Neural dynamics of visual working memory representation during sensory distraction

Jonas Karolis Degutis[1,2,3], Simon Weber[1,4], Joram Soch[1,5,6,7], John-Dylan Haynes[1,2,3,4,8,9]

1. Bernstein Center for Computational Neuroscience Berlin and Berlin Center for Advanced Neuroimaging, Charité Universitätsmedizin Berlin, corporate member of the Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany.
2. Max Planck School of Cognition, Stephanstrasse 1a, Leipzig, Germany.
3. Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany.
4. Research Training Group “Extrospection” and Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Berlin, Germany.
5. Institute of Psychology, Otto von Guericke University, Mageburg, Germany.
6. Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany.
7. German Center for Neurodegenerative Diseases, Göttingen, Germany.
8. Research Cluster of Excellence “Science of Intelligence”, Technische Universität Berlin, Berlin, Germany.
9. Collaborative Research Center “Volition and Cognitive Control”, Technische Universität Dresden, Dresden, Germany.

Correspondence to: j.karolis.degutis@maxplanckschools.de

Keywords: visual working memory, dynamic coding, neural subspace, sensory distractor, multiplexing

Abstract

Recent studies have provided evidence for the concurrent encoding of sensory percepts and visual working memory contents (VWM) across visual areas; however, it has remained unclear how these two types of representations are concurrently present. Here, we reanalyzed an open-access fMRI dataset where participants memorized a sensory stimulus while simultaneously being presented with sensory distractors. First, we found that the VWM code in several visual regions did not generalize well between different time points, suggesting a dynamic code. A more detailed analysis revealed that this was due to shifts in coding spaces across time. Second, we collapsed neural signals across time to assess the degree of interference between VWM contents and sensory distractors, specifically by testing the alignment of their encoding spaces. We find that VWM and feature-matching sensory distractors are encoded in separable coding spaces. Together, these results indicate a role of dynamic coding and temporally stable coding spaces in helping multiplex perception and VWM within visual areas.

Introduction

To successfully achieve behavioral goals, humans rely on the ability to remember, update, and ignore information. Visual working memory (VWM) allows for a brief maintenance of visual stimuli that are no longer present within the environment [1–3]. Previous studies have revealed that the contents of VWM are present throughout multiple visual areas, starting from V1 [4–12]. These findings raised the question of how areas that are primarily involved in visual perception can also maintain VWM information without interference between the two contents. Recent studies that had participants remember a stimulus while simultaneously being presented with sensory stimuli during the delay period have found supporting evidence that both VWM contents and sensory percepts are multiplexed in occipital and parietal regions [8,13,14]. However, the mechanism employed in order to segregate bottom-up visual input from VWM contents remains poorly understood.

One proposed mechanism to achieve the separation between sensory and memory representations is dynamic coding [25–27]: the change of the population code encoding VWM representations across time. Recent work has shown that the format of VWM might not be as persistent and stable throughout the delay as previously thought [28,29]. Frontal regions display dynamic population coding across the delay during the maintenance of category [30] and spatial contents in the absence of interference [31,32], and also shows dynamic recoding of the memoranda after sensory distraction [33,34]. The visual cortex in humans displays dynamic coding of contents during high load trials [35] and during a spatial VWM task [36]. However, it is not yet clear whether dynamic coding of VWM might help evade sensory distraction in human visual areas.

Another line of evidence suggests that perception could potentially be segregated from VWM representations using stable non-overlapping coding spaces [15]. For example, evidence from neuroanatomy indicates that the sensory bottom-up visual pathway primarily projects to the cytoarchitectonic Layer 4 in V1, while feedback projections culminate in superficial and deep layers of the cortex [16]. Functional results are in line with neuroanatomy by showing that VWM signals preferentially activate the superficial and deep layers in humans [17] and non-human primates [18], while perceptual signals are more prevalent in the middle layers [19]. In addition to laminar separation, regional multiplexing of multiple items could potentially rely on rotated representations, as seen in memory and sensory representations orthogonally coded in the auditory cortex [20] and in the storage of a sequence of multiple spatial locations in the prefrontal cortex (PFC) [21]. Non-overlapping orthogonal representations have also been seen in both humans and trained recurrent neural networks as a way of segregating attended and unattended VWM representations [22–24].

Here we investigated whether the concurrent presence of VWM and sensory information is compatible with predictions offered by dynamic coding or by stable non-aligned coding spaces. For this, we reanalyzed an open-access fMRI dataset by Rademaker et al. [8] where participants performed a delayed-estimation VWM task with and without sensory distraction. To investigate dynamic coding we employed a temporal cross-decoding analysis that assessed how well the multivariate code encoding VWM generalizes from one time point to another [32,37–39], and a

temporal neural subspace analysis that examined a sensitive way of looking at alignment of neural populations coding for VWM at different time points. To assess the non-overlapping coding hypothesis, we used neural subspaces [20,31,36] to see whether temporally stable representations of the VWM target and the sensory distractor are coded in separable neural populations. Finally, we examined the multivariate VWM code changes during distractor trials when compared to the no-distractor VWM format.

Results

Temporal cross-decoding in distractor and no-distractor trials

In the previously published study [8] participants completed a VWM task where on a given trial they were asked to remember an orientation of a grating, which they had to then recall at the end of the trial. The delay period was either left blank (no-distractor) or a noise or randomly oriented grating distractor was presented (**Fig. 1a**). To investigate the dynamics of the VWM code, we examined how the multivariate pattern of activity encoding VWM memoranda changed across the duration of the delay period. To do so, we ran a temporal cross-decoding analysis where we trained a decoder (periodic support vector regression, see [40]) on the target orientation, separately for each time point and tested on all time points in turn in a cross-validated fashion. If the information encoding VWM memoranda were to have the same code, the trained decoder would generalize to other time points, indicated by similar decoding accuracies on the diagonal and off-diagonal elements of the matrix. However, if the code exhibited dynamic properties, despite information about the memonda being present (above-chance decoding on the diagonal of the matrix), both off-diagonal elements corresponding to a given on-diagonal element would have lower decoding accuracies (**Fig. 1b**). Such off-diagonal elements are considered an indication of a dynamic code.

We ran the temporal cross-decoding analysis for the three VWM delay conditions: no-distractor, noise distractor and orientation distractor (feature-matching distractor). First, we examined each element of the cross-decoding matrix to test whether decoding accuracies were above chance. In all three conditions and throughout all ROIs, we found clusters where decoding was above chance (**Fig. 1c-e**, black outline; nonparametric cluster-permutation test against null; all clusters $p < 0.05$) from as early as 4 s after the onset of the delay period. We found that decoding on the diagonal was highest during no-distractor compared to noise and orientation distractor trials in most regions of interest (ROI; **Fig. 4a**).

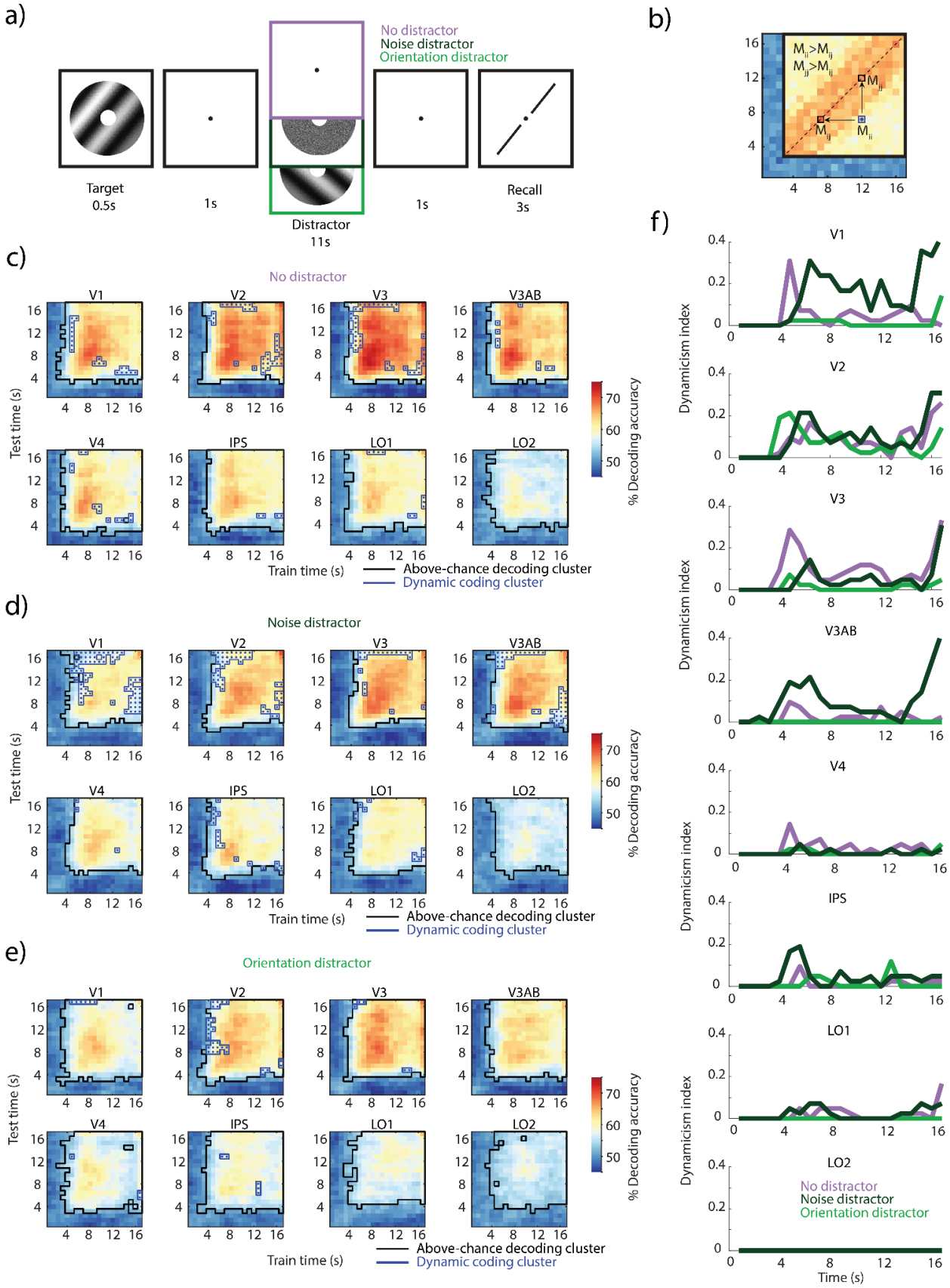

a)
No distractor
Noise distractor
Orientation distractor
Target
0.5s
1s
Distractor
11s
1s
Recall
3s
b)
$M_{ii}>M_{ij}$
$M_{jj}>M_{ij}$
M_{jj}
M_{ij}
M_{ii}
c)
No distractor
d)
Noise distractor
e)
Orientation distractor
V1
V2
V3
V3AB
V4
IPS
LO1
LO2
Test time (s)
Train time (s)
% Decoding accuracy
Above-chance decoding cluster
Dynamic coding cluster
f)
Dynamicism index
Time (s)
No distractor
Noise distractor
Orientation distractor

Figure 1. Task and temporal cross-decoding. a) On each trial an oriented grating was presented for the 0.5 s followed by a delay period of 13 s [8]. In a third of the trials a noise distractor was presented for 11 s during the middle of the delay; in another third another orientation grating was presented; one third of trials had no distractor during the delay. **b)** Illustration of dynamic coding elements. An off-diagonal element had to have a lower decoding accuracy compared to both corresponding diagonal elements (see Methods for details). **c)** Temporal generalization of the multivariate code encoding VWM representations in three conditions across occipital and parietal regions. Across-participant mean temporal cross-decoding of no-distractor trials. Black outlines: matrix elements showing above-chance decoding (cluster-based permutation test; $p < 0.05$). Blue outlines with dots: dynamic coding elements; parts of the cross-decoding matrix where the multivariate code fails to generalize (off-diagonal elements having lower decoding accuracy than their corresponding two diagonal elements; conjunction between two cluster-based permutation tests; $p < 0.05$). **d)** Same as c), but noise distractor trials. **e)** Same as c), but orientation distractor trials. **f)** Dynamicism index; the proportion of dynamic coding elements across time. High values indicate a dynamic non-generalizing code, while low values indicate a generalizing code. Time indicates the time elapsed since the onset of the delay period.

Second, we examined off-diagonal elements to assess whether there was any indication that they reflected a non-generalizing dynamic code (see Methods for full details). Despite a high degree of temporal generalization, we found dynamic coding clusters in all three conditions. Some degree of dynamic coding was observed in all ROIs but LO2 in the noise distractor and no-distractor trials, while it was only present in V1, V2, V3, V4, and IPS in the orientation distractor condition (**Fig. 1c-e**, blue outline). The difference between noise and orientation distractor conditions could not be explained by the amount of information present in each ROI, as the decoding accuracy of the diagonal was similar across all ROIs in both the noise and orientation distractor conditions (**Fig. 4a**). We saw a nominally larger number of dynamic coding elements in V1, V2 and V3AB during the noise distractor condition and in V3 during the no-distractor condition (**Fig. 1d**).

To qualitatively compare the amount of dynamic coding in the three conditions across the delay period, we calculated a dynamicism index [32] (**Fig. 1e**; see Methods), which measured the multivariate code's uniqueness at each time point; more precisely, the proportion of dynamic elements corresponding to each diagonal element. High values indicate dynamic code and low values indicate a generalizing code. Across all conditions, most dynamic elements occurred between the encoding and early delay periods (4-8 s), and the late delay and retrieval (14.4-16.8 s). Interestingly, during the noise distractor trials in V1 we also saw dynamic coding during the middle of the delay period; the multivariate code not only changed during the onset and offset of the noise stimulus, but also during its presentation and throughout the extent of the delay.

Dynamics of VWM neural subspaces across time

The temporal cross-decoding analysis revealed more dynamic coding in the early visual cortex primarily during the early and late delay phase and a more generalized coding throughout the delay in higher-order regions. In order to understand the nature of these effects in more detail, we conducted a separate series of analyses that directly assessed the neural subspaces in which the orientations were encoded and how these potentially changed across time. Specifically, we followed a previous methodological framework [36] and applied a principal component analysis (PCA) to the high-dimensional activity patterns at each time point to identify the two axes that explained maximal variance across orientations (see **Fig. 2** and Methods).

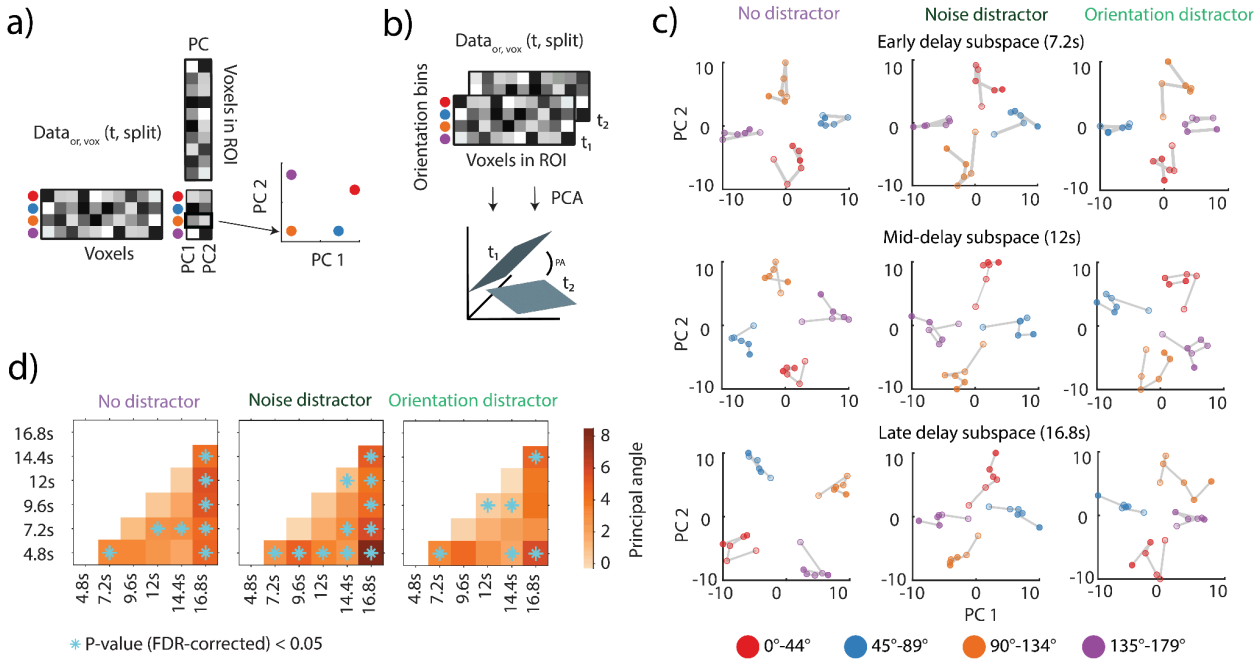


Figure 2. Assessing the dynamics of neural subspaces in V1-V3AB. a) Schematic illustration of the neural subspace analysis. A given data matrix (voxels x orientation bins) was subjected to a principal components analysis and the first two dimensions were used to define a neural subspace onto which a left-out test data matrix was projected. This resulted in a matrix of two coordinates for each orientation bin and was visualized (see right). The x and y axes indicate the first two principal components. Each color depicts an angular bin. **b)** Schematic illustration of the calculation of an above-baseline principal angle (aPA). A principal angle (PA) is the angle between the 2D PCA-based neural subspaces (as in **a**) for two different time points t_1, t_2. A small angle would indicate alignment of coding spaces; an angle of above-baseline would indicate a shift in the coding space. The above-baseline principle angle (aPA) is the angle for a comparison between two time points (t_1, t_2) minus the angle between cross-validated pairs of the same time points. **c)** Each row shows a projection that was estimated for one of two time ranges (middle and late delay) and then applied to all time points (using independent, split-half cross-validated data). Opacity increases from early to late time

points. For visualization purposes the subspaces were estimated on a participant-aggregated ROI [36]. **Fig. S1** depicts the same projections as neural trajectories. **d)** aPA between all pairwise time point comparisons (nonparametric permutation test against null; FDR-corrected $p < 0.05$) averaged across 1,000 split-half iterations. Corresponding *p*-values found in **Supplementary Table 1**.

First, we visualized the consistency of the neural subspaces across time. For this, we computed low-dimensional 2D neural subspaces for a given time point and projected left-out data from six time points during the delay onto this subspace [20,36]. A projection of data from a single time point resulted in four orientation bin values placed within the subspace (**Fig. 2a**, colored circles indicate orientation). Taking into account projected data from all timepoints, if the VWM code were generalizing, we would see a clustering of orientation points in a subspace; however, if orientation points were scattered around the neural subspace, this would show a non-generalizing code.

We examined the projections in a combined ROI spanning V1-V3AB aggregated across participants. We projected left-out data from all six time point bins onto subspaces generated from the early (7.2 s), middle (12 s), and late (16.8 s) time point data for each of the three conditions. Overall, the results showed generalization across time with some exceptions (**Fig. 2c, Fig. S1**). The clustering of orientation bins in the no-distractor condition was most pronounced (**Fig. 4a**). In contrast, the noise distractor trials showed a resemblance of some degree of dynamic coding, as seen by less variance explained by early time points projected onto the middle subspace and the early and middle time points projected onto late subspace (**Fig. 2c, Fig. S1**).

To quantify the visualized changes, we measured the alignment between each pair of subspaces by calculating the above-baseline principal angle (**Fig. 2b**) within the combined V1-V3AB ROI. The above-baseline principal angle (aPA) measures the alignment between the 2D subspaces encoding the VWM representations: the higher the angle, the smaller the alignment between two subspaces and an indication of a changed neural coding space. Unlike in the projection of data from time points, the aPA was calculated participant-wise. Using a split-half approach, we measured the aPA between each split-pair of subspaces and subtracted the angles measured within each of the subspaces with the latter acting as a null baseline.

All three conditions showed significant aPAs (**Fig, 2d**; cyan stars; permutation test; $p < 0.05$, FDR-corrected). Corresponding to the results from the cross-decoding analysis, the early (4.8s) and late (16.8) delay subspaces showed the highest number of significant pairwise aPAs in all conditions, with noise distractor trials having all pairwise aPAs including the early and late subspaces being significant. The three conditions each had two significant aPAs between timepoints in the middle of the delay period.

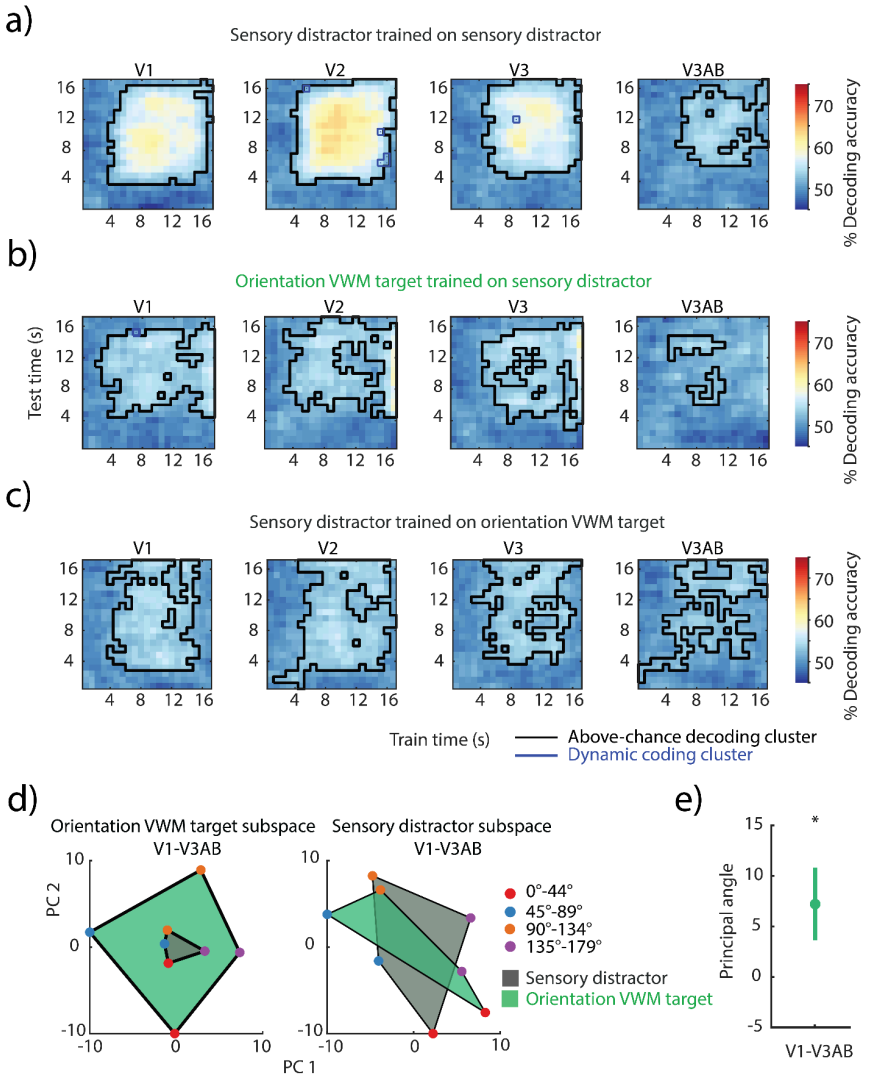


Figure 3. Generalization between target and distractor codes in orientation distractor VWM trials in V1-V3AB. a) Across-participant mean temporal cross-decoding of the sensory distractor. Black outlines: matrix elements showing above-chance decoding (cluster-based permutation test; $p < 0.05$). Blue outlines with dots: dynamic coding element (conjunction between two cluster-based permutation tests; $p < 0.05$). **b)** Same as a), but the decoder was trained on the target and tested on the sensory distractor in orientation VWM trials. **c)** Same as a), but trained on the sensory distractor and tested on the target. See **Fig. S2** for ROIs from V4-LO2. **d)** Left: projection of left-out target (green) and sensory distractor (gray) onto an orientation VWM target neural subspace. Right: same as left, but the projections are onto the sensory distractor subspace. **e)** Principal angle between the sensory distractor and orientation VWM target subspaces ($p = 0.0297$, one-tailed permutation test of sample mean). Average across 1,000 split-half iterations. Errorbars indicate ± SEM across participants.

Alignment between distractor and target subspaces in orientation distractor trials

Next, we assessed any similarity in encoding between the memorized orientation targets and the orientation distractors by focusing on those trials where both occurred. First, we examined whether the encoding of the sensory distractor is stable across its entire presentation duration (1.5 s - 12.5 s after target onset) using the same approach as for the VWM target (**Fig. 1e**). We found stable coding of the distractor in all ROIs with only a few dynamic elements in V2 and V3 (**Fig. 3a**, **Fig. S2**). We then assessed whether the sensory distractor had a similar code to the VWM target by examining whether the multivariate code across time generalizes from the target to the distractor and vice versa. When cross-decoded, the sensory distractor (**Fig. 3c**) and target orientation (**Fig. 3b**) had lower decoding accuracies in the early visual cortex compared to when trained and tested on the same label-type, indicative of a non-generalizing code. Such a difference was not seen in higher-order visual regions, as the decoding of the sensory distractor was low to begin with (**Fig. S2**).

Since we found minimal dynamics in the encoding of the distractor (**Fig. 3a**) and target (**Fig. 1e**), we focused on temporally stable neural subspaces that encoded the target and sensory distractor. We computed stable neural subspaces where we disregarded the temporal variance by averaging across the whole delay period and binned the trials either based on the target orientation (**Fig. 3d**, left subpanel) or the distractor orientation (**Fig. 3d**, right subpanel). We then projected left-out data binned based on the target (**Fig. 3d**, green quadrilateral) or the distractor (**Fig. 3d**, gray quadrilateral). This projection provided us with both a baseline (as when training and testing on the same label) and a cross-generalization. Unsurprisingly, the target subspace explained the left-out target data well (**Fig. 3d**, left subpanel, green quadrilateral); however, the target subspace explained less variance of the left-out distractor data (**Fig. 3d**, left subpanel, gray quadrilateral), as qualitatively seen from the smaller spread of the sensory distractor orientations. A similar but less pronounced dissociation between projections was seen in the distractor subspace (**Fig. 3d**, left, quadrilateral in green) with the distractor subspace better explaining the left-out distractor data. We quantified the difference between the target and distractor subspaces and found a significant aPA between them (p = 0.0297, one-tailed nonparametric permutation test; **Fig. 3e**). These results provide evidence for the presence of separable stable neural subspaces that might enable the multiplexing of VWM and perception across the extent of the delay period.

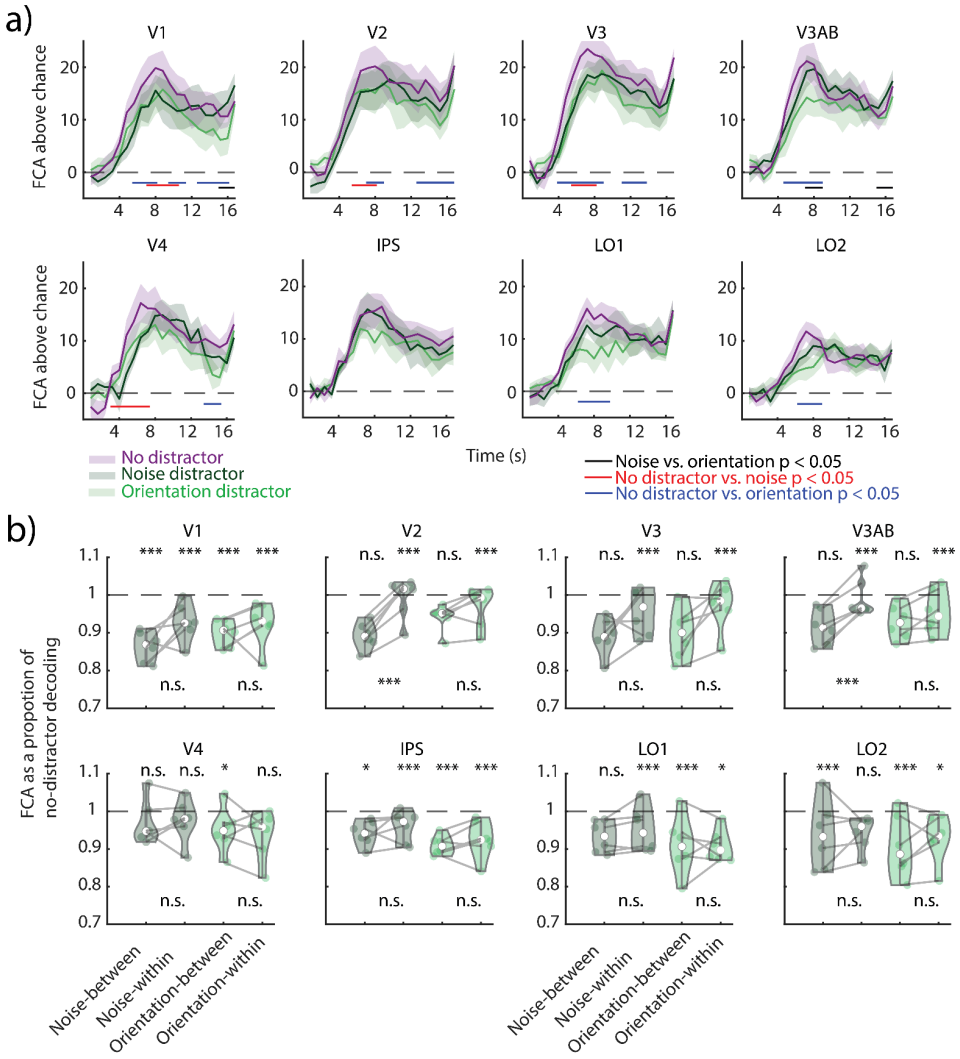


Figure 4. Cross-decoding between distractor and no-distractor conditions. a) Decoding accuracy (feature continuous accuracy; FCA) across time for train and test on no-distractor trials (purple), train and test on noise distractor trials (dark green) and train and test on orientation distractor trials (light green). Horizontal lines indicate clusters where there is a difference between two time courses (all clusters $p < 0.05$; nonparametric cluster permutation test, see color code on the right). **b)** Decoding accuracy as a proportion of no-distractor decoding estimated on the averaged delay period (4-16.8s). Nonparametric permutation tests compared the decoding accuracy of each analysis to the no-distractor decoding baseline (indicated as a dashed line) and between a decoder trained and tested on distractor trials (noise- or orientation-within) and a decoder trained on no-distractor trials and tested on distractor trials (noise or orientation-cross). FDR-corrected across ROIs. * $p < 0.05$, *** $p < 0.001$. Corresponding *p*-values found in **Supplementary Table 2**.

Impact of distractors on VWM multivariate code

To further assess the impact of distractors on the available VWM information, we examined the decoding accuracies of distractor and no-distractor trials across time. Decoding accuracy was higher in the no-distractor trials compared to both orientation and noise distractor trials across all ROIs, but IPS (**Fig. 4a**, red and blue lines, $p < 0.05$, cluster permutation test) across several stages of the delay period. To further assess how distractors affected the delay period information, we increased sensitivity by collapsing across it, because time courses were comparable in all conditions (**Fig. 4a**). To assess to which degree VWM encoding generalized from no-distractor to distractor trials, we trained a decoder on no-distractor trials and tested it on both types of distractor trials (**Fig. 4b** noise- and orientation-cross). We expressed the decoding accuracy of each distractor condition as a proportion of the decoding accuracy in the no-distractor condition. Values close to one indicate comparable information, while values below one mean the decoder does not generalize well. We found that the cross-decoding accuracies were significantly lower than the no-distractor in all ROIs but V4 (in both noise and orientation) and LO2 (only noise). Thus, in most areas the decoder did not generalize well from the no-distractor to distractor conditions. However, the total amount of information in distractor trials was generally slightly lower (**Fig. 4a**). Thus, we also compared the generalization to a decoder trained and tested on the same distractor condition (**Fig. 4b** noise- and orientation-within), which might thus be able to extract more information. We found that indeed information recovered in areas V2 and V3AB in the noise distractor condition (**Fig. 4b**, pairwise permutation test). Thus, there was more information in the noise distractor condition, but it was not accessible to a decoder trained only on no-distractor trials. Additionally, a temporal cross-decoding analysis where all training time points were no-distractor trials had less dynamic coding in early visual regions (**Fig. S3**) when compared to the temporal cross-decoding matrix when trained and tested on noise distractor trials (**Fig. 1d**). These results indicate a change in the VWM format between the noise distractor and no-distractor trials.

Discussion

We examined the dynamics of visual working memory (VWM) with and without distractors and explored the impact of sensory distractors on the coding spaces of VWM contents in visual areas by reanalyzing previously published data [8]. Participants completed a task during which they had to maintain an orientation stimulus in VWM. During the delay period either no distractor, an orientation distractor, or a noise distractor were presented. We assessed two potential mechanisms that could help concurrently maintain the superimposed sensory and memory representations. First, we examined whether changes were observable in the multivariate code for memory contents across time, which we term dynamic coding. For this we used two different analyses: temporal cross-classification and a direct assessment of angles between coding spaces. We found evidence for dynamic coding in all conditions, but there were differences in these dynamics between conditions and regions. Dynamic coding was most pronounced during the noise distractor trials in early visual regions. Second, we assessed the complementary question of temporally stable coding spaces. We computed the stable neural

subspaces by averaging across the delay period. We saw that coding of the VWM target and concurrent sensory distractors occurred in different stable neural subspaces. Finally, we observed that the format of the multivariate VWM code during the noise distraction differs from the VWM code when distractors were not present.

Dynamic encoding of VWM contents has been previously repeatedly examined. Temporal cross-decoding analyses have been used in a number of non-human primate electrophysiology and human fMRI studies [25,32,33,35,36,38,41]. Spaak et al. [32] found dynamic coding in the non-human primate PFC during a spatial VWM task. They observed a change in the multivariate code between different stages; specifically a first shift between the encoding and maintenance periods, and also a second shift between the maintenance and retrieval periods. The initial transformation between the encoding and maintenance periods might recode the percept of the target into a stable VWM representation, whereas the second might transform the stable memoranda into a representation suited for initiation of motor output. A similar dynamic coding pattern was also observed in human visual regions using neuroimaging [36]. In this study, in all three conditions we find a comparable pattern of results, where the multivariate code changes between the early delay and middle delay, and middle delay and late delay periods.

When noise distractors are added to the delay period we find evidence of additional coding shifts in V1 during the middle of the delay. Previous research in non-human primates has shown that the presentation of a distractor induces a change in multivariate encoding for VWM in lateral PFC (lPFC) [33]. More precisely, a lack of generalization was observed between the population code encoding VWM before the presentation of a distractor (first half of the delay) and after its presentation (the second half). Additionally, continuous shifts in encoding have been observed in the extrastriate cortex throughout the extent of the delay period when decoding multiple remembered items at high VWM load [35]. The dynamic code has been interpreted to enable multiplexing of representations when the visual cortex is overloaded by the maintenance of multiple stimuli at once. Future research could examine how properties of the distractor and of the target stimulus could interact to lead to dynamic coding. One intriguing hypothesis is that distractors that perturb the activity of feature channels that are used to encode VWM representations induce changes in its coding space over time. It is important to note that in this experiment, the activation of the encoded target features was highest for the noise stimulus. Thus the shared spatial frequencies between noise distractor and the VWM contents potentially contribute to a more pronounced dynamic coding effect.

In a complementary analysis we directly assessed subspaces in which orientations were encoded in VWM. We defined the subspaces for three different time windows, early, middle and late. We find no evidence that the identity of orientations is confusable across time, e.g. we do not observe 45° at one given time point being recoded as 90° from a different time point. Such dynamics have been previously observed in the rotation of projected angles within a fixed neural subspace [20,22]. Rather, we find a decreased generalization between neural subspaces at different time points, as previously observed in a spatial VWM task [36]. These results suggest that the temporal dynamics across the VWM trial periods are driven by changes in the coding subspace of VWM. We do observe a preservation of the topology of the projected angles, as more similar angles remained closer together (e.g. the bin containing 45° was always closer to

the bin containing 0° and 90°). Such a topology has been seen in V4 during a color perception task [42].

We also find evidence that the VWM contents are encoded in a different way depending on whether a noise distractor is presented or not. The decoder trained on no-distractor trials does not generalize well, presumably because it fails to fully access all the information present in noise distractor trials. If the decoders are trained directly on the distractor conditions the VWM related information is much higher. Additionally, we see that the code generalizes better across time when training on no-distractor trial time points and testing on noise distractor trials. This may imply that by training our decoder on the no-distractor trials we are able to uncover an underlying stable population code encoding VWM in noise distractor trials. Consistent with this finding, Murray et al. [31] demonstrated that subspaces derived on the delay period could still generalize to the more dynamic encoding and retrieval periods, albeit not perfectly.

Interestingly, we found limited dynamic coding in the orientation distractor condition; primarily a change in the code between the early delay and middle delay periods was observed. Nonetheless, we find distinct temporally stable coding spaces in which sensory distractors and memory targets are encoded. These results correspond to prior research demonstrating a rotated format between perception and memory representations [20], attended and unattended VWM representations in both humans and recurrent-neural networks trained on a 2-back VWM [22] and serial retro-cueing tasks [23,43]. Additionally, similar rotation dynamics have been observed between multiple spatial VWM locations stored in the non-human primate lPFC [21]. Considering the consistency of these results across different paradigms, we speculate that separate coding spaces might be a general mechanism of how feature-matching items can be concurrently multiplexed within visual regions. With growing evidence of the relationship between VWM capacity and neural resources available within the visual cortex [44–46], further research could examine the number of feature-matching items that can be stored in non-aligned coding spaces.

It remains to be seen whether the degree of change or rotation between subspaces correlates with behavior. In this experiment, we do not observe a behavioral deficit in the feature-matching orientation distractor trials [8]. Yet there is evidence from behavioral and neural studies that show interactions between perception and VWM: feature-matching distractors behaviorally bias retrieved VWM contents [47,48]; VWM representations influence perception [49–52]; neural visual VWM representations in the early visual cortices are biased towards distractors [53]; and the fidelity of VWM neural representations within the visual cortex negatively correlates with behavioral errors when recalling VWM during a sensory distraction task [54]. In cases where a distractor does induce a drop in recall accuracy or biases the recalled VWM target, VWM and the sensory distractor neural subspaces might overlap more.

To our surprise, we did not observe a significant difference in the coding format of VWM between orientation distractor and no-distractor trials. Our initial expectation was that the VWM coding might undergo changes due to the target representation avoiding the distractor stimulus. However, the presence of a generalizing code between no-distractor and orientation distractor trials, along with the non-aligned coding spaces between the target and distractor in the orientation trials, suggests an alternative explanation. We suggest that the sensory distractor

stimulus occupies a distinct coding space throughout its presentation during the delay, while the coding space of the target remains the same in both orientation and no-distractor trials. Layer-specific coding differences in perception and VWM might explain these findings [17,19,55]. Specifically, the sensory distractor neural subspace might predominantly reside in the bottom-up middle layers of early visual cortices, while the neural subspace encoding VWM might primarily occupy the superficial and deep layers.

We provide evidence for two types of mechanisms found in visual areas during the presence of both VWM and sensory distractors. First, our findings show dynamic coding of VWM within the human visual cortex during sensory distraction and indicate that such activity is not only present within the lPFC. Second, we find that VWM and feature-matching sensory distractors are encoded in shifted coding spaces. Taking into account previous findings, we posit that different coding spaces within the same region might be a more general mechanism of segregating feature-matching stimuli. In sum, these results provide possible mechanisms of how VWM and perception are concurrently present within visual areas.

Methods

Participants, stimuli, procedure, and preprocessing

The following section is a brief explanation of parts of the methods covered in Rademaker et al. [8]. Readers may refer to that paper for details. We reanalyzed data from Experiment 1.

Six participants performed two tasks while in the scanner: a VWM task and a perceptual localizer task. In the perceptual localizer task, either a donut-shaped or a circle-shaped grating was presented in 9 second blocks. The participants had to respond whenever the grating dimmed. There were a total of 20 donut-shaped and 20 circle-shaped gratings in one run. Participants completed a total of 15-17 runs.

The visual VWM task began with the presentation of a colored 100% valid cue which indicated the type of trial: no-distractor, orientation distractor, or noise distractor. Following the cue, the target orientation grating was presented centrally for 500 ms, followed by a 13 s delay period. In the trials with the distractor, a stimulus of the same shape and size as the target grating was presented centrally for 11 s in the middle of the delay period (**Fig. 1a**). The orientation and noise distractors reversed contrast at 4 Hz. At the end of the delay, a probe stimulus bar appeared at a random orientation. The participants had to align the bar to the target orientation and had to respond in 3 s.

The orientations for the VWM sample were pseudo-randomly chosen from six orientation bins each consisting of 30 orientations. The orientation distractor and sample were counterbalanced in order not to have the same orientation presented as a distractor. Each run consisted of four trials of each condition. Across three sessions participants completed 27 runs of the task resulting in a total of 108 trials per condition.

The data were acquired using a simultaneous multi-slice EPI sequence with a TR of 800 ms, TE of 35 ms, flip angle of 52°, and isotropic voxels of 2 mm. The data were preprocessed using FreeSurfer and FSL and time-series were z-scored across time for each voxel.

Voxel selection

We used the same regions of interest (ROI) as in Rademaker et al. [8], which were derived using retinotopic mapping. In contrast to the original study, we reduced the size of our ROIs by selecting voxels that reliably responded to both the donut-shaped orientation perception task and the no-distractor VWM task. In order to select reliably activating voxels, we calculated four tuning functions for each voxel: two from the perceptual localizer and two from the no-distractor VWM task. The tuning functions spanned the continuous feature space in bins of 30°. Thus, to calculate the tuning functions, we ran a split-half analysis using stratified sampling where we binned all trials into six bins (of 30°). For both halves, tuning functions were estimated using a GLM that included six orientation regressors (one for each bin) and assumed an additive noise component independent and identically distributed across trials. We calculated Pearson correlations between the no-distractor memory and the perception tuning functions across the six parameter estimates extracted from the GLM, thus generating one memory-memory and one perception-perception correlation coefficient for each voxel.

The same analysis was additionally performed 1,000 times on randomly permuted orientation labels to generate a null distribution for each participant and each ROI. These distributions were used to check for the reliability of voxel activation to perception and no-distractor VWM. After performing Fisher z-transformation on the correlations, we selected voxels that had a value above the 75th percentile of the null distributions in both the memory-memory and perception-perception correlations. This population of voxels was then used for all subsequent analyses. IPS included reliable voxels from retinotopically derived IPS0, IPS1, and IPS2.

Periodic support vector regression

We used periodic support vector regression (pSVR) to predict the target orientation from the multivariate BOLD activity [40]. PSVR uses a regression approach to estimate the sine and cosine components of a given orientation independently and therefore accounts for the circular nature of stimuli. In order to have a proper periodic function, orientation labels from the range [0°, 180°) were projected into the range [0, 2π).

We used the support vector regression algorithm using a non-linear radial basis function (RBF) kernel implemented in LIBSVM [56] for orientation decoding. Specifically, sine and cosine components of the presented orientations were predicted based on multivariate fMRI signals from a set of voxels at specific time points within a trial (see *Temporal Generalization*). In each cross-validation fold, we rescaled the training data voxel activation into the range [0, 1] and applied the training data parameters to rescale the test data. For each participant we had a total of three iterations in our cross-validation, where we trained on two thirds (i.e. two sessions) and

tested on one third of the data (i.e. the left-out session). We selected three iterations in order to mitigate training and test data leakage (see *Temporal Generalization*).

After pSVR-based analysis, reconstructed orientations were obtained by plugging the predicted sine and cosine components into the four-quadrant inverse tangent:

$$\theta_p = \mathrm{atan2}(x_p, y_p)$$

where x_p and y_p are pSVR outputs in the test set. Prediction accuracy was measured as the trial-wise absolute angular deviation between predicted orientation and actual orientation:

$$\Delta x = \left|(\theta - \theta_p)_{\mathrm{circ}}\right|$$

where θ is the labeled orientation and θ_p is the predicted orientation. This measure was then transformed into a trial-wise feature continuous accuracy (FCA) [57] as follows:

$$\mathrm{FCA} = \frac{\pi - \Delta \mathrm{x}}{\pi} \cdot 100$$

The final across-trial accuracy was the mean of the trial-wise FCAs. Mean FCA was calculated across predicted orientations from all test sets after cross-validation was complete. The FCA is an equivalent measurement to standard accuracy measured in decoding analyses falling into the range between 0 and 100%, but extended to the continuous domain. In the case of random guessing, the expected angular deviation is $\pi/2$, resulting in chance-level FCA at 50%.

Temporal cross-decoding

To determine the underlying stability of the VWM code, we ran a temporal cross-decoding analysis using pSVR (**Fig. 1**). We trained on data from a given time point and then predicted orientations for all time points, using the presented targets as labels. We trained on two-thirds of the trials per iteration and tested on the left-out third. Training and test data were never taken from the same trials, both when testing on the same and different time points.

We used a cluster-based approach to test for significance for above-chance decoding clusters [58]. To determine whether the size of the cluster of the above-chace values was significantly larger than chance, we calculated a summed t-value for each cluster. We then generated a null distribution by randomly permuting the sign of the estimated above-chance accuracy (each FCA value was subtracted by 50%, such that 0 corresponds to chance level) of all components within the temporal cross-decoding matrix. We calculated the summed t-value for the largest randomly occurring above-chance cluster. This procedure was repeated 1000 times to estimate a null distribution. The empirical summed t-value of each cluster was then compared to the null distribution to determine significance ($p < 0.05$; without control of multiple cluster comparisons).

Dynamic coding clusters were defined as elements within the temporal cross-decoding matrix where the multivariate code at a given time point did not fully generalize to another time point; in

other words, an off-diagonal element was significantly smaller in accuracy compared to its two corresponding on-diagonal elements ($a_{ij} < a_{ii}$ and $a_{ij} < a_{jj}$, **Fig. 1b**). In order to test for significance of these clusters, we ran two cluster-permutation tests as done in previous studies to define dynamic clusters [32,36]. In each test, we subtracted one or the other corresponding diagonal elements from the off-diagonal elements ($a_{ij} - a_{ii}$ and $a_{ij} - a_{jj}$). We then ran the same sign permutation test as for the above-chance decoding cluster for both comparisons. An off-diagonal element was deemed dynamic, if both tests were significant ($p < 0.05$) and it was part of the above-chance decoding cluster.

Following [32], we also computed the dynamicism index as a proportion of elements across time that were dynamic. Specifically, we calculated the proportion of (off-diagonal) dynamic elements corresponding to a diagonal time point in both columns (corresponding to the test time points) and rows (corresponding to the train time points) of the temporal cross-decoding matrix.

Neural subspaces

We adapted the method from [36] to calculate two-dimensional neural subspaces encoding VWM information at a given time point. To do so, we used principal component analysis (PCA). To maximize power, we binned trial-wise fMRI activations into four equidistant bins of 45 degrees and averaged the signal across all trials within a bin (**Fig. 2a**). The data matrix **X** was defined as a $p \times v$ matrix where $p = 4$ was the four orientation bins, and v was the number of voxels. We mean-centered the columns (i.e. each voxel) of the data matrix.

This analysis focused on the time points from 4 s to 17.6 s after delay onset. The first TRs were not used since the temporal cross-decoding results showed no above-chance decoding. We averaged across every three TRs leading to six non-overlapping temporal bins resulting in six **X** matrices. We calculated the principal components (PCs) using eigendecomposition of the covariance matrix for each **X** and defined the matrix **V** using the two largest eigenvalues as a $v \times 2$ matrix, resulting in six neural subspaces, one for each non-overlapping temporal bin.

Neural subspaces across time

For visualization purposes, we used three out of the total of six neural subspaces from the following time points: early (7.2 s), middle (12 s), and late (16.8 s). Following the aforementioned procedure, these subspaces were calculated on half of the trials, as we projected the left-out data onto the subspaces. The left-out data were binned into six temporal bins between 4 s and 17.6 s after target onset with no overlap just like in the calculation of the six subspaces. The projection resulted in a $p \times 2$ matrix **P** for each projected time bin (resulting in a total of six **P** matrices). We use distinct colors to plot the temporal trajectories of each orientation bin across time in a 2D subspace flattened (**Fig. 2c**) and not flattened (**Fig. S1**) across the time dimension. Importantly, the visualization analysis was done on a combined

participant-aggregated V1-V3AB region, which included all reliable voxels across the four regions and all six participants (see *Voxel Selection*).

To measure the alignment between coding spaces at different times, we calculated an above-baseline principal angle (aPA) between all subspaces (**Fig. 2c**). We used the MATLAB function `subspace` for an implementation of the method proposed by [59] to measure the angle between two **V** matrices. This provided us with a possible principal angle between 0-90°; the higher the angle, the larger the difference between the two subspaces. In order to avoid overfitting and as in the visualization analysis, we used a split-half approach to compute the aPA between subspaces. Half of the binned trials were used to calculate $\mathbf{V}_{i,A}$ and $\mathbf{V}_{j,A}$ and half for $\mathbf{V}_{i,B}$ $\mathbf{V}_{j,B}$, where **A** and **B** refer to the two halves of the split and and **i** and **j** refer to the two time bins compared. For significance testing, the within-subspace angle (the angle between two splits of the data within a given temporal bin (i.e. $\mathbf{V}_{i,A}$ and $\mathbf{V}_{i,B}$)) was subtracted from the between-subspace PA (the angle between two different temporal bins (e.g. $\mathbf{V}_{i,A}$ and $\mathbf{V}_{j,B}$)). Unlike the visualization analysis, the PA was calculated per participant 1,000 times using different splits of the data on a combined V1-V3AB region that included the reliable voxels across the four regions (see *Voxel Selection*). The final aPA value was an average across all iterations for each participant.

Sensory distractor and orientation VWM target neural subspaces

For the orientation VWM target and sensory distractor neural subspace, we followed the aforementioned subspace analysis, but instead of calculating subspaces on six temporal bins, we averaged across the 4-17.6 s delay period and calculated a single subspace. As in the previous analysis, we split the orientation VWM trials in half. We then binned the trials either based on the target orientation or the sensory distractor. For visualization purposes, we projected the left-out data averaged based on the sensory distractor and the target onto subspaces derived from both the sensory distractor and target subspaces. As in the previous visualization, the analysis was run on a participant-aggregated V1-V3AB region.

To calculate the aPA we had the following subspaces: $\mathbf{V}_{Target,A}$, $\mathbf{V}_{Dist,A}$, $\mathbf{V}_{Target,B}$ and $\mathbf{V}_{Dist,B}$, where the subspaces were calculated on trials binned either based on the target orientation or the sensory distractor. The aPA was calculated by subtracting the within-subspace angle ($\mathbf{V}_{Target,A}$ and $\mathbf{V}_{Target,B}$, $\mathbf{V}_{Dist,A}$ and $\mathbf{V}_{Dist,B}$) from the sensory distractor and working memory angle ($\mathbf{V}_{Target,A}$ and $\mathbf{V}_{Dist,B}$, $\mathbf{V}_{Target,B}$ and $\mathbf{V}_{Dist,A}$). The split-half aPA analysis was performed 1,000 times and the final value was an average across these iterations for each participant.

Data availability

The data is shared open-access https://osf.io/dkx6y/. The analysis scripts will be shared open-access https://osf.io/XXX/ when published.

Acknowledgements

J.K.D. was funded by the Max Planck Society and BMBF (as part of the Max Planck School of Cognition). J.D.H. was supported by the Deutsche Forschungsgemeinschaft (DFG, Exzellenzcluster Science of Intelligence); SFB 940 “Volition and Cognitive Control”; and SFB-TRR 295 “Retuning dynamic motor network disorders using neuromodulation”. S.W. was supported by Deutsche Forschungsgemeinschaft (DFG) Research Training Group 2386 451 and EXC 2002/1 “Science of Intelligence.” We thank Rosanne Rademaker, Chaipat Chunharas, and John Serences for collecting and sharing their data open access, without which this reanalysis would not have been possible. We also thank Rosanne Rademaker, Michael Wolff, Amir Rawal, and Maria Servetnik for extensive discussions of the results. We also thank Vivien Chopurian and Thomas Christophel for their feedback on the manuscript.

Author contributions

Conceptualization, J.K.D.; Methodology, J.K.D, S.W., J.S., J.-D.H.; Formal Analysis, J.K.D.; Software, J.K.D, S.W., J.S.; Visualization, J.K.D.; Funding Acquisition, J.K.D., J.-D.H.; Writing - Original Draft Preparation, J.K.D.; Writing – Review & Editing, J.K.D., S.W., J.S., J.-D.H. Supervision, J.-D.H.

Declaration of interests

The authors declare no competing interests.

References

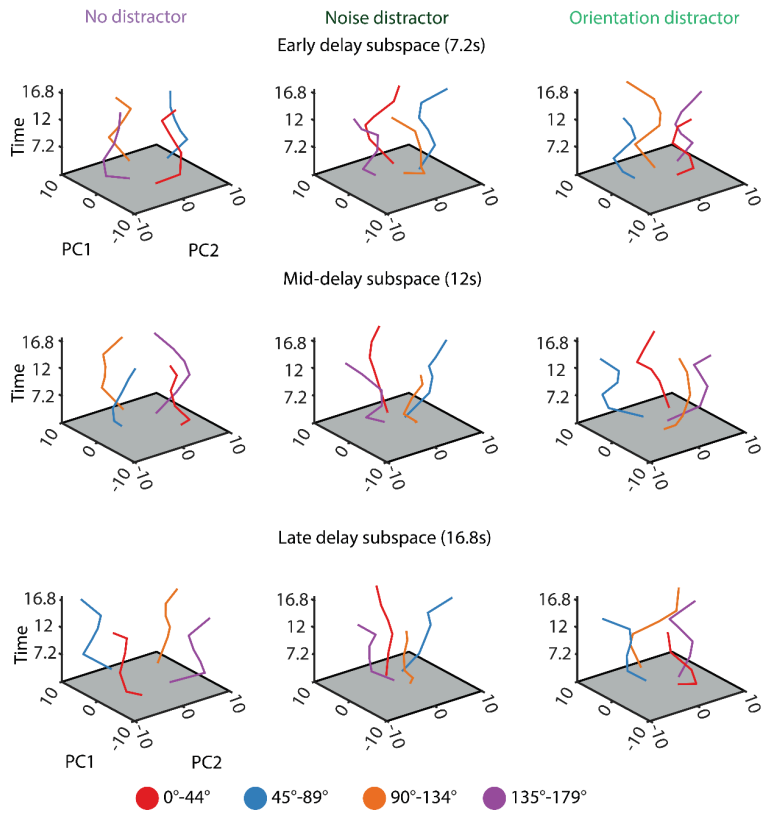
1. Curtis, C. E. & D’Esposito, M. Persistent activity in the prefrontal cortex during working memory. *Trends Cogn. Sci.* **7**, 415–423 (2003).
2. Goldman-Rakic, P. S. Cellular basis of working memory. *Neuron* **14**, 477–485 (1995).
3. D’Esposito, M. & Postle, B. R. The Cognitive Neuroscience of Working Memory. *Annu. Rev. Psychol.* **66**, 115–142 (2015).
4. Fuster, J. M. & Alexander, G. E. Neuron Activity Related to Short-Term Memory. *Science* **173**, 652–654 (1971).
5. Harrison, S. A. & Tong, F. Decoding reveals the contents of visual working memory in early visual areas. *Nature* **458**, 632–635 (2009).
6. Christophel, T. B., Hebart, M. N. & Haynes, J.-D. Decoding the Contents of Visual Short-Term Memory from Human Visual and Parietal Cortex. *J. Neurosci.* **32**, 12983–12989 (2012).
7. Ester, E. F., Sprague, T. C. & Serences, J. T. Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. *Neuron* **87**,

893–905 (2015).
8. Rademaker, R. L., Chunharas, C. & Serences, J. T. Coexisting representations of sensory and mnemonic information in human visual cortex. *Nat. Neurosci.* **22**, 1336–1344 (2019).
9. Riggall, A. C. & Postle, B. R. The Relationship between Working Memory Storage and Elevated Activity as Measured with Functional Magnetic Resonance Imaging. *J. Neurosci.* **32**, 12990–12998 (2012).
10. Serences, J. T., Ester, E. F., Vogel, E. K. & Awh, E. Stimulus-Specific Delay Activity in Human Primary Visual Cortex. *Psychol. Sci.* **20**, 207–214 (2009).
11. Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R. & Haynes, J.-D. The Distributed Nature of Working Memory. *Trends Cogn. Sci.* **21**, 111–124 (2017).
12. Curtis, C. E. & Sprague, T. C. Persistent Activity During Working Memory From Front to Back. *Front. Neural Circuits* **15**, 696060 (2021).
13. Iamshchinina, P., Christophel, T. B., Gayet, S. & Rademaker, R. L. Essential considerations for exploring visual working memory storage in the human brain. *Vis. Cogn.* **29**, 425–436 (2021).
14. Bettencourt, K. C. & Xu, Y. Decoding the content of visual short-term memory under distraction in occipital and parietal areas. *Nat. Neurosci.* **19**, 150–157 (2016).
15. Lorenc, E. S., Mallett, R. & Lewis-Peacock, J. A. Distraction in Visual Working Memory: Resistance is Not Futile. *Trends Cogn. Sci.* **25**, 228–239 (2021).
16. Felleman, D. J. & Van Essen, D. C. Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cereb. Cortex* **1**, 1–47 (1991).
17. Lawrence, S. J. D. *et al.* Laminar Organization of Working Memory Signals in Human Visual Cortex. *Curr. Biol.* **28**, 3435-3440.e4 (2018).
18. van Kerkoerle, T., Self, M. W. & Roelfsema, P. R. Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. *Nat. Commun.* **8**, 13804 (2017).
19. Lawrence, S. J. D., Norris, D. G. & de Lange, F. P. Dissociable laminar profiles of concurrent bottom-up and top-down modulation in the human visual cortex. *eLife* **8**, e44422 (2019).
20. Libby, A. & Buschman, T. J. Rotational dynamics reduce interference between sensory and memory representations. *Nat. Neurosci.* **24**, 715–726 (2021).
21. Xie, Y. *et al.* Geometry of sequence working memory in macaque prefrontal cortex. *Science* **375**, 632–639 (2022).
22. Wan, Q., Menendez, J. A. & Postle, B. R. Priority-based transformations of stimulus representation in visual working memory. *PLOS Comput. Biol.* **18**, e1009062 (2022).
23. Wan, Q., Ardalan, A., Fulvio, J. M. & Postle, B. R. *Representing Context and Priority in Working Memory*. http://biorxiv.org/lookup/doi/10.1101/2023.10.24.563608 (2023) doi:10.1101/2023.10.24.563608.
24. Van Loon, A. M., Olmos-Solis, K., Fahrenfort, J. J. & Olivers, C. N. Current and future goals are represented in opposite patterns in object-selective cortex. *eLife* **7**, e38677 (2018).
25. Stokes, M. G., Muhle-Karbe, P. S. & Myers, N. E. Theoretical distinction between functional states in working memory and their corresponding neural states. *Vis. Cogn.* **28**, 420–432 (2020).
26. Stokes, M. G. ‘Activity-silent’ working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* **19**, 394–405 (2015).
27. Stroud, J. P., Duncan, J. & Lengyel, M. The computational foundations of dynamic coding in

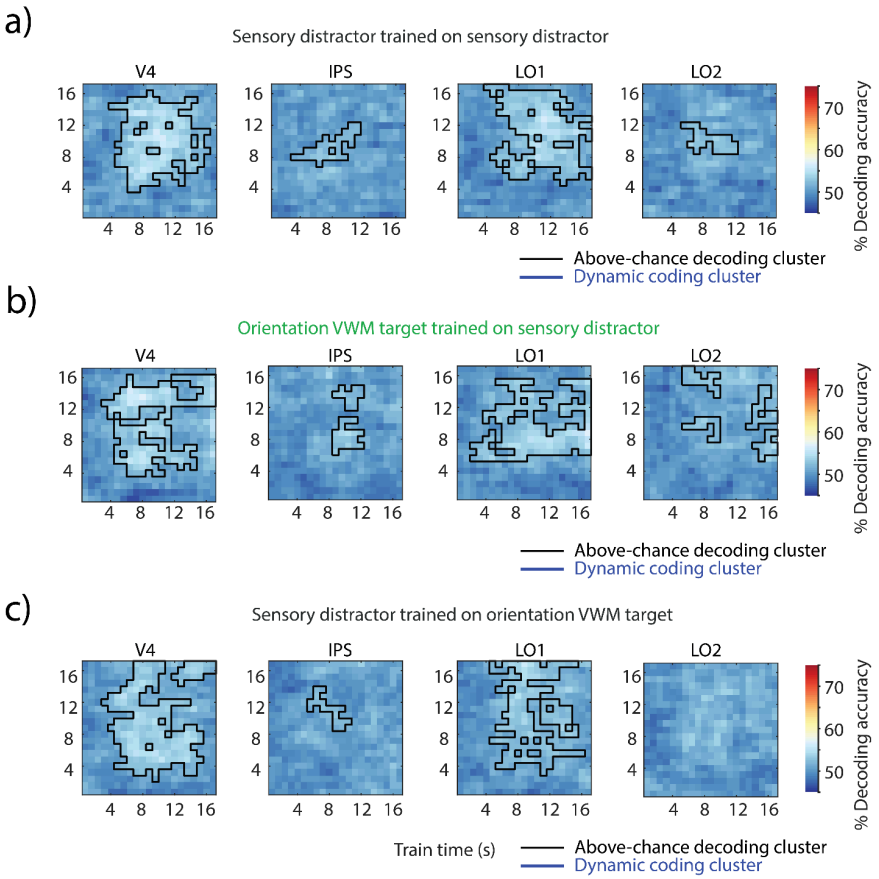
working memory. *Trends Cogn. Sci.* S1364661324000536 (2024) doi:10.1016/j.tics.2024.02.011.
28. Miller, E. K., Lundqvist, M. & Bastos, A. M. Working Memory 2.0. *Neuron* **100**, 463–475 (2018).
29. Sreenivasan, K. K., Curtis, C. E. & D'Esposito, M. Revisiting the role of persistent neural activity during working memory. *Trends Cogn. Sci.* **18**, 82–89 (2014).
30. Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K. & Poggio, T. Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex. *J. Neurophysiol.* **100**, 1407–1419 (2008).
31. Murray, J. D. *et al.* Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl. Acad. Sci.* **114**, 394–399 (2017).
32. Spaak, E., Watanabe, K., Funahashi, S. & Stokes, M. G. Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. *J. Neurosci.* **37**, 6503–6516 (2017).
33. Parthasarathy, A. *et al.* Mixed selectivity morphs population codes in prefrontal cortex. *Nat. Neurosci.* **20**, 1770–1779 (2017).
34. Parthasarathy, A. *et al.* Time-invariant working memory representations in the presence of code-morphing in the lateral prefrontal cortex. *Nat. Commun.* **10**, 4995 (2019).
35. Sreenivasan, K. K., Vytlacil, J. & D'Esposito, M. Distributed and Dynamic Storage of Working Memory Stimulus Information in Extrastriate Cortex. *J. Cogn. Neurosci.* **26**, 1141–1153 (2014).
36. Li, H.-H. & Curtis, C. E. Neural population dynamics of human working memory. *Curr. Biol.* **33**, 3775-3784.e4 (2023).
37. Stokes, M. G. *et al.* Dynamic Coding for Cognitive Control in Prefrontal Cortex. *Neuron* **78**, 364–375 (2013).
38. Degutis, J. K. *et al. Dynamic Layer-Specific Processing in the Prefrontal Cortex during Working Memory*. http://biorxiv.org/lookup/doi/10.1101/2023.10.27.564330 (2023) doi:10.1101/2023.10.27.564330.
39. Anders, S., Heinzle, J., Weiskopf, N., Ethofer, T. & Haynes, J.-D. Flow of affective information between communicating brains. *NeuroImage* **54**, 439–446 (2011).
40. Weber, S., Christophel, T., Görgen, K., Soch, J. & Haynes, J. Working memory signals in early visual cortex are present in weak and strong imagers. *Hum. Brain Mapp.* **45**, e26590 (2024).
41. Cavanagh, S. E., Towers, J. P., Wallis, J. D., Hunt, L. T. & Kennerley, S. W. Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nat. Commun.* **9**, 3498 (2018).
42. Brouwer, G. J. & Heeger, D. J. Decoding and Reconstructing Color from Responses in Human Visual Cortex. *J. Neurosci.* **29**, 13992–14003 (2009).
43. Piwek, E. P., Stokes, M. G. & Summerfield, C. A recurrent neural network model of prefrontal brain activity during a working memory task. *PLOS Comput. Biol.* **19**, e1011555 (2023).
44. Cohen, M. A., Konkle, T., Rhee, J. Y., Nakayama, K. & Alvarez, G. A. Processing multiple visual objects is limited by overlap in neural channels. *Proc. Natl. Acad. Sci.* **111**, 8955–8960 (2014).

45. Franconeri, S. L., Alvarez, G. A. & Cavanagh, P. Flexible cognitive resources: competitive content maps for attention and memory. *Trends Cogn. Sci.* **17**, 134–141 (2013).
46. Sprague, T. C., Ester, E. F. & Serences, J. T. Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load. *Curr. Biol.* **24**, 2174–2180 (2014).
47. Rademaker, R. L., Bloem, I. M., De Weerd, P. & Sack, A. T. The impact of interference on short-term memory for visual orientation. *J. Exp. Psychol. Hum. Percept. Perform.* **41**, 1650–1665 (2015).
48. Mallett, R., Mummaneni, A. & Lewis-Peacock, J. A. Distraction biases working memory for faces. *Psychon. Bull. Rev.* **27**, 350–356 (2020).
49. Teng, C. & Kravitz, D. J. Visual working memory directly alters perception. *Nat. Hum. Behav.* **3**, 827–836 (2019).
50. Kang, M.-S., Hong, S. W., Blake, R. & Woodman, G. F. Visual working memory contaminates perception. *Psychon. Bull. Rev.* **18**, 860–869 (2011).
51. Gayet, S., Paffen, C. L. E. & Van Der Stigchel, S. Information Matching the Content of Visual Working Memory Is Prioritized for Conscious Access. *Psychol. Sci.* **24**, 2472–2480 (2013).
52. Gayet, S. *et al.* Visual Working Memory Enhances the Neural Response to Matching Visual Input. *J. Neurosci.* **37**, 6638–6647 (2017).
53. Lorenc, E. S., Sreenivasan, K. K., Nee, D. E., Vandenbroucke, A. R. E. & D'Esposito, M. Flexible Coding of Visual Working Memory Representations during Distraction. *J. Neurosci.* **38**, 5267–5276 (2018).
54. Hallenbeck, G. E., Sprague, T. C., Rahmati, M., Sreenivasan, K. K. & Curtis, C. E. Working memory representations in visual cortex mediate distraction effects. *Nat. Commun.* **12**, 4714 (2021).
55. Iamshchinina, P. *et al.* Perceived and mentally rotated contents are differentially represented in cortical depth of V1. *Commun. Biol.* **4**, 1069 (2021).
56. Chang, C.-C. & Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 1–27 (2011).
57. Pilly, P. K. & Seitz, A. R. What a difference a parameter makes: A psychophysical comparison of random dot motion algorithms. *Vision Res.* **49**, 1599–1612 (2009).
58. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).
59. Bjorck, A. & Golub, G. H. Numerical Methods for Computing Angles Between Linear Subspaces. *Math. Comput.* **27**, 579–594 (1973).

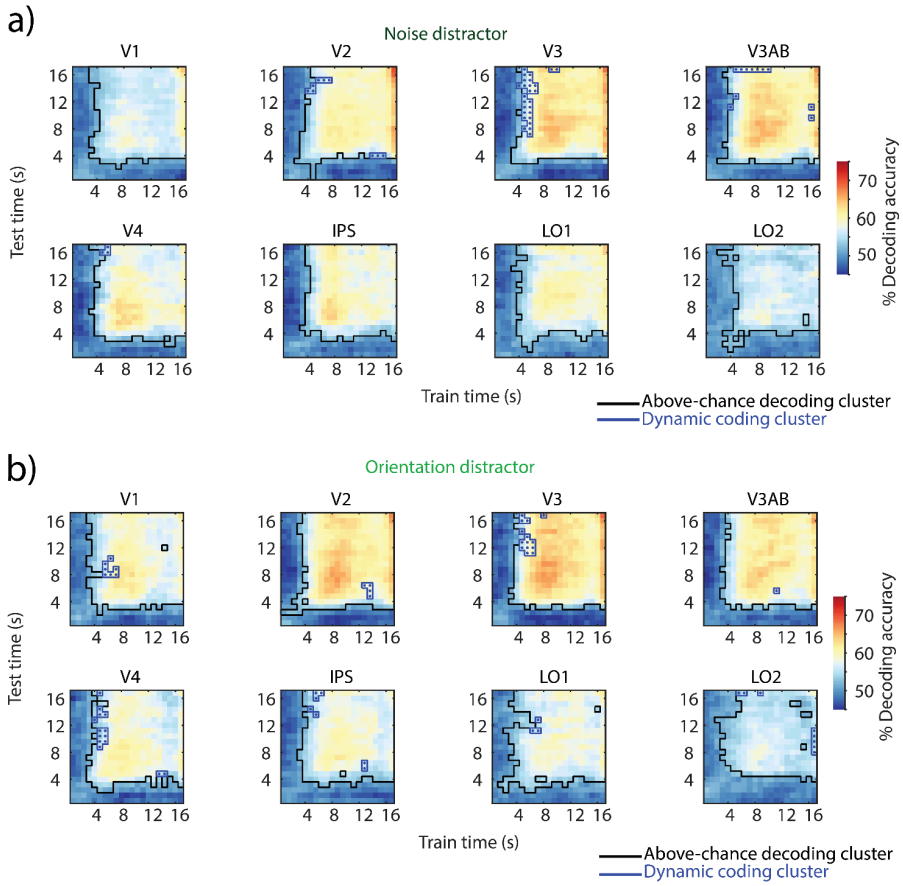
Supplementary Figures



Supplementary Figure 1. Neural trajectories across time. Same as Figure 2c), but the time dimension is on the z-axis.



Supplementary Figure 2. Extension of Figure 3 for V4-LO2.



Supplementary Figure 3. Temporal cross-decoding generalization between distractor and no-distractor VWM trials. a) Across-participant mean temporal cross-decoding of noise distractor trials when trained on no-distractor trials. **b)** Same as a), but orientation distractor trials trained on no-distractor trials.

Supplementary Table 1. FDR-corrected *p*-values corresponding to Figure 2d.

Time points	No distractor	Noise distractor	Orientation distractor
4.8 - 7.2s	0	0.032	0.031
4.8 - 9.6s	0.081	0	0.052
4.8 - 12s	0.092	0	0.052
4.8 - 14.4s	0.081	0	0
4.8 - 16.8s	0.031	0	0
7.2 - 9.6s	0.078	0.523	0.375
7.2 - 12s	0.031	0.067	0.2004
7.2 - 14.4s	0	0.031	0.081
7.2 - 16.8s	0.031	0	0.092
9.6 - 12s	0.289	0.0667	0
9.6 - 14.4s	0.158	0.067	0.031
9.6 - 16.8s	0.031	0	0.081
12 - 14.4	0.289	0	0.648
12 - 16.8s	0.031	0	0.067
14.4s - 16.8	0	0	0

Supplementary Table 2. FDR-corrected *p*-values corresponding to Figure 4b.

Test	V1	V2	V3	V3AB	V4	IPS	LO1	LO2
Noise-between baseline	0	0	0	0	0.167	0	0	0.085
Noise-within baseline	0	0.388	0.139	0.337	0.2001	0.027	0.079	0
Orientation-between baseline	0	0	0	0	0.0598	0	0.0258	0.0454
Orientation-within baseline	0	0.176	0.2002	0.084	0.0258	0	0	0
Noise generalization	0.092	0	0.092	0	0.454	0.246	0.107	0.378
Orientation generalization	0.246	0.092	0.0896	0.118	0.763	0.246	0.551	0.251