reconstructions in the main task (see C–E). We adjusted difficulty on a run-by-run basis to maintain vigilance and equate performance across participants and sessions (73.738% ± 1.819% accuracy, mean ± SEM).

(B) We presented the mapping stimulus at each of 36 positions arrayed across a 6 × 6 square grid (one trial per position per run).

(C) To estimate spatial sensitivity profiles for each voxel, we predicted the response of each of 36 hypothetical "information channels" (spatial filters) to each stimulus used in the training runs [16]. Then, we took the measured response of each voxel and the predicted hypothetical channel responses to each stimulus position and used ordinary least-squares linear regression to estimate the contribution of each information channel to the signal observed in each voxel. This step is performed on each voxel independently (see the Supplemental Experimental Procedures, Equation 3).

(D) For each collection of voxels for which we computed reconstructions (ROIs, Figures 3 and 4; all voxels from all ROIs, Figure 4) we computed a mapping from voxel space into channel space (Supplemental Experimental Procedures, Equation 4). In contrast to "population receptive field" analyses [22], this step is multivariate and must be performed using all voxels that contribute to the image reconstruction. Using the computed linear mapping, the measured activation pattern across all voxels is transformed into "information space"—the amount each channel must have been active in order to produce the measured voxel activation pattern. A "raw" reconstruction can be computed for any single observation (e.g., one fMRI volume from area V1) by computing a sum of the spatial filters that define the information channels weighted by the estimated channel responses (right panel).

(E) When computing average reconstructions across all trials (Figures 4C and S2B), we coregistered different target positions on each trial to a common location by first rotating the spatial filters around the fixation point such that the target lies along the Cartesian x axis, then shifting the filter centers horizontally such that the target is positioned 3.25° from fixation along the x axis (white dot in reconstructions shown in Figures 4C and S2B). For R0 and R2 trials, this is done for each remembered target, and the coregistered reconstructions aligned to each target are averaged. Importantly, this coregistration procedure enables us to average the representations of spatial WM targets that appeared at different positions in the display on different trials.

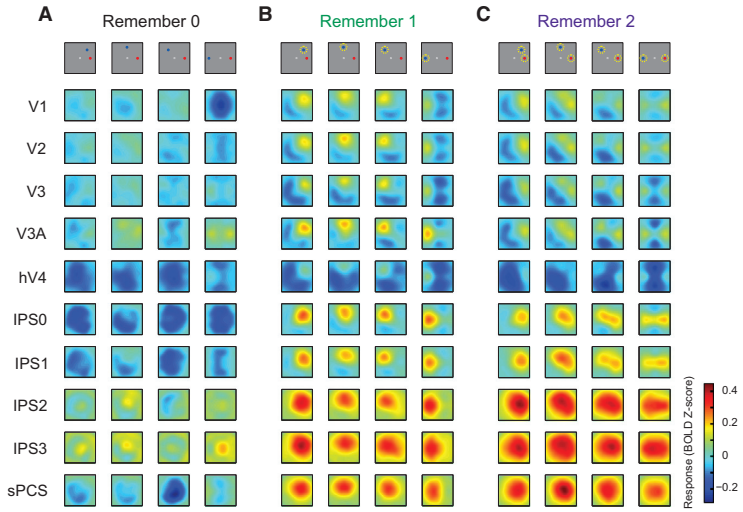


Figure 3. Reconstructed Contents of Spatial WM Measured using Delay-Period Patterns of Activation

Image reconstructions for all target position arrangements during remember zero (A), remember one (B), and remember two (C) conditions from each ROI. Each reconstruction is computed using spatial filters that have been rotated around the fixation point and flipped over the horizontal meridian such that there are four possible target arrangements (top panel; dashed yellow circles indicate remembered target[s]). Targets appeared uniformly within each of these four windows. Early visual (V1–hV4) and parietal (IPS0–IPS1) ROIs carry precise target representations over the delay interval of a single remembered position (remember one; B) or both remembered positions (remember two; C). Reconstructions from anterior parietal (IPS2–IPS3) and frontal (sPCS) ROIs carry moderately precise target representations when a single position is maintained in WM, but they are not as disjoint when both positions are simultaneously held in WM (IPS, intraparietal sulcus; sPCS, superior precentral sulcus, human homolog to macaque frontal eye fields [18, 19]). Additionally, despite a significant reduction in average BOLD response during the delay period in occipital ROIs (Figure S1), reconstructions contain robust representations of remembered stimuli. The color map is identical across all panels. See also Figure S1, Figure S2A for spatial reconstructions from each individual participant, and Movie S1 for temporal unfolding of reconstructions across the duration of the trial.

across memory load conditions and performing an identical coregistration and quantification procedure as that used in Figure 4. These simulations determined that fit target representation size is artificially inflated by 8.62% on average due to the coregistration and averaging procedure. Importantly, our empirically observed size expansion in these regions (IPS2, 24.8%; IPS3, 32.7%; sPCS, 19.6%) was substantially larger than that induced by the analysis procedure itself (see Figure S4E and the Supplemental Experimental Procedures), suggesting that there are still important changes in target representation size across memory load conditions.

The amplitude of best-fit surfaces decreased with increasing memory load in striate and extrastriate occipital (V1–hV4) and posterior parietal (IPS0–IPS1) ROIs, consistent with predictions from a model in which increasing memory load results in lower gain of population-level representations of remembered stimuli [14, 15]. In contrast, fit amplitude trended toward increasing, with greater memory load in anterior parietal (IPS2–IPS3) and frontal (sPCS) ROIs (trend defined as $p < 0.05$, uncorrected for multiple comparisons). This latter result is consistent with previous demonstrations that average delay-period activation levels increase in frontoparietal ROIs with memory load [6, 20, 21] (Figure S1). Furthermore, simulations confirm that the fit amplitude parameter captures changes in the amplitude of the target representation and is independent of changes in baseline or size (Figure S4).

Finally, the nonspatial baseline parameter significantly increased with memory load in posterior parietal ROIs (IPS0–IPS1). The fact that nonspatial baseline levels increased only in IPS0–IPS1 with greater memory load suggests that previously documented univariate BOLD response increases in the more anterior parietal and frontal ROIs (Figure S1A; IPS2–IPS3 and sPCS) most likely correspond to a spatially focal change in target representation amplitude as opposed to spatially uninformative baseline modulations.

We observed population codes for remembered spatial positions in all of the ROIs that we examined, and the representations of remembered locations within these reconstructed images changed in different ways with increasing memory load (Figures 3 and 4). However, the activation pattern across all these ROIs may provide additional information above and beyond the activation pattern within any individual ROI, and reconstructions computed using all these across-ROI modulations may be more closely associated with behavioral memory load effects than reconstructions computed from individual ROIs alone (on the assumption that mnemonic fidelity is a function of information represented across multiple brain regions). We tested this by computing reconstructions as before (Figure 2), but using all voxels from the ten ROIs in each participant (importantly, because this is a multivariate analysis, this is not equivalent to averaging reconstructions across all ROIs; see the Supplemental Experimental Procedures). Comparison of target representations within these WM reconstructions computed using the combined ROI (Figures 4C and 4D, "all voxels combined") across memory load conditions revealed each of the significant results found in the ROIs when analyzed individually (Figures 4C and 4D): size broadened, amplitude decreased, and baseline increased when two items were remembered compared to when one item was remembered (all $p < 0.001$, resampling test). As an additional exploratory analysis, we evaluated how these target representations (Figures 4C and 4D) were related to behavioral performance by computing and quantifying target representations within WM reconstructions as described above using data from each participant, ROI, and memory load individually. These results are presented and discussed in Figures S2B and S2C.

Discussion

Here, we employed an image reconstruction approach implemented using a multivariate inverted encoding model [8, 16, 25–28] to reconstruct the contents of spatial WM based on activation patterns in occipital, parietal, and frontal regions of human cortex. Prior studies have used measures like

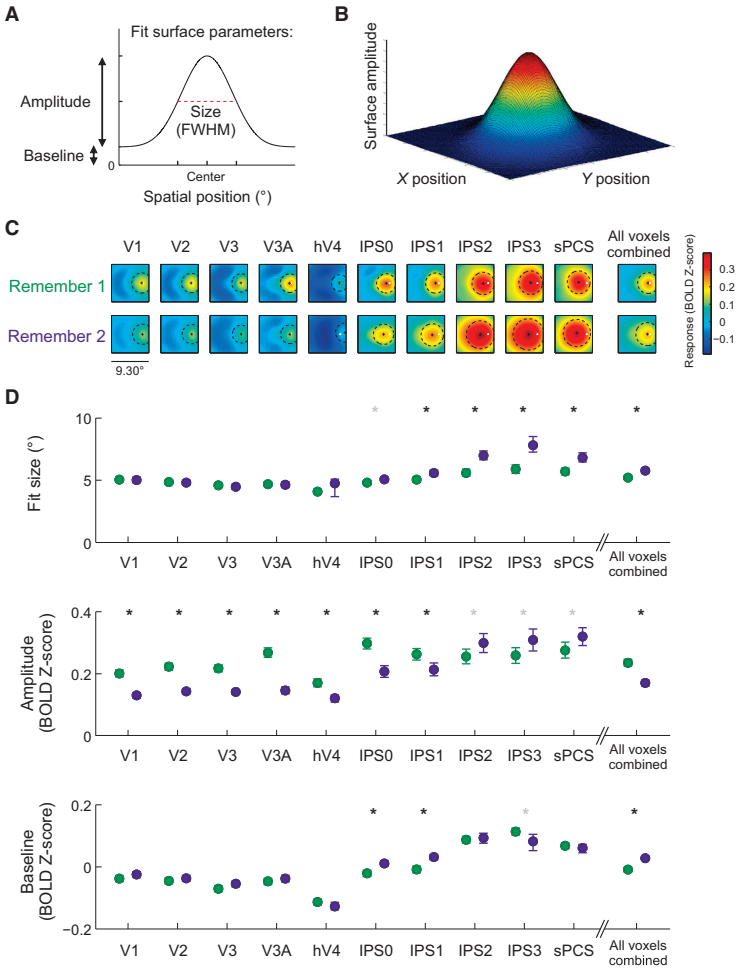


Figure 4. Target Representations within WM Reconstructions Are Less Informative with Greater Memory Load

(A) To quantify the topography of the reconstructed images averaged across trials within each memory load condition, we fit a surface to the average reconstruction that was centered at its global maximum by allowing the size (FWHM), amplitude, and baseline of the surface to freely vary.

(B) Example surface used for fitting.

(C) All reconstructions from each ROI and memory condition (remember one and remember two), rotated and shifted such that the exact target position is aligned to the small white dot (see Figure 2E). We combined trials across participants and resampled all trials, with replacement, from each memory condition and ROI and quantified the averaged reconstruction on each resampling iteration (see Figure S2 for reconstructions and best-fit parameters for each participant individually). The + and dotted circle indicate the average best-fit smooth surface to the target representation within the reconstruction (+ indicates the center, and the dashed line is drawn at the FWHM of the fit surface). For remember two, representations of each target are averaged together before fitting. See Movie S2 for temporal evolution of coregistered reconstructions across the duration of the trial.

(D) Parameters describing best-fit surfaces to target representations from each ROI and memory condition. Target representation size remains constant in early visual areas (V1–hV4), but amplitude decreases with larger memory load, suggestive of a less informative population code (Figure S3). Anterior parietal and frontal ROIs have larger target representations with increasing memory load, as well as trends toward higher amplitude representations, though some size increases are introduced during the coregistration and averaging procedure (see the Supplemental Experimental Procedures, Simulating and fitting target representations with known parameters; Figure S4). The spatially nonselective baseline parameter remains largely constant across memory load conditions, except in IPS0 and IPS1. Black asterisks indicate significant differences at $p < 0.05$, Bonferroni-corrected within each parameter across ten comparisons (ROIs). Gray asterisks indicate trends defined as $p < 0.05$, uncorrected for multiple comparisons. All tests performed using resampling procedures (see the Experimental Procedures). Error bars indicate 95% confidence intervals computed via resampling of data pooled across participants. See Table S1 for p values.

See also Figures S2–S4, Table S1, and Movie S2.

classification accuracy to correlate behavioral performance with the discriminability of neural activation patterns [6, 13]. Although these analyses have many advantages due to a relative lack of model assumptions, changes in decoding accuracy may result from many different types of neural response pattern modulation [25, 29]. In contrast, by assuming a set of spatial basis functions, our method allows us to assess whether each region encoded information about the location of a remembered stimulus (e.g., [5, 30]), as well as to visualize and quantify the characteristics of these covert representations of target locations and relate different aspects of these quantified representations to behavioral performance (e.g., [8, 16, 25, 27]). In addition, these findings reinforce the importance of measuring the effect of cognitive manipulations on population-level estimates of mnemonic representations rather than on particular properties of the underlying neural generators, as these population-level representations can be more closely associated with cognition and behavior than activity changes in single neurons or voxels [8, 16, 25–29, 31–33].

These image reconstruction and quantification analyses revealed lower amplitude and, in some anterior parietal and frontal ROIs, broader target representations with increasing memory load (Figure 4). From an information-theoretic perspective, response variability (i.e., intertrial variability in the reconstructed images) has two components: signal entropy, which is variability associated with experimental manipulations (remembered location), and noise entropy, which is variability not associated with experimental manipulations. The decrease in target representation amplitude under increased memory load should lead to less variability that is related to the remembered location(s) and thus to a decrease in the signal entropy and information about the remembered location. An increase in target representation size should also decrease signal entropy, as increased size leads to

more overlap between target representations for different locations, which would decrease the ability of the population code to discriminate between locations. In contrast, baseline shifts should not strongly influence information content as an additive shift in the entire reconstruction does not change signal entropy [14, 16, 34] (Figure S3). Thus, the observation of higher amplitude target representations corresponds to higher information content of population codes about a spatial position [14–16, 32–34] (Figure S3) and may be a consequence of changes in delay-period neural gain associated with neurons tuned to remembered locations [14]. In addition, modest increases in target representation size in anterior IPS and sPCS may reflect poorer mnemonic fidelity within particular ROIs, echoing previous results that the dispersion (analogous to size here) of reconstructed profiles of remembered *features* (e.g., orientation) correlates with behavioral performance [8, 25, 27]. However, future work using larger spatial stimulus arrays may help to more accurately disentangle and characterize multiple WM representations in anterior IPS and sPCS.

We were able to reconstruct the covert contents of spatial WM not only in occipital [4, 6–10, 13] and posterior parietal regions [10, 13], but also in anterior parietal and frontal cortex [5, 11]. These widespread modulations raise the possibility that distributed WM representations can be optimized to differentially contribute to complementary sensory (e.g., target localization) and motor (e.g., eye movements, reaches) behaviors. Consistent with this idea, a recent demonstration that induced alpha oscillations (which are often thought to reflect synchronized activity of large-scale cortical networks [35]) measured with scalp EEG can be used to reconstruct remembered orientations also suggests that long-range, interacting representations across much of human cortex support the maintenance of information in WM [27]. The successive representations of spatial position reported here may thus allow for a common coordinate system with which low-level stimulus features (such as spatial position and color) that are represented in occipital cortex are bound with spatial motor plans (such as eye movements and arm reaches [36]) that are more closely associated with representations in parietal and frontal cortex.

Experimental Procedures

Functional Magnetic Resonance Imaging

We scanned each participant for four sessions, each lasting 2 hr. Each session included runs of the spatial WM task (Figure 1), an independent spatial "mapping" task (Figures 2A and 2B; Supplemental Experimental Procedures) [16], and a visual localizer task (5 min each).

Encoding Model: Reconstructing Contents of Spatial WM

We modeled the response of each voxel as a linear combination of 36 spatially selective information channels (see [16]; Figure 2; Supplemental Experimental Procedures). Using a separate set of training data during which we presented a flickering checkerboard "mapping" stimulus at different locations on the screen (Figures 2A and 2B), we estimated the relative contribution of all 36 information channels to the observed signal in each voxel using ordinary least-squares regression (Figure 2C). Then, using all of these measured "channel weights" across a given ROI, combined with the multivariate pattern of activation measured from that ROI during performance of the main spatial WM task (Figure 1A), we computed the channel responses that were most likely to produce the measured pattern of activation (Figure 2D). We combined these computed channel responses and the spatial filters (information channels) to produce reconstructed images of the spatial WM contents within each ROI for each measured pattern of activation (Figures 3 and 4, activation patterns measured 6.75–9 s after target onset; Movies S1 and S2, activation patterns measured at each time point during the trial).

Quantifying Target Representations in WM Reconstructions

We fit a surface to each reconstruction that was allowed to vary in its size, amplitude, and baseline (Figures 4A and 4B). Its center was constrained to be the position, in visual field coordinates, with the highest local reconstruction amplitude (local average within a 0.5° radius).

Statistics

For group-level analyses (Figures 1C–1E, 4D, and S1B), we combined data from all participants within a given ROI and memory load condition and resampled all trials with replacement and computed a mean measurement value for that resampling iteration (Figure 1C, behavioral recall error; Figure 1D, behavioral accuracy; Figure 1E, response time; Figure 4D, target representation fit parameters; Figure S1B, mean BOLD signal). We repeated this procedure 1,000 times to produce a resampled distribution of each measured value for each memory load condition. We computed p values for each ROI and each parameter as the two-tailed probability of observing an effect in the opposite direction of the mean effect observed. Comparisons are Bonferroni corrected across ROIs for each parameter (Figure 4D, ten comparisons) or across all comparisons performed (Figure S1B, 30 pairwise comparisons). All error bars are 95% confidence intervals derived from these resampled distributions unless indicated otherwise (Figure S1A).

For exploratory individual-participant analyses (Figure S2C), we performed an identical procedure but resampled only across each participant's data when computing confidence intervals.

Supplemental Information

Supplemental Information includes Supplemental Experimental Procedures, four figures, one table, and two movies and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2014.07.066.

Author Contributions

T.C.S., E.F.E., and J.T.S. developed the experiment protocol and wrote the manuscript. T.C.S. and E.F.E. collected data. T.C.S. analyzed data. J.T.S. supervised the project.

Acknowledgments

We thank Miranda Scolari and Mary Smith for assistance in developing parietal cortex mapping protocols, Anna Byers for assistance with data collection, Sirawaj Itthipuripat, Vy Vo, and Alexander Heitman for discussion, and Sirawaj Itthipuripat, Vy Vo, and Stephanie Nelli for comments on the manuscript. This work was supported by a NSF Graduate Research Fellowship to T.C.S., NIH T32-MH020002-12 to E.F.E., and NIH R01 MH-092345 to J.T.S. Data from each cortical region of interest and scripts required to produce results as presented in this manuscript are available upon request.

Received: April 30, 2014
Revised: June 13, 2014
Accepted: July 25, 2014
Published: September 4, 2014

References

1. Kane, M.J., and Engle, R.W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: an individual-differences perspective. Psychon. Bull. Rev. *9*, 637–671.
2. Park, S., and Holzman, P.S. (1992). Schizophrenics show spatial working memory deficits. Arch. Gen. Psychiatry *49*, 975–982.
3. Luck, S.J., and Vogel, E.K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends Cogn. Sci. *17*, 391–400.
4. Serences, J.T., Ester, E.F., Vogel, E.K., and Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. Psychol. Sci. *20*, 207–214.
5. Jerde, T.A., Merriam, E.P., Riggall, A.C., Hedges, J.H., and Curtis, C.E. (2012). Prioritized maps of space in human frontoparietal cortex. J. Neurosci. *32*, 17382–17390.
6. Emrich, S.M., Riggall, A.C., Larocque, J.J., and Postle, B.R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. J. Neurosci. *33*, 6516–6523.

7. Riggall, A.C., and Postle, B.R. (2012). The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. J. Neurosci. *32*, 12990–12998.
8. Ester, E.F., Anderson, D.E., Serences, J.T., and Awh, E. (2013). A neural measure of precision in visual working memory. J. Cogn. Neurosci. *25*, 754–761.
9. Harrison, S.A., and Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. Nature *458*, 632–635.
10. Christophel, T.B., Hebart, M.N., and Haynes, J.-D. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. J. Neurosci. *32*, 12983–12989.
11. Funahashi, S., Bruce, C.J., and Goldman-Rakic, P.S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. J. Neurophysiol. *61*, 331–349.
12. Pasternak, T., and Greenlee, M.W. (2005). Working memory in primate sensory systems. Nat. Rev. Neurosci. *6*, 97–107.
13. Albers, A.M., Kok, P., Toni, I., Dijkerman, H.C., and de Lange, F.P. (2013). Shared representations for working memory and mental imagery in early visual cortex. Curr. Biol. *23*, 1427–1431.
14. Bays, P.M. (2014). Noise in neural populations accounts for errors in working memory. J. Neurosci. *34*, 3632–3645.
15. Ma, W.J., Husain, M., and Bays, P.M. (2014). Changing concepts of working memory. Nat. Neurosci. *17*, 347–356.
16. Sprague, T.C., and Serences, J.T. (2013). Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nat. Neurosci. *16*, 1879–1887.
17. Awh, E., Jonides, J., and Reuter-Lorenz, P.A. (1998). Rehearsal in spatial working memory. J. Exp. Psychol. Hum. Percept. Perform. *24*, 780–790.
18. Srimal, R., and Curtis, C.E. (2008). Persistent neural activity during the maintenance of spatial position in working memory. Neuroimage *39*, 455–468.
19. Paus, T. (1996). Location and function of the human frontal eye-field: a selective review. Neuropsychologia *34*, 475–483.
20. Todd, J.J., and Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. Nature *428*, 751–754.
21. Xu, Y., and Chun, M.M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. Nature *440*, 91–95.
22. Dumoulin, S.O., and Wandell, B.A. (2008). Population receptive field estimates in human visual cortex. Neuroimage *39*, 647–660.
23. Andersen, R.A., Essick, G.K., and Siegel, R.M. (1985). Encoding of spatial location by posterior parietal neurons. Science *230*, 456–458.
24. Mohler, C.W., Goldberg, M.E., and Wurtz, R.H. (1973). Visual receptive fields of frontal eye field neurons. Brain Res. *61*, 385–389.
25. Anderson, D.E., Ester, E.F., Serences, J.T., and Awh, E. (2013). Attending multiple items decreases the selectivity of population responses in human primary visual cortex. J. Neurosci. *33*, 9273–9282.
26. Brouwer, G.J., and Heeger, D.J. (2009). Decoding and reconstructing color from responses in human visual cortex. J. Neurosci. *29*, 13992–14003.
27. Anderson, D.E., Serences, J.T., Vogel, E.K., and Awh, E. (2014). Induced α rhythms track the content and quality of visual working memory representations with high temporal precision. J. Neurosci. *34*, 7587–7599.
28. Garcia, J.O., Srinivasan, R., and Serences, J.T. (2013). Near-real-time feature-selective modulations in human cortex. Curr. Biol. *23*, 515–522.
29. Serences, J.T., and Saproo, S. (2012). Computational advances towards linking BOLD and behavior. Neuropsychologia *50*, 435–446.
30. Tong, F., and Pratte, M.S. (2012). Decoding patterns of human brain activity. Annu. Rev. Psychol. *63*, 483–509.
31. Pouget, A., Dayan, P., and Zemel, R.S. (2003). Inference and computation with population codes. Annu. Rev. Neurosci. *26*, 381–410.
32. Ma, W.J., Beck, J.M., Latham, P.E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nat. Neurosci. *9*, 1432–1438.
33. Graf, A.B.A., Kohn, A., Jazayeri, M., and Movshon, J.A. (2011). Decoding the activity of neuronal populations in macaque primary visual cortex. Nat. Neurosci. *14*, 239–245.
34. Saproo, S., and Serences, J.T. (2010). Spatial attention improves the quality of population codes in human visual cortex. J. Neurophysiol. *104*, 885–895.
35. Nunez, P.L., Reid, L., and Bickford, R.G. (1978). The relationship of head size to alpha frequency with implications to a brain wave model. Electroencephalogr. Clin. Neurophysiol. *44*, 344–352.
36. Sereno, M.I., and Huang, R.S. (2014). Multisensory maps in parietal cortex. Curr. Opin. Neurobiol. *24*, 39–46.

Supplemental Data

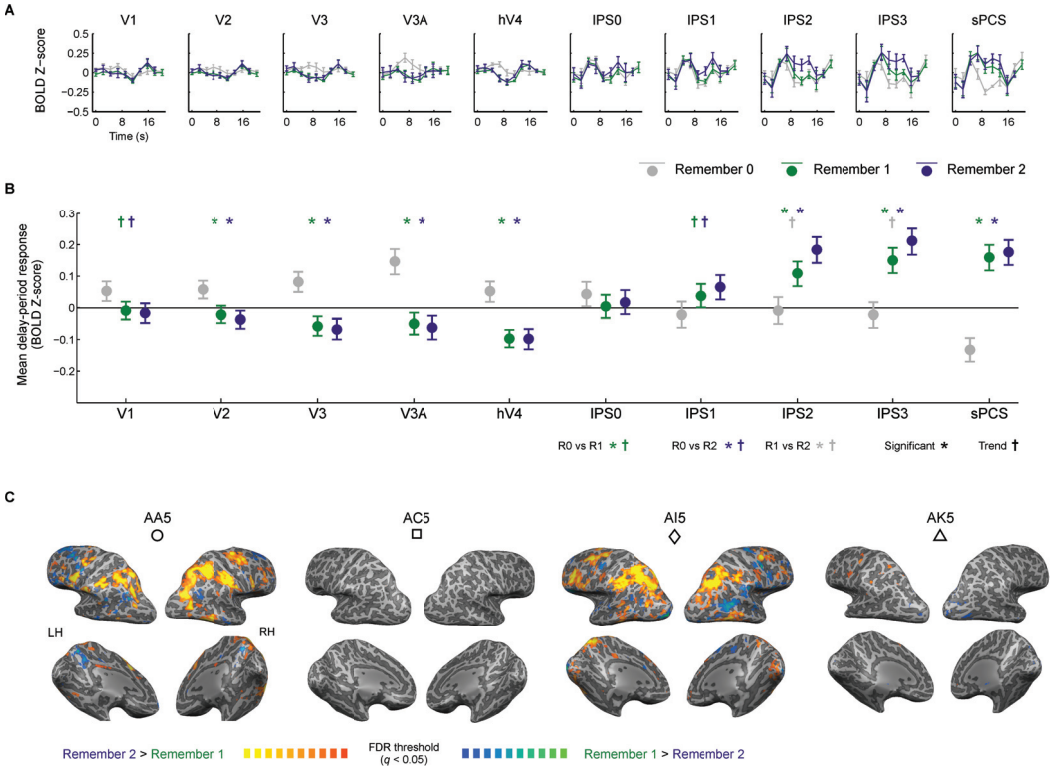


Figure S1 Mean BOLD signal depends on WM maintenance, related to Figures 1 and 3.

For each participant and each ROI we extracted the target-locked hemodynamic response function (HRF; event-related average) for each memory load condition. Then, we averaged HRFs across participants. Error bars ±1 SEM (across participants, $n = 4$, sessions 1 & 2). Time courses qualitatively replicate previous findings [S1] in which early IPS visual field maps (e.g., IPS0) show more transient activation, while later IPS visual field maps (e.g., IPS3) show more sustained activation during delay periods. (**B**) Average delay-period activation is significantly reduced during WM maintenance in occipital ROIs V2-V3A and hV4 (R1 < R0 and R2 < R0, $p < 0.001$, resampling test, see Supplemental Experimental Procedures). In contrast, delay-period activation is higher during WM maintenance in parietal and frontal ROIs (IPS1-sPCS, R1 > R0 and R2 > R0, $p < 0.024$). In IPS2 and IPS3, we observed a trend (defined at uncorrected $\alpha = 0.05$) towards greater delay-period activation during R2 than during R1 trials ($p = 0.012$ and 0.03, respectively). Despite these decreases in mean BOLD response with WM performance, we can still reconstruct robust spatial representations of 1 or 2 remembered stimuli (Fig. 3B-C). All significant p-values pass a Bonferroni-corrected threshold for 30 comparisons of $\alpha = 0.0017$ (10 ROIs, 3 comparisons for each ROI: R2 vs. R0, R1 vs. R0, R2 vs. R1) and are indicated by asterisks in **B**. Trends, defined at uncorrected $\alpha = 0.05$, are indicated by †. P-values presented in Table S1. Error bars 95% CI over resampled distribution pooled across participants. (**C**) For each participant we analyzed neuroimaging data using a

univariate general linear model (GLM) with 3 predictors corresponding to the 3 memory load conditions (Remember 0, Remember 1, and Remember 2). Here, we show significant voxels for the contrast Remember 2 > Remember 1, corrected within each participant for multiple comparisons using the false discovery rate as implemented in BrainVoyager 2.6.1 ($q < 0.05$). Within 2 of 4 individual observers, we observe data similar to previous reports that increasing the number of remembered items results in increased BOLD responses in parietal and frontal cortex [S2–S4]. Note that even though there are few significant voxels in the remaining two participants, patterns of activation for both Remember 1 and Remember 2 conditions can be used to reconstruct spatial WM contents (albeit with lower amplitude, see Fig. S2). Symbols match those used in Figure 1C-E and Figure S2.

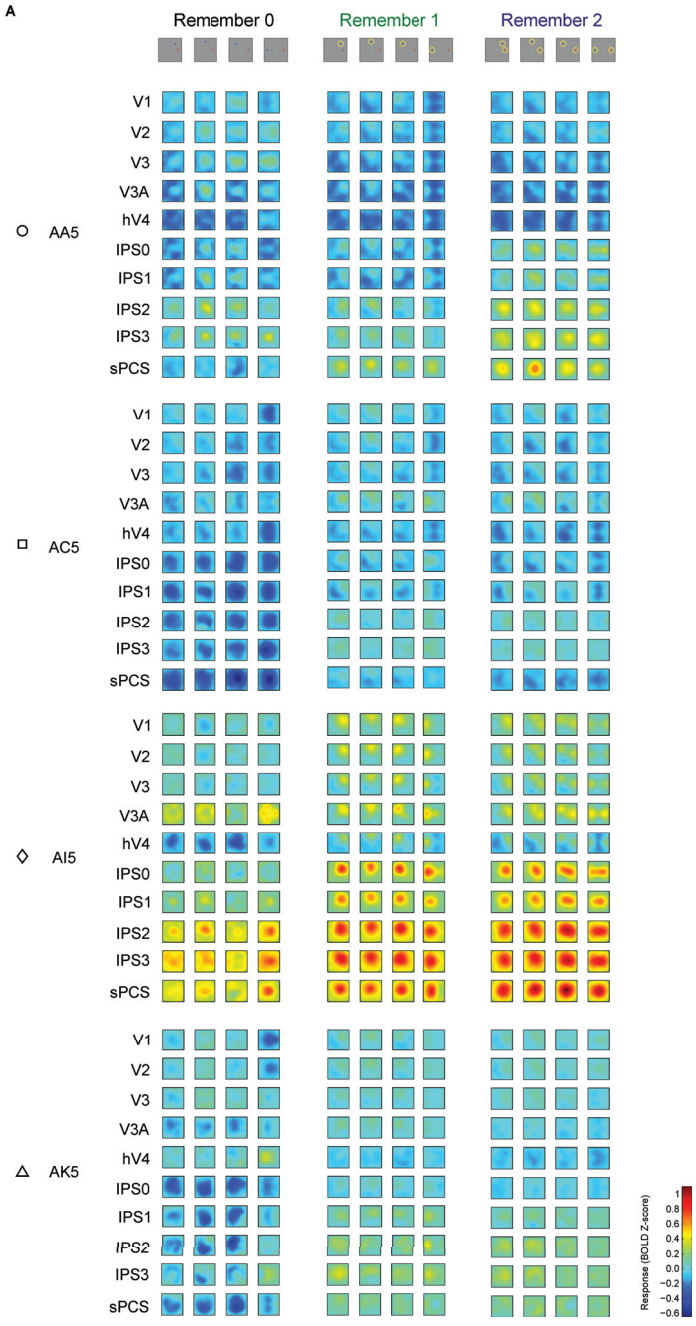


Figure S2 Individual-participant WM reconstructions and fit surface parameters compared to behavioral recall error, related to Figures 3-4 (continued on next page)

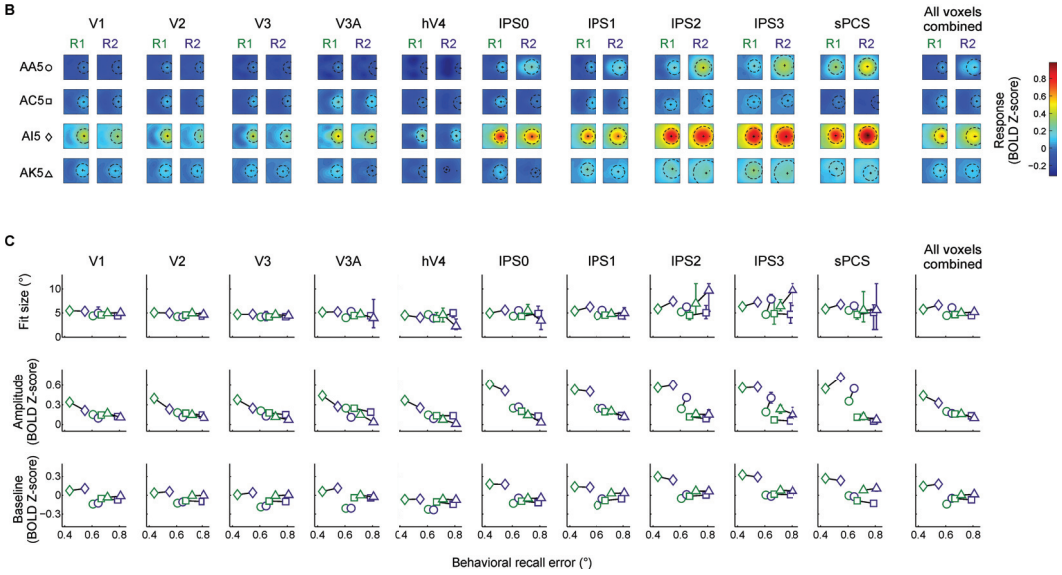


Figure S2 (continued) Individual-participant WM reconstructions and fit surface parameters compared to behavioral recall error, related to Figures 3-4

(**A**) Data as in Figure 3 displayed for each participant individually. Symbols match those used in Figure 1C-E and Figure S1. Color scale is the same for all participants, memory load conditions, and ROIs within this panel. (**B**) Coregistered spatial WM reconstructions for each participant for each memory load condition (R1: Remember 1, R2: Remember 2) for each ROI reported in Figures 3 and 4 (coregistered as in Figure 4C). Black dot indicates the target position; black + is mean centroid of best-fitting surface; black dashed circle is drawn around the mean centroid at the mean full-width half-maximum (FWHM) of best-fitting surface. Color scale is the same for all participants, memory load conditions, and ROIs within this panel. (**C**) Best-fit size, amplitude, and non-spatial baseline surface parameters to target representations for each participant and each memory load condition. Error bars 95% CI computed via resampling all trials per condition within each participant and ROI. Though we do not have adequate statistical power to identify whether an effect is present, for any given ROI, these scatterplots suggest that target representation amplitude, more so than size or baseline, is best correlated with behavioral recall performance (especially IPS0 and All voxels combined) such that high representation amplitude is associated with better behavioral recall performance, both within and across participants. These analyses imply that the amplitude of representations across the visual hierarchy provides the primary constraint on behavioral performance in our spatial WM task, suggesting that the amplitude of population-level representations of remembered locations are more closely related to their fidelity as indexed by corresponding measures of behavioral performance than are other parameters, like their size [S5].

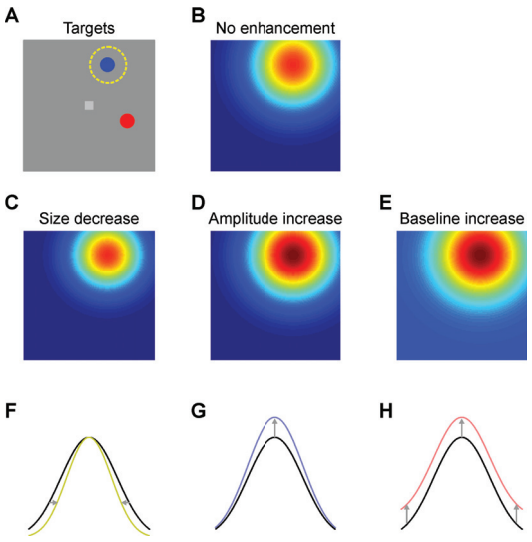


Figure S3 Effects of different parameters on the information content of a target representation, related to Figure 4

Population-level codes for a remembered spatial position can change in several ways, each of which may have different consequences on the information content of the population code for the remembered position. For a given remembered target position (**A**), a brain region carries a representation of the remembered position as a bivariate Gaussian-like representation (**B**). This representation could be modulated in several ways. If the size of the target representation decreased (**C**, **F**), this would reflect a more precise representation of spatial position, and could lead to more accurate localization of a remembered target. However, depending on the noise amplitude relative to the amplitude of the target representation, this type of modulation could be more susceptible to noise across trials, resulting in poor localization. Instead, an increase in the amplitude of the target representation (**D**, **G**) would increase the signal-to-noise ratio (SNR) of the population code even if the spatial precision (size) of the representation remained fixed. Because such an amplitude modulation would increase the SNR, this type of representational modulation would be more robust to high levels of cross-trial neural noise in the population code [S6, S7]. Alternatively, the baseline response level in the WM reconstruction could increase (**E**, **H**), which would not change the information content of the target representation given most reasonable noise models. For example, the absolute effect of a baseline shift on the information content of the population code depends on the across-trial noise distribution. If noise scales with the mean (e.g. Poisson noise) then a pure baseline shift would increase the noise level without a corresponding change in the amplitude of the target representation over baseline, which would decrease the information content of the neural code (i.e. noise would go up, but the dynamic range of stimulus-locked signals would remain the same). Under conditions of independent and identically distributed (IID) noise across positions and trials, a baseline shift would have no effect on the information content of the population code, as the dynamic range of the target representation amplitude over reconstruction baseline compared to the noise level would not change. However, note that under an unlikely scenario where noise *decreases* with the signal amplitude, a baseline shift would be beneficial, as increasing signal amplitude would decrease the corrupting influence of noise on the population-level target representation. All figures are modeled using the fit surface function (Equation 5) and adjustments either to the size (**C**, **F**), amplitude (**D**, **G**), or baseline (**E**, **H**) parameters. (**F-H**) slices through the center of

the representation. Note that combinations of these modulations are possible and these are meant only as illustrative examples.

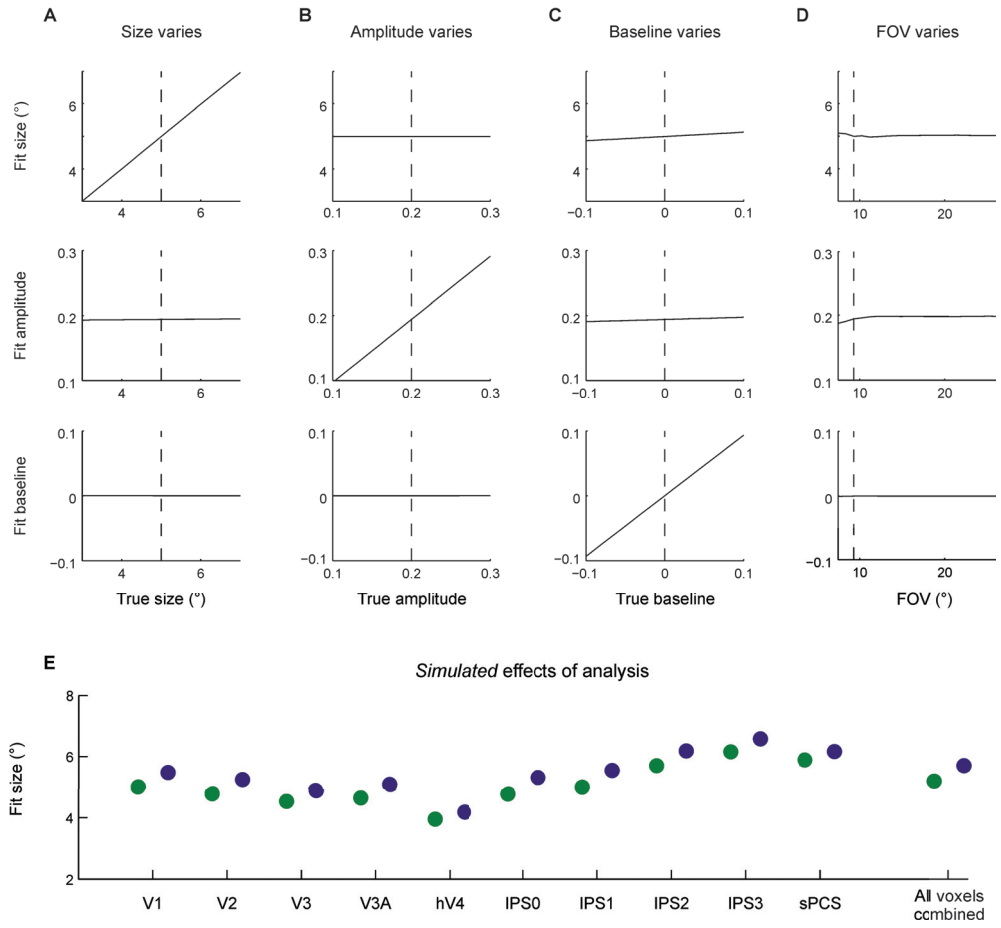


Figure S4 Fits to simulated surfaces demonstrate sensitivity and specificity of fitting approach, related to Figure 4

It may be the case that our fitting procedure is unable to assess changes in a given fit surface parameter independent of changes in another parameter. For example, changes in amplitude of the "true" target representation may be incorrectly attributed to changes in baseline. To address this concern, we generated simulated target representations using known parameters (vertical dashed lines in A-C, size: 5°; amplitude: 0.2, baseline: 0, approximately the values from R1 condition in the 'All voxels combined' analysis, Fig. 4D), then varied one parameter at a time while keeping all others constant. After generating simulated representations, we performed an identical surface fitting procedure to that performed using experimental data (see Supplemental Experimental Procedures: Simulating and fitting representations with known parameters; Fig. 4). We plot each fit parameter as a function of the target representation parameter that was varied. (**A**) Size was varied; (**B**) amplitude was varied; (**C**) baseline was varied. For each of these manipulations, only the fit parameter matching that which was manipulated changes as a function of parameter value; all others remain constant at the value indicated by the vertical dashed lines in their respective panels. An additional concern is that our narrow field of

view (FOV; 9.30°) over which we reconstruct WM contents and fit a surface to target representations may result in erroneous estimates of size, amplitude or baseline. To address this issue, we kept the simulated representation constant, but expanded the FOV over which we simulated target representations and fit surfaces (**D**). Best-fit size, amplitude, and baseline are largely constant across a wide range of FOVs. Importantly, we choose to maintain our original FOV, as that is the window over which we "trained" the spatial encoding model during the spatial mapping runs (Fig. 2A-C; Supplemental Experimental Procedures: Spatial encoding model). Allowing reconstructions to evaluate to 0 at high-eccentricity positions (as would be the case with an artificially-enlarged FOV) may not capture properties of the actual target representations, and so we intentionally avoid this. Additionally, it might be that the coregistration and averaging procedure (Fig. 2E) may artificially inflate estimates of fit size on R2 trials relative to R1 trials, even if there were no "true" change in target representation size. If this were the case, our observations of size increases in anterior parietal and frontal ROIs (IPS2-3, sPCS) might be partially due to the analysis procedure. Importantly, this may disproportionately mask results for ROIs in which target representations are not readily separable during R2 trials (Fig. 3C). To evaluate this possibility, we simulated reconstructions from both R1 and R2 trials under the null assumption that representations of the remembered target(s) do not change as a function of memory load. The representation parameters used to generate R2 reconstructions from a given ROI were taken from the best-fit parameter values reported in Figure 4D. Results from this analysis are presented in **E**. Using actual target positions remembered by participants and an identical analysis procedure, fit size indeed increased by an average of 8.62% across all ROIs for R2 compared to R1, even when the "true" simulated representations were constant in size. However, this effect is small when compared to the actual (significant) size increases observed in IPS2 (simulated: 10.62%, measured: 24.93% [18.85% 31.59%], mean [95% CI]), IPS3 (simulated: 11.73%, measured: 32.7% [23.28% 44.58%]), and sPCS (simulated: 8.25%, measured: 19.6% [13.50% 26.53%]). Because the size increases from simulations seeded with R1 representation parameters in these ROIs lie outside the 95% CIs of the observed size increases, the observed size changes are above and beyond those that would be expected given the analysis procedure alone. Finally, we also simulated the effect of increasing spatial overlap (via simulating reconstructions of 2 targets with increasing sizes) in order to evaluate whether increased overlap alone can lead to overestimates of fit size. We observed a decrease in the size overestimation once we exceeded simulated representation sizes of 5.8°. Though the size of R1 target representations in IPS2 (5.591°), IPS3 (5.889°), and sPCS (5.696°) hover near this maximum bias point, the size increase we observe in those regions for R2 remains substantially larger than that introduced by the analysis procedure alone. In all simulations performed using realistic parameter values, we never observed a decrease in representation amplitude as a result of the coregistration and fitting procedure.

Table S1 P values for resampled tests reported in Figure 4D and Figure S1B.

P-values derived from resampling tests described in Supplemental Experimental Procedures: Statistical methods. A value of 0 indicates $p < 0.001$ (due to 1,000 sampling iterations). Bold indicates significant test after correcting for multiple comparisons within parameter using Bonferroni's method (Figure 4D, 10 comparisons, $\alpha = 0.005$; Figure 4D, All voxels combined, 1 comparison, $\alpha = 0.05$; Figure S1B, 30 comparisons, $\alpha = 0.0017$). Italics indicate trends at $p < 0.05$, uncorrected for multiple comparisons.

	V1	V2	V3	V3A	hV4	IPS0	IPS1	IPS2	IPS3	sPCS	All voxels combined
Figure 4D											
Size	0.824	0.682	0.362	0.73	0.244	*0.022*	**0.004**	**0**	**0**	**0**	**0**
Amplitude	**0**	**0**	**0**	**0**	**0**	**0**	**0.002**	*0.02*	*0.02*	*0.03*	**0**
Baseline	0.064	0.196	*0.02*	0.234	0.08	**0**	**0**	0.582	*0.008*	0.402	**0**
Figure S1B											
R0 vs R1	*0.004*	**0**	**0**	**0**	**0**	0.148	*0.024*	**0**	**0**	**0**	
R0 vs R2	*0.002*	**0**	**0**	**0**	**0**	0.344	*0.004*	**0**	**0**	**0**	
R1 vs R2	0.720	0.428	0.682	0.628	0.938	0.650	0.298	*0.012*	*0.030*	0.584	

Supplemental Experimental Procedures

Participants

We used 4 healthy participants from the UCSD community (aged 24-31 yrs, 2 female, all right-handed). All participants were experienced psychophysical observers, and data from these participants has been reported previously using the same participant identifiers [S5]. One participant (AA) was an author (TCS). All participants gave informed consent as approved by the UCSD Institutional Review Board and were compensated for their time ($20/hr fMRI sessions, $10/hr behavioral sessions).

Stimulus & Task

All participants took part in 4 fMRI testing sessions and 2-4 behavioral testing sessions. The size of the stimulus display was fixed across sessions. During each trial, two target stimuli were presented at pseudorandomly chosen positions on the screen for 500 ms, followed by a post-cue (fixation point changing color) and 8,000 ms delay interval. The color of the post-cue during this delay interval instructed participants to precisely remember the position of one target (matching the color of the fixation point; Remember 1), remember the position of both targets (purple fixation point; Remember 2), or to passively fixate and wait for the next trial to begin (only in some fMRI sessions; Fig. 1A; Remember 0).

Each target was presented within one of 8 discs with 0.6° radius evenly spaced around a 3.25° circle, offset from the horizontal and vertical meridians. Both targets were presented in different discs. On every trial, the exact target position within the area of the disc was randomly chosen with uniform density to discourage alternative coding strategies (e.g., verbally labeling the location, such as “up and to the left”; “8 o’clock”, etc.).

During fMRI scanning sessions, a probe stimulus appeared at the end of the delay period (500 ms). The probe stimulus was always presented near to and in the color of the remembered target during Remember 1 trials, and with equal probability in either color/location during Remember 2 trials (matching the color of the corresponding target). The probe stimulus was in exactly the same position as the target stimulus on 50% of trials, and an offset position in a random direction on the other 50% of trials. The magnitude of the offset was uniformly chosen from the range 0.1° - 1.5° at 0.2° steps. Participants performed a two-alternative forced-choice task at the end of each Remember 1 or Remember 2 trial, comparing whether the probe stimulus was in exactly the same or in a slightly different position relative to the corresponding target position.

In the behavioral recall task performed outside the scanner, we instructed participants to use the mouse to click as accurately as possible at the remembered target position on the screen. For Remember 2 trials, the response was cued by a change in fixation color from purple to red or blue.

During the first 2 fMRI sessions acquired for each participant we included Remember 0 trials, but these were dropped from the final 2 sessions in an effort to maximize data acquired during memory maintenance. Remember 0 trials were omitted from the behavioral recall task.

In order to independently estimate a spatial encoding model for each voxel, participants also performed a “spatial mapping” task (Fig. 2A-B), which is similar to the Spatial WM condition in our previous work [S5]. Briefly, they remembered the exact position of a target dot (500 ms) over a brief delay interval (3,000 ms) and made a same/different 2AFC judgment on a probe dot presented after the delay (500 ms). The principal difference between this task and the main spatial WM task described above (Fig. 1A) is the presentation of a high-contrast flickering checkerboard (1.163° radius) around the remembered target position (Fig. 2A). Also, these memory targets and interstitial delay-period checkerboards were presented on a 6 × 6 grid (spaced by 1.163°, Fig. 2B). This stimulus and spatial arrangement allowed us to drive a strong BOLD response and efficiently estimate the spatial sensitivity of each voxel across the entire visual field. Task

difficulty was adjusted between runs to ensure sufficient task engagement (mean ± standard error of accuracy: 73.738 ± 1.819 %).

Finally, observers also performed a spatial WM task while a large full-hemifield flickering checkerboard was presented. Data from this task was used to independently identify which voxels were significantly modulated by visual input from our stimulus display for inclusion in further analyses (see *ROI definition*).

Behavioral analysis

For each participant, we combined data across all behavioral recall sessions. We discarded trials in which errant clicks were made (responses > 7.0° eccentricity or within 1.0° of fixation; these typically were a result of accidentally pressing the mouse before moving the cursor or errant mouse movements; 11 of 3,264 trials discarded across all participants). Behavioral recall error was defined as the mean Euclidean distance between the response position and the correct target position across all trials for each condition within a participant.

fMRI scanning

We scanned all 4 participants for 4 sessions, with each session lasting 2 hrs. Each session included 3 types of runs: spatial mapping runs (Fig. 2A-B, 5 min each, 3-4 per session), used for encoding model estimation (Fig 2C), spatial WM main task runs (Fig. 1A; 2-3 per session, each subdivided over 4 shorter runs lasting 4-6 min each depending on whether Remember 0 trials were included), and visual localizer runs for identifying visually-responsive voxels involved in WM maintenance (6 min each, 1-3 per session). These participants were additionally scanned for another 1.5-2 hr session in order to map retinotopically organized IPS subregions IPS0-IPS3 using methods described previously (Methods and IPS retinotopic maps for 3 of 4 participants can be found in [S5]).

We scanned all participants on a 3 T research-dedicated GE MR750 scanner at the UCSD Keck Center for Functional Magnetic Resonance Imaging using a 32 channel send/receive head coil (Nova Medical, Wilmington, MA). We acquired functional data using a gradient echo planar imaging (EPI) pulse sequence [S5] (19.2 × 19.2 cm field of view, 96 × 96 matrix size, 31 3-mm-thick slices with 0-mm gap, obliquely-oriented through occipital, parietal & dorsal frontal cortex, TR = 2,250 ms, TE = 30 ms, flip angle = 90°, voxel size 2 × 2 × 3 mm, xyz).

During each session we also acquired a high resolution anatomical scan (FSPGR T1-weighted sequence, TR/TE = 11/3.3 ms, TI = 1,100 ms, 172 slices, flip angle = 18°, 1 mm^3 resolution). Functional images were coregistered to a separate anatomical scan collected during a different session by aligning each session's functional images to the respective session's anatomical scan, and then aligning the anatomical scan to the target anatomical scan. Images were preprocessed as described previously [S5] using FSL (Oxford, UK) and BrainVoyager 2.3 (BrainInnovations). Preprocessing included unwarping the EPI images using routines provided by FSL, slice-time correction, three-dimensional motion correction (six-parameter affine transform), temporal high-pass filtering (to remove first-, second- and third-order drift), transformation to Talairach space and normalization of signal amplitudes by converting to Z-scores separately for each run. We did not perform any spatial smoothing beyond the smoothing introduced by resampling during the co-registration of the functional images, motion correction and transformation to Talairach space. All subsequent analyses were computed using custom code written in Matlab (version 2012a, The Mathworks, Inc).

ROI definition

All reported ROIs except sPCS were defined using standard retinotopic mapping procedures [S5, S8, S9]. For analysis, voxels from ROIs in the left and right hemispheres which showed a significant response to either hemifield during the visual localizer runs (across sessions, corrected using the false discovery rate (FDR, [S10]) across all measured voxels within each participant, $q < 0.05$) were concatenated to produce bilateral ROIs. We concatenated voxels from the dorsal

and ventral aspects of V2 and V3. The sPCS ROI was defined using voxels with significant activation localized in the posterior part of the superior precentral sulcus during the visual localizer runs, FDR-corrected for multiple comparisons. The "All voxels combined" ROI presented in Figure 4C was defined by concatenating all voxels from all 10 ROIs.

fMRI analysis

For each trial of the spatial mapping runs, we extracted the Z-scored BOLD signal from each voxel averaged over the two TRs occurring 6.75 – 9.00 s after target onset as the observed signal in that voxel for that trial. For each trial of the main spatial WM task runs, we extracted signal from each TR following the target onset and computed spatial representations for each TR independently.

Spatial encoding model

We implemented an inverted spatial encoding model [S5] to compute spatial WM reconstructions using the pattern of activation over subsets of voxels during each time point (TR) of the WM delay period. This method assumes (1) the BOLD signal reflects an approximately linear combination of neural responses within each voxel, (2) at least some voxels within each ROI have non-uniform responses to stimuli presented at different positions on the screen during the "training" phase of the analysis (Fig. 2A-C), and (3) a voxel's estimated encoding model during the training phase is maintained during the main task runs.

For this method we first estimated the spatial sensitivity profile of each voxel (Fig. 2A-C). Then, with novel data, we used the pattern of activation across all voxels of interest and the independently estimated spatial sensitivity profile estimated for each of those voxels to compute the reconstruction carried by the activity across all voxels in a ROI (Fig. 2D). In this implementation of the inverted encoding model we used entirely different datasets for estimating the encoding model and for reconstructing spatial WM contents (i.e., there is no need for a "leave-one-out" procedure). We trained and tested the encoding model using data from each session independently, then combined resulting reconstructions across sessions. Thus, we used 3-4 runs of "training" data from spatial mapping runs at a time for encoding model estimation, and 8-12 runs of "testing" data for computing reconstructions.

We modeled the response of each voxel as a linear combination of a grid of 36 spatial filters (or information channels; Fig. 2C) where $\boldsymbol{B_1}$ is the observed signal in each voxel on each trial (m voxels × n trials), $\boldsymbol{C_1}$ is the predicted responses for each channel on every trial (k channels × n trials; see below), and $\boldsymbol{W}$ is a matrix describing the mapping between "channel space" and "voxel space" (m voxels × k channels):

$$\boldsymbol{B_1 = WC_1} \qquad \text{(Equation 1)}$$

During the training phase, we estimated the spatial sensitivity profile of each voxel by first filtering all training stimuli (Fig. 2A-B) by a "basis set" of 36 spatial filters, each of the form:

$$f(r) = 0.5 + 0.5\cos(r\pi/s)^7 \text{ for } r < s, \text{ 0 elsewhere} \quad \text{(Equation 2)}$$

Where r is the distance from the center of the basis function, and s is a size constant which corresponds to the distance from the basis function center at which the function reaches zero (same function used in [S5]). Spatial filter centers are spaced by 1.86° and have a full-width at half-maximum (FWHM) of 2.31° (the corresponding size constant, s, is 5.82°).

The resulting filtered predicted channel responses ($\boldsymbol{C_1}$, k channels × n trials) enter into an ordinary least-squares regression, along with the observed signal in each voxel ($\boldsymbol{B_1}$) to estimate the contribution of each spatial information channel to the observed response in each voxel across all training trials ($\widehat{\boldsymbol{W}}$, m voxels × k channels):

$$\widehat{\boldsymbol{W}} = \boldsymbol{B_1 C_1^T}\left(\boldsymbol{C_1 C_1^T}\right)^{-1} \qquad \text{(Equation 3)}$$

This step amounts to a simple general linear model of the form commonly implemented in fMRI analyses, and is performed on each voxel individually (i.e., it is a univariate analysis).

Once the spatial sensitivity profile ("channel weights", $\widehat{\boldsymbol{W}}$) was estimated for each voxel using all spatial mapping runs within a session, we computed the pseudoinverse of the channel weights to obtain a mapping from the observed pattern of activation across all voxels within an ROI ("voxel space") into estimated channel responses ($\widehat{\boldsymbol{C}}_{\boldsymbol{2}}$, $k \times n$, "information space"):

$$\widehat{\boldsymbol{C}}_{\boldsymbol{2}} = \left(\widehat{\boldsymbol{W}}^{\boldsymbol{T}}\widehat{\boldsymbol{W}}\right)^{-\boldsymbol{1}}\widehat{\boldsymbol{W}}^{\boldsymbol{T}}\boldsymbol{B}_{\boldsymbol{2}} \qquad \text{(Equation 4)}$$

This mapping is computed using all voxels assigned to a given ROI (e.g., all V1 voxels, Figs. 3-4, or all voxels across all ROIs, Fig. 4), and is different for any given combination of voxels (i.e., it is a multivariate operation).

The resulting estimated channel responses reflect the response of each information channel that is most likely to have given rise to the observed pattern of activation across all voxels within an ROI, given our linear model of observed BOLD responses as a function of information channel activation.

To compute spatial WM reconstructions from each vector of channel responses (one 36-dimensional channel response vector for each fMRI data volume for each ROI) we computed a sum of the basis functions, each weighted by the corresponding estimated channel response for each channel response vector (Fig. 2D). For Figures 2-4 and Figure S2 we averaged reconstructions from 6.75 – 9.00 s (2 volumes during the delay period). For all Supplemental Movies, we did not average reconstructions across time. Each frame of the movies corresponds to a single fMRI data volume (averaged across all trials and participants).

Critically, because we used data from an independent task (Fig. 2A-B) to estimate the encoding model which was then used to generate reconstructions during the memory task (Fig. 1A), we can be confident that stimulus-specific idiosyncrasies in the data (i.e., overfitting noise in a leave-one-out cross-validation design) were not responsible for our ability to reconstruct the contents of spatial WM.

Coregistering reconstructions

Because the representations of remembered targets within WM reconstructions are rather weak (especially compared to stimulus representations of flickering checkerboard stimuli, [S5]) we implemented several different coregistration procedures in order to visualize different aspects of the WM reconstructions. Because the pattern of BOLD activation was mapped into an information space, we can manipulate the functions describing information space in a manner that allows us to average target representations within the WM reconstructions from trials in which remembered visual stimuli appeared in different positions by effectively "rotating" the WM reconstructions in order to match target positions. That is, we can combine the target representations when the remembered stimulus was on the top left of the screen with the target representation when the remembered stimulus was at the top right of the screen. This removes any potential inhomogeneities in representations that are dependent upon particular screen positions and allows us to ascertain how target representations generally change across the visual system and across task demands independent of the exact position of a remembered stimulus. This is a unique ability afforded by this analysis method – it is otherwise very challenging to determine how one would average responses to stimuli at different parts of the screen in "voxel space".

First, we sought to qualitatively evaluate WM reconstructions for different stimulus arrangements across the 3 memory load conditions. We collapsed all trials in which the 2 target stimuli were an equal average angular distance apart. To do this, we rotated all reconstructions clockwise such that the non-probed target (for Remember 1, this was the non-cued target; for Remember 2, this was the target which was not probed at the end of the trial) appeared along the positive *x* axis,

and flipped reconstructions in which the probed stimulus was below the *x* axis across the horizontal midpoint such that there are four possible target arrangements (the four columns in Fig. 3, Fig. S2A, Movies S1A-C).

Second, we sought to coregister reconstructions so that the remembered target was always centered at exactly the same position (along the *x* axis, 3.25° from fixation). We accomplished this by rotating each trial's reconstruction by the angular distance between the target and the horizontal axis, then horizontally shifting the reconstruction to remove any remaining jitter (Fig. 2E). For Remember 0 and Remember 2 trials, we did this for each target in turn and averaged the resulting reconstructions (Fig. 4; Fig. S2B; Movie S2). Note that this procedure was identical across all ROIs, so any changes in WM reconstructions resulting directly from the coregistration procedure would be similar across all ROIs. Because the effects of memory load on target representations within WM reconstructions we observed are different across different ROIs, those effects must be due to changes in fidelity of the target representation rather than artifacts of the analysis procedure (see Fig. S4).

Surface fitting

For each exactly-coregistered WM reconstruction (aligned to optimally reveal the target representation) we fit a surface (Fig. 4A-B) parameterized by its position, size (*s*, distance from center at which surface reaches zero), amplitude (*a*) and baseline (*b*, non-spatially-selective) shift:

$$f(r) = b + a(0.5 + 0.5\cos(r\pi/s))^7 \text{ for } r < s, \quad 0 \text{ elsewhere} \qquad \text{(Equation 5)}$$

To ensure robust fits, we restricted the position of the best-fit function to be the point on the reconstruction with the largest local average. The local average was computed across a disc 0.5° in radius (via convolution of each reconstruction with a circular disc). All other parameters (*s*, *a*, and *b*) were allowed to freely vary.

We computed fits using fminsearch in MATLAB (R2012a, The Mathworks, Inc), which implements the Nelder-Mead algorithm. For every fit, we began from 10 different randomly selected initial values and chose the fit that minimized the sum of squared errors between the surface function and the coregistered WM reconstruction.

Statistical procedures

To evaluate behavioral effects of the memory load manipulation (Fig. 1) we performed a resampling test in which we resampled all valid trials across all participants (with a response, for RT/accuracy in the scanner, Fig. 1C-D; within a reasonable spatial response window, behavioral recall task, Fig. 1E) with replacement for each memory load condition 1,000 times and computed the distribution of differences between R2 and R1 for inside-scanner accuracy, RT and outside-scanner recall error. We defined *p*-values as the percentage of resampling iterations in which an effect was observed in the opposite direction of the mean effect and multiplied by two, as we did not make any *a priori* predictions about the direction of the effect. Each of these comparisons were tested against a threshold of $\alpha = 0.05$.

When comparing best-fit parameters to target representations with WM reconstructions across memory load conditions (Fig. 4), we implemented an across-participant resampling procedure. We combined data across all 4 participants into one large pool of 1,248 trials per condition. Within each condition, we resampled across all trials, with replacement, 1,000 times and computed an averaged WM reconstruction. We then implemented the surface fitting procedure described above to quantify the target representation on each resampling iteration, resulting in a resampled distribution of 1,000 best fit values for each parameter (size, amplitude, and baseline). Error bars on figures are 95% confidence intervals derived from this resampled fit parameter distribution. *P*-values are computed as the percentage of resampled iterations in which a difference opposite to the mean difference was observed (e.g., if Remember 2 had a larger value on average than Remember 1, *p* would correspond to the percentage of resampled iterations in which Remember 2 had a smaller value than Remember 1). All *p*-values were doubled (because we made no *a priori* hypothesis about the direction of the effect)

and compared against an alpha threshold corrected for multiple comparisons across 10 ROIs for each parameter using Bonferroni's method (corrected $\alpha = 0.005$), and trends are defined by *p*-values below an uncorrected threshold $\alpha = 0.05$ and are indicated using gray asterisks in Figure 4D (see also Table S1 for *p*-values for all comparisons in Fig. 4D). Because the results from the "All voxels combined" ROI include data from each of the other 10 ROIs, we performed statistics separately for the 10 constituent ROIs and for the combined ROI (that is, we corrected for 10 comparisons within each parameter when evaluating the statistical significance of each ROI separately, then performed no corrections when evaluating the statistical significance of the combined ROI). However, as can be seen in Table S1, correcting for an additional comparison would not change which tests are found to be significant. Since we computed 1,000 iterations in this resampling procedure, we only report *p-values* as < 0.001 if we do not observe an effect opposite the mean in the resampled distribution. It is possible that with more iterations we could see a result, and so it is inappropriate to report $p = 0$.

When evaluating statistical significance of mean delay-period BOLD responses (Fig. S1B), we resampled the average BOLD signal during the delay period (6.75 – 9.00 s after target onset, as in Figs. 3-4, S2) on each trial for each set size condition pooled across all participants (1,000 resampling iterations, resampling all trials of a given set size condition with replacement). Because we only included Remember 0 trials during Sessions 1 and 2, we only included those sessions in this analysis. Then, we compiled 3 resampled distributions corresponding to the difference between each pair of set size conditions (R0 vs. R1, R0 vs. R2, R1 vs. R2) and computed *p*-values as the percentage of resampled iterations in which a difference opposite to the mean difference was observed, doubled (as described above). To evaluate significance for this exploratory analysis, we corrected for multiple comparisons using Bonferroni's method across all 30 computed *p*-values (3 pairwise comparisons for each of 10 ROIs; $\alpha = 0.0017$, Table S1).

Simulating and fitting target representations with known parameters

To evaluate the accuracy and sensitivity of our fitting procedure (Fig. 4A-B), we simulated WM reconstructions containing a target representation (or multiple target representations) with known parameters, then implemented the fitting procedure in the same way as used on the actual data (see *Surface Fitting*, above). We simulated target representations as a single surface (Equation 5) with fixed *s*, *a*, and *b* parameters, centered at (3.25°, 0°) and computed over a square field of view (FOV) from -4.65° to 4.65° along the *x* and *y* axes, sampled at a grid of 151 × 151 points (along each axis). Then, we varied a single parameter while keeping the others constant and plotted best-fit parameters as a function of the value of the single parameter that we varied (Fig. S4A-C). Additionally, we performed this same procedure while allowing the FOV to vary (Fig. S4D), but keeping all target representation parameters constant, in order to evaluate how allowing the representation to artificially return to "baseline" at distant edges of the reconstruction might influence the fitting routine.

Finally, to quantify how the coregistration procedure (Fig. 2E) might result in changes in best-fit parameter estimates without any "true" changes in parameter values of the underlying target representations, we generated a simulated dataset in which we centered representations at the actual target positions participants remembered during scanning, performed an identical coregistration procedure as to that in the main analysis, and fit the resulting coregistered and averaged representations. We did this using the actual best-fit parameters for the R1 condition in each ROI plotted in Figure 4D (Fig. S4E).

Supplemental References

S1. Jerde, T. A., Merriam, E. P., Riggall, A. C., Hedges, J. H., and Curtis, C. E. (2012). Prioritized Maps of Space in Human Frontoparietal Cortex. J. Neurosci. *32* , 17382–17390.

S2. Todd, J. J., and Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. Nature *428*, 751–4.

S3. Xu, Y., and Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. Nature *440*, 91–5.

S4. Emrich, S. M., Riggall, A. C., Larocque, J. J., and Postle, B. R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. J. Neurosci. *33*, 6516–23.

S5. Sprague, T. C., and Serences, J. T. (2013). Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nat. Neurosci. *16*, 1879–87.

S6. Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nat. Neurosci. *9*, 1432–8.

S7. Graf, A. B. A., Kohn, A., Jazayeri, M., and Movshon, J. A. (2011). Decoding the activity of neuronal populations in macaque primary visual cortex. Nat. Neurosci. *14*, 239–245.

S8. Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E.-J., and Shadlen, M. N. (1994). fMRI of human visual cortex. Nature *369*, 525.

S9. Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., Rosen, B. R., and Tootell, R. B. H. (1995). Borders of Multiple Visual Areas in Humans Revealed by Functional Magnetic Resonance Imaging. Science *268*, 889–893.

S10. Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann. Stat. *29*, 1165–1188.

Chapter 3, in full, is a reprint of the material as it appears in a report entitled "Reconstructions of information in visual spatial working memory degrade with memory load" published in Current Biology 2014. Sprague, Thomas C.; Ester, Edward F.; Serences, John T., Cell Press, 2014. The dissertation author was the primary author of the manuscript. We thank Miranda Scolari and Mary Smith for assistance in developing parietal cortex mapping protocols, Anna Byers for assistance with data collection, Sirawaj Itthipuripat, Vy Vo, and Alexander Heitman for discussion, and Sirawaj Itthipuripat, Vy Vo, and Stephanie Nelli for comments on the manuscript. This work was supported by a NSF Graduate Research Fellowship to T.C.S., NIH T32-MH020002-12 to E.F.E., and NIH R01 MH-092345 to J.T.S.

Chapter 4:

Restoring latent visual working memory representations in human cortex

4.1: Introduction

Even in the context of simple visual tasks, such as tracking multiple moving objects or identifying the colors associated with basic shapes, an observer's ability to accurately represent sensory information declines rapidly as the complexity of the scene increases (Franconeri et al., 2013; Tsubomi et al., 2013). These processing limits are highlighted in working memory (WM) tasks, which require the maintenance and manipulation of a subset of sensory information that is no longer physically present in the environment (Baddeley and Hitch, 1974; Bays, 2015; Curtis and D'Esposito, 2003; D'Esposito and Postle, 2014; D'Esposito, 2007; Gazzaley and Nobre, 2012; Luck and Vogel, 2013; Ma et al., 2014; Postle, 2015; Sreenivasan et al., 2014; Stokes, 2015). A classic finding in WM is that increasing the amount of information stored leads to impaired behavioral performance when recalling visual features (Bays and Husain, 2008; Bays, 2015, 2014; Keshvari et al., 2013; Ma et al., 2014; Zhang and Luck, 2008). Classically, WM representations are thought to be maintained by sustained spiking activity (Funahashi et al., 1989; Fuster and Alexander, 1971) accessible using fMRI activation patterns (Harrison and Tong, 2009; Serences et al., 2009) and the pattern of EEG alpha-band potentials across the scalp (Foster et al., 2015) during the delay period of a task. Accordingly, impaired performance with increasing set sizes is accompanied by a decrease in spike rates related to relevant memoranda in macaques, or by a decrease in the ability to differentiate fMRI activation patterns tied to different remembered items in humans (Buschman et al., 2011; Emrich et al., 2013; Landman et al., 2003a; Matsushima and Tanaka, 2014; Sprague et al., 2014). Importantly, the fidelity of activity patterns in these areas of visual cortex as measured with fMRI is tied to behavioral performance on WM tasks (Albers et al., 2013; Emrich et al., 2013; Ester et al., 2013; Reinhart et al., 2012; Sprague et al., 2014), which suggests that load-based modulations of WM-related activity patterns play a role in guiding behavior.

One general mechanism that might mediate this relationship between WM load effects and behavioral load effects is inter-item competition manifest as divisive normalization whereby the presence of each item's representation suppresses all other representations, resulting in degraded spiking representations for each item (Bays, 2015, 2014; Carandini and Heeger, 2012; Franconeri et al., 2013). Such a mechanism would result in an irreversible loss of information encoded via active spiking representations because diminished spiking representations become permanently corrupted by noise as spike rates are reduced (Bays, 2014). This loss of information is irreversible and cannot recover with any type of additional processing (Cover and Thomas, 1991; Saproo and Serences, 2014, 2010; Shannon, 1948; Sprague et al., 2015).

However, the notion that increasing the number of items in WM leads to an obligatory and irreversible degradation of neural representations is complicated by results from behavioral experiments in which participants are given an informative retrospective cue (a "valid retro-cue") during the delay-period that indicates which element in a memory array is going to be eventually relevant for behavior. Participants respond more quickly and more accurately on cued trials compared to a baseline condition using non-informative "neutral" retro-cues (Griffin and Nobre, 2003; Landman et al., 2003b; Makovsik and Jiang, 2007; Matsukura et al., 2007). While these results hint that neural WM representations improve following retrocues, it could potentially be the case that a retro-cue prevents representations from decaying further, but does not enhance their fidelity, or that WM representations remain unchanged but a cue improves access to static representations.

To address this discrepancy, we hypothesized that behavioral retro-cue benefits are observed because *active* WM representations – those that are reflected in elevated firing rates and/or BOLD fMRI responses – can be augmented using information encoded via within-trial *latent* or "activity-silent" codes (Stokes, 2015; Stokes et al., 2013; Wolff et al., 2015). For

example, transient information might be encoded via subthreshold membrane potential depolarization, changes in synaptic efficacy (Briggs et al., 2013), item-related fluctuations of pre-synaptic calcium concentration (Mongillo et al., 2008), or some combination thereof. Importantly, these hypothesized latent codes would be effective within the current trial only, and are not directly related to the transfer of information into a stable form of long term memory (LTM) that is supported by a largely separable neural substrate involving the redistribution of receptors and/or protein synthesis which operates at a slower timescale than is relevant for representing information within a single trial (Milner et al., 1998; Squire and Wixted, 2011).

In this framework, when participants are validly cued that one of several items in WM is relevant, the active representation of the cued item might be recovered based on these other potential transient sources of information that are each invisible to common neural measures such as spike rate or BOLD activation level. For example, a set of neurons carrying a transient latent WM representation in the form of elevated subthreshold membrane potential without a change in spike rate could be activated by input from another neuron or brain region, allowing the latent representation to contribute to the fidelity of an active (spiking) representation. While previous work has identified initial evidence for such transient latent representations of category-level information (LaRocque et al., 2013; Lepsien and Nobre, 2007; Lewis-Peacock and Postle, 2012; Lewis-Peacock et al., 2012), it remains unknown how the relative fidelity of each item's representation is updated after presentation of a retro-cue, and how those representations are related to behavioral performance on a task requiring high-precision maintenance of feature values.

The hypothesis that latent transient codes can bolster active neural representations makes several predictions. First, in line with existing data, behavioral performance and the fidelity of active neural WM representations should become degraded with increasing memory

load. However, a retro-cue indicating that only 1 item (of several) is relevant should lead to recovery of behavioral performance (e.g., LaRocque et al., 2015). Second, the recovery of mnemonic precision following a retro-cue should be accompanied by a corresponding recovery of an already-degraded active neural WM representation. Additionally, if latent representations contribute to behavior, the degree to which latent information facilitates the restoration of active neural representations following a retro-cue may co-vary with behavioral performance. Critically, direct evaluation of the fidelity of active WM representations in a stimulus-referred feature space (Sprague et al., 2015) is necessary to distinguish this hypothesis that retro-cues enable the restoration of degraded active representations to a more informative state from an alternative account whereby retro-cues enhance access to otherwise stable representations.

We tested each of these predictions using a task where participants precisely maintained the spatial positions of 1 or 2 items in visual WM. On some trials, we presented a retro-cue midway through the delay interval validly cueing which item was relevant for behavior; on the remainder of trials we presented a non-informative neutral retro-cue. Consistent with previous results, we identified degraded behavioral performance and neural WM representations when more items were remembered (Emrich et al., 2013; Sprague et al., 2014). However, when participants were cued during the delay which item was relevant, behavioral performance and neural WM representations each substantially improved. Together, these results demonstrate that degraded WM representations can recover, implicating the existence of information within latent neural codes that can support improved behavioral performance.

4.2: Results

We tested the fidelity of WM representations using a mixture of behavioral and neural measures while participants performed a retro-cued spatial recall task. Participants held 1 (Remember 1, R1) or 2 (Remember 2, R2) items from a 2-item display in spatial WM as indicated by the initial color of the central fixation point over a 16 s delay period. On some Remember 2 trials, we changed the color of the fixation point to provide either an informative "valid" (R2-valid) or an uninformative "neutral" (R2-neutral) retro-cue at the end of the first half of the delay interval that indicated which item(s) might be cued for recall at the end of the entire delay interval (Fig. 4-1A). We used fully reliable retro-cues to ensure that participants utilized the cue to optimize behavior. At the end of the delay period, participants recalled the exact horizontal or vertical position of one of the items by adjusting vertical or horizontal response bar, respectively (Fig. 4-1A, response coordinate randomly assigned on each trial). On Remember 1 and Remember 2-valid trials, the probed item was the only item that required active maintenance in WM during the second delay period, and on Remember 2-neutral trials, we randomly selected which of the two remembered items would be queried for recall. Note that Remember 2-neutral and Remember 2-valid trials were identical during the first delay period, and differed only during the second delay period when participants were cued to remember either 1 or 2 items (Fig. 4-1A). The Remember 2-valid condition allowed us to assess the manner in which an informative retro-cue influences behavioral performance and neural representational fidelity compared to performance when both items were remembered in the Remember 2-neutral condition. We pseudo-randomly chose target positions from an array of 6 invisible discs that were spaced equally along an imaginary circle 3.5° from fixation and which were rotated around fixation on every trial (Fig. 4-1B). Targets were randomly positioned within each disc so that discrete or verbal encoding strategies (e.g. "up and to the right" or "8 o'clock") would not be adequate to support sufficiently accurate recall

performance on the task (Sprague et al., 2014). Each participant (n = 5) completed three 2-hr fMRI scan sessions (324 to 378 total trials per participant). We indexed behavioral performance on all trials after the end of the second delay period based on the distance between the response bar and the relevant target at the conclusion of a 3 s response period. We also collected additional behavioral data outside the scanner (1-3 sessions per participant), but do not include these data in our behavioral analyses due to a response artifact where participants performed worse on horizontal recall trials than vertical recall trials, likely due to the wider aspect ratio (width:height) of the LCD monitor used for behavioral testing (16:9) compared to the rear projection display used during scanning (4:3).

Behavioral performance improves with a valid retro-cue

Participants performed more poorly, as indicated by higher average recall error, on Remember 2-neutral trials as compared to Remember 1 trials (Fig. 4-1C; R1 vs. R2-neutral: $p < 0.001$, resampling test, see Experimental Procedures). This drop in recall accuracy is consistent with degraded neural representations that accompany increasing set sizes, and replicates previous findings (Sprague et al., 2014; see also: Bays and Husain, 2008; Bays, 2014; Emrich et al., 2013; Zhang and Luck, 2008). When one item was cued midway through the delay interval (Remember 2-valid trials), behavioral performance improved as compared to Remember 2-neutral trials (R2-neutral vs. R2-valid: $p = 0.016$). Performance was slightly worse on Remember 2-valid trials compared to Remember 1 trials (R1 vs. R2-valid: $p = 0.024$), suggesting substantial but imperfect recovery of WM representations with valid cues.

Reconstructing WM representations

Changes in behavioral performance are consistent with changes in the quality of neural representations following a valid retro-cue. For example, increased response conflict

could result in poorer access to otherwise stable and robust neural WM representations when 2 items are maintained as compared to one item. In order to isolate and assess the information content of WM representations from these other potential intervening factors, we implemented an inverted encoding model (IEM) to reconstruct images of the contents of spatial WM based on blood oxygenation level dependent (BOLD) fMRI activation patterns (Brouwer and Heeger, 2009; Ester et al., 2015, 2013; Sprague and Serences, 2013; Sprague et al., 2015, 2014). We computed reconstructions in each of 10 independently-identified regions of interest (ROIs): retinotopic occipital visual areas (V1-hV4; V3A), retinotopic posterior parietal areas (IPS0-IPS3), and the superior precentral sulcus (sPCS; thought to be a human homolog to macaque frontal eye fields; the sPCS ROI was identified using an independent spatial WM localizer task; see Experimental Procedures and (Srimal and Curtis, 2008)). We also assayed representations encoded by the joint activation pattern across all these regions after concatenating voxels from all areas before computing reconstructions ("All ROIs combined").

The first step of the reconstruction analysis involved acquiring an independent set of mapping scans in which we presented a 1.083°-radius circular flickering checkerboard stimulus at each location on a triangular grid subtending the full extent of the stimulus display while participants performed a demanding spatial WM task near the stimulus location (following Sprague and Serences, 2013; Sprague et al., 2014; see Experimental Procedures and Figure 4-5A). We used responses to these mapping stimuli to estimate the spatial sensitivity of each voxel in each ROI by modelling voxel activation levels as a weighted sum of spatial information channels (spatial filters) arrayed across a hexagonal window on the visual field in a triangular grid (Fig. 4-4A; each filter can be thought of as hypothetical neural subpopulations with different spatial RF positions). Using the measured activation level in response to mapping stimuli presented in each location, we estimated the contribution each spatial channel made to the signal measured from each voxel – that is, the *encoding model* for

each voxel. This procedure is univariate, and was implemented using a general linear model (GLM; see Experimental Procedures: fMRI analysis: inverted encoding model).

Then, using the estimated spatial sensitivity profile (or the pattern of channel weights) from all voxels within an ROI, we computed a mapping that transforms multivariate activation patterns measured during the WM task (Fig. 4-1A) from voxel space (1 response level from each voxel) into channel space (1 response level for each modeled channel). The resulting channel activation patterns (Fig. 4-4B) reflect the activation level in each modeled information channel that accounts for the observed activation pattern across all voxels in a given ROI, given our independently-estimated encoding model of the spatial selectivity of each voxel (Fig. 4-4A). Finally, we used these channel activation patterns to compute a sum of spatial filters, weighted by the corresponding activation level of each spatial channel. This produces a smooth image and yields a reconstruction of the contents of spatial WM in units of BOLD activation (due to the linear signal transformation) plotted in visual field coordinates. Thus, we obtain reconstructions of the entire visual field, and we refer to the light spots in these reconstructions appearing at position(s) held in spatial WM as “target representations”.

Because target positions are unique on every trial of the experiment, we next rotated all WM reconstructions so that target positions were aligned across trials (see Sprague et al., 2014; Experimental Procedures, Figure 4-5C). Due to this rotation that respects only the position of a WM target on each trial, any substantial target representation in the WM reconstructions at the aligned position reflects only activation that follows the position of the WM target across trials in channel information space (visual field coordinates). This is a key part of the analysis procedure, because across trials the only consistent feature in these aligned reconstructions is the location of the WM target. This approach is thus far more powerful than analyses performed in native voxel (brain) space, where translating and coregistering activation patterns across different stimulus locations would be challenging or impossible due

to difficulties associated with co-registering target representations encoded by the univariate responses of individual voxels that are idiosyncratically embedded in ROIs that vary in shape and size across participants. Moreover, the transformation of data from univariate native voxel space to an *a priori* defined information space allows for all voxels within each ROI to jointly constrain the estimated fidelity of the target representations (i.e. the target representations reflect a multivariate estimate that can exploit the information contained in the pattern of all of the individual voxel responses within a given ROI).

Reconstructions track the dynamic contents of spatial WM

First, we evaluated whether target representations in WM reconstructions track the remembered position(s) maintained by participants. We plotted WM reconstructions computed using activation patterns from each time point during the trial (0-20.25 s) averaged over all trials with similar WM target arrangements within each WM condition (colored discs in Fig. 1B; note that reconstruction time courses are not adjusted for the hemodynamic delay, so changes in reconstructions lag changes in stimuli/WM contents by ~6 s). We combined trials with similar relative target arrangements (±60°, ±120°, ±180°), and rotated reconstructions to align all similar trials (see Experimental Procedures, Figure 4-5). On Remember 1 trials, reconstructions computed using an early time point (4.50 s) contain representations of both targets (example target arrangement condition shown in Fig. 4-6A). However, shortly thereafter, only the relevant target (yellow dashed circle) remains visible (6.75-18.00 s). While the target representation becomes less visually pronounced over the duration of the trial following the initial encoding transient, it remains visible throughout the late delay interval. The same pattern holds for Remember 2 trials with a neutral cue (Fig. 4-6B): representations of items maintained in WM persist in reconstructions after the initial sensory transient through the late delay period, though target representations are weaker than those in Remember 1

trials, especially during the late delay period. On Remember 2 trials with a valid cue (Fig. 4-6C), we observed a transition from two simultaneous target representations (early delay) to one target representation (late delay) following the cue, confirming that these spatial reconstructions track the dynamically changing contents of WM over extended delay intervals. Furthermore, the representation of the cued item during the late delay period appears to be enhanced compared to each of the 2 target representations earlier in the delay period (after the encoding transient subsides at ~9.00 s).

For several subsequent analyses of WM reconstructions, we focused on average reconstructions during the first delay period (Delay 1; 6.75-9.00 s after target onset) and during the second delay period (Delay 2; 15.75-18.00 s after target onset) after accounting for the hemodynamic delay. These delay-period target representations tracked the position(s) of item(s) in WM. When we binned trials by the relative position of WM targets (Fig. 4-1B), target representations always appeared nearby and only in the position(s) corresponding to the remembered item(s) during that condition and delay period (Fig. 4-7). Additionally, the quality of target representations always exhibited the same pattern across delay periods regardless of target arrangement – during the first delay period, representations degraded when 2 items were maintained (Fig. 4-7A compared to Fig. 4-7C,E), and during the second delay period, a valid cue restored the cued representation to a high-fidelity state (Fig. 4-7E-F).

Fidelity of WM target representations

To quantify whether or not a target representation was robust in a given ROI, we computed reconstructions over an annulus around fixation (2.9-4.1°) and averaged these reconstructions across eccentricity, resulting in a 1-d reconstruction as a function of polar angle position (Fig. 4-4C; Experimental Procedures: Quantifying WM representations).

First, we plotted these rotated and aligned 1-d reconstructions as a function of time to evaluate the relative fidelity of WM representations over the entire trial duration (Fig. 4-8A). On Remember 1 and Remember 2-neutral trials, an initially high-fidelity representation during WM encoding subsides, but remains present in many ROIs throughout Delay 2 (e.g., V3A; IPS0). On Remember 2-valid trials, the cued item is robust even at late time points during Delay 2, often increasing in fidelity following the cue (Remember 2-valid, compare early and late time points, e.g. V1).

To determine the strength of a WM representation in these 1-d polar angle reconstructions, we developed a "representational fidelity" metric. We define representational fidelity as the vector mean of a set of unit vectors pointing in each polar angle direction weighted by the reconstruction activation for the corresponding polar angle, projected on a unit vector pointing in the "correct" direction (here, always 0° polar angle, because we rotate all 1-d reconstructions to a common center; Fig. 4-4C; Experimental Procedures: Eq 5). If this metric is reliably greater than zero (statistically evaluated using a resampling procedure, see Experimental Procedures: Quantifying WM representations), then there is a consistently identifiable WM target representation in the corresponding reconstruction. In contrast, if the reconstruction has a uniform activation profile, then this metric will result in a value indistinguishable from zero. To visualize the dynamics of representational fidelity, we plotted the time course of this metric over the entire trial (Fig. 4-8B). The fidelity of the target representations gradually decreases over time on Remember 1 and Remember 2 trials, in which WM contents are held constant, but substantially recovers following the valid cue on Remember 2-valid cue trials when one item is cued to be relevant.

Next, we compared 1-d polar angle reconstructions and representational fidelity during each delay period (Fig. 4-9). Importantly, we found significant representational fidelity in all ROIs across both delay intervals on Remember 1 and Remember 2-valid cue conditions

($p \leq 0.001$; one-tailed resampling test against 0, FDR corrected for multiple comparisons, see Experimental Procedures; all *p*-values available in Table 4-2). On Remember 2 trials with neutral cues we found representations in all ROIs during Delay 1 ($p < 0.001$), and all ROIs except IPS1, IPS2, and sPCS during Delay 2 (Fig. 4-9A; significant ROIs all $p \leq 0.039$, maximum p-value V3A; non-significant ROIs all $p \geq 0.351$, minimum *p*-value IPS1).

Finally, we compared representational fidelity between each delay period within each cue condition (Fig. 4-9B). Representational fidelity significantly declined from Delay 1 to Delay 2 in V1-hV4, IPS0 and All ROIs combined on Remember 1 trials ($p < 0.001$; FDR-corrected) and in V1-hV4, IPS0-IPS2, sPCS, and All ROIs combined on Remember 2-neutral trials ($p \leq 0.018$, maximum *p*-value in hV4; two-tailed resampling test of differences in representational fidelity between Delay 1 & 2 against 0; all *p*-values in Table 4-3). In contrast, representational fidelity did not decline between delay periods on Remember 2-valid trials, and in fact fidelity was significantly higher during Delay 2, after the valid cue, than during Delay 1 in several occipital and parietal ROIs (V1, V3, IPS0-IPS3, and All ROIs combined; $p \leq 0.022$, maximum *p*-value in V3). In sum, these analyses identify reliable WM representations on Remember 2-neutral trials, even when they are not easy to visualize in the reconstructed WM images (Fig. 4-7), and quantify a significant enhancement of representations on Remember 2-valid trials following the cue (Fig. 4-9B).

Relative target activation through time

As an alternate means of evaluating whether WM reconstructions accurately track the contents of WM and the recovery of WM representations following a valid retro-cue, we also extracted the activation time course from each trial's reconstruction over a narrow spatial window surrounding the exact target position on each trial. Then, we computed the difference between the extracted activation for the target queried at the end of the trial and for the non-

queried target. These difference timecourses index the relative activation of each target representation across all voxels of each ROI. On Remember 1 trials, the single target representation should be consistently maintained across the entire trial, and its activation should be higher as compared to the non-remembered target representation. We observe this predicted pattern in all ROIs measured (Figure 4-10A). On Remember 2-neutral trials, the difference between target representation activation should be negligible (until the response period begins, at which time a visual transient occurs near the queried target's position). No representation activation differences deviate from zero on these trials (Figure 4-10B). On Remember 2-valid trials we predicted a transition from no target activation differences to a substantial target activation difference following the cue (8.0 s), reflecting the relative enhancement of the cued target representation over the non-cued target representation. Many ROIs, especially retinotopic visual ROIs V1-V3A and posterior parietal regions (IPS0-1) show evidence of this transition in relative representation activation (Figure 4-10C).

Quantifying properties of target representations

Next, we sought to quantify how target representations change across memory loads and cueing conditions. For example, when multiple items are remembered representations could be lower-fidelity because they are "dimmer" above noisy background signals, as indexed by a lower amplitude over baseline, or because they are less spatially precise, as indexed by a larger size (Sprague et al., 2015, 2014). First, in each ROI during each delay interval for each WM condition, we precisely aligned all reconstructions across trials such that the target position was at an exactly known position (red dots, Fig. 4-11A,E; Figure 4-5D). Then, we fit a surface, defined by its size (reflecting the spatial precision of the representation), amplitude (reflecting the magnitude of the representation over spatially-global baseline levels in the reconstructions), and baseline (reflecting a spatially-global offset in the

reconstruction unrelated to WM target position), to each coregistered reconstruction using a resampling procedure (see Experimental Procedures; Fig 4-4D; Figure 4-5E). Because fits to a null representation (i.e., one with representational fidelity indistinguishable from 0, Figs 4-8 and 4-9, Table 4-2) are impossible to interpret, we only consider and report comparisons of fit parameters between pairs of conditions in which each condition has non-zero representational fidelity. However, for completeness, we plot these fits in Figure 4-11 and Figure 4-13 and report statistical comparisons in Tables 4-4 and 4-5.

First delay: increasing memory load degrades representation amplitude

During the first delay, averaged reconstructions qualitatively appeared higher in amplitude during Remember 1 trials than Remember 2 trials under either cueing condition (both of which have identical WM contents during this delay period; Fig. 4-11A). Replicating previous results (Sprague et al., 2014), target representation amplitude during the first delay was higher on Remember 1 trials as compared to both Remember 2-neutral and Remember 2-valid trials in visual (V1-V3A and hV4, all $p < 0.001$; Fig. 4-11C) and most parietal (IPS0-2, $p \leq 0.014$; maximum p-value IPS2, R1 vs R2-vaild) ROIs, as well as in reconstructions computed using all ROIs combined ($p < 0.001$; all comparisons of fit parameters based on resampling test between condition pairs and FDR-corrected for multiple comparisons within fit parameter, see Statistical Procedures and Table 4-4 for all p-values). As expected given the identical nature of the trials during the first delay interval, no ROIs exhibited unequal representation amplitude between Remember 2-neutral and Remember 2-valid conditions during Delay 1 (all $p \geq 0.158$, minimum p-value in V3A). Fit baseline was significantly greater on Remember 2-neutral and -valid trials as compared to Remember 1 trials in V3, V3A, IPS0, and in reconstructions computed from all ROIs combined (Fig. 4-11D, $p \leq 0.018$; maximum p-value All ROIs combined, R1 vs R2-neutral). In V1 and V2, a significantly

greater baseline was seen when comparing Remember 2-valid to Remember 1 trials ($p \leq 0.002$). In V1 we also observed a significant difference between fit baseline for Remember 2-neutral and Remember 2-valid representations ($p = 0.006$). Finally, WM representation size was significantly smaller on Remember 2-neutral trials as compared to Remember 1 trials in hV4 ($p < 0.001$, Fig. 4-11B). Though this finding is unexpected (a smaller representation is consistent with a more precisely-maintained spatial position; Sprague et al., 2015, 2014), this region remains an outlier in its observed size decrease with memory load. This may be because this is the only ventral stream region examined in this study, in part due to placement of the imaging volume to optimize coverage of parietal and dorsal occipital cortex. Future studies examining the fidelity of spatial WM representations in the ventral stream may help identify contributions of these regions to WM maintenance.

These first delay results closely replicate our previous report in which we characterized how WM representations change as set size is manipulated from 1 to 2 items (Sprague et al., 2014). In that report, we found extensive evidence for decreases in WM representation amplitude with increasing set size across visual and posterior parietal cortex, which we fully reproduced here (Fig. 4-11C). Slight deviations between our results here and reported previously are likely due to refinements in the spatial mapping task and analysis procedure (Figure 4-5) between studies.

Second delay: cued representation enhanced via amplitude changes

During the second delay period, target representations appear weaker, though they are still identifiably present in many ROIs (Fig. 4-11E). Mean best-fit surfaces always appear near the true target position, indicating spatially-specific target-related signal in WM reconstructions. Because our fitting procedure did not restrict the best-fit position of surfaces to be near the "correct" position, the identification of WM representations nearby the true

target position suggests the presence of a WM target representation (see also Fig. 4-9A). WM representation amplitude was significantly higher during Remember 2-valid trials than Remember 1 trials in V1, V2, V3, V3A, sPCS, and All ROIs combined (Fig. 4-11G, $p \leq 0.014$, maximum p-value sPCS), and was higher than representation amplitude in Remember 2-neutral trials in all individual ROIs with WM representations during these conditions ($p \leq 0.01$, maximum p-value IPS3; all p-values in Table 4-4). Additionally, several ROIs showed a similar set size effect for representation amplitude during the second delay as during the first delay, such that Remember 1 amplitude was significantly greater than Remember 2-neutral amplitude (V2, V3, V3A, hV4, IPS0, and All ROIs combined, $p \leq 0.012$; maximum p-value hV4). Importantly, WM representation size during the second delay was always similar between Remember 1 and Remember 2-valid conditions, during which participants are maintaining the same number of items in WM (all p*'s* ≥ 0.07, minimum p-value in IPS3). Finally, fit baseline was greater during Remember 2-valid trials than Remember 1 and Remember 2-neutral trials in parietal and frontal cortex (Fig. 6h, R2-valid > R1: IPS0-IPS3, sPCS, All ROIs combined; R2-valid > R2-neutral: IPS3, All ROIs combined; all $p < 0.001$). We also observed a set size effect in which Remember 2-neutral baseline was greater than Remember 1 baseline in IPS0 and IPS3 ($p \leq 0.016$).

Improvements in WM representations of the cued item during the second delay of Remember 2-valid trials are primarily found in the representation amplitude, with additional increases in spatially-global reconstruction baseline levels. The former amplitude increases reflect increased information content about the cued target position over a noisy baseline (Saproo and Serences, 2014, 2010; Sprague et al., 2015, 2014), and the latter reflect non-spatially-specific increased mean activation levels in these regions following an informative cue (see also Figure 4-3).

Behavioral performance varies with WM representation amplitude

Finally, we examined whether the fidelity of the cued target representation on Remember 2-valid trials was related to participants' behavioral recall error on each trial. Because WM representations appear to most consistently vary in their amplitude across manipulations of memory load both across and within trials (Fig. 4-11; and see Sprague et al., 2014), we anticipated that changes in target representation amplitude would be most closely related to behavioral performance. Additionally, since behavioral performance is likely related to the overall fidelity of WM representations across many brain regions, we focused our analysis linking behavior and WM representations on reconstructions computed using voxels concatenated across ROIs (see Figure 4-13 for this analysis performed on each ROI individually).

We separated trials into low- and high-recall error groups based on the median recall error within each condition, session, and participant to ensure that each participant is represented equally in each behavioral performance bin. During the first delay period, we saw no differences in the amplitude, size, or baseline offset of WM representations between low- and high-recall error trials for any WM condition (Fig. 4-12A-B, minimum $p = 0.082$ for R1 baseline, all p-values listed in Table 4-5). However, during the second delay period, cued WM target representations were qualitatively improved (Fig. 4-12C), and quantitatively they were significantly higher in amplitude on low recall error trials compared to high recall error trials (Fig. 4-12D, $p < 0.001$). This observation suggests that participant performance is related to the signal-to-noise ratio (i.e. amplitude over baseline) of the validly-cued WM representation. When each ROI is examined individually, we never observed a significant change in any parameter except amplitude (V3, Remember 2-valid, delay 2; $p < 0.001$), which was in the same direction as reported for All ROIs combined (Fig. 4-12D; amplitude greater for low recall error than high recall error).

4.3: Discussion

Behavioral judgments about sensory stimuli rely on population-level neural representations, and these representations decrease in fidelity as the amount of information increases (Drew et al., 2012, 2011; Tsubomi et al., 2013). When performing a demanding task in which stimuli that are used to guide behavior are no longer present in the display, only sustained internal representations held in working memory (WM) can be used, as no further information can be gathered from the environment. We used an image reconstruction technique (Fig. 4-4) to compare the fidelity of region-level WM representations across memory load conditions and replicated previous findings that behavioral performance (Fig. 4-1) and neural representations (Figs. 4-8, 4-11) degrade with increasing load (Buschman et al., 2011; Emrich et al., 2013; Landman et al., 2003a; Sprague et al., 2014). However, upon presentation of an informative cue indicating which WM representation was relevant for behavior, the fidelity of a degraded representation substantially recovered (Figs. 4-7 through 4-11), and the quality of this recovered representation was related to behavioral performance on the task (Fig. 4-12). These data challenge the notion that WM representations rely primarily on active codes (e.g., spiking activity), for which degraded representations resulting from mutual suppression are permanently impaired (Bays, 2015, 2014). Instead, these data suggest that WM is supported by additional 'spike-silent' information that is manifest in a latent state inaccessible to typical measurement techniques (single unit firing rates or fMRI activation), but can be reinvigorated to an accessible, active state when task demands require an updated representation. Furthermore, this reinstantiation process appears nearly lossless, as behavioral performance and neural representational fidelity recover nearly fully.

Our demonstration that a valid retro-cue enhances the fidelity of WM representations primarily via an increase in their amplitude bears a striking similarity to the effects of spatial attention on visual representations as measured neurally and behaviorally (Gazzaley and

Nobre, 2012; Itthipuripat and Serences, 2015; Lepsien and Nobre, 2007; Nobre et al., 2004; Saproo and Serences, 2014, 2010; Sprague and Serences, 2013; Sprague et al., 2015). However, in these experiments information used to improve neural representations and performance on the task is directly accessible in the sensory input to the visual system. As such, it is not possible to make strong inferences about the ability of neural codes to store "latent" information that can augment degraded representations, as information is still available in the environment during the performance of the task. By using a visual WM task, in which all information a participant can use to perform the task is necessarily represented in the nervous system and not in the display, we were able to demonstrate directly that latent information sources must be present in the brain to bolster neural representations above and beyond an initially degraded state which can then support improved behavioral performance.

Sources of recovered information

Both our behavioral (Fig. 4-1C) and neural (Figs. 4-7 through 4-12) results suggest that the fidelity of neural representations can improve following the presentation of an informative retro-cue. This raises an important issue: what was the format of this information before the cue appeared? In information theory, the data processing inequality theorem provides the strong constraint that the total information about one variable given the observed state of another variable (i.e. mutual information) can never increase with additional processing; it can at best remain constant (Cover and Thomas, 1991; Quian Quiroga and Panzeri, 2009; Saproo and Serences, 2014, 2010; Shannon, 1948; Sprague et al., 2015). Accordingly, we can conclude that the information used to complete the behavioral recall task more accurately following the presentation of a retro-cue must be, in some way, present in the system before the cue appears. However, WM item representations in fMRI-based image reconstructions before the retro-cue is presented were already degraded by this point in the

trial (Fig. 4-9; Sprague et al., 2014). This suggests that the restored representation results from neural response features inaccessible to or masked from our BOLD activation pattern measurements.

One potential source of the restored representational fidelity is WM-related patterns of sub-threshold membrane potential and/or changes in short-term synaptic efficacy, as suggested by prior theoretical and computational modeling efforts (Barak and Tsodyks, 2014; Mongillo et al., 2008; Stokes, 2015; Stokes et al., 2013). Both of these mechanisms render a circuit dynamically sensitive to input as a function of WM contents, and both processes may be difficult to detect with typical electrophysiological or neuroimaging techniques in animals or humans. Consistent with this view, a recent study found that motion-sensitive visual area MT did not carry information about the memorized stimulus over a brief delay interval via stimulus-related changes in spike counts (Mendoza-Halliday et al., 2014). However, changes in local field potentials (LFP) did co-vary with changes in the contents of WM. Such LFP changes may reflect changes in the membrane potentials of nearby neural populations, which could enable more robust mnemonic encoding following the re-allocation of attention (Griffin and Nobre, 2003; Landman et al., 2003b; Lepsien et al., 2011; Makovsik and Jiang, 2007) or a sweep of non-specific activity across the network (Mongillo et al., 2008; Stokes, 2015; Stokes et al., 2013; Wolff et al., 2015). In fact, Mendoza-Haliday and colleagues found evidence for such top-down control of LFP representations by identifying spike-field coherence between prefrontal spikes and LFP activity in the beta band (Mendoza-Halliday et al., 2014), and a recent study that decoded WM representations from EEG scalp potentials found evidence that nonspecific visual input can reveal such hidden states in visual WM (Wolff et al., 2015). Future experiments measuring membrane potentials of single cells while animals perform demanding WM tasks under varying load conditions (Buschman et al., 2011; Kornblith et al., 2015; Landman et al., 2003a; Lara and Wallis, 2014, 2012) may reveal how such non-spiking

sources of neural information can augment more traditional neural population codes that are typically described solely in terms of spiking activity (Bays, 2014; Ma et al., 2006; Tan et al., 2014).

Accordingly, degradation in visual WM representations with greater load could be related to a neural normalization process (Bays, 2015, 2014; Carandini and Heeger, 2012; Franconeri et al., 2013; Ma et al., 2014) whereby each of several simultaneously-held representations mutually suppresses the spiking output of (but not the synaptic input to) all other WM representations. This could allow for latent information encoded via short-term synaptic plasticity of inputs or subthreshold membrane potentials to exert an influence on spiking activity of cells after the presentation of an informative cue (e.g. the mid-delay retro-cue on R2-valid trials in the present study), perhaps by removing the suppressive influence of the irrelevant item on other representations. Then, depolarized membrane potentials caused by continued synaptic input, which is "latent" in this case because it does not cause spiking while both representations are present, would now enable reinvigoration of active neural representations as measured by spike rates due to reduced inhibitory drive. A similar normalization account has also been used to predict attentional modulations as a function of the spatial extent of items attended (Reynolds and Heeger, 2009), which is supported experimentally by EEG and behavioral measurements of representational fidelity (Herrmann et al., 2010; Itthipuripat et al., 2014). Accordingly, normalization of simultaneous representations may reflect a general neural constraint on representing information for the guidance of behavior.

Fidelity of feature representations in WM

Several previous studies cued participants to focus on a single item among multiple items maintained in WM. Lepsien and colleagues (2007) post-cued participants to remember

either a face or scene after both types of stimuli were encoded at the beginning of each trial, and Lewis-Peacock, LaRocque and colleagues cued participants during the delay period to focus on one from among two different stimulus categories presented at the start of each trial (LaRocque et al., 2013; Lewis-Peacock and Postle, 2012; Lewis-Peacock et al., 2012). These studies found evidence for enhanced representations of the cued item category by either comparing mean signal amplitude in different category-selective ROIs (Lepsien and Nobre, 2007) or comparing multivariate classifier evidence for each item category during the delay interval before and after the post-cue (LaRocque et al., 2013; Lewis-Peacock and Postle, 2012; Lewis-Peacock et al., 2012). These studies suggest that cueing one of several items in WM can effectively trigger a switch in the focus of attention to internal category-level representations and accordingly enhance information about the relevant category (LaRocque et al., 2013; Lepsien and Nobre, 2007; Lewis-Peacock and Postle, 2012; Lewis-Peacock et al., 2012). However, these studies did not evaluate the fidelity of WM representations of the category members themselves (i.e., are the retro-cued face representations in FFA more informative about which face is in WM?). Moreover, they do not establish if behavioral benefits following a retro-cue are due to improved representational fidelity of precise feature information in WM or due to more efficient selection of available information at the end of the delay period.

In contrast, we show here that latent information can be revealed by (1) cueing participants to one of several items of the same category (spatial positions) and (2) quantitatively evaluating the feature-specific information content of WM representations carried by fMRI activation patterns throughout the trial. Our results thus conceptually replicate the general finding that the contents of visual WM are dynamic and can be modulated by delay-period cues (Fig. 4-6). However, we show here that such cues can directly enhance the

fidelity with which an individual cued item is represented via the use of latent information (Figs. 4-9 and 4-11) in a manner related to behavioral performance (Fig. 4-12).

Working memory vs recall from long-term memory

It could be the case that the improved behavioral performance and restored representational fidelity following a valid retrocue are a result of recalling precise spatial positions from long-term memory (LTM) rather than performing computations on items held in active maintenance within WM. Recent behavioral studies have found that high-fidelity feature representations can be recalled from LTM (Brady et al., 2013; Sutterer and Awh, 2015) in tasks in which participants first study a set of item-feature pairs where the items are photographs or drawings of distinguishable objects, then must recall the precise feature value (e.g., color) associated with each item in a later test. Performance on these tasks is nearly as robust as when maintaining an item in WM, suggesting the possibility that participants may transfer spatial positions to LTM during our long delay intervals and recall feature values when given a valid cue.

However, we argue that there are several reasons this is not likely the case. First, these studies use unique stimulus-feature value pairs on each trial such that each stimulus can be used as a "tag" for retrieving a given feature value associated with that stimulus. In our task, all stimuli are essentially identical (red and blue dots), and so there are no distinguishing features from trial to trial except the relevant feature values on that trial (their spatial positions). As such, representations from each trial would likely become intermingled and difficult to disentangle by a long-term memory system. Second, in previous work recall from LTM was poorer than WM for a single item (Brady et al., 2013), suggesting that this information would not likely augment performance in a substantial way. However, it remains possible that information encoded into LTM is divergent or complementary to that actively

maintained within WM, in which case the restoration of information in WM could be possible (that is, “noise” in each representation could be “averaged out”). Further study is necessary to evaluate this possibility.

4.4: Conclusions

By reconstructing and quantifying representations of each item held in visual spatial WM over an extended delay interval, we show that post-cuing an item accessible only in WM can enhance the fidelity of its item-specific representation. In many areas, this post-cue leads to the recovery of representational fidelity similar to that observed in trials during which only a single item was maintained. Furthermore, behavioral performance is related to the amplitude of the cued representation, demonstrating that our observed representations support behavioral performance. Information theoretic constraints preclude spike-based models from accounting for these post-cue effects because spike-based models predict that a loss of spiking integrity should be irreversible. Thus, these data suggest the maintenance of additional information about the cued item in a latent, high-fidelity state that can restore degraded active representations in response to changing behavioral demands. Finally, representations of information in neural activity patterns may more broadly rely on such sub-threshold components that are not typically assayed in neuroimaging or electrophysiological experiments.

4.5: Acknowledgments

Chapter 4, in full, is a manuscript entitled “Restoring latent visual working memory representations in human cortex” submitted for publication. Sprague, Thomas C.; Ester, Edward F.; Serences, John T. The dissertation author was the primary investigator and author of the manuscript. We thank Edward Awh, Brad Postle, Sirawaj Itthipuripat, Stephanie Nelli, Samantha Scudder, and Vy Vo for helpful comments on a draft of this manuscript, and Haider Al-Hakeem for helpful discussions and assistance with data collection. Funded by NIH R01-MH092345 and a James S. McDonnell Foundation Scholar Award to J.T.S., NIH T32-MH20002 to T.C.S. and E.F.E., and an NSF Graduate Research Fellowship to T.C.S.

4.6: Experimental Procedures

Participants

Five participants (4 female; aged 23-29 yrs) naïve to the purpose of the experiment were recruited from the UC San Diego community. We used a small sample size, but acquired substantial data from each participant to maximize sensitivity to subtle WM representations, similar to our previous report (Sprague et al., 2014). Participant identifiers are identical to those used in previous reports to facilitate comparison of data across experiments (Ester et al., 2015; Sprague and Serences, 2013; Sprague et al., 2014). Participants AI and AL participated in the experiments reported in Sprague & Serences (2013). Participant AI participated in the experiments reported in Sprague et al (2014). Participants AI, AL and AP participated in Ester et al (2015). Participants gave written informed consent as approved by the UCSD Institutional Review Board and were compensated for their time ($20/hr for fMRI sessions, $10/hr for behavioral sessions).

Spatial WM retro-cueing task

All participants underwent 3 fMRI scanning sessions and 1 retinotopic mapping scanning session, each lasting 2 hrs. Participants also completed 2-4 behavioral sessions, each lasting 1-1.5 hrs. The size of the stimulus display was fixed across all behavioral and scanning sessions. However, the size of the screen, which constantly contained a gray background, differed (inside scanner: 18.18° × 13.64°, aspect ratio 4:3; outside scanner: 44.71° × 25.15°, aspect ratio 16:9).

We adapted a spatial WM task reported previously (Sprague et al., 2014). On each trial, we presented 2 target stimuli (a red and a blue dot, 0.15° diameter) for 500 ms at pseudorandom locations 3.5° from fixation on average. Following target presentation, the

fixation point (square, 0.2°/side) immediately changed color to either red, blue, or purple. A red or blue fixation cue (1/3 of trials) indicated the target to be maintained in WM over the delay interval (Remember 1). A purple fixation cue (2/3 of trials) indicated both targets should be maintained in WM (Remember 2). After an 8,000 ms delay interval (Delay 1), the fixation cue changed color once again. On Remember 1 trials, the cue always changed to black, indicating participants should maintain the encoded target in WM over the subsequent second delay interval. On ½ of Remember 2 trials (1/3 of trials overall), the fixation cue turned black, providing a neutral cue as to which target was relevant for behavior (Remember 2-neutral condition). On the remaining ½ of Remember 2 trials (1/3 of trials overall), the fixation cue changed from purple to either red or blue, cueing the participants with 100% validity to remember only one of the targets (Remember 2-valid condition). Following this cue change, participants continued to maintain 1 or 2 items in WM over an additional 8,000 ms delay interval (Delay 2).

At the end of each trial after both delay intervals, participants recalled the exact horizontal or vertical coordinate of the item cued by the color of the fixation point. The response coordinate was randomly chosen on every trial so that participants could not implement a uni-dimensional encoding scheme (i.e., encode only *x* or *y* coordinate). Participants responded by adjusting the position of a gray horizontal bar up or down (for *y* coordinate trials) or a vertical bar left or right (for *x* coordinate trials) using either a computer keyboard or an MR-compatible button box (bar thickness: 0.02°). We took the adjusted bar coordinate at the end of a 3,000 ms response window as the participant's response.

Target locations were drawn from an isoeccentric ring 3.5° from fixation at 60° polar angle intervals along the ring, where the starting angle was jittered by up to ±15° on each trial. The position of the second target relative to the first target was always offset from the first target by 60°, 120°, or 180° in either direction (clockwise or counterclockwise, see colored

discs in Fig. 1B). This resulted in a minimum target separation distance of 2.3° and a maximum separation distance of 8.2°. By using random target positions on each trial, we ensured that participants maintained precise spatial locations in WM rather than using alternative coding strategies, like verbally labeling the location(s). Additionally, constraining relative target positions within one of several discs allowed for comparison of data from trials with similar target arrangements (Figs. 3-4).

We counterbalanced trials for target position (1 of 6 discs), relative target position (1 of 3 relative angular separation distances, Fig. 1B), and memory condition (Remember 1, Remember 2-neutral, or Remember 2-valid), resulting in 54 trials per full counterbalanced repetition. Each full set of trials (or "super-run") was broken up into 3 runs, each with 18 trials, each 19.5 s long. Trials were separated by a random inter-trial interval chosen from a uniform distribution from 3 to 6 s.

Spatial mapping task

Inside the scanner, participants completed 4 runs per session of a spatial mapping task used to estimate voxel-level encoding models for reconstruction analyses described below. On each trial, participants remembered the exact position of a single target stimulus over a 3,000 ms delay interval during which a flickering checkerboard disc (6 Hz, full-field flicker, 1.083° radius, 1.474 cycles/°; Figure S3A) was presented nearby the memorized location. Following checkerboard presentation, participants indicated whether a probe stimulus (black dot) was either to the left or right or above or below the remembered stimulus position, as cued by an oriented bar at fixation (horizontal bar: respond left vs. right; vertical bar: respond above vs. below; probe and response bar presented for 750 ms). Participants could respond until the beginning of the next trial (after 2,000 – 4,500 ms inter-trial interval, uniform distribution). We maintained performance at ~75% correct by adjusting the target-probe separation distance

between runs, but due to a programming error, accuracy was computed incorrectly during scanning ("null" trials were counted as incorrect responses, so actual accuracy on task trials was ~89%, not ~75%, Figure S3 caption). To ensure participants did not just encode one target coordinate dimension (*x* or *y*), the irrelevant coordinate was jittered on each trial by a small amount, preventing a scenario in which the presentation of the probe stimulus added certainty to the position maintained in WM. Each run included 6 null trials (no target/mapping stimulus/probe presented) during which participants passively fixated until the subsequent trial began.

During each run of this spatial mapping task the checkerboard stimuli were presented at each of 36 positions arrayed along a hexagonal grid (see Figure S3B-C) and the target position was randomly chosen from a uniform disc centered at the checkerboard position with radius 0.542°. On each run, we rotated the angular orientation of the entire hexagonal grid by 15° polar angle (Figure S3C). Across sessions, we rotated the "baseline" angular orientation of the grid by 5° polar angle. This resulted in $4 \times 3 \times 36 = 432$ unique stimulus positions across all scanning sessions. We used different grid orientations (and thus stimulus positions) on each scanning run to maximize the number of unique stimulus positions so that we could estimate as robust a spatial encoding model as possible (see below), as well as to ensure our model was not identifying peculiarities specific to a given set of mapping stimulus positions.

Localizer task

To focus our neuroimaging analyses to voxels responsive during spatial WM maintenance over the area subtended by our display setup, we scanned each participant on 6-8 runs (AI: 6, AL: 7, AS: 8, AR: 7, AP: 7) of a visual spatial WM localizer task similar to one we have described before (Sprague et al., 2014). On each trial we presented a flickering radial checkerboard annulus in one visual hemifield extending from 0.8° to 6.0° from fixation (1.25

cycles/° from fixation, 12° per polar angle cycle, 6 Hz contrast-reversing) for 10 s. During the stimulus interval, we presented 2 spatial WM trials in which participants remembered the precise position of 1 red dot over a 3 s delay interval. At the end of each delay interval, participants responded whether a green probe stimulus was to the left or to the right, or above or below, the remembered target position as indicated by a horizontal or vertical bar at fixation, respectively. WM targets could only appear within the stimulated hemifield. We maintained performance at ~75% by adjusting the task difficulty (target/probe separation distance) across trials. Stimulus epochs were separated by 3 – 5 s ITIs (uniform distribution). Each run contained 4 null trials that were the same duration as normal trials but did not contain checkerboard stimuli.

Behavioral analysis

For the main WM task, we defined behavioral recall error as the absolute distance along the relevant coordinate dimension (x or y) between the position of the response bar at the conclusion of the response window and the actual coordinate of the recalled target. We averaged all recall errors across all trials from scanning sessions within each participant.

In fMRI analyses in which we split trials based on behavioral performance, we computed the median recall error within each WM condition (R1, R2-neutral, R2-valid) within each scanning session. Trials with recall error greater than or equal to the median value were labeled “high recall error” and trials with recall error less than the median value were labeled “low recall error” (Fig. 8 and Figure S5).

fMRI acquisition

We scanned all participants on a 3 T research-dedicated GE MR750 scanner located at the UCSD Keck Center for Functional Magnetic Resonance Imaging with a 32 channel

send/receive head coil (Nova Medical, Wilmington, MA) using identical sequences to those we have reported previously (Sprague and Serences, 2013; Sprague et al., 2014). We acquired functional data using a gradient echo planar imaging (EPI) pulse sequence (19.2 × 19.2 cm field of view, 96 × 96 matrix size, 31 3-mm-thick slices with 0-mm gap, obliquely-oriented through occipital, parietal & dorsal frontal cortex, TR = 2,250 ms, TE = 30 ms, flip angle = 90°, voxel size 2 × 2 × 3 mm, xyz).

To anatomically coregister images across sessions, and within each session, we also acquired a high resolution anatomical scan during each scanning session (FSPGR T1-weighted sequence, TR/TE = 11/3.3 ms, TI = 1,100 ms, 172 slices, flip angle = 18°, 1 mm^3 resolution). For all sessions but one, anatomical scans were acquired with ASSET acceleration. For the remaining session, we used an 8 channel send/receive head coil and no ASSET acceleration to acquire anatomical images with minimal signal inhomogeneity near the coil surface, which enabled improved segmentation of the gray-white matter boundary. We transformed these anatomical images to Talairach space and then reconstructed the gray/white matter surface boundary in BrainVoyager 2.6.1 (BrainInnovations) which we used for identifying ROIs.

fMRI preprocessing

We preprocessed fMRI data similarly to our previous report (Sprague et al., 2014). We coregistered functional images to a common anatomical scan across sessions (used to identify gray/white matter surface boundary as described above) by first aligning all functional images within a session to that session's anatomical scan, then aligning that session's scan to the common anatomical scan. We performed all preprocessing using FSL (Oxford, UK) and BrainVoyager 2.6.1 (BrainInnovations). Preprocessing included unwarping the EPI images using routines provided by FSL, then slice-time correction, three-dimensional motion correction (six-parameter affine transform), temporal high-pass filtering (to remove first-,

second- and third-order drift), transformation to Talairach space (resampling to 2 × 2 × 2 mm resolution) in BrainVoyager, and finally normalization of signal amplitudes by converting to Z-scores separately for each run using custom MATLAB scripts. We did not perform any spatial smoothing beyond the smoothing introduced by resampling during the co-registration of the functional images, motion correction and transformation to Talairach space. All subsequent analyses were computed using custom code written in MATLAB (release 2014b, The Mathworks, Inc).

One participant (AS) changed positions inside the scanner substantially during one session. As a result, the field inhomogeneities estimated with the field map scan used for unwarping were only accurate for half of the runs during this session and could not be used to unwarp the other half of scans. To mitigate this problem with the raw data, we did not perform unwarping on any session for this participant in order to maintain consistency in the analysis procedure across sessions for this participant. This did not appear to affect any aspect of their results.

Identifying regions of interest (ROIs)

We identified 10 ROIs using independent scanning runs from those used for all analyses reported in the text. For retinotopic ROIs (V1-V3, hV4, V3A, IPS0-IPS3), we utilized a combination of retinotopic mapping techniques. Each participant completed several scans of meridian mapping in which we alternately presented flickering checkerboard "bowties" along the horizontal and vertical meridians. Additionally, each participant completed several runs of an attention-demanding polar angle mapping task in which they detected brief contrast changes of a slowly-rotating checkerboard wedge (described in detail in Sprague and Serences, 2013). We used a combination of maps of visual field meridians and polar angle preference for each voxel to identify retinotopic ROIs (Engel et al., 1994; Swisher

et al., 2007). Polar angle maps computed using the attention-demanding mapping task for all participants are available in previous publications (AI: Sprague & Serences, 2013; AL and AP: Ester et al., 2015) or in Figure S6 (AR and AS). We combined left- and right-hemispheres for all ROIs, as well as dorsal and ventral aspects of V2 and V3 for all analyses by concatenating voxels.

We defined superior precentral sulcus (sPCS) by plotting voxels active during either the left or right conditions of the localizer task described above on the reconstructed gray/white matter boundary of each participant's brain and manually identifying clusters appearing near the superior portion of the precentral sulcus, following previous reports (Srimal and Curtis, 2008).

The "All ROIs combined" region reported throughout the text consists of all voxels from all 10 individual ROIs concatenated together, and so all multivariate analyses involving this ROI reflect the net information content of the entire set of regions studied (see also Sprague et al., 2014).

fMRI analysis: univariate

For all ROI analyses, we used data from the localizer scans to identify voxels significantly active during checkerboard stimulus presentation and WM maintenance (FDR corrected, $q = 0.05$) for inclusion in further analyses. All analyses include only those voxels.

We computed BOLD time series by extracting signal at each time point averaged over all voxels within an ROI on each trial from 0 to 24.75 s (0 to 11 TRs) after the beginning of the first delay (rounded to the nearest TR), then averaging time series over all trials. We extracted mean activation levels for each delay period by averaging the TRs 6.75-9.00 s after probe onset for Delay 1 and 15.75-18.00 s after probe onset for Delay 2.

fMRI analysis: inverted encoding model

To reconstruct images of spatial WM contents, we implemented an inverted encoding model (IEM) for spatial position. This analysis involves first estimating an encoding model (sensitivity profile over the relevant feature dimension(s) as parameterized by a small number of modeled information channels) for each voxel in a region using a "training set" of data reserved for this purpose. Then, the encoding models across all voxels within a region are inverted to estimate a mapping used to transform novel activation patterns from a "test set" into activation patterns in a modeled set of information channels.

We built an encoding model for spatial position based on a linear combination of spatial filters (Sprague and Serences, 2013; Sprague et al., 2015, 2014). Each voxel's response was modeled as a weighted sum of each of 37 identical spatial filters arrayed in a hexagonal grid (Fig. 2A). Centers were spaced by 2.293° and each filter was a Gaussian-like function with full-width half-maximum of 2.523°:

Equation 1: $$f(r) = \left(0.5 + 0.5\cos\frac{2\pi r}{s}\right)^7 \text{ for } r < s\text{; 0 otherwise}$$

Where r is the distance from the filter center and s is a "size constant" reflecting the distance from the center of each spatial filter at which the filter returns to 0. Values greater than this are set to 0, resulting in a single smooth round filter at each position along the hexagonal grid (s = 6.349°; see Fig. 2A, Figure S3E for illustration of filter layout and shape; see also Sprague and Serences, 2013; Sprague et al., 2014). Each filter's sensitivity ranges from 0 to 1.

This hexagonal grid of filters forms the set of information channels for our analysis. Each mapping task stimulus is converted from a contrast mask (1's for each pixel subtended by the stimulus, 0's elsewhere) to a set of filter activation levels by taking the dot product of the vectorized stimulus mask and the sensitivity profile of each filter. This results in each mapping stimulus being described by 37 filter activation levels rather than 1,024 × 768 =

786,432 pixel values. Once all filter activation levels are estimated, we normalize so that the maximum filter activation is 1.

We model the response in each voxel as a weighted sum of filter responses (which can loosely be considered as hypothetical discrete neural populations, each with spatial RFs centered at the corresponding filter position).

Equation 2: $$B_1 = C_1 W$$

Where B_1 (n trials × m voxels) is the observed BOLD activation level of each voxel during the spatial mapping task (averaged over 6.75 – 9.00 s after WM target onset; Figure S3A), C_1 (n trials × k channels) is the modeled response of each spatial filter, or information channel, on each trial of the mapping task (normalized from 0 to 1), and W is a weight matrix (k channels × m voxels) quantifying the contribution of each information channel to each voxel. Because we have more stimulus positions than modeled information channels, we can solve for W using ordinary least-squares linear regression:

Equation 3: $$\widehat{W} = (C_1^T C_1)^{-1} C_1^T B_1$$

This step is univariate and can be computed for each voxel in the brain independently. Next, we used all estimated voxel encoding models within a ROI ($\widehat{W}$) and a novel pattern of activation from the WM task (each TR from each trial, in turn) to compute an estimate of the activation of each channel ($\hat{C}_2$, n trials × k channels) which gave rise to that observed activation pattern across all voxels within that ROI (B_2, n trials × m voxels):

Equation 4: $$\hat{C}_2 = B_2 \widehat{W}^T \left(\widehat{W}\widehat{W}^T\right)^{-1}$$

The Moore-Penrose pseudoinverse of the estimated weight matrix from the training set ($\widehat{W}$) is the *inverted* part of the IEM: all encoding models across all voxels are used, and this step is multivariate. This analysis is only feasible when more voxels are measured than information channels are modeled. The Moore-Penrose pseudoinverse acts as a linear mapping

from data measured in voxel space (B_2) into channel space ($\hat{C}_2$), and accordingly stretches, scales and skews voxel activation patterns during this transformation, but importantly does not result in any nonlinear transformations. This analysis can be considered a directed form of dimensionality reduction in which activation patterns are transformed from an idiosyncratic activation pattern across voxels (unique to each individual participant and ROI, and thus difficult to directly compare) to a common information space, common across ROIs and participants, which allows for direct manipulation, quantification, and comparison of activation patterns in an intuitive and stimulus-referred coordinate space.

Once channel activation patterns are computed (Equation 4), we compute spatial reconstructions by weighting each filter's spatial profile by the corresponding filter's reconstructed activation level and summing all weighted filters together. This step aids in visualization, quantification, and coregistration of trials across WM target positions, but does not confer additional information.

We analyzed all data within each session: the 4 mapping task runs for a given session were used to estimate the encoding model for each voxel, then that encoding model was inverted and used to reconstruct WM representations during all main WM task runs within that same session. Then, we averaged reconstructions over sessions within each participant.

Because WM target positions were unique on each and every trial, direct comparison of WM reconstructions on each trial is not possible without coregistration of reconstructions so that WM targets appeared at a common position across trials. To accomplish this, we adjusted the center position of the spatial filters on each trial such that we could rotate (and sometimes translate) the resulting reconstruction. For Figures 3-4, we rotated each trial such that one target (the target not queried at the end of each trial) was on average centered at x = 3.5° and y = 0° and the other target was in the upper visual hemifield (which required flipping ½ of reconstructions across the horizontal meridian). For Figures 7 and 8 and Figure S5, we

coregistered each trial so that the queried target position was always centered at exactly x = 3.5° and y = 0° by first rotating the reconstruction so that the target was aligned along the positive x Cartesian axis, then horizontally translating it so that its x coordinate was exactly 3.5° (Figure S3D).

Because we carefully designed our task such that we presented an equal number of trials for each target separation condition (+60°, +120°, +180°, -60°, -120°, and -180° polar angle) in order to minimize the potential for participants to discover geometric regularities in the target arrangements, there was an overabundance of trials at ±180° polar angle separation distance, which led to a non-uniform distribution of positions for the non-coregistered target (that is, there were double the number of trials with non-coregistered targets at 180° polar angle from the coregistered target as there were for ±60° and ±120°). As a result, we excluded the second half of 180° separation condition trials from each super-run from all reconstruction-based analyses. When the other half of these trials is included, there is often a noticeable "bump" along the negative x axis corresponding to the greater number of trials in which a non-coregistered target appeared near that position, which renders quantification of target representations via curvefitting methods (see below) suboptimal.

Quantifying WM representations

We took three approaches to WM representation quantification. First, we evaluated the "representational fidelity" for each reconstruction by determining the extent to which its target representation was reliably present. To accomplish this, we first reduced the reconstruction from a 2-d image to a 1-d line plot by averaging over each of 220 evenly-spaced polar angle arms subtending 2.9-4.1° eccentricity (subset illustrated in Fig. 2C). The resulting 1-dimensional reconstruction reflects the average profile along an annulus around fixation. A target representation in these reconstructions would be a "bump" near 0° after the

reconstructions have been rotated to a common center (where 0° corresponds to the actual target polar angle). To reduce these 1-d reconstructions to a single number which could be used to quantify the presence of target representations (F), we computed a vector mean of the 1-d reconstruction ($r(\theta)$, where θ is the polar angle of each point and $r(\theta)$ is the reconstruction activation) when plotted as a polar plot, as projected along the *x* axis (because the reconstructions were rotated such that the target was presented at 0°; Fig. 2C):

Equation 5: $$F = mean(r(\theta)\cos\theta)$$

If *F* is reliably greater than zero, over a resampling procedure (see Statistical Procedures), this quantitatively demonstrates that the net activation over the entire reconstruction carries information above chance about the target position. This measure is independent of baseline activation level in the reconstruction, as the mean of $r(\theta)$ is removed by averaging over the full circle. We computed timecourses of representational fidelity (Fig. 5B), as well as representational fidelity for each delay period (Fig. 6). To determine whether the cue on Remember 2-valid trials restores representations, we compared *F* between Delay 2 and Delay 1 for each ROI.

Additionally, we sought to evaluate the size, amplitude, and baseline of the WM target representation(s) from each WM condition and WM delay interval to establish how the information content of the population code changed across conditions. We followed procedures developed previously (Sprague et al., 2014) whereby we resampled with replacement all trials concatenated across all sessions from all participants from a condition 1,000 times and computed a single mean coregistered reconstruction (Figs. 7-8, Figure S5) on each resampling iteration. Then, we fit the mean reconstruction with a round Gaussian-like function parameterized by its center position, size, amplitude, and baseline:

Equation 6: $$f(r) = b + a\left(0.5 + 0.5\cos\frac{2\pi r}{s}\right)^7 \text{ for } r < s;\ 0$$ otherwise

Where r is the distance from the center of the surface, s is the size constant (as in Eq. 1), and a and b are the amplitude and baseline, respectively. Because there are many free parameters and some reconstructions are noisy, we adopted several heuristics to constrain our optimization problem. First, we found the maximum point on the entire reconstruction and used this as the center position (Sprague et al., 2014). Then, we performed a search through different sizes of fit surface function (FWHM: 0.099° to 9.934° in 0.099° steps). At each search iteration, we used ordinary least squares linear regression to find the amplitude and baseline which minimized residual errors between the reconstruction and the fit function. Finally, we used the best-fit amplitude, baseline, and size parameters from this search procedure and the global maximum position on the reconstruction as seed values for a constrained nonlinear optimization fitting algorithm (Matlab's fmincon function) subject to several constraints: position could not deviate more than one reconstruction "pixel size" (0.235° × 0.235°) from the global maximum position; size could not surpass the range used in the grid search procedure (0.099° to 9.934°), and amplitude/baseline could each not go below -5 or above 10 (BOLD Z-score units). This entire curvefitting procedure was repeated on each resampling iteration, for each condition described in the text (R1, R2-neutral, R2-valid broken down by Delay 1 and Delay 2 for Fig. 7, each of those broken down by High and Low recall error for Fig. 8 and Figure S5), resulting in 1,000 resampled estimates of each fit parameter on each condition for each ROI. Average resampled reconstructions over all resampling iterations are shown in Figure 7A,E, Figure 8A,C and Figure S5.

As a third means of quantifying the integrity of WM representations, we evaluated the relative strengths of each target representation at each time point of the trial by extracting the average reconstruction activation within a 0.5° radius circle centered at each target position.

Then, we took the difference between the reconstructed target representation activation of the target probed at the end of each trial and that of the target which was not probed at the end of each trial (on R1 trials, the probed target was always the remembered target; on R2-neutral trials, the probed target was the target queried at the end of the trial; on R2-valid trials, the probed target was always the remaining target following the valid retro-cue; Figure S4). This allowed us to directly compare the strength of the representation through time for each target in a manner which did not require fitting a surface with many free parameters.

Statistical procedures

All statistical statements reported in the text are based on resampling procedures in which a variable of interest is computed over 1,000 iterations. In each iteration, all single-trial variables from a given condition are resampled with replacement and averaged, resulting in 1,000 resampled averages for a given condition. We then subjected these distributions of resampled averages to pairwise comparisons by computing the distribution of differences between one resampled distribution (e.g., R1) and another resampled distribution (e.g., R2), yielding a new distribution of 1,000 difference values. We tested whether these difference distributions significantly differed from 0 in either direction by performing two one-tailed tests (p = proportion of values greater than or less than 0; null hypothesis that difference between conditions = 0) and doubling the smaller p value. For tests in which we compared whether representations were present in 1-d reconstructions using the representational fidelity measure, we performed one-tailed tests (null hypothesis that $F \leq 0$). Because we performed 1,000 iterations of this analysis, we cannot identify p values less than 0.001, so all comparisons in which resampled difference distributions were all greater than or less than 0 are reported as $p < 0.001$. Because we performed many pairwise comparisons, we corrected all repeated tests within an analysis using the false discovery rate (Benjamini and Yekutieli,

2001) and a threshold of $q = 0.05$ (except for tests of behavioral performance, which were corrected using Bonferroni's method due to the small number of comparisons performed). All p-values for all tests are reported in Supplementary Tables. All error bars reflect 95% confidence intervals as estimated using this resampling procedure.

4.7: References

Albers, A.M., Kok, P., Toni, I., Dijkerman, H.C., de Lange, F.P., 2013. Shared representations for working memory and mental imagery in early visual cortex. Curr. Biol. 23, 1427–31. doi:10.1016/j.cub.2013.05.065

Baddeley, A.D., Hitch, G., 1974. Working memory. Psychol. Learn. Motiv. 8, 47–89. doi:10.1016/S0079-7421(08)60452-1

Barak, O., Tsodyks, M., 2014. Working models of working memory. Curr. Opin. Neurobiol. 25, 20–4. doi:10.1016/j.conb.2013.10.008

Bays, P.M., 2015. Spikes not slots: noise in neural populations limits working memory. Trends Cogn. Sci. 19, 431–8. doi:10.1016/j.tics.2015.06.004

Bays, P.M., 2014. Noise in neural populations accounts for errors in working memory. J. Neurosci. 34, 3632–45. doi:10.1523/JNEUROSCI.3204-13.2014

Bays, P.M., Husain, M., 2008. Dynamic shifts of limited working memory resources in human vision. Science (80-.). 321, 851–4. doi:10.1126/science.1158023

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29, 1165–1188.

Brady, T.F., Konkle, T., Gill, J., Oliva, A., Alvarez, G.A., 2013. Visual long-term memory has the same limit on fidelity as visual working memory. Psychol. Sci. 24, 981–90. doi:10.1177/0956797612465439

Briggs, F., Mangun, G.R., Usrey, W.M., 2013. Attention enhances synaptic efficacy and the signal-to-noise ratio in neural circuits. Nature 499, 476–80. doi:10.1038/nature12276

Brouwer, G., Heeger, D., 2009. Decoding and Reconstructing Color from Responses in Human Visual Cortex. J. Neurosci. 29, 13992–14003.

Buschman, T.J., Siegel, M., Roy, J.E., Miller, E.K., 2011. Neural substrates of cognitive capacity limitations. Proc. Natl. Acad. Sci. U. S. A. 108, 11252–5. doi:10.1073/pnas.1104666108

Carandini, M., Heeger, D.J., 2012. Normalization as a canonical neural computation. Nat. Rev. Neurosci. 13, 51–62. doi:10.1038/nrn3136

Cover, T., Thomas, J., 1991. Elements of information theory. Wiley, New York.

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

D'Esposito, M., 2007. From cognitive to neural models of working memory. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 362, 761–72. doi:10.1098/rstb.2007.2086

D'Esposito, M., Postle, B.R., 2014. The Cognitive Neuroscience of Working Memory. Annu. Rev. Psychol. 66, 115–42. doi:10.1146/annurev-psych-010814-015031

Drew, T., Horowitz, T.S., Wolfe, J.M., Vogel, E.K., 2012. Neural measures of dynamic changes in attentive tracking load. J. Cogn. Neurosci. 24, 440–50. doi:10.1162/jocn_a_00107

Drew, T., Horowitz, T.S., Wolfe, J.M., Vogel, E.K., 2011. Delineating the neural signatures of tracking spatial position and working memory during attentive tracking. J. Neurosci. 31, 659–68. doi:10.1523/JNEUROSCI.1339-10.2011

Emrich, S.M., Riggall, A.C., Larocque, J.J., Postle, B.R., 2013. Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. J. Neurosci. 33, 6516–23. doi:10.1523/JNEUROSCI.5732-12.2013

Engel, S.A., Rumelhart, D.E., Wandell, B.A., Lee, A.T., Glover, G.H., Chichilnisky, E.-J., Shadlen, M.N., 1994. fMRI of human visual cortex. Nature 369, 525.

Ester, E.F., Anderson, D.E., Serences, J.T., Awh, E., 2013. A Neural Measure of Precision in Visual Working Memory. J. Cogn. Neurosci. 754–761. doi:10.1162/jocn_a_00357

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, 893–905. doi:10.1016/j.neuron.2015.07.013

Foster, J.J., Sutterer, D.W., Serences, J.T., Vogel, E.K., Awh, E., 2015. The topography of alpha-band activity tracks the content of spatial working memory. J. Neurophysiol. jn.00860.2015. doi:10.1152/jn.00860.2015

Franconeri, S.L., Alvarez, G.A., Cavanagh, P., 2013. Flexible cognitive resources: competitive content maps for attention and memory. Trends Cogn. Sci. 17, 134–41. doi:10.1016/j.tics.2013.01.010

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

Fuster, J.M., Alexander, G.E., 1971. Neuron activity related to short-term memory. Science 173, 652–4.

Gazzaley, A., Nobre, A.C., 2012. Top-down modulation: bridging selective attention and working memory. Trends Cogn. Sci. 16, 129–35. doi:10.1016/j.tics.2011.11.014

Griffin, I.C., Nobre, A.C., 2003. Orienting attention to locations in internal representations. J. Cogn. Neurosci. 15, 1176–94. doi:10.1162/089892903322598139

Harrison, S.A., Tong, F., 2009. Decoding reveals the contents of visual working memory in early visual areas. Nature 458, 632–635. doi:10.1038/nature07832

Herrmann, K., Montaser-Kouhsari, L., Carrasco, M., Heeger, D.J., 2010. When size matters: attention affects performance by contrast or response gain. Nat. Neurosci. 13, 1554–1559.

Itthipuripat, S., Garcia, J.O., Rungratsameetaweemana, N., Sprague, T.C., Serences, J.T., 2014. Changing the spatial scope of attention alters patterns of neural gain in human cortex. J. Neurosci. 34, 112–23. doi:10.1523/JNEUROSCI.3943-13.2014

Itthipuripat, S., Serences, J.T., 2015. Integrating Levels of Analysis in Systems and Cognitive Neurosciences: Selective Attention as a Case Study. Neuroscientist. doi:10.1177/1073858415603312

Keshvari, S., van den Berg, R., Ma, W.J., 2013. No evidence for an item limit in change detection. PLoS Comput. Biol. 9, e1002927. doi:10.1371/journal.pcbi.1002927

Kornblith, S., Buschman, T.J., Miller, E.K., 2015. Stimulus Load and Oscillatory Activity in Higher Cortex. Cereb. Cortex. doi:10.1093/cercor/bhv182

Landman, R., Spekreijse, H., Lamme, V.A.F., 2003a. Set size effects in the macaque striate cortex. J. Cogn. Neurosci. 15, 873–82. doi:10.1162/089892903322370799

Landman, R., Spekreijse, H., Lamme, V.A.F., 2003b. Large capacity storage of integrated objects before change blindness. Vision Res. 43, 149–64.

Lara, A.H., Wallis, J.D., 2014. Executive control processes underlying multi-item working memory. Nat. Neurosci. 17, 876–83. doi:10.1038/nn.3702

Lara, A.H., Wallis, J.D., 2012. Capacity and precision in an animal model of visual short-term memory. J. Vis. 12. doi:10.1167/12.3.13

LaRocque, J., Lewis-Peacock, J., Drysdale, A., Oberauer, K., Postle, B.R., 2013. Decoding attended information in short-term memory: An eeg study. J. Cogn. Neurosci. 25, 127–142.

LaRocque, J.J., Eichenbaum, A.S., Starrett, M.J., Rose, N.S., Emrich, S.M., Postle, B.R., 2015. The short- and long-term fates of memory items retained outside the focus of attention. Mem. Cognit. 43, 453–68. doi:10.3758/s13421-014-0486-y

Lepsien, J., Nobre, A.C., 2007. Attentional modulation of object representations in working memory. Cereb. Cortex 17, 2072–83. doi:10.1093/cercor/bhl116

Lepsien, J., Thornton, I., Nobre, A.C., 2011. Modulation of working-memory maintenance by directed attention. Neuropsychologia 49, 1569–77. doi:10.1016/j.neuropsychologia.2011.03.011

Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., Postle, B.R., 2012. Neural evidence for a distinction between short-term memory and the focus of attention. J. Cogn. Neurosci. 24, 61–79. doi:10.1162/jocn_a_00140

Lewis-Peacock, J.A., Postle, B.R., 2012. Decoding the internal focus of attention. Neuropsychologia 50, 470–8. doi:10.1016/j.neuropsychologia.2011.11.006

Luck, S.J., Vogel, E.K., 2013. Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends Cogn. Sci. 17, 391–400. doi:10.1016/j.tics.2013.06.006

Ma, W.J., Beck, J.M., Latham, P.E., Pouget, A., 2006. Bayesian inference with probabilistic population codes. Nat. Neurosci. 9, 1432–8. doi:10.1038/nn1790

Ma, W.J., Husain, M., Bays, P.M., 2014. Changing concepts of working memory. Nat. Neurosci. 17, 347–56. doi:10.1038/nn.3655

Makovsik, T., Jiang, Y. V, 2007. Distributing versus focusing attention in visual short-term memory. Psychon. Bull. Rev. 14, 1072–8.

Matsukura, M., Luck, S.J., Vecera, S.P., 2007. Attention effects during visual short-term memory maintenance: protection or prioritization? Percept. Psychophys. 69, 1422–34.

Matsushima, A., Tanaka, M., 2014. Different neuronal computations of spatial working memory for multiple locations within versus across visual hemifields. J. Neurosci. 34, 5621–6. doi:10.1523/JNEUROSCI.0295-14.2014

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, 1255–62. doi:10.1038/nn.3785

Milner, B., Squire, L.R., Kandel, E.R., 1998. Cognitive neuroscience and the study of memory. Neuron 20, 445–68.

Mongillo, G., Barak, O., Tsodyks, M., 2008. Synaptic theory of working memory. Science 319, 1543–6. doi:10.1126/science.1150769

Nobre, A.C., Coull, J.T., Maquet, P., Frith, C.D., Vandenberghe, R., Mesulam, M.M., 2004. Orienting attention to locations in perceptual versus mental representations. J. Cogn. Neurosci. 16, 363–73. doi:10.1162/089892904322926700

Postle, B.R., 2015. The cognitive neuroscience of visual short-term memory. Curr. Opin. Behav. Sci. 1, 40–46. doi:10.1016/j.cobeha.2014.08.004

Quian Quiroga, R., Panzeri, S., 2009. Extracting information from neuronal populations: information theory and decoding approaches. Nat. Rev. Neurosci. 10, 173–85. doi:10.1038/nrn2578

Reinhart, R.M.G., Heitz, R.P., Purcell, B.A., Weigand, P.K., Schall, J.D., Woodman, G.F., 2012. Homologous mechanisms of visuospatial working memory maintenance in

macaque and human: properties and sources. J. Neurosci. 32, 7711–22. doi:10.1523/JNEUROSCI.0215-12.2012

Reynolds, J.H., Heeger, D.J., 2009. The normalization model of attention. Neuron 61, 168–185.

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, 12990–8. doi:10.1523/JNEUROSCI.1892-12.2012

Saproo, S., Serences, J.T., 2014. Attention Improves Transfer of Motion Information between V1 and MT. J. Neurosci. 34, 3586–3596. doi:10.1523/JNEUROSCI.3484-13.2014

Saproo, S., Serences, J.T., 2010. Spatial Attention Improves the Quality of Population Codes in Human Visual Cortex . J. Neurophysiol. 104 , 885–895. doi:10.1152/jn.00369.2010

Serences, J.T., Ester, E.F., Vogel, E.K., Awh, E., 2009. Stimulus-Specific Delay Activity in Human Primary Visual Cortex. Psychol. Sci. 20 , 207–214. doi:10.1111/j.1467-9280.2009.02276.x

Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423. doi:10.1145/584091.584093

Sprague, T.C., Ester, E.F., Serences, J.T., 2014. Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load. Curr. Biol. doi:10.1016/j.cub.2014.07.066

Sprague, T.C., Saproo, S., Serences, J.T., 2015. Visual attention mitigates information loss in small- and large-scale neural codes. Trends Cogn. Sci. 19, 215–26. doi:10.1016/j.tics.2015.02.005

Sprague, T.C., Serences, J.T., 2013. Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nat. Neurosci. 16, 1879–87. doi:10.1038/nn.3574

Squire, L.R., Wixted, J.T., 2011. The cognitive neuroscience of human memory since H.M. Annu. Rev. Neurosci. 34, 259–88. doi:10.1146/annurev-neuro-061010-113720

Sreenivasan, K.K., Curtis, C.E., D'Esposito, M., 2014. Revisiting the role of persistent neural activity during working memory. Trends Cogn. Sci. 18, 82–9. doi:10.1016/j.tics.2013.12.001

Srimal, R., Curtis, C.E., 2008. Persistent neural activity during the maintenance of spatial position in working memory. Neuroimage 39, 455–468.

Stokes, M.G., 2015. "Activity-silent" working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn. Sci. 19, 394–405. doi:10.1016/j.tics.2015.05.004

Stokes, M.G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., Duncan, J., 2013. Dynamic coding for cognitive control in prefrontal cortex. Neuron 78, 364–75. doi:10.1016/j.neuron.2013.01.039

Sutterer, D.W., Awh, E., 2015. Retrieval practice enhances the accessibility but not the quality of memory. Psychon. Bull. Rev. doi:10.3758/s13423-015-0937-x

Swisher, J.D., Halko, M.A., Merabet, L.B., McMains, S.A., Somers, D.C., 2007. Visual topography of human intraparietal sulcus. J. Neurosci. 27, 5326–5337.

Tan, A.Y.Y., Chen, Y., Scholl, B., Seidemann, E., Priebe, N.J., 2014. Sensory stimulation shifts visual cortex from synchronous to asynchronous states. Nature 509, 226–9. doi:10.1038/nature13159

Tsubomi, H., Fukuda, K., Watanabe, K., Vogel, E.K., 2013. Neural limits to representing objects still within view. J. Neurosci. 33, 8257–63. doi:10.1523/JNEUROSCI.5348-12.2013

Wilken, P., Ma, W.J., 2004. A detection theory account of change detection. J. Vis. 4, 1120–35. doi:10.1167/4.12.11

Wolff, M.J., Ding, J., Myers, N.E., Stokes, M.G., 2015. Revealing hidden states in visual working memory using electroencephalography. Front. Syst. Neurosci. 9, 123. doi:10.3389/fnsys.2015.00123

Zhang, W., Luck, S.J., 2008. Discrete fixed-resolution representations in visual working memory. Nature 453, 233–235. doi:10.1038/nature06860

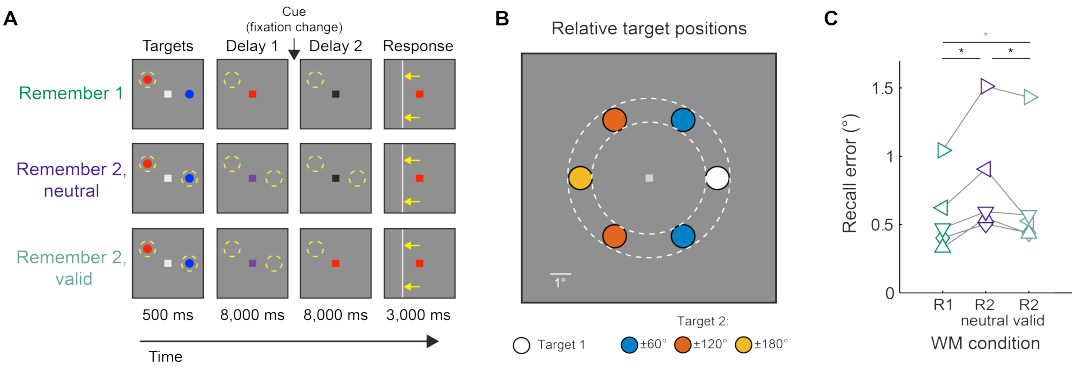


Figure 4-1: An informative cue enables behavioral performance to recover on a visual spatial WM task

We tested how cueing one of two items within spatial WM changed behavioral performance and neural WM representations. **(A)** On each trial, participants viewed 2 target stimuli (red and blue dots) for 500 ms and maintained the spatial position of either one or both targets as precisely as possible in spatial WM as cued by a change in fixation color during a WM delay interval (16 s total). On 33% of trials, participants were cued to maintain one of the two positions over the entire delay interval (fixation became red or blue, Remember 1; R1). On the remaining 67% of trials, the fixation point became purple, demanding participants maintain both locations (Remember 2; R2). This set of trials was further divided in half: on "neutral cue" trials, we gave no further information about which item was relevant (fixation point became black; R2-neutral condition); on "valid cue" trials the fixation point became red or blue after 8 s, reliably informing participants which target to recall at the end of the trial (R2-valid condition). Participants responded by adjusting the position of a response bar to match the position of the target cued by the fixation color as precisely as possible. Cartoon stimuli shown (not to scale, see Experimental Procedures). Dashed yellow circles indicate positions maintained in WM and did not appear on the display. Yellow arrows indicate movement of response bar and did not appear on the display. **(B)** The two targets appeared at positions uniformly drawn from two circular discs, each with a 0.6° radius centered 3.5° from fixation. Targets never appeared within the same disc; they appeared ±60° (blue), ±120° (orange), or ±180° (yellow) of polar angle apart on each trial. We randomly rotated the entire target arrangement on each trial so that across trials the targets could appear anywhere within the dashed annulus (see Experimental Procedures). **(C)** All participants (n = 5) performed more poorly on R2-neutral trials than R1 trials, as indicated by higher recall error ($p < 0.001$, resampling test, see Experimental Procedures), demonstrating a robust memory load effect on recall precision (Bays and Husain, 2008; Wilken and Ma, 2004; Zhang and Luck, 2008). However, when we cued one of two positions in WM (R2-valid), performance improved as compared to R2-neutral trials for all participants ($p = 0.016$, resampling test), indicating that participants could improve the fidelity of WM representations as indexed behaviorally. Performance differed only marginally between R1 and R2-valid conditions ($p = 0.024$). Black asterisks indicate significant difference as determined by pairwise resampling test, corrected for 3 comparisons using Bonferroni's method; gray asterisks indicate trends defined as $p \leq 0.05$, uncorrected. Each symbol in (C) is a single participant, and like symbols are used throughout all figures in which single-participant data is shown. See Figure S1 for recall error histograms for each condition and participant individually.

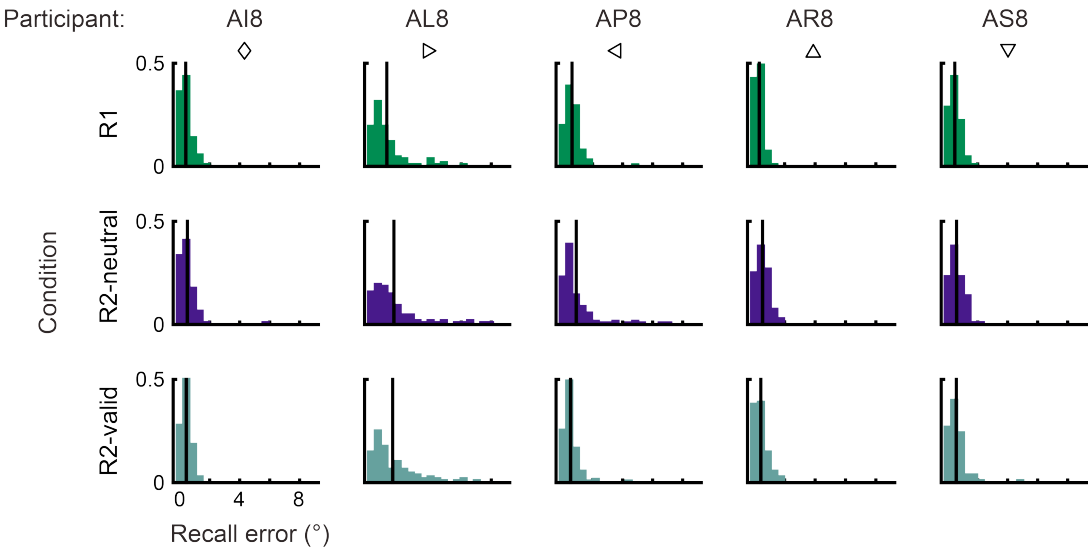


Figure 4-2: Recall performance recovers when one of two items is cued (histograms) Histograms of recall error across all trials for each participant and condition for data presented in Fig. 1C. Y axis indicates "proportion of trials". Same participant identifiers used as in previous reports to facilitate comparison of data across experiments (Ester et al., 2015; Sprague and Serences, 2013; Sprague et al., 2014).

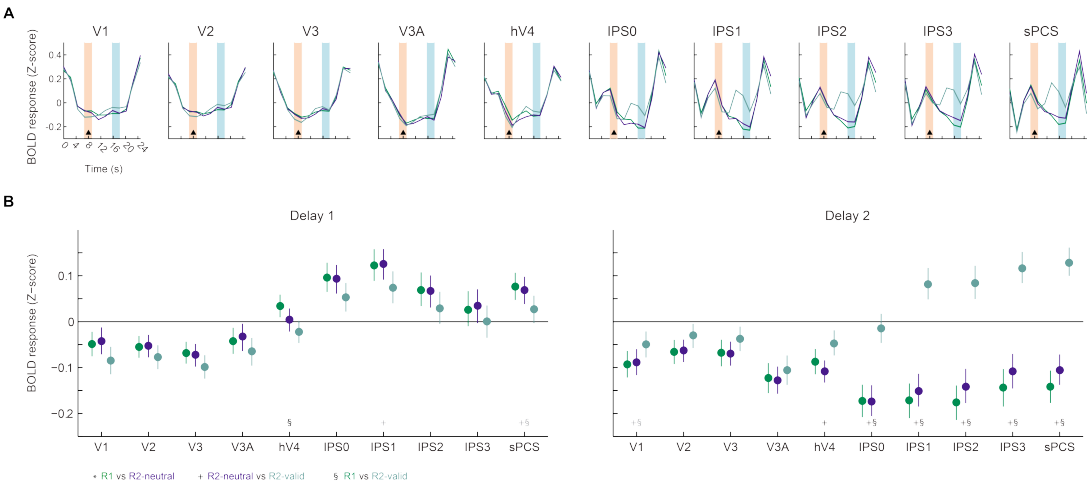


Figure 4-3: Univariate BOLD responses from all ROIs considered
(A) Mean BOLD activation timecourse (event-related average, time-locked to beginning of WM delay periods) averaged across all trials, all participants, and all voxels within each ROI. Replicating previous work (Emrich et al., 2013; Harrison and Tong, 2009; Riggall and Postle, 2012; Serences et al., 2009; Sprague et al., 2014), we observe no substantial activation in occipital ROIs (V1-hV4) in the univariate BOLD signal. For subsequent analyses, we identified time points primarily corresponding to the delay period before the cue (Delay 1, 6.75-9.00 s; orange box), and the delay period after the cue (Delay 2; 15.75-18.00 s; blue box). **(B)** Mean delay-period activation during Delay 1 (left) and Delay 2 (right) as a function of WM condition. During Delay 1, we found trends towards increased activation with set size (R2-neutral>R1 and/or R2-valid>R1) in sPCS. We also observed significantly higher activation during R2-valid trials in hV4 as compared to R1, but not R2-neutral, trials. During Delay 2, we observed significant cue-related activation (R2-valid>R1 and/or R2-valid>R2-neutral) in hV4, IPS0-IPS3, and sPCS, as well as trends towards this effect in V1. Significant tests reflect FDR-correction for all comparisons. Trends defined as $p \leq 0.05$, uncorrected for multiple comparisons. Error bars 95% confidence intervals via resampling all trials, with replacement, 1,000 times (see Experimental Procedures: statistical procedures). All *p*-values for this analysis presented in Table 4-1.

Figure 4-4: Inverted encoding model (IEM) for reconstructing and quantifying spatial WM representations

To evaluate whether fMRI-based measurements of spatial WM representations are modulated throughout the trial, we implemented an inverted encoding model (IEM) of visual space (Brouwer and Heeger, 2009; Ester et al., 2015; Sprague and Serences, 2013; Sprague et al., 2015, 2014). **(A)** To estimate voxel-level encoding models, we modeled the response of each voxel as a weighted sum of 37 information channels, each defined as a round smooth spatial filter, spanning a hexagonal spatial grid. We used measured activation levels across all trials to estimate the contribution of each information channel to each voxel using a standard general linear model (GLM). This procedure results in a set of 37 weights for each voxel, each describing the contribution of the associated modeled information channel to the observed signal in that voxel. **(B)** Inverting the encoding models across all voxels enables reconstruction. After estimating encoding models for all voxels within an ROI, we used the pattern of encoding models across all voxels in an ROI to compute an IEM. Once activations are represented in our modeled information space, we compute a sum of spatial filters weighted by their estimated activation, resulting in a reconstructed image of the visual field which must have been maintained in WM in order to observe the measured activation pattern, given the measured voxel-level encoding models from the mapping task in that region. The "bright" (yellow) region in the reconstruction (right) is spatially consistent with the position held in WM (left, dashed circle) on this example trial, and we call these areas of elevated activation in WM reconstructions "target representations". We reconstructed images at each time point in the trial (0 s to 24.75 s), and spatially coregistered all reconstructions across trials (see Experimental Procedures, Figure 4-5D) so that targets were centered at known positions, enabling us to average over trials in which different spatial positions were maintained in WM. **(C)** In order to assess whether a WM representation of a target was present in a reconstruction, we computed a "representational fidelity" metric by first extracting a 1-d reconstruction as a function of polar angle by computing the mean reconstruction activation from 2.9-4.1° from fixation (inside dashed black lines). Then, we used this 1-d polar angle reconstruction to compute a vector mean of a circular set of unit vectors, each weighted by its corresponding activation. We projected this vector mean onto a unit vector pointing in the polar angle direction of the WM target to generate a single-parameter metric of representational fidelity (subset of unit vectors shown as colored lines; vector mean shown as black arrow; polar angle reconstruction rotated so that 0° corresponds to target direction). On the polar plot, each radial ring corresponds to 0.2 units of BOLD Z-score. (**D**) We quantified several parameters of WM representations (amplitude, size, and spatially-nonspecific baseline offset) by fitting a 2-d surface to average coregistered reconstructions (Figure 4-5D) on each of 1,000 resampling iterations (Figs. 4-11, 4-12; Sprague et al., 2014). To assess statistical significance, we compared distributions of best-fit parameters between conditions (Fig. 4-11) or behavioral performance bins (Figure 4-12 and Figure 4-13). See Experimental Procedures for more details on identification and quantification of WM representations.

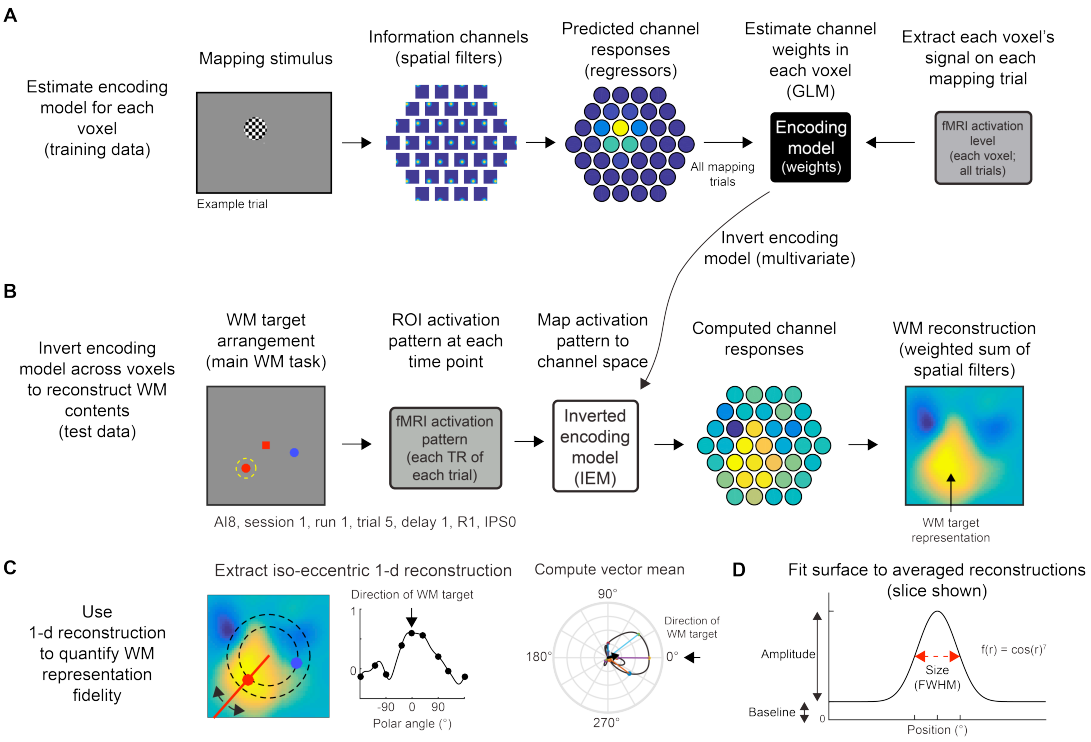
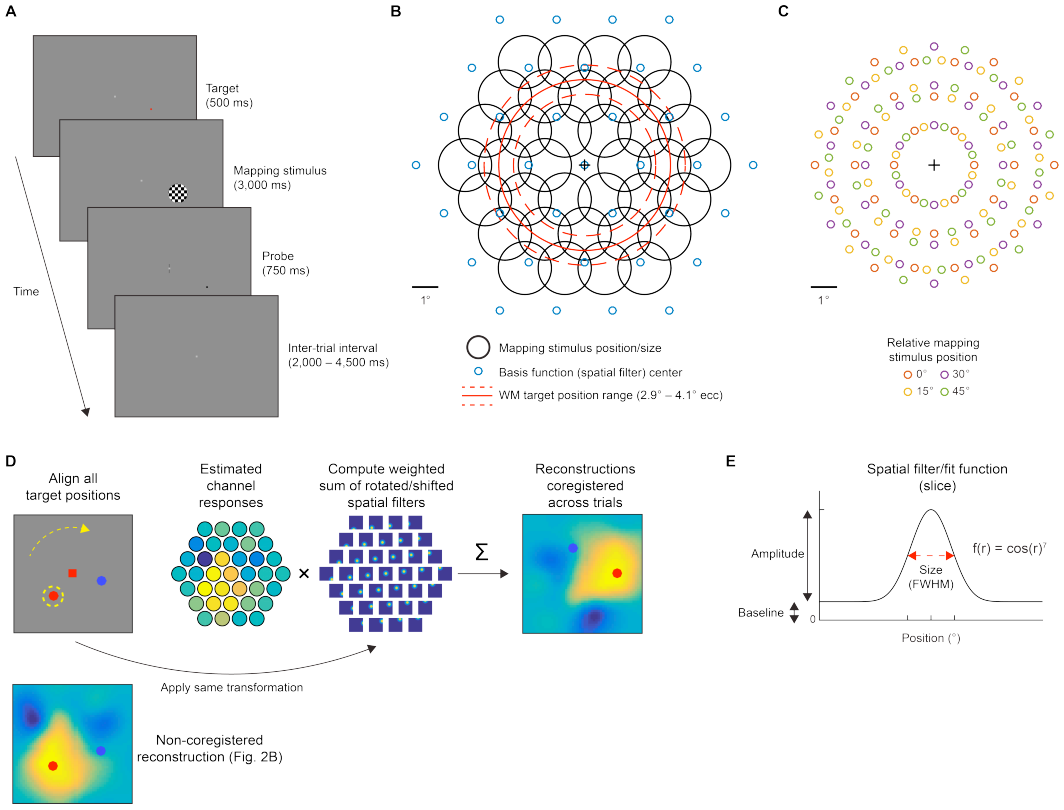

A
Estimate encoding model for each voxel (training data)
Mapping stimulus
Example trial
Information channels (spatial filters)
Predicted channel responses (regressors)
All mapping trials
Estimate channel weights in each voxel (GLM)
Encoding model (weights)
Extract each voxel's signal on each mapping trial
fMRI activation level (each voxel; all trials)
Invert encoding model (multivariate)
B
Invert encoding model across voxels to reconstruct WM contents (test data)
WM target arrangement (main WM task)
ROI activation pattern at each time point
fMRI activation pattern (each TR of each trial)
Map activation pattern to channel space
Inverted encoding model (IEM)
Computed channel responses
WM reconstruction (weighted sum of spatial filters)
WM target representation
AI8, session 1, run 1, trial 5, delay 1, R1, IPS0
C
Use 1-d reconstruction to quantify WM representation fidelity
Extract iso-eccentric 1-d reconstruction
Direction of WM target
Polar angle (°)
Compute vector mean
90°
180°
0°
270°
Direction of WM target
D
Fit surface to averaged reconstructions (slice shown)
Amplitude
Baseline
Size (FWHM)
f(r) = cos(r)'
Position (°)

Figure 4-5: IEM procedures: mapping task, stimulus layout, and reconstruction coregistration

(A) Participants performed 4 runs of a spatial mapping task during each fMRI scanning session. On each trial, we presented a single WM target stimulus (red dot) for 500 ms, followed immediately by a flickering checkerboard (1.083° radius; 6 Hz full-field flicker) overlapping the WM target location. After 3,000 ms, a probe stimulus (black dot) appeared slightly offset to either the left or right, or above or below, the remembered position (distance varied across runs to equate difficulty) for 750 ms. Simultaneously, a horizontal or vertical bar appeared at fixation, indicating the participant must make a 2AFC "left/right" or "above/below" judgment in response to the question "was the probe dot [left/above] or [right/below] [of] the remembered position?" before the end of the inter-trial interval (2-4.5 s). All stimulus features are drawn to scale. Participants performed on average 87.69% correct (target/probe separation distance adjusted across runs to maintain sufficient task difficulty). **(B)** The position of the mapping stimulus varied on each trial along a hexagonal grid (black circles), both inside and outside the range of eccentricities used for the main WM task (red ring). This enabled us to reconstruct images of the contents of spatial WM across the entire visual field subtended by the projector screen inside the scanner (Fig. 4-4), despite only remembering items from a small range of positions in the WM task (Fig. 4-1). Blue dots indicate the center of spatial filters used for image reconstruction (Fig. 4-4). **(C)** On each run of the spatial mapping task, we rotationally offset the position of the mapping stimuli by a fixed angular amount. Across sessions, we adjusted the "baseline" angle by 5° (session 1 arrangement shown here). **(D)** On each trial of the primary WM task (Fig. 4-1), the WM targets appeared pseudo-randomly within the red dashed ring in (B). To align data across trials in "information space", we rotated basis functions so as to zero-out the polar angle component of the WM target coordinate (1-d reconstructions & representational fidelity analyses; Figs. 4-8 and 4-9). Then, for analyses in which we precisely aligned target positions (Figs. 4-11, 4-12 and 4-13), we also shifted them horizontally to precisely align the target position to the coordinate x = 3.5°, y = 0° (see red dot, Fig 4-9A, D). For example, if a target appeared at 42° polar angle (up and to the right) and 3.7° eccentricity, we first rotated each basis function by 42° polar angle clockwise, then shifted all basis functions horizontally 0.2° to the left, before computing reconstructions. This means that we used a slightly different set of basis functions for computing each trial's reconstructions (same set of basis functions used for each time point of each trial), eliminating any potential idiosyncrasies caused by the exact set of filter centers we used. **(E)** Once we averaged coregistered reconstructions from all trials (on each resampling iteration, see Experimental Procedures: Statistical procedures), we fit a surface function (slice shown), which was shaped identically to each spatial filter, to the mean reconstruction. We allowed the function to vary in its size, baseline, and amplitude, and its position was constrained to be nearby the maximum pixel of the average reconstruction (see Experimental Procedures).


A
Target
(500 ms)
Mapping stimulus
(3,000 ms)
Probe
(750 ms)
Time
Inter-trial interval
(2,000 – 4,500 ms)
B
1°
Mapping stimulus position/size
Basis function (spatial filter) center
WM target position range (2.9° – 4.1° ecc)
C
1°
Relative mapping
stimulus position
0°
30°
15°
45°
D
Align all
target positions
Estimated
channel
responses
Compute weighted
sum of rotated/shifted
spatial filters
Reconstructions
coregistered
across trials
Apply same transformation
Non-coregistered
reconstruction (Fig. 2B)
E
Spatial filter/fit function
(slice)
Amplitude
Baseline
Size
(FWHM)
f(r) = cos(r)′
Position (°)

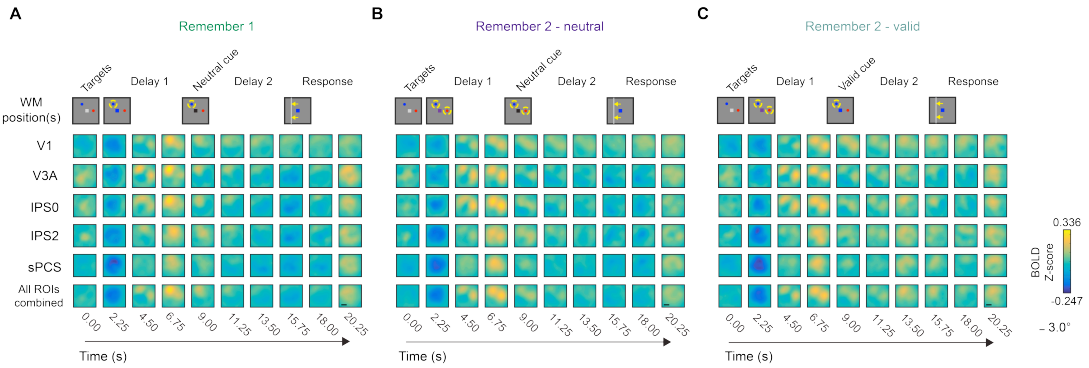


Figure 4-6: Delay-period image reconstructions reflect dynamic contents of WM
We reconstructed the contents of spatial WM at each point in time during the trial using activation patterns from several visual, parietal, and frontal ROIs defined using independent localizers (subset shown for brevity). Here we show reconstructions from an example target arrangement condition in which the WM targets were separated by an average of ~120° polar angle (top row). Trials in which the targets were presented at different positions are all rotated to match the cartoons and averaged over trials and participants (n = 5, 3 2-hr scanning sessions each). Cartoons are shown at approximate times of trial events; see Fig. 1A for exact timings. Yellow dashed circles in the stimulus cartoons indicate position(s) held in WM at each point in time. Each image portrays the reconstructed contents of spatial WM using activation patterns at the indicated timepoint (column) after the beginning of the trial from each ROI (row). Reconstructions have not been adjusted for hemodynamic delay, so reconstructions lag changes in contents of WM by ~6 s. All images represent a 12° × 12° square visual field aperture and are plotted on the same colorscale. IPS: intraparietal sulcus; sPCS: superior precentral sulcus. **(A)** On Remember 1 trials, stable WM representations emerge ~6-9 s following the first delay cue and remain throughout the entire 16 s delay interval, though appear less pronounced at later timepoints. **(B)** On Remember 2-neutral trials, stable WM representations are preserved over the entire 16 s delay interval, though are substantially weaker than those on R1 trials. **(C)** On Remember 2-valid trials, there is a transition from 2 representations during the first delay to a single representation during the second delay, tracking the contents of WM following the informative cue. Timepoint labels reflect time of each imaging volume relative to beginning of WM delay periods.

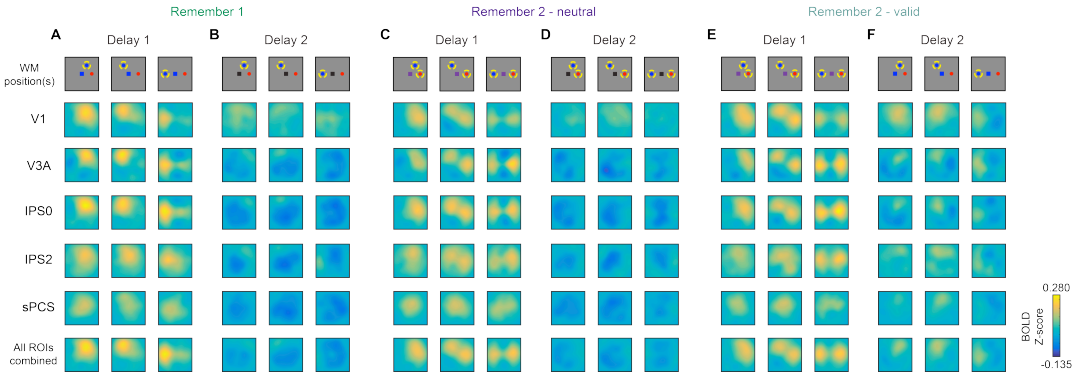


Figure 4-7: WM reconstructions track target positions
WM reconstructions computed and plotted as in Fig. 4-6, for each target arrangement condition (see cartoons, top row) and averaged over 2 timepoints during each delay period (Delay 1: 6.75 and 9.00 s; Delay 2: 15.75 and 18.00 s). **(A)** During Delay 1, Remember 1 trials do not show evidence for WM target representations of the non-remembered target (red dot in this cartoon), and thus are unlikely to be contaminated by sensory transients (see also Sprague et al., 2014). Remember 2-neutral **(C)** and Remember 2-valid **(E)** trials have the same contents of WM during Delay 1, and WM reconstructions look qualitatively very similar (compare (C) and (E)). During Delay 2, WM representations on Remember 1 **(B)** and Remember 2-neutral (**D**) trials appear weaker than those during Delay 1, but continue to track the relevant positions in WM. During the second delay on Remember 2-valid trials **(F)**, the non-cued position no longer appears visible, and the remaining position appears more strongly than during Delay 1 (C,E). All reconstructions plotted on same color scale to facilitate comparison between conditions.

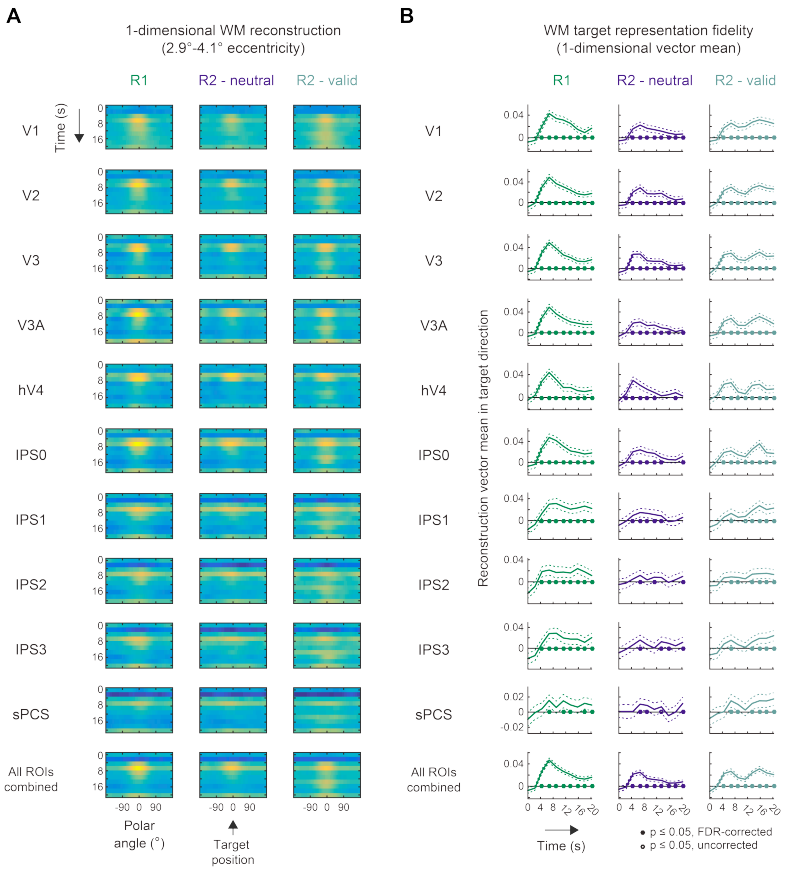


Figure 4-8: Target representations persist across entire delay interval
We computed reconstructions along radial vectors spanning the full circle and averaged reconstruction activation from 2.9-4.1° eccentricity, then rotated all reconstructions such that the probed target appeared at 0° (Fig. 4-4C; arrow in panel A). **(A)** Each plot shows the reconstructed target representations for a single ROI and WM condition throughout all time points of the trial, averaged over all participants. In some ROIs (e.g., V1, IPS0, and "All ROIs combined"), representations are qualitatively stronger during later time points than earlier time points on R2-valid trials, suggestive that the valid cue enhances WM representations. **(B)** To quantify whether WM representations were statistically present in each ROI during each delay interval of each WM condition, we computed a "representational fidelity" metric (see Experimental Procedures). If this score is reliably positive over a resampling procedure (corrected for multiple comparisons), we consider a WM representation to be present in the reconstruction. Plotted is WM representational fidelity computed for each time point. Although representational fidelity weakens later in the trial on Remember 1 and Remember 2-neutral trials, representations can still be quantitatively identified. On Remember 2-valid trials, representational fidelity increases following the informative cue (rather than remaining constant for the remainder of the trial), indicating that the cue enables the remaining representation to be bolstered. Filled symbols at $y = 0$ indicate significant WM representations, corrected for multiple comparisons via FDR ($q = 0.05$; across all ROIs, WM conditions and time points); open symbols indicate non-significant trends at $p \leq 0.05$; error bars mark 95% confidence intervals via resampling procedure. See also Figure 4-10 for an alternative means of evaluating the strength of WM representations and Figure 4-9 for quantitative comparison of representations across delay periods.

Figure 4-9: Valid cue recovers degraded WM representations
(A) 1-d polar angle reconstructions as in Fig. 4-8A, averaged over each delay period. Black asterisks indicate significant WM representations (FDR-corrected); gray asterisks indicate non-significant trends ($p \leq 0.05$; uncorrected; see Table 4-2 for all p-values from this analysis); error bars mark 95% confidence intervals via resampling procedure. Nearly all ROIs exhibit reliable WM representations during both delay periods, even though representations in some ROIs are difficult to visualize in reconstruction images (Fig. 4-7; e.g., R2-neutral, V3A, Delay 2). **(B)** To evaluate whether WM representations significantly change in fidelity across time, we directly compared delay-period representational fidelity for each ROI and condition (as in Fig. 4-7C). After a neutral cue (R1 and R2-neutral), the fidelity of representations substantially fades in many ROIs (as in Fig. 4-8B). In contrast, a valid cue significantly enhances WM representations in V1, V3, IPS0-IPS3 and All ROIs combined Asterisks indicate significant differences between delay periods, two-tailed, FDR-corrected for multiple comparisons (q = 0.05). Error bars mark 95% confidence intervals via resampling procedure. See Table 4-3 for all p-values from this analysis.

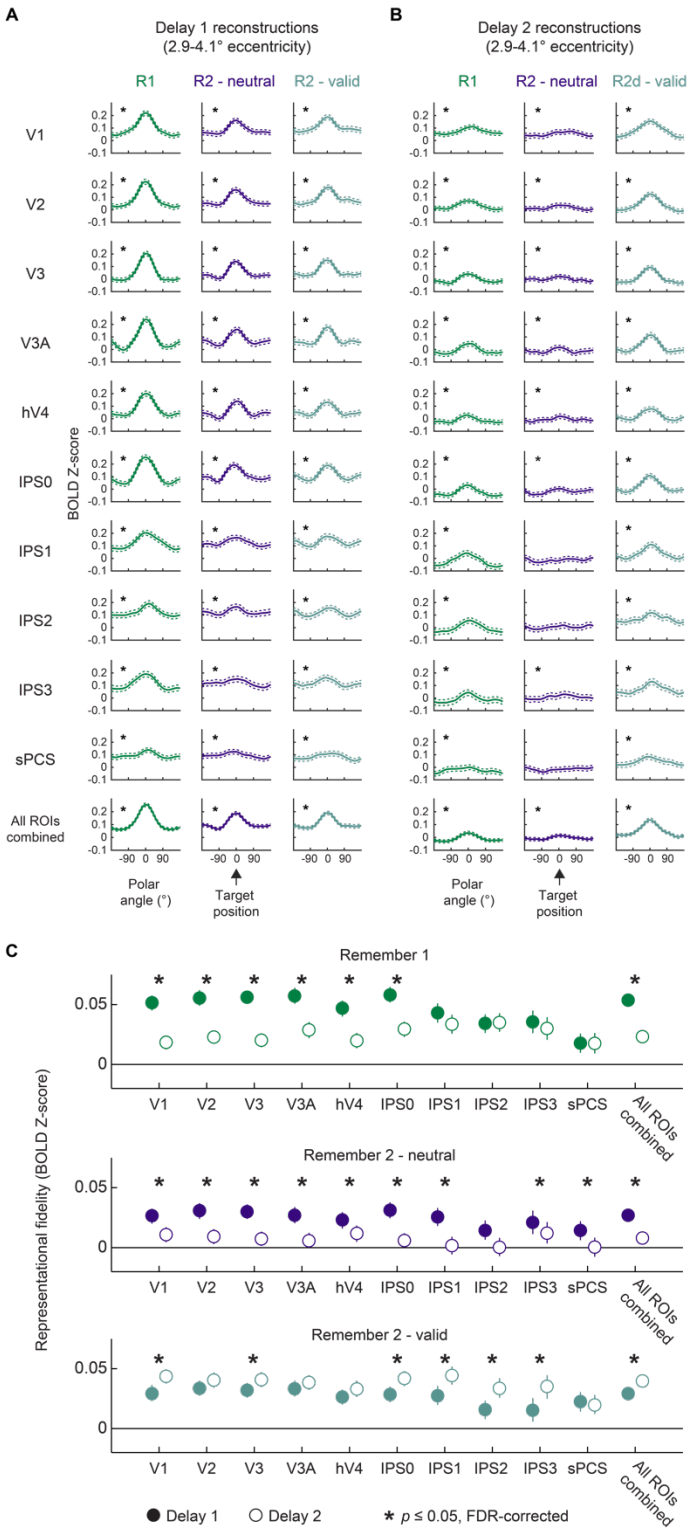
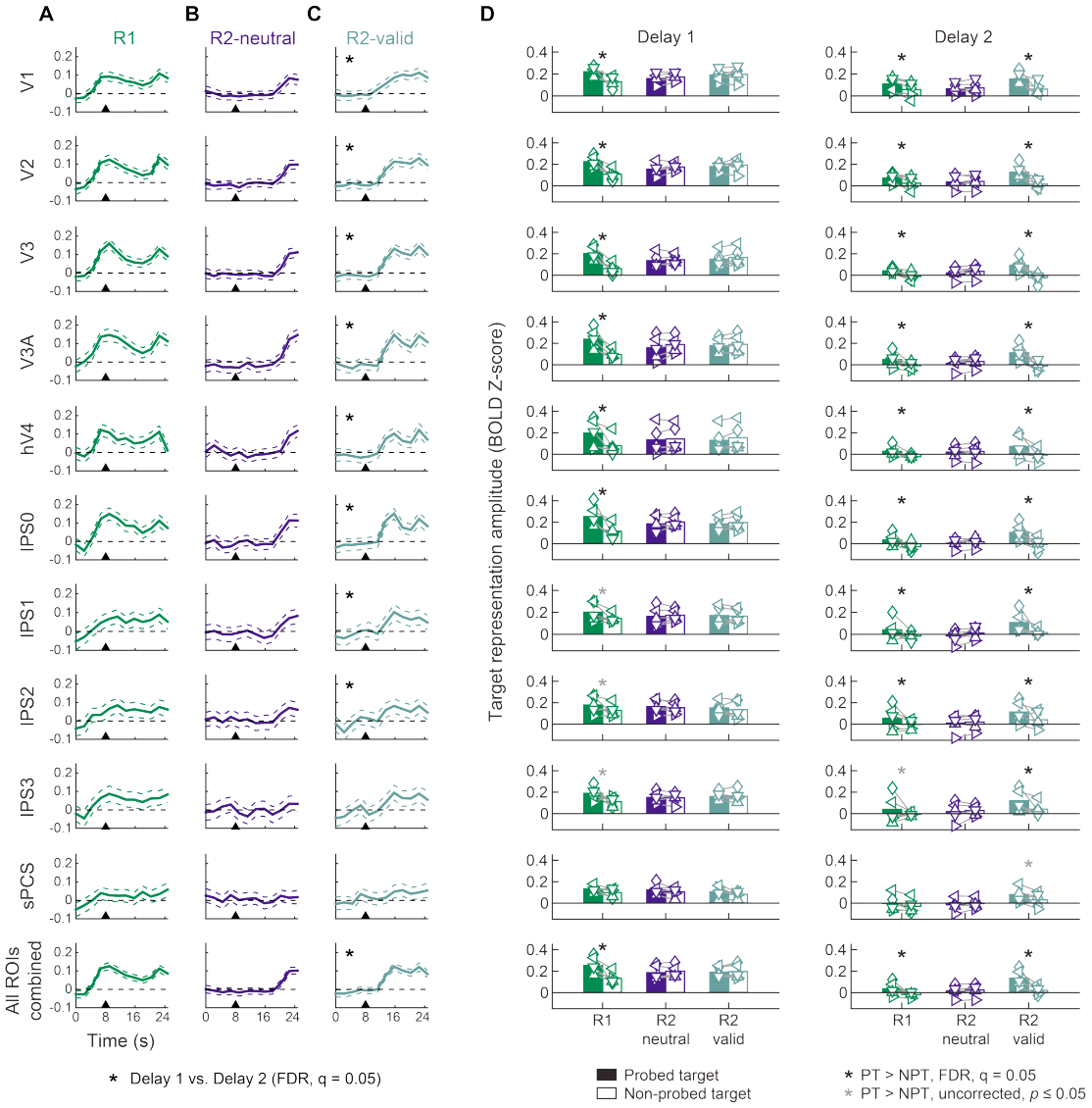

A
Delay 1 reconstructions
(2.9-4.1° eccentricity)
R1
R2 - neutral
R2 - valid
B
Delay 2 reconstructions
(2.9-4.1° eccentricity)
R1
R2 - neutral
R2d - valid
V1
V2
V3
V3A
hV4
IPS0
IPS1
IPS2
IPS3
sPCS
All ROIs combined
BOLD Z-score
0.2
0.1
0
-0.1
-90 0 90
Polar angle (°)
Target position
C
Remember 1
Remember 2 - neutral
Remember 2 - valid
Representational fidelity (BOLD Z-score)
0.05
0
V1 V2 V3 V3A hV4 IPS0 IPS1 IPS2 IPS3 sPCS All ROIs combined
Delay 1
Delay 2
* p ≤ 0.05, FDR-corrected

Figure 4-10: Informative cue shifts target representations from R2- to R1-like state
As an alternative visualization of the time course of WM target representations to those shown in Figures 4-6 and 4-8, we extracted the activation from each reconstruction within a 0.5° radius circular aperture centered at the exact target positions for each trial. We call this signal the "reconstruction activation", as it reflects BOLD activation patterns transformed into visual field coordinates and extracted at the relevant visual field position. Then, we computed the difference between the activation at the probed target location and the non-probed target location (on R1 trials, the probed target was always the target in WM, on R2-neutral trials, the probed target was the one queried at the end of the trial; on R2-valid trials, the probed target was the validly-cued target in WM). **(A)** On Remember 1 trials, the remembered target representation shows elevated activation relative to the non-remembered target representation throughout the entire 16 s delay interval, despite the weakening target representations as visualized in reconstructions in Figs 4-6 and 4-7. **(B)** On Remember 2-neutral trials, both target representations are equal throughout the delay period, with the queried target representation becoming stronger once the response period begins (16.0 s). (**C**) On Remember 2-valid trials, we see a transition from Remember 2-like target representations (both are equal, and so the difference is near zero) during the first delay period to Remember 1-like target representations (the remaining target representations recover) during the second delay period. Black triangle at 8.0 s indicates beginning of second delay interval. Units are BOLD Z-score. Dashed lines mark 95% CI via resampling, see Experimental Procedures: Statistical procedures. (**D**) We also computed mean delay-period reconstruction activation separately for probed (filled bars) and non-probed (open bars) target positions for each participant individually (each symbol reflects a single participant, as in Fig. 4-1C; Figure 4-1). Asterisks in panels A-C indicate a significant change between Delay 1 and Delay 2 (two-tailed); asterisks in panel d indicate that the probed target representation activation is greater than the non-probed target representation activation (one-tailed). All statistics computed using a resampling procedure (see Experimental Procedures: Statistical procedures). Black asterisks indicate a significant difference after FDR-correction for multiple comparisons ($q = 0.05$); gray asterisks indicate a non-significant trend defined using an uncorrected threshold of $\alpha = 0.05$). All *p*-values from this analysis available in Table 4-6.


A
B
C
D
R1
R2-neutral
R2-valid
V1
V2
V3
V3A
hV4
IPS0
IPS1
IPS2
IPS3
sPCS
All ROIs combined
Time (s)
Delay 1
Delay 2
Target representation amplitude (BOLD Z-score)
R1
R2 neutral
R2 valid
Probed target
Non-probed target
* Delay 1 vs. Delay 2 (FDR, q = 0.05)
* PT > NPT, FDR, q = 0.05
* PT > NPT, uncorrected, p ≤ 0.05

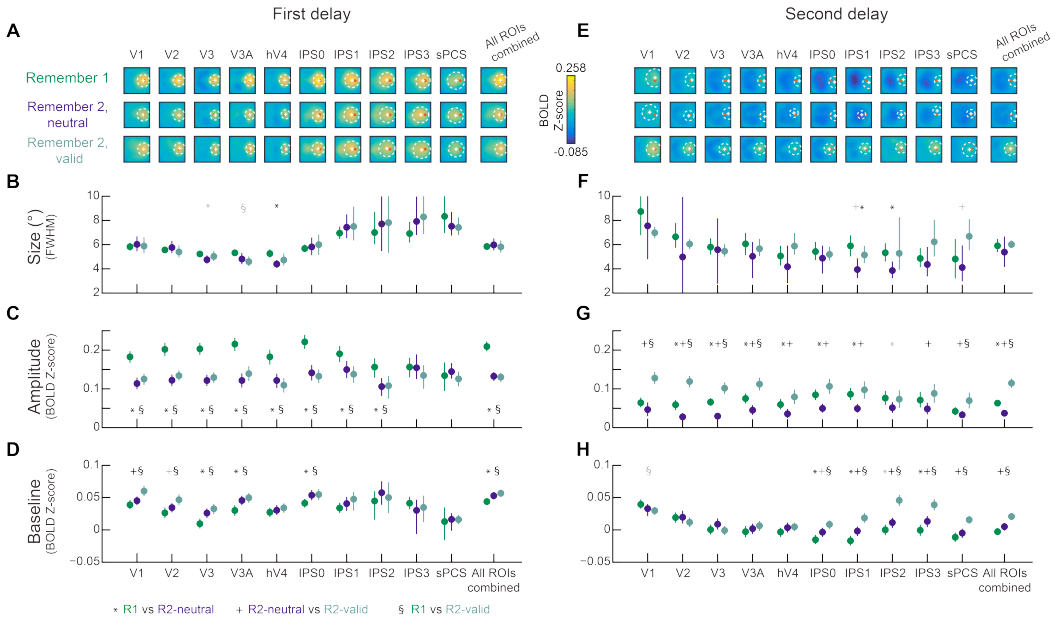


Figure 4-11: Target representations degrade with memory load and recover with valid retro-cue primarily through amplitude changes

To quantify WM target representations, we coregistered reconstructions from each trial so that all targets appeared at the same position (red circle in a; see Figure 4-5D). We resampled all trials within each condition, with replacement, 1,000 times, computed an average reconstruction from the resampled trials, and fit a surface allowed to vary in its size (full-width half-maximum; FWHM), amplitude, and baseline constrained to the position with maximum reconstruction activation for that resampling iteration (see Experimental Procedures; Fig. 4-4D). **(A)** Average reconstructions over all resampling iterations with best-fit surfaces. Mean best-fit position and size are plotted on each reconstruction as a white + surrounded by a dashed circle drawn at the surface FWHM (note that these fits reflect the mean of best-fit parameters over resampling iterations, not the fit to the average reconstructions shown in (A)). **(B)** Best-fit parameters from surface fitting for each condition. We computed pair-wise *p*-values between all condition pairs (R1 vs. R2-neutral, R2-neutral vs. R2-valid, R1 vs. R2-valid) within each ROI, delay-period, and parameter via resampling (see Experimental Procedures). Black symbols indicate significant pairwise differences after FDR correction for all comparisons within a fit parameter ($q = 0.05$). Gray symbols indicate trends, defined as $p \leq 0.05$, uncorrected for multiple comparisons. All correction for multiple comparisons considered all 10 individual ROIs because the "All ROIs combined" region is not independent of the others. Error bars indicate 95% confidence intervals obtained from the distribution of best-fit parameters to resampled reconstructions. All *p*-values from this analysis are shown in Table 4-4.

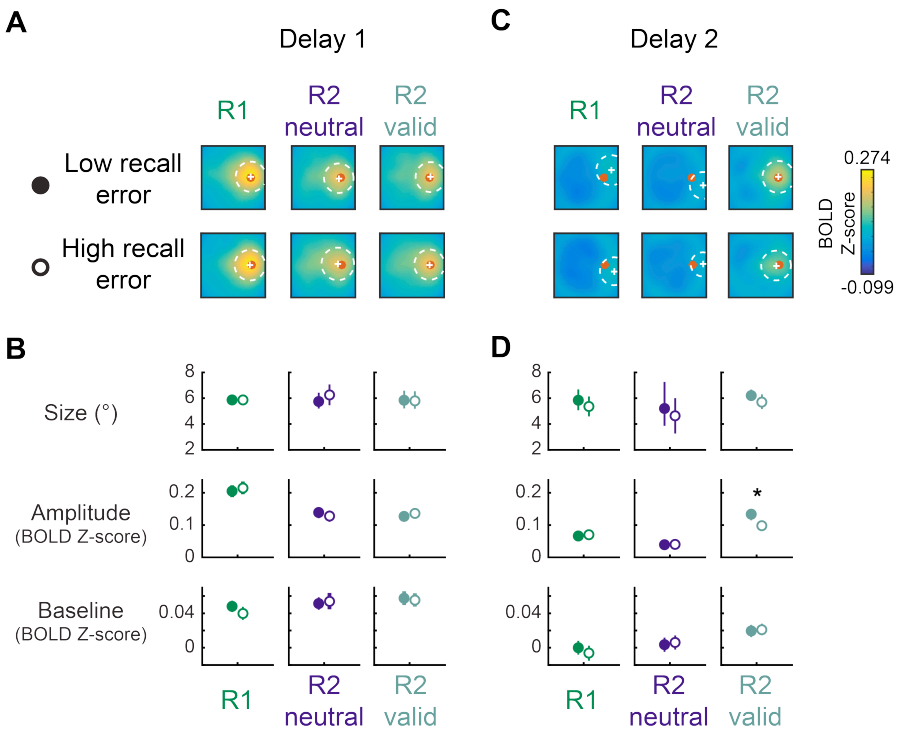
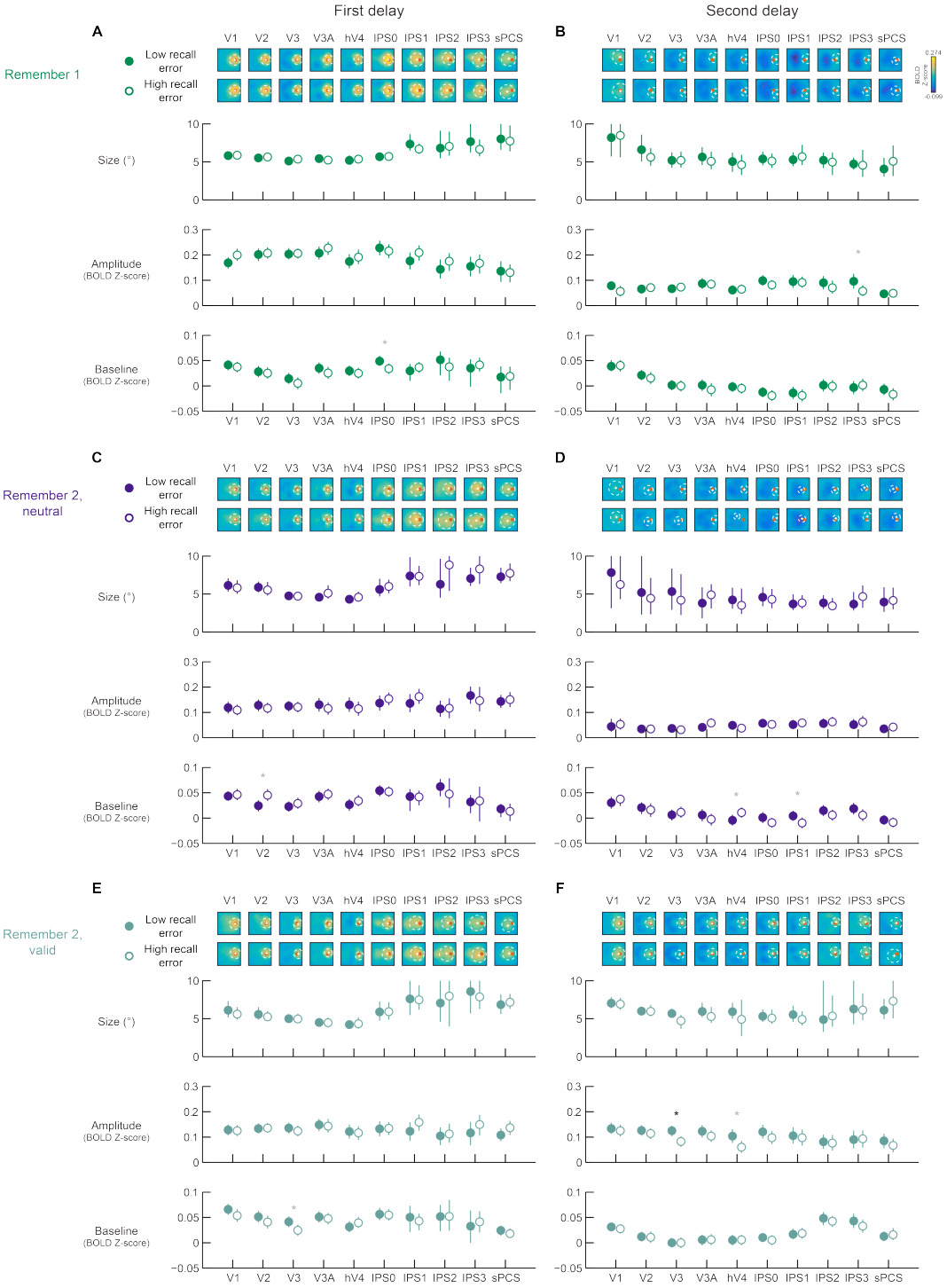


Figure 4-12: Amplitude of recovered representation on valid-cue trials indexes behavioral performance

Within each participant, session, and WM condition, we performed a median split on trials based on recall error, then quantified low- and high-error reconstructions separately via a resampling procedure (as in Fig. 4-11). All data shown here are from reconstructions computed from all ROIs, concatenated ("All ROIs combined" in previous figures). For this analysis on each ROI independently, see Figure 4-13. **(A)** During the first delay, reconstructions were similar across recall error conditions. White plus and dashed white circle indicate mean fit position and mean size (full-width at half maximum; FWHM). Red circle indicates exact target position. **(B)** Quantified WM representations did not differ across recall error group (all $p \geq 0.082$, resampling test; see Table 4-5 for p-values for all comparisons). **(C)** During the second delay, the cued representation on R2-valid trials is visibly more robust on low- compared to high-error trials. **(D)** The cued WM representation is related to behavioral performance selectively via representation amplitude: on trials when participants performed more accurately, cued representation amplitude was higher ($p < 0.001$). All other WM conditions and parameters showed no differences across behavioral performance bins ($p \geq 0.136$). Error bars mark 95% CI of fit parameters to resampled reconstructions (some lie behind circles). Asterisk indicates significant differences after FDR correction for multiple comparisons ($q = 0.05$).

Figure 4-13: Amplitude of recovered representation on valid-cue trials indexes behavioral performance in V3

Data plotted as in Figure 4-11, with trials sorted based on behavioral recall performance. All resampling and fitting procedures are identical to those used for Figure 4-12. **(A)** Remember 1 trials, Delay 1. IPS0 baseline trended to be larger on low recall error trials ($p = 0.014$) **(B)** Remember 1 trials, Delay 2. IPS3 amplitude trended to be larger for low recall error compared to high recall error trials ($p = 0.028$). **(C)** Remember 2-neutral trials, first delay. V2 baseline trended to be smaller for low recall error trials ($p = 0.002$). **(D)** Remember 2-neutral trials, second delay. In hV4, baseline trended to be smaller for low recall error trials ($p = 0.046$). In IPS1, Baseline trended to be larger for low recall error trials ($p = 0.04$). **(E)** Remember 2-valid trials, first delay. In V3, baseline trended to be larger for low recall error trials ($p = 0.032$). **(F)** Remember 2-valid trials, second delay. In V3, amplitude was significantly larger on low error trials ($p < 0.001$), and in hV4, amplitude trended towards being larger on low recall error trials ($p = 0.008$). All error bars 95% confidence intervals over resampled fitting iterations. Black asterisks indicate significant difference between low- and high-recall error trials for that WM condition, delay period, and fit parameter, FDR-corrected for multiple comparisons within each parameter ($q = 0.05$). Gray asterisks are trends, thresholded at $\alpha = 0.05$, uncorrected for multiple comparisons. All p-values available in Table 4-5.


First delay
Second delay
A
B
C
D
E
F
Remember 1
Remember 2, neutral
Remember 2, valid
Low recall error
High recall error
V1 V2 V3 V3A hV4 IPS0 IPS1 IPS2 IPS3 sPCS
Size (°)
Amplitude (BOLD Z-score)
Baseline (BOLD Z-score)

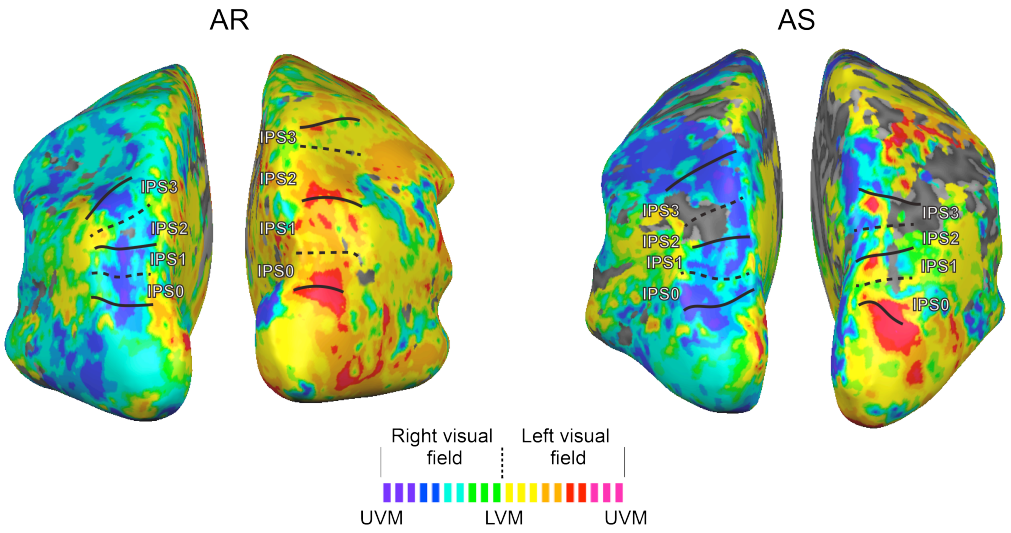


Figure 4-14: Retinotopic maps used to define IPS subregions for participants AR and AS Cortical topography of polar angle preference for each vertex, thresholded so that only vertices with power at the stimulus rotation frequency (1/36 Hz) greater than 0.005 the maximum across all voxels are shown. Solid and dashed lines mark upper and lower vertical meridian representations (UVM; LVM), respectively, used to define IPS0-IPS3. See Sprague and Serences, 2013 for retinotopic maps for participant AI; Ester et al., 2015 for participants AL and AP.

Table 4-1: Statistical comparisons for mean delay-period activation
P-values for comparisons of mean delay-period activation over all voxels within each ROI between WM conditions (two-tailed). All *p*-values reflect pair-wise comparisons between conditions (R1 vs R2-neutral, R2-neutral vs. R2-valid, and R1 vs. R2-valid). For all tables, a p-value of 0 indicates $p < 0.001$, the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons ($q = 0.05$, all comparisons and all individual ROIs). Italicized numbers indicate trends, defined using $\alpha = 0.05$, uncorrected. Significant comparisons and trends are shown graphically in Figure 4-3. FDR threshold for V1-sPCS is $p \leq 0.006$

Figure 4-3	**Delay 1**			**Delay 2**		
Comparison:	R1 vs R2-neutral	R2-neutral vs. R2-valid	R1 vs R2-valid	R1 vs R2-neutral	R2-neutral vs. R2-valid	R1 vs R2-valid
V1	0.772	0.056	0.098	0.85	*0.048*	*0.026*
V2	0.872	0.162	0.228	0.87	0.056	0.068
V3	0.854	0.118	0.084	0.902	0.086	0.152
V3A	0.646	0.114	0.304	0.814	0.306	0.486
hV4	0.106	0.136	**0.006**	0.288	**0**	0.064
IPS0	0.938	0.07	0.076	0.98	**0**	**0**
IPS1	0.886	*0.04*	0.056	0.464	**0**	**0**
IPS2	0.956	0.154	0.124	0.216	**0**	**0**
IPS3	0.74	0.212	0.354	0.226	**0**	**0**
sPCS	0.752	0.044	*0.024*	0.14	**0**	**0**

Table 4-2: Statistical comparisons for significant delay-period representational fidelity
P-values for comparisons between representational fidelity computed separately within each WM delay (one-tailed, against the null hypothesis that representational fidelity ≤ 0). P-value of 0 indicates $p < 0.001$, the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons ($q = 0.05$, all conditions and all individual ROIs, and separately for "All ROIs combined" across all conditions, see Experimental Procedures). Italicized numbers indicate trends, defined using $\alpha = 0.05$, uncorrected. Significant comparisons and trends are shown graphically in Figure 4-9A-B. FDR threshold for V1-sPCS is $p \leq 0.039$, and for All ROIs combined is $p < 0.001$.

Representational fidelity (Fig. 4-9A-B)	Delay 1 (one-tailed)			Delay 2 (one-tailed)		
	R1	R2-neutral	R2-valid	R1	R2-neutral	R2-valid
V1	**0**	**0**	**0**	**0**	**0.001**	**0**
V2	**0**	**0**	**0**	**0**	**0.002**	**0**
V3	**0**	**0**	**0**	**0**	**0.009**	**0**
V3A	**0**	**0**	**0**	**0**	**0.039**	**0**
hV4	**0**	**0**	**0**	**0**	**0.002**	**0**
IPS0	**0**	**0**	**0**	**0**	**0.023**	**0**
IPS1	**0**	**0**	**0**	**0**	0.351	**0**
IPS2	**0**	**0**	**0**	**0**	0.476	**0**
IPS3	**0**	**0**	**0.001**	**0**	**0.003**	**0**
sPCS	**0**	**0**	**0**	**0**	0.454	**0**
All ROIs combined	**0**	**0**	**0**	**0**	**0**	**0**

Table 4-3: Statistical comparisons for significant differences between Delay 1 and Delay 2 representational fidelity

P-values for comparisons of representational fidelity between Delay 1 and Delay 2 (two-tailed). P-value of 0 indicates $p < 0.001$, the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons ($q = 0.05$, all conditions and all individual ROIs, and separately for "All ROIs combined" across all conditions, see Experimental Procedures). Italicized numbers indicate trends, defined using $\alpha = 0.05$, uncorrected. Significant comparisons and trends are shown in Figure 4-9C. FDR threshold for V1-sPCS is $p \leq 0.022$ and for All ROIs combined is $p < 0.001$.

Representational fidelity (Fig. 4-9C)	Delay 1 vs. Delay 2		
	R1	R2-neutral	R2-valid
V1	**0**	**0**	**0**
V2	**0**	**0**	0.084
V3	**0**	**0**	**0.022**
V3A	**0**	**0**	0.244
hV4	**0**	**0.018**	0.116
IPS0	**0**	**0**	**0.002**
IPS1	0.088	**0**	**0**
IPS2	0.9	**0.01**	**0.004**
IPS3	0.402	0.162	**0.006**
sPCS	1	**0.014**	0.656
All ROIs combined	**0**	**0**	**0**

Table 4-4: Statistical comparisons for best-fit surface parameters between condition pairs within each delay period

P-values for comparisons of parameters (size, amplitude, baseline) for best-fit surfaces to coregistered reconstructions between conditions for each delay period individually (two-tailed). All p-values reflect pair-wise comparisons between conditions (R1 vs R2-neutral, R2-neutral vs. R2-valid, and R1 vs. R2-valid). P-value of 0 indicates $p < 0.001$, the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons ($q = 0.05$, all conditions and all individual ROIs, and separately for "All ROIs combined" across all conditions, see Experimental Procedures). Italicized numbers indicate trends, defined using $\alpha = 0.05$, uncorrected. Significant comparisons and trends are shown graphically in Figure 4-11.

Fig. 4-11	Param:	**Size**			**Amplitude**			**Baseline**		
Delay	ROI	R1 vs R2-neutral	R2-neutral vs R2-valid	R1 vs R2-valid	R1 vs R2-neutral	R2-neutral vs R2-valid	R1 vs R2-valid	R1 vs R2-neutral	R2-neutral vs R2-valid	R1 vs R2-valid
1	V1	0.598	0.796	0.93	**0**	0.272	**0**	0.226	**0.006**	**0**
1	V2	0.562	0.304	0.514	**0**	0.238	**0**	0.126	*0.032*	**0.002**
1	V3	*0.038*	0.322	0.426	**0**	0.424	**0**	**0.002**	0.206	**0**
1	V3A	0.088	0.53	*0.008*	**0**	0.158	**0**	**0.006**	0.422	**0**
1	hV4	**0**	0.276	0.092	**0**	0.326	**0**	0.548	0.488	0.19
1	IPS0	0.738	0.712	0.494	**0**	0.546	**0**	**0.012**	0.832	**0.01**
1	IPS1	0.402	0.992	0.492	**0.004**	0.42	**0**	0.278	0.366	0.094
1	IPS2	0.632	0.956	0.582	**0.002**	0.874	**0.014**	0.302	0.656	0.712
1	IPS3	0.212	0.742	0.118	0.866	0.338	0.256	0.438	0.9	0.532
1	sPCS	0.454	0.9	0.396	0.688	0.164	0.702	0.978	0.816	0.926
1	All ROIs	0.632	0.65	0.874	**0**	0.77	**0**	**0.018**	0.314	**0**
2	V1	0.548	0.808	0.088	0.128	**0**	**0**	0.282	0.496	*0.04*
2	V2	0.57	0.784	0.27	**0.002**	**0**	**0**	0.968	0.236	0.154
2	V3	0.986	0.816	0.446	**0**	**0**	**0**	0.124	0.082	0.83
2	V3A	0.166	0.394	0.506	**0.004**	**0**	**0**	0.414	0.464	0.104
2	hV4	0.35	0.084	0.186	**0.012**	**0**	0.096	0.21	0.792	0.118
2	IPS0	0.388	0.628	0.624	**0**	**0**	0.054	**0.016**	*0.024*	**0**
2	IPS1	**0.002**	*0.042*	0.188	**0**	**0**	0.422	**0.01**	**0**	**0**
2	IPS2	**0.002**	0.09	0.788	*0.034*	0.124	0.842	*0.044*	**0**	**0**
2	IPS3	0.458	0.052	0.07	0.076	**0.01**	0.276	**0.016**	**0**	**0**
2	sPCS	0.51	*0.018*	0.08	0.202	**0**	**0.014**	0.222	**0**	**0**
2	All ROIs	0.434	0.32	0.776	**0**	**0**	**0**	0.058	**0**	**0**
FDR thresh:	V1-sPCS	0.002			V1-sPCS	0.014		V1-sPCS	0.016	
	All ROIs	<0.001			All ROIs	< 0.001		All ROIs	0.018	

Table 4-5: Statistical comparisons for best-fit surface parameters between low- and high-recall error trials

P-values for comparisons of parameters (size, amplitude, baseline) for best-fit surfaces to coregistered reconstructions between low recall error and high recall error trials (two-tailed, always equal number of trials in each bin per participant and session). P-value of 0 indicates $p < 0.001$, the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons within each parameter ($q = 0.05$, all conditions and all individual ROIs, and separately for "All ROIs combined" across all conditions, see Experimental Procedures; FDR thresholds indicated at bottom of table). Italicized numbers indicate trends, defined using $\alpha = 0.05$, uncorrected. Significant comparisons and trends are shown graphically in Figure 4-12 and Figure 4-13. For the All ROIs combined comparisons, use of a threshold derived with Bonferroni's method produces identical significant comparisons.

	Param:	**Size**			**Amplitude**			**Baseline**		
Delay	ROI	R1	R2-neutral	R2-valid	R1	R2-neutral	R2-valid	R1	R2-neutral	R2-valid
1	V1	0.818	0.6	0.522	0.052	0.534	0.876	0.586	0.696	0.116
1	V2	0.632	0.526	0.576	0.744	0.42	0.864	0.684	0.002	0.228
1	V3	0.326	0.904	0.918	0.886	0.8	0.458	0.272	0.4	0.032
1	V3A	0.592	0.286	0.89	0.218	0.378	0.718	0.238	0.472	0.72
1	hV4	0.634	0.548	0.872	0.35	0.408	0.762	0.53	0.302	0.244
1	IPS0	0.926	0.532	0.962	0.522	0.396	0.944	*0.014*	0.804	0.854
1	IPS1	0.272	0.972	0.93	0.15	0.212	0.124	0.406	0.92	0.61
1	IPS2	0.824	0.188	0.698	0.16	0.898	0.792	0.28	0.37	0.982
1	IPS3	0.31	0.356	0.662	0.63	0.478	0.3	0.616	0.77	0.684
1	sPCS	0.8	0.548	0.728	0.804	0.696	0.172	0.986	0.686	0.312
1	All ROIs	0.934	0.29	0.938	0.408	0.324	0.44	0.082	0.634	0.702
2	V1	0.874	0.562	0.812	0.12	0.608	0.524	0.856	0.392	0.552
2	V2	0.322	0.924	0.914	0.634	0.998	0.354	0.474	0.6	0.872
2	V3	1	0.548	0.06	0.602	0.604	**0**	0.864	0.55	0.942
2	V3A	0.488	0.416	0.272	0.89	0.156	0.262	0.294	0.316	0.978
2	hV4	0.612	0.432	0.328	0.822	0.334	*0.008*	0.658	0.046	0.968
2	IPS0	0.708	0.74	0.642	0.214	0.742	0.168	0.332	0.17	0.412
2	IPS1	0.698	0.808	0.376	0.838	0.498	0.714	0.588	*0.04*	0.792
2	IPS2	0.752	0.494	0.586	0.312	0.606	0.826	0.804	0.238	0.456
2	IPS3	0.824	0.294	0.982	*0.028*	0.476	0.898	0.61	0.124	0.284
2	sPCS	0.482	0.802	0.364	0.804	0.476	0.41	0.168	0.466	0.606
2	All ROIs	0.438	0.662	0.136	0.686	0.946	**0**	0.294	0.642	0.78
FDR thresh:	V1-sPCS	<0.001			V1-sPCS	< 0.001		V1-sPCS	<0.001	
	All ROIs	.0083 (Bonferroni)		All ROIs	0.0083		All ROIs	0.0083		

Table 4-6: Statistical comparisons for target activation differences between Delay 1 and Delay 2

P-values for comparisons of target activation differences (probed target – non-probed target) between Delay 1 and Delay 2 (two-tailed). P-value of 0 indicates $p < 0.001$, the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons ($q = 0.05$, all conditions and all individual ROIs, and separately for "All ROIs combined" across all conditions, see Experimental Procedures). Italicized numbers indicate trends, defined using $\alpha = 0.05$, uncorrected. Significant comparisons and trends are shown in Figure 4-10A-C. FDR thresholds for V1-sPCS and for All ROIs combined are $p < 0.001$. Identical comparisons remain significant when correcting with Bonferroni's method.

	V1	**V2**	**V3**	**V3A**	**hV4**	**IPS0**	**IPS1**	**IPS2**	**IPS3**	**sPCS**	**All ROIs**
R1	0.326	0.948	0.672	0.774	0.722	0.486	0.166	0.182	0.55	0.96	0.954
R2-neutral	0.772	0.39	0.63	0.836	0.71	0.416	0.624	0.426	0.428	0.5	0.888
R2-valid	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	*0.018*	0.218	**0**

Table 4-7: Statistical comparisons between target activation for probed target and non-probed target within each delay

P-values for comparisons between probed target (PT) activation and non-probed target (NPT) activation computed separately within each WM delay (one-tailed, against the null hypothesis that PT ≤ NPT). P-value of 0 indicates $p < 0.001$, the minimum p-value achievable per our resampling procedure with 1,000 iterations. Bold numbers indicate significant differences after FDR correction for all comparisons ($q = 0.05$, all conditions and all individual ROIs, and separately for "All ROIs combined" across all conditions, see Experimental Procedures). Italicized numbers indicate trends, defined using $\alpha = 0.05$, uncorrected. Significant comparisons and trends are shown in Figure 4-10D. FDR thresholds for V1-sPCS and for All ROIs combined are $p < 0.001$.

Condition:	**Remember 1**		**Remember 2 - neutral**		**Remember 2-valid**	
Delay:	Delay 1	Delay 2	Delay 1	Delay 2	Delay 1	Delay 2
V1	**0**	**0**	0.905	0.821	0.787	**0**
V2	**0**	**0**	0.915	0.577	0.803	**0**
V3	**0**	**0**	0.638	0.88	0.908	**0**
V3A	**0**	**0**	0.99	0.982	0.969	**0**
hV4	**0**	**0**	0.667	0.826	0.969	**0**
IPS0	**0**	**0**	0.99	0.905	0.911	**0**
IPS1	*0.018*	**0**	0.844	0.958	0.665	**0**
IPS2	*0.006*	**0**	0.346	0.745	0.428	**0**
IPS3	*0.001*	**0.002**	0.091	0.39	0.257	**0**
sPCS	0.076	0.085	0.252	0.561	0.255	*0.016*
All ROIs combined	**0**	**0**	0.928	0.932	0.864	**0**