Report

Current Biology

Neural population dynamics of human working memory

Highlights

- Both stable and dynamic neural codes support visual spatial working memory (WM)
- Surprisingly, WM dynamics are greater in visual compared to frontoparietal cortex
- Neural dynamics were made interpretable by modeling population-level activity
- Reformatting of WM representations drives neural dynamics

Authors

Hsin-Hung Li, Clayton E. Curtis

Correspondence

clayton.curtis@nyu.edu

In brief

Li and Curtis use fMRI to investigate neural codes that support working memory (WM). Stability of WM varies across the brain, with early visual cortex showing the strongest dynamics. In early visual cortex, memorized target locations are reformatted into line-like patterns resembling memory-guided saccade trajectories, which drive neural dynamics.

Li & Curtis, 2023, Current Biology *33*, 3775–3784
September 11, 2023
https://doi.org/10.1016/j.cub.2023.07.067

CellPress

Current Biology

CellPress

Report

Neural population dynamics of human working memory

Hsin-Hung Li[1,2] and Clayton E. Curtis[1,2,3,*]
[1]Department of Psychology, New York University, New York, NY 10003, USA
[2]Center for Neural Science, New York University, New York, NY 10003, USA
[3]Lead contact
*Correspondence: clayton.curtis@nyu.edu
https://doi.org/10.1016/j.cub.2023.07.067

SUMMARY

The activity of neurons in macaque prefrontal cortex (PFC) persists during working memory (WM) delays, providing a mechanism for memory.[1–11] Although theory,[11,12] including formal network models,[13,14] assumes that WM codes are stable over time, PFC neurons exhibit dynamics inconsistent with these assumptions.[15–19] Recently, multivariate reanalyses revealed the coexistence of both stable and dynamic WM codes in macaque PFC.[20–23] Human EEG studies also suggest that WM might contain dynamics.[24,25] Nonetheless, how WM dynamics vary across the cortical hierarchy and which factors drive dynamics remain unknown. To elucidate WM dynamics in humans, we decoded WM content from fMRI responses across multiple cortical visual field maps.[26–48] We found coexisting stable and dynamic neural representations of WM during a memory-guided saccade task. Geometric analyses of neural subspaces revealed that early visual cortex exhibited stronger dynamics than high-level visual and frontoparietal cortex. Leveraging models of population receptive fields, we visualized and made the neural dynamics interpretable. We found that during WM delays, V1 population initially encoded a narrowly tuned bump of activation centered on the peripheral memory target. Remarkably, this bump then spread inward toward foveal locations, forming a vector along the trajectory of the forthcoming memory-guided saccade. In other words, the neural code transformed into an abstraction of the stimulus more proximal to memory-guided behavior. Therefore, theories of WM must consider both sensory features and their task-relevant abstractions because changes in the format of memoranda naturally drive neural dynamics.

RESULTS

To facilitate direct comparisons with the existing macaque neurophysiological studies (e.g., Funahashi et al., Constantinidis et al., Spaak et al., Murray et al., and Wimmer et al.[3,5,8,20,21,49]), we used a memory-guided saccade task to study neural dynamics during spatial working memory (WM). In each trial, a brief target dot was presented in the periphery, followed by a 12-s delay. The polar angle of the target spanned the full circle pseudo-randomly. After the delay, participants reported the remembered location with a memory-guided saccade (Figure 1A). Participants were able to make precise memory reports close to the target (Figure 1B). Participants also underwent a pRF (population receptive field) mapping session,[50,51] allowing us to define four retinotopic visual (V1, V2, V3, and V3AB) and four parietal (IPS0, IPS1, IPS2, and IPS3) areas as the regions of interest (ROIs).

Coexistence of stable and dynamic WM codes

We first characterized the dynamics of the WM neural code by a temporal generalization analysis.[20,52,53] We found a stable WM code in all the ROIs (pink dashed lines in Figure 1C), quantified by the above-chance decoding performance among the off-diagonal elements of the cross-decoding matrices. That is, the location of the WM target remained decodable even when the training data and the testing data came from different time points. In addition, shortly after the delay onset (~2 s), WM content was decodable in all ROIs and remained so throughout the delay. As expected, we observed robust decoding performance along the on-diagonal elements in all ROIs. We next asked if WM representations changed over time by testing if the off-diagonal elements showed significantly worse decoding performance than their corresponding diagonal elements. In most ROIs, we observed clusters of off-diagonal elements exhibiting significantly reduced decoding performance, indicating a dynamic WM neural code (light blue lines in Figure 1C), which coexisted with the stable code.

The decodable signals we observed throughout the delay do not merely reflect a slow decay of sensory-evoked hemodynamic responses, but instead rely on WM maintenance. We conducted a passive viewing experiment in which a high-contrast flickering peripheral "WM target," treated as an irrelevant stimulus by the subjects, was present continuously throughout the delay and thus did not require WM (Figure S1A). In this case, we only observed a stable code without significant dynamics, and the peripheral WM target was only decodable for a much shorter period of time early in the delay without persisting through the delay (Figure S1B). In addition,

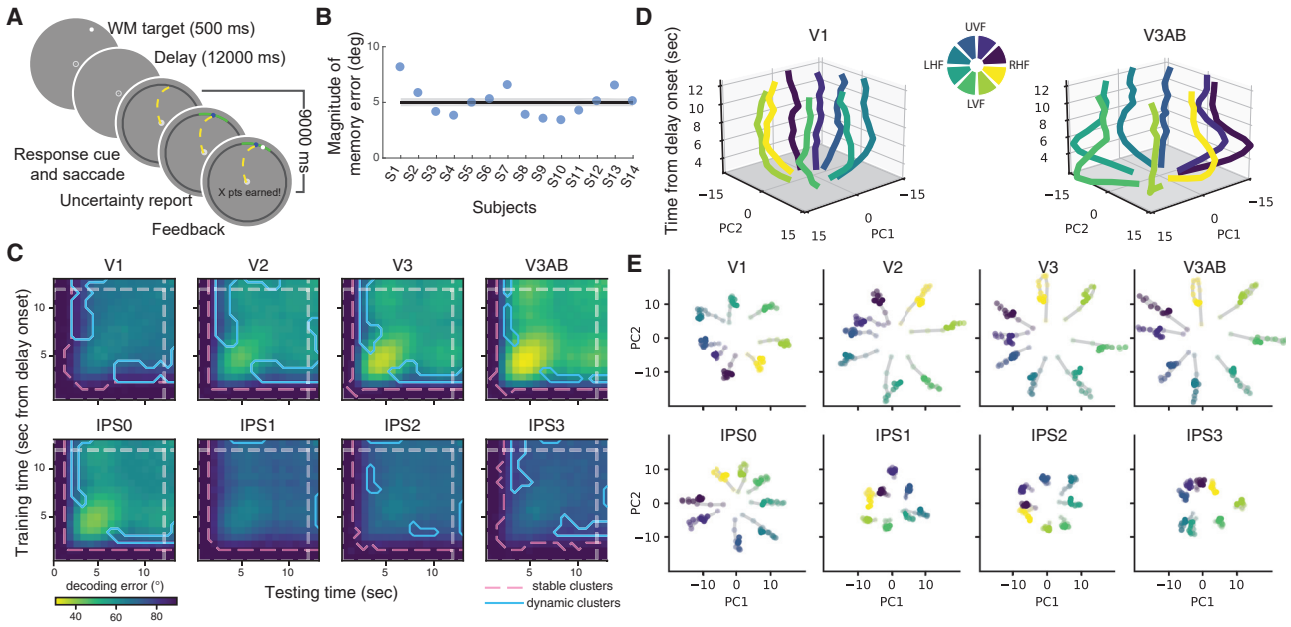

Figure 1. Task, behavior, temporal generalization, and stable subspaces

(A) Task. Participants maintained fixation while remembering the location of a target dot presented at a pseudorandom polar angle and at 12° eccentricity from fixation. At the end of each trial, participants made memory-guided saccades to report their memory and adjusted the length of an arc to report uncertainty.

(B) The mean magnitude (absolute value) of memory error of each participant. The horizontal line represents the group mean with ±1 SEM.

(C) The temporal generalization analysis. Decoders were trained to decode the target location from the fMRI response. The decoders trained with voxel activity patterns of each time point were tested with the data of all the time points. Pink dashed lines: the stable clusters, the cluster that exhibited above-chance decoding performance. Blue solid lines: the dynamic clusters, the off-diagonal elements that exhibited lower decoding performance than (both of) their corresponding diagonal elements. Gray dashed lines: the onset of the response cue.

(D) V1 and V3AB response at each time point during the delay period was projected into the stable subspace, the top two principal components (x axis and y axis) obtained from PCA. The z axis represents time from delay onset. Each curve represents one bin of target location. The legend indicates the location, in the visual field, represented by each of the eight colors.

(E) Similar to (D) but without visualizing the time axis (z axis); therefore, it is equivalent to the bird's eye view of (D). Each dot represents data from one time point during the delay, with more saturated colors representing later time points.

See also Figures S1A, S1B, and S4B.

with the same dataset, we had previously shown that the BOLD signals in the WM experiment can predict behavioral memory error and reported memory uncertainty, supporting that the representations we analyzed are related to WM-guided behaviors.[46]

Neural subspaces

We used principal component analysis (PCA) to visualize the neural subspaces that represent memorized locations. As an approach complementary to temporal generalization, PCA allows us to further (1) quantify how target locations are encoded topologically, reflecting their relationships in the visual field space; (2) estimate the geometry underlying the dynamics; and (3) compare the stability across brain regions.

A coding subspace within which target locations maintain their relative positions throughout the delay would be indicative of a stable code. We characterized such a stable subspace by applying PCA on time-averaged BOLD responses, discarding the dynamical aspect of the data.[21] We projected the data of each single time point during the delay into this stable subspace and found that in this subspace, target locations were topologically organized, preserving their spatial relationships across time during the delay (Figures 1D and 1E).

We next investigated the dynamic aspects of the neural code by dividing the 12-s delay into three—the early, middle, and late—time windows with equal length and computing their corresponding neural subspaces. In a type of rotational dynamic, the same neural subspace is fixed but targets rotate over time on that space.[54–56] We did not observe such rotations when projecting data of individual time points (Figure 2A), or each time window (Figures 2B, 2C, and S2A), into the early, middle, or late subspace. Instead of rotating on a subspace, the spread of the target (or variance explained) shrank when projecting the data of one time point into the subspace at the other time points. These results indicate that WM dynamics are driven by a changing neural subspace (Figure 2D), where different neural populations represent target locations at different times.

To compare the stability of WM codes across brain regions, we quantified how much the neural subspace changed over time by the principal angle between the subspaces of the early and late time windows within each ROI (STAR Methods).[57–59] We found that the stability of WM representations, quantified by principal angles, varied across ROIs ($F(7, 91) = 3.88$, $p < 0.00$; Figure 2E). We further conducted a trend analysis to investigate how WM stability varies across cortical hierarchy. We found a significant linear trend ($\beta = -5.54$, $p < 0.05$),

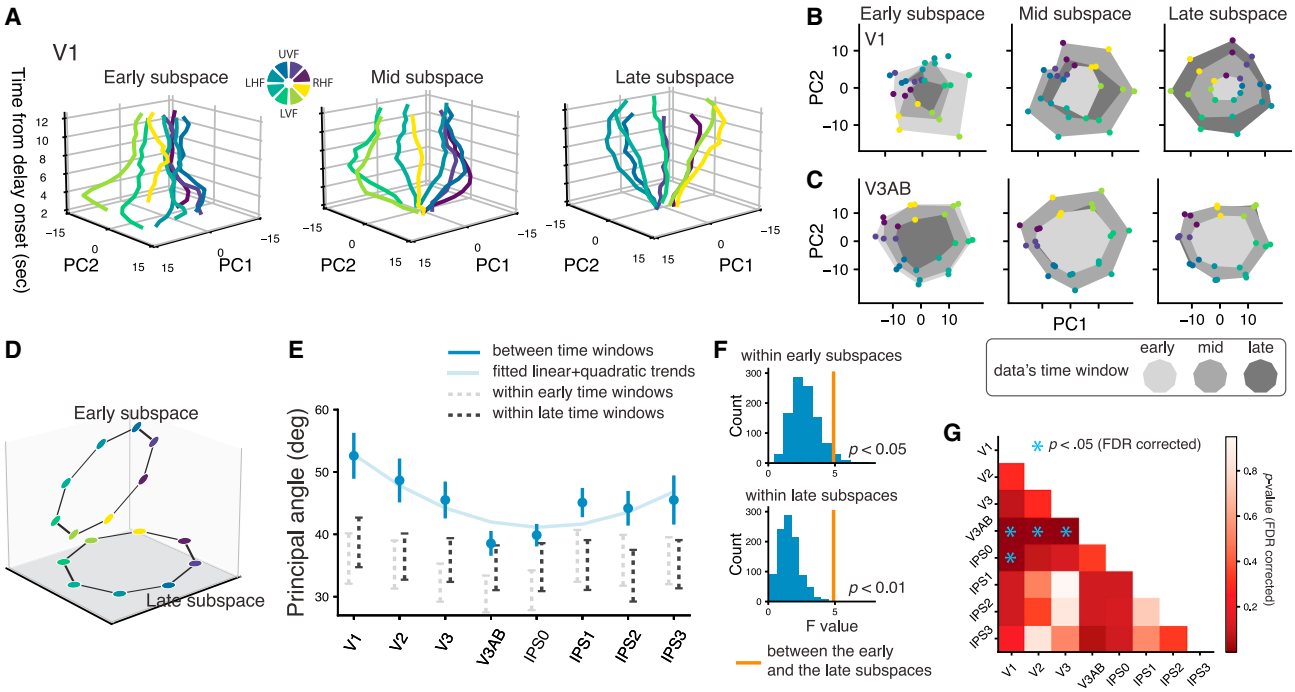

Figure 2. Dynamic subspaces

Early, middle, and late neural coding subspaces were identified by applying PCA to voxel activity patterns in three different time windows.

(A) V1 response of each time point during the delay was projected into the early, mid, and late subspaces, where the top two principal components (x axis and y axis) were derived from PCA. The z axis represents the time from delay onset. Each curve represents one bin of the target locations.

(B) Projections of the early, mid, and late (data points connected with areas with three different gray levels) voxel activity patterns of V1 into the early, mid, or late subspaces (from the late to right panel).

(C) Same as (B) but for V3AB. See Figure S2A for the projections of other ROIs.

(D) Cartoon illustration of two subspaces from different time points that encode the WM target locations in the high-dimensional space.

(E) The dark blue data points represent the principal angle, between the early and late subspaces, of each ROI (group mean with ±1 SEM). The dashed lines represent 95% confidence intervals of the distribution of the principal angles computed between the two subspaces estimated by resampling the data from either the same early (light gray dashed lines) or the late (dark gray dashed lines) time windows.

(F) The main effect of ROI on the principal angle between the late and the early subspaces is larger than that predicted by the principal angle within the same time window. Top: by a bootstrapping procedure, in each iteration, we resampled the data, estimated two early subspaces, and computed the principal angle between them. We then computed the main effect of ROI on the principal angle by ANOVA, leading to a bootstrapped distribution of the F values (blue histograms). p value was computed by comparing the bootstrapped distributions with the empirical F value computed using the principal angle between the early and the late subspaces (orange vertical lines). Bottom: same as the top figure, but the bootstrapped distribution was computed by resampling the late subspaces.

(G) Pairwise comparisons of the principal angles. After computing the principal angle of each ROI, we tested whether the angle of an ROI was smaller/larger than that of the other.

See also Figure S2.

indicating increasing stability with cortical hierarchy, and a quadratic trend ($\beta = 8.92$, $p < 0.01$), explaining that the principal angle exhibited a dip around V3AB and IPS0. Pairwise comparisons between individual ROIs showed significant differences when comparing early visual cortex against V3AB and IPS0 (Figure 2G). We obtained similar results when comparing the stability using the ratio of variance explained (Figure S2B). We also computed the principal angles between two subspaces that were estimated by the data resampled within the same time window (dashed lines in Figure 2E; also see Figure S2C, in which data were split into two halves for comparing between-time and within-time principal angles). As expected, these baseline values were smaller than the angles computed between time windows, and they did not show variations across ROIs like what we observed for the principal angles computed across time (Figures 2E and 2F). Thus, the stability of WM differed across ROIs, which cannot be explained by factors such as the reliability of subspace estimation.

Factors driving WM dynamics

Although previous electrophysiological studies in macaques[15–21] and EEG studies in humans[24,25] provided descriptive evidence for neural dynamics during WM, the reasons for such dynamics remain unknown. To understand and interpret the WM dynamics we observed, we leveraged models of each voxel's receptive field to compute activation maps by projecting voxel activity patterns into the coordinates of two-dimensional visual field space (STAR Methods). Here, we focus on two ROIs—V1 and V3AB—where we observed the strongest and the weakest dynamics (see Figure S3 for other ROIs).

In V1, the spatial pattern of the population neural response showed clear changes across time. The response first emerged

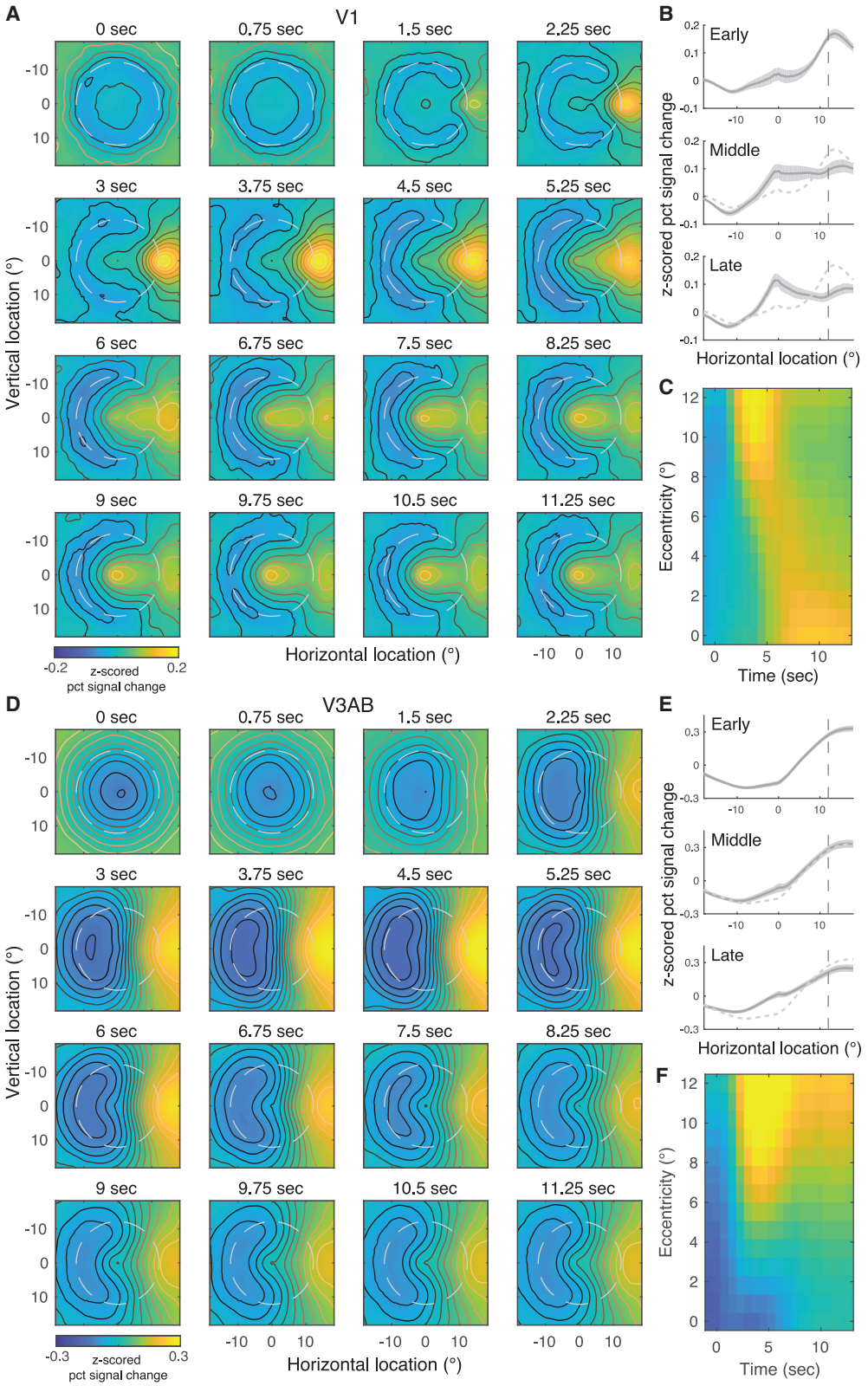

Figure 3. Visualization of WM dynamics in V1 and V3AB

(A) V1's activation maps visualize the projection of voxel activity patterns onto two-dimensional visual field space. Each image represents the activation map reconstructed for each time sample (TR) during the delay.

(legend continued on next page)

at the target's polar angle and eccentricity, consistent with previous fMRI studies that visualized neural responses in the visual field space.[40,48,60–62] After about 4.5 s, the response at the target's eccentricity declined and spread inward across the visual field in a line between the target and foveal locations (Figures 3A and 3B). Toward the end of the delay, activity peaked at the foveal region with a tail pointing toward the target. We observed sequential activity when binning the locations on the activation maps (at target's polar angle) by eccentricity—response at the target's (far) eccentricity exhibited the earliest peak, followed by intermediate locations, while the foveal locations peaked at the latest time points (Figure 3C). Overall, the neural WM code changed from a bump at the target into a line-like pattern along the trajectory of the planned memory-guided saccade. In V3AB, the ROI with the greatest stability, we found that the peak of activation remained at the target's peripheral location over the course of the trial (Figures 3D–3F). Critically, when visualizing the activation maps for V1 during the passive viewing experiment, we did not observe dynamics like those in the WM experiment. The response emerged at the target's peripheral location at the early time points and diminished in the middle of the delay (Figure S1C). This confirms that the population dynamics we observed in V1 were not due to sluggish hemodynamics.

We considered two factors that might contribute to the neural dynamics observed in V1. First, the response cue in the experiment consisted of a change of color of the fixation point and the onset of a ring (Figure 1A). V1's response in the late time window might reflect the anticipation of the response cue at the fixation. We conducted the same experiment but with a response cue consisting only of the peripheral ring without changes on the fixation point. We observed the same results when visualizing V1's population neural response (Figure S1D). Thus, the anticipation of the response cue at fixation was not required to drive the neural dynamics we observed. Second, even when maintaining fixation during the delay, gaze positions may still exhibit bias toward the target locations.[63] By sorting the trials based on the relative position between the target and the gaze position, we found that the neural response in V1 exhibited the same neural dynamics—an emergence of the line-like patterns over time—regardless of whether the gaze position exhibited bias toward or away from the target (Figures S1E–S1G).

Besides eccentricity, changes of neural population's selectivity for polar angle could also contribute to WM dynamics. We computed polar angle response functions by collapsing the two-dimensional activation maps across eccentricity (Figures 4A and 4B). We fitted von Mises functions to the response function at each time point with the center, gain, and width as the free parameters. All ROIs had response functions that centered around the polar angle of the target. For both V1 and V3AB, the gain of the response function peaked at about 4–5 s and remained above zero throughout the delay. The width of the functions showed different dynamics between the two ROIs (Figure 4E). In V3AB, the width remained stable once it reached the lowest level (similar patterns were observed in V3 and all the ROIs in IPS; Figure S4A). In contrast, the width in V1 first reached a lower value than V3AB, which is expected as V1 has a smaller pRF when measured by retinotopic mapping procedures,[51,64] but then increased and became similarly wide as V3AB. These dynamics may indicate that the neural activity observed in V1 reflects feedback signals during the later time points of the delay (see discussion).

Neural code during WM is stable in PFC

Previous studies of the neural population codes during WM in nonhuman primates have largely focused on the neural activity in prefrontal cortex (PFC).[11] When extending our analyses to two frontal regions, iPCS (inferior precentral sulcus) and sPCS (superior precentral sulcus), the regions where we previously observed topographic organization in the pRF mapping session,[51] we found decodable WM content in both frontal regions, but their decoding errors were larger than other ROIs and we only observed stable codes without significant dynamic clusters (Figure S4B). Projecting the voxel activity pattern of each time point into the stable subspace extracted by PCA, we found that the locations of the targets remained largely stable and separable within the stable subspace (Figure S4C). However, their spatial topology was not as organized as those observed in the visual and the parietal cortices.

DISCUSSION

To summarize, by temporal generalization, dimensionality reduction, subspace geometric analyses, and pRF reconstruction of neural population response, we found converging evidence of coexisting stable and dynamic WM codes. The stability of WM varied across the cortical hierarchy in the human brain, with early visual cortex exhibiting the strongest dynamics. Our visualization techniques allowed us to interpret and reveal the latent factors driving neural dynamics during WM. The V1 population, which initially encoded a narrowly tuned bump of activation centered on the peripheral target, later represented a vector along the trajectory of the forthcoming memory-guided saccade. Thus, neural dynamics in V1 resulted from the format of the WM representation changing into a behaviorally relevant abstraction of the stimulus.

It might be tempting to dismiss these dynamics as simply the delayed hemodynamic response to the retinal stimulation caused by the target during encoding mixed with the responses during WM maintenance. Both theoretical and empirical reasons, however, suggest these dynamics are robust, meaningful, and important. The results from our temporal generalization and

(B) The horizontal slice over the activation map at vertical location = 0° for each time window (5 TRs per time window; TR at 0 s was not included). For comparison, the activity of the early time window was plotted as dashed curves in the middle and late time windows. Data represent mean ± SEM.
(C) Population response for different eccentricities as a function of time. Response is computed by averaging over the activation maps in (B) constrained within a sector spanning ±30° centered at the target's polar angle. To visualize response as a function of eccentricity, the sector is further segmented into bins from far to near eccentricity in steps of 1°.
(D and E) Same as (A)–(C), but for V3AB.
See also Figures S1 and S3.

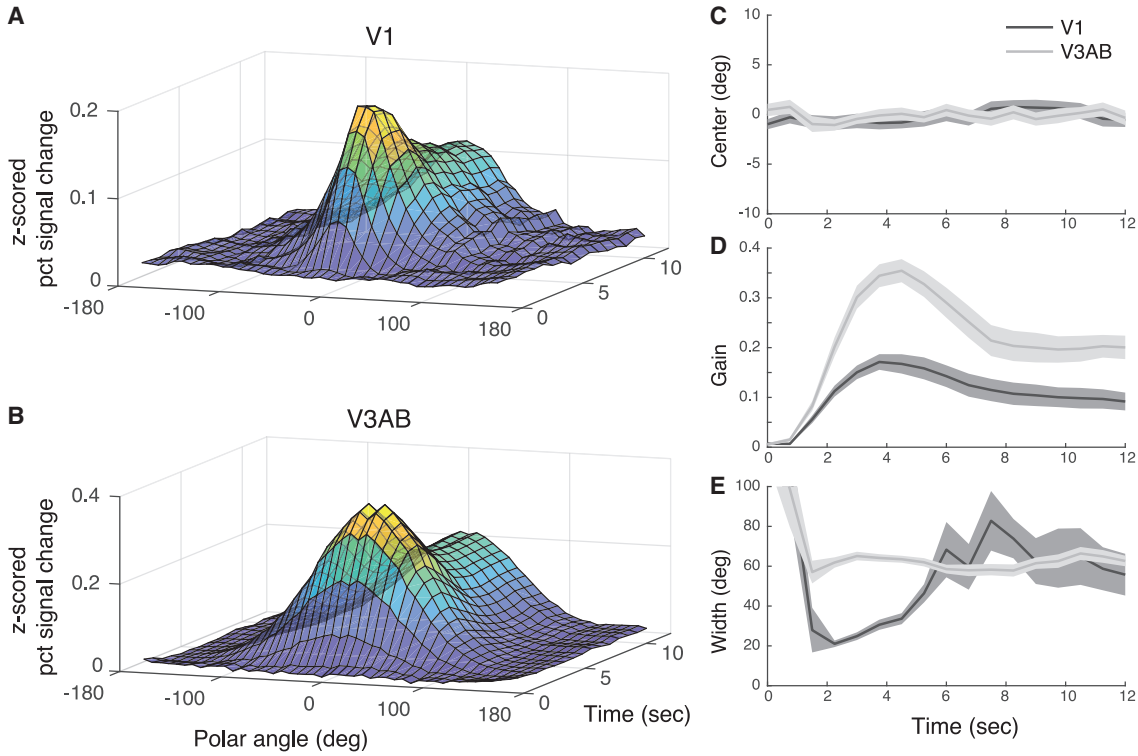

Figure 4. Polar angle response functions

(A) V1's polar angle response functions as a function of time from delay onset.

(B) Same as (A), but for V3AB.

(C) The center of the response functions. Data represent mean ± SEM.

(D) The gain of the response functions. Data represent mean ± SEM.

(E) The width of the response function. Data represent mean ± SEM.

See also Figure S4A.

dimensionality reduction analyses show striking similarities to those observed in electrophysiological recordings from macaque PFC during memory-guided saccade WM tasks. Applying the same temporal generalization analysis to data from Watanabe and Funahashi,[65] Spaak et al.[20] identified brief dynamics limited to the stimulus presentation and the beginning ~1 s of the delay period, which was then followed by stable WM code that persisted during the delay. Similarly, in a reanalysis of Constantinidis et al.[66] and Romo et al.,[7] Murray et al.[21] used dimensionality reduction to identify a dynamical subspace that only outperformed a stable subspace in representing stimulus locations during the stimulus presentation and early part of the delay. Therefore, one common source of dynamics identified in previous electrophysiological studies and the present study involves the transition from perceptual encoding to WM maintenance. Current theoretical and computational models of WM do not account for this transformation. Overall, these results explain why classifiers trained on fMRI patterns during visual stimulation often show lower decoding performance during WM maintenance.[27,47,48] Moreover, WM representations remain decodable in the presence of visual distractors,[42,45,47] suggesting that because these representations do not cause catastrophic interference, they perhaps have different representational formats.[48,54,67,68]

By projecting voxel activity in V1 into the visual field space, we found that the stimulus is reformatted into a representation that is more proximal to the behavior guided by the memory: in this case, activity at fixation with a line along the trajectory of the future memory-guided saccade (Figures 3A–3C). Although the current results do not explain the mechanisms underlying the line-like representation we observed in V1, we can offer some intriguing possibilities. The location of saccade targets can be decoded from delay period activity in the human superior colliculus (SC) indicating that participants are rehearsing the forthcoming memory-guided saccade.[43] The SC represents saccade targets as the error between the current eye position and the target, where a hill of activity in SC propagates along a path from the saccade target to fixation as the eye moves.[69–71] Feedback signals carrying a corollary discharge of the rehearsed saccades could be routed through the putamen to target V1, producing a line along the rehearsed trajectory. Although effects are unknown in V1, saccades also produce traveling waves of neural activity that propagate from fovea to the saccade target in V4,[72] while classical receptive fields remap to future receptive fields by sweeping across the visual field in LIP.[73] In addition, neurons in SC show build-up activity during saccade preparation, and many of these cells are selective for the direction, but not specific endpoint, of saccades.[69] These possibilities can

be investigated in future research. Interestingly, our findings in early visual cortex are consistent with recent neurophysiological studies in mice motor cortex[74] and monkey's PFC,[75] showing that neural activity for motor planning emerged during WM delay and drove neural dynamics.

Although the hypothesis that the line encoded in the population response in V1 represents the forthcoming saccade trajectory is the simplest explanation, we considered others. First, central support for our reformatting hypothesis is our finding that the line begins at the target location, then propagates inward to the foveal part of the map. A sluggish hemodynamic response to the target stimulus cannot account for these dynamics. In our control experiment (Figures S1A–S1C), when a memory-guided saccade was not planned to the peripheral target, we found no evidence for a line-like representation despite the salience of the flickering peripheral stimulus and the attention-demanding task at central fixation. Second, previous studies have found that neurons' receptive fields show a "convergent shift" toward saccade targets.[76–79] Even though one could interpret our findings in V1 as a form of receptive field shift, there are major differences between the two sets of findings: the perisaccadic shift of receptive fields was found to be time-locked to saccade onsets with a narrow time window (±100 ms), whereas the dynamics here were observed several seconds before the response cue. In addition, convergent receptive field shift occurred to all the neurons regardless of their receptive field centers relative to the saccade target[76,78,79]; the effects here only involved the voxels between the fixation and the saccade target. Last, we did not exclude that the neural dynamics we observed are related to the dynamics of spatial attention as WM and attention are often considered to be intertwined.[80,81] However, simply adding an attentional focus[82–84] at the fixation cannot explain the line-like pattern we observed and how neural activity propagated between the target and fixation. Instead, during the delay, attention might spread across the fixation and the target, leading to a pattern similar to the upcoming saccade trajectory. In general, this interpretation is consistent with the findings that saccade preparation and attention are partially coupled.[85–89]

We acknowledge potential limitations in using fMRI to study neural dynamics. Modeling fast and more complicated neural dynamics is likely beyond the scope of fMRI given the sluggishness of the hemodynamics. However, if one uses slow, event-related paradigms with long delay periods combined with fast BOLD acquisitions (750 ms time sample [TR]) like that used here, certain dynamics can still be measured with caution. We confirmed this by simulating neural dynamics similar to those reported in Spaak et al. and Murray et al. and found that such dynamics can still be characterized after considering the slowness of hemodynamic response function and the temporal resolution of data acquisition here (Figure S4D).

Feedback signals to early visual cortex likely underlie the WM representations we measured. When quantifying the polar angle response function evoked by the WM target, the widths of the functions in early visual cortex widened over time. In contrast, the widths in higher-level cortical regions were constant throughout the delay. These results suggest that WM-related activation in early visual cortex may rely on feedback signals originating from the downstream regions. During target encoding, polar angle response functions in early visual cortex were narrow, owing to the small receptive field sizes in these areas. Feedback signals originating from higher-level cortical areas, with larger receptive fields, might broaden the widths of the response functions in early visual cortex during WM. A shift from bottom-up sensory signals to top-down WM-related feedback may contribute to the dynamics in early visual cortex. These results are consistent with laminar recording in macaque V1[90] and fMRI measurements of human V1[40,48,86] showing that the persistent activity during WM in V1 has top-down origins. Moreover, WM representations in early visual cortex are surprisingly flexible. They can, for instance, represent WM target locations that were never retinally stimulated, but have been geometrically remapped.[40,86] More broadly, when compared to the precision during stimulus encoding, both mental imagery[91] and episodic memory[62] evoke wider topographic responses in early visual cortex consistent with feedback signals originating from higher cortical areas.

Distinct from the neurophysiological results in macaque PFC, we find evidence for coexisting stable and dynamic WM codes in early visual cortex, not PFC. While we cannot know whether these differences are due to differences in species or in measurements of neural activity, they parallel a vast literature in humans showing the importance of early visual cortex for WM.[11] Our results revealed that even during the simplest of WM tasks, different brain regions represent WM content in different formats. Neural responses in higher-level visual cortex and parietal cortex almost exclusively represented the memorized locations. In contrast, representations in early visual cortex appear to concurrently reflect both a retrospective code (i.e., the target location) and a reformatted prospective code (i.e., memory-guided saccade trajectory).

Recently, we demonstrated that voxel activity patterns in visual cortex are recoded into a line-like spatial coding scheme when subjects are asked to remember the orientation of a grating or direction of moving dots.[48] Previous neurophysiological studies on monkeys' PFC observed neural dynamics of WM when a neural subspace that represents motor preparation emerged late in the delay during memory-guided saccades,[75] or when memorized items were projected into a different neural subspace once they were selected to guide visual search.[10] These findings, together with our current findings that visual targets are transformed into representations resembling saccade trajectories, indicate that WM representations are surprisingly agile and efficiently adapt to behavioral demands (also see Myers[92]). By implementing different task demands, future studies will solve how the dynamics of WM neural code are affected by the level of abstraction or reformation of the perceptual inputs required in memory-guided behaviors.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2023.07.067.

ACKNOWLEDGMENTS

We thank New York University's Center for Brain Imaging for technical support. This research was supported by the National Eye Institute (R01 EY-016407, R01 EY-027925, and R01 EY-033925 to C.E.C.). H.-H.L. was supported by the Swartz Foundation Postdoctoral Fellowship.

AUTHOR CONTRIBUTIONS

H.-H.L. and C.E.C. conceptualized and designed the research. H.-H.L. conducted the experiments and analyzed the data. H.-H.L. and C.E.C. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no conflict of interest.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: February 27, 2023
Revised: June 20, 2023
Accepted: July 31, 2023
Published: August 17, 2023

REFERENCES

1. Fuster, J.M., and Alexander, G.E. (1971). Neuron activity related to short-term memory. Science *173*, 652–654. https://doi.org/10.1126/science.173.3997.652.
2. Kubota, K., and Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance in monkeys. J. Neurophysiol. *34*, 337–347.
3. 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.
4. Miller, E.K., Erickson, C.A., and Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. J. Neurosci. *16*, 5154–5167.
5. Funahashi, S., Bruce, C.J., and Goldman-Rakic, P.S. (1993). Dorsolateral prefrontal lesions and oculomotor delayed-response performance: evidence for mnemonic "scotomas.". J. Neurosci. *13*, 1479–1497.
6. Funahashi, S., Chafee, M.V., and Goldman-Rakic, P.S. (1993). Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. Nature *365*, 753–756.
7. Romo, R., Brody, C.D., Hernández, A., and Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. Nature *399*, 470–473.
8. Constantinidis, C., Franowicz, M.N., and Goldman-Rakic, P.S. (2001). The sensory nature of mnemonic representation in the primate prefrontal cortex. Nat. Neurosci. *4*, 311–316.
9. Wasmuht, D.F., Spaak, E., Buschman, T.J., Miller, E.K., and Stokes, M.G. (2018). Intrinsic neuronal dynamics predict distinct functional roles during working memory. Nat. Commun. *9*, 3499.
10. Panichello, M.F., and Buschman, T.J. (2021). Shared mechanisms underlie the control of working memory and attention. Nature *592*, 601–605.
11. Curtis, C.E., and Sprague, T.C. (2021). Persistent activity during working memory from front to back. Front. Neural Circuits *15*, 696060.
12. Riley, M.R., and Constantinidis, C. (2015). Role of prefrontal persistent activity in working memory. Front. Syst. Neurosci. *9*, 181.
13. Compte, A., Brunel, N., Goldman-Rakic, P.S., and Wang, X.J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cereb. Cortex *10*, 910–923.
14. Wang, X.-J. (2021). 50 years of mnemonic persistent activity: quo vadis? Trends Neurosci. *44*, 888–902.
15. Brody, C.D., Hernández, A., Zainos, A., and Romo, R. (2003). Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. Cereb. Cortex *13*, 1196–1207.
16. Shafi, M., Zhou, Y., Quintana, J., Chow, C., Fuster, J., and Bodner, M. (2007). Variability in neuronal activity in primate cortex during working memory tasks. Neuroscience *146*, 1082–1108.
17. Harvey, C.D., Coen, P., and Tank, D.W. (2012). Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature *484*, 62–68.
18. Baeg, E.H., Kim, Y.B., Huh, K., Mook-Jung, I., Kim, H.T., and Jung, M.W. (2003). Dynamics of population code for working memory in the prefrontal cortex. Neuron *40*, 177–188.
19. Barak, O., Tsodyks, M., and Romo, R. (2010). Neuronal population coding of parametric working memory. J. Neurosci. *30*, 9424–9430.
20. Spaak, E., Watanabe, K., Funahashi, S., and Stokes, M.G. (2017). Stable and dynamic coding for working memory in primate prefrontal cortex. J. Neurosci. *37*, 6503–6516. https://doi.org/10.1523/jneurosci.3364-16.2017.
21. Murray, J.D., Bernacchia, A., Roy, N.A., Constantinidis, C., Romo, R., and Wang, X.-J. (2017). Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. Proc. Natl. Acad. Sci. USA *114*, 394–399.
22. Parthasarathy, A., Herikstad, R., Bong, J.H., Medina, F.S., Libedinsky, C., and Yen, S.-C. (2017). Mixed selectivity morphs population codes in prefrontal cortex. Nat. Neurosci. *20*, 1770–1779.
23. Cavanagh, S.E., Towers, J.P., Wallis, J.D., Hunt, L.T., and Kennerley, S.W. (2018). Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. Nat. Commun. *9*, 3498.
24. Wolff, M.J., Ding, J., Myers, N.E., and Stokes, M.G. (2015). Revealing hidden states in visual working memory using electroencephalography. Front. Syst. Neurosci. *9*, 123.
25. Wolff, M.J., Jochim, J., Akyürek, E.G., Buschman, T.J., and Stokes, M.G. (2020). Drifting codes within a stable coding scheme for working memory. PLoS Biol. *18*, e3000625.
26. 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.
27. Harrison, S.A., and Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. Nature *458*, 632–635.
28. 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.

29. 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.

30. 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.

31. 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.

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

33. Xing, Y., Ledgeway, T., McGraw, P.V., and Schluppeck, D. (2013). Decoding working memory of stimulus contrast in early visual cortex. J. Neurosci. *33*, 10301–10311.

34. Sprague, T.C., Ester, E.F., and Serences, J.T. (2014). Reconstructions of information in visual spatial working memory degrade with memory load. Curr. Biol. *24*, 2174–2180.

35. Sreenivasan, K.K., Vytlacil, J., and D'Esposito, M. (2014). Distributed and dynamic storage of working memory stimulus information in extrastriate cortex. J. Cogn. Neurosci. *26*, 1141–1153.

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

37. Sprague, T.C., Ester, E.F., and Serences, J.T. (2016). Restoring latent visual working memory representations in human cortex. Neuron *91*, 694–707.

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

39. Christophel, T.B., Klink, P.C., Spitzer, B., Roelfsema, P.R., and Haynes, J.-D. (2017). The distributed nature of working memory. Trends Cogn. Sci. *21*, 111–124.

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

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

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

43. Rahmati, M., DeSimone, K., Curtis, C.E., and Sreenivasan, K.K. (2020). Spatially specific working memory activity in the human superior colliculus. J. Neurosci. *40*, 9487–9495.

44. Brissenden, J.A., Tobyne, S.M., Halko, M.A., and Somers, D.C. (2021). Stimulus-specific visual working memory representations in human cerebellar lobule VIIb/VIIIa. J. Neurosci. *41*, 1033–1045.

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

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

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

48. Kwak, Y., and Curtis, C.E. (2022). Unveiling the abstract format of mnemonic representations. Neuron *110*, 1822–1828.e5.

49. Wimmer, K., Nykamp, D.Q., Constantinidis, C., and Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat. Neurosci. *17*, 431–439.

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

51. Mackey, W.E., Winawer, J., and Curtis, C.E. (2017). Visual field map clusters in human frontoparietal cortex. eLife *6*, e22974, https://doi.org/10.7554/eLife.22974.

52. Stokes, M.G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., and Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. Neuron *78*, 364–375.

53. King, J.-R., Pescetelli, N., and Dehaene, S. (2016). Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. Neuron *92*, 1122–1134.

54. Libby, A., and Buschman, T.J. (2021). Rotational dynamics reduce interference between sensory and memory representations. Nat. Neurosci. *24*, 715–726.

55. Wan, Q., Menendez, J.A., and Postle, B.R. (2022). Priority-based transformations of stimulus representation in visual working memory. PLoS Comput. Biol. *18*, e1009062.

56. Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., and Shenoy, K.V. (2012). Neural population dynamics during reaching. Nature *487*, 51–56.

57. Björck, Å., and Golub, G.H. (1973). Numerical methods for computing angles between linear subspaces. Math. Comput. *27*, 579–594.

58. Gallego, J.A., Perich, M.G., Naufel, S.N., Ethier, C., Solla, S.A., and Miller, L.E. (2018). Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. Nat. Commun. *9*, 4233.

59. Xie, Y., Hu, P., Li, J., Chen, J., Song, W., Wang, X.-J., Yang, T., Dehaene, S., Tang, S., Min, B., and Wang, L. (2022). Geometry of sequence working memory in macaque prefrontal cortex. Science *375*, 632–639.

60. Zhou, Y., Curtis, C.E., Sreenivasan, K.K., and Fougnie, D. (2022). Common neural mechanisms control attention and working memory. J. Neurosci. *42*, 7110–7120.

61. 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.

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

63. van Ede, F., Chekroud, S.R., and Nobre, A.C. (2019). Human gaze tracks attentional focusing in memorized visual space. Nat. Hum. Behav. *3*, 462–470.

64. Wandell, B.A., and Winawer, J. (2015). Computational neuroimaging and population receptive fields. Trends Cogn. Sci. *19*, 349–357.

65. Watanabe, K., and Funahashi, S. (2007). Prefrontal delay-period activity reflects the decision process of a saccade direction during a free-choice ODR task. Cereb. Cortex *17*, i88–i100.

66. Constantinidis, C., Franowicz, M.N., and Goldman-Rakic, P.S. (2001). Coding specificity in cortical microcircuits: a multiple-electrode analysis of primate prefrontal cortex. J. Neurosci. *21*, 3646–3655.

67. Kamiński, J., Sullivan, S., Chung, J.M., Ross, I.B., Mamelak, A.N., and Rutishauser, U. (2017). Persistently active neurons in human medial frontal and medial temporal lobe support working memory. Nat. Neurosci. *20*, 590–601.

68. Kornblith, S., Quian Quiroga, R., Koch, C., Fried, I., and Mormann, F. (2017). Persistent single-neuron activity during working memory in the human medial temporal lobe. Curr. Biol. *27*, 1026–1032.

69. Munoz, D.P., and Wurtz, R.H. (1995). Saccade-related activity in monkey superior colliculus. I. Characteristics of burst and buildup cells. J. Neurophysiol. *73*, 2313–2333.

70. Munoz, D.P., and Wurtz, R.H. (1995). Saccade-related activity in monkey superior colliculus. II. Spread of activity during saccades. J. Neurophysiol. *73*, 2334–2348.

71. Munoz, D.P., Pélisson, D., and Guitton, D. (1991). Movement of neural activity on the superior colliculus motor map during gaze shifts. Science *251*, 1358–1360.

72. Zanos, T.P., Mineault, P.J., Nasiotis, K.T., Guitton, D., and Pack, C.C. (2015). A sensorimotor role for traveling waves in primate visual cortex. Neuron *85*, 615–627.

73. Wang, X., Fung, C.C.A., Guan, S., Wu, S., Goldberg, M.E., and Zhang, M. (2016). Perisaccadic receptive field expansion in the lateral intraparietal area. Neuron *90*, 400–409.

74. Inagaki, H.K., Fontolan, L., Romani, S., and Svoboda, K. (2019). Discrete attractor dynamics underlies persistent activity in the frontal cortex. Nature *566*, 212–217.

75. Tang, C., Herikstad, R., Parthasarathy, A., Libedinsky, C., and Yen, S.-C. (2020). Minimally dependent activity subspaces for working memory and motor preparation in the lateral prefrontal cortex. eLife *9*, e58154, https://doi.org/10.7554/eLife.58154.

76. Tolias, A.S., Moore, T., Smirnakis, S.M., Tehovnik, E.J., Siapas, A.G., and Schiller, P.H. (2001). Eye movements modulate visual receptive fields of V4 neurons. Neuron *29*, 757–767.

77. Burr, D.C., Ross, J., Binda, P., and Morrone, M.C. (2010). Saccades compress space, time and number. Trends Cogn. Sci. *14*, 528–533.

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

79. Zirnsak, M., Steinmetz, N.A., Noudoost, B., Xu, K.Z., and Moore, T. (2014). Visual space is compressed in prefrontal cortex before eye movements. Nature *507*, 504–507.

80. Chun, M.M. (2011). Visual working memory as visual attention sustained internally over time. Neuropsychologia *49*, 1407–1409.

81. Kiyonaga, A., and Egner, T. (2013). Working memory as internal attention: toward an integrative account of internal and external selection processes. Psychon. Bull. Rev. *20*, 228–242.

82. Kastner, S., Pinsk, M.A., De Weerd, P., Desimone, R., and Ungerleider, L.G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. Neuron *22*, 751–761.

83. Gandhi, S.P., Heeger, D.J., and Boynton, G.M. (1999). Spatial attention affects brain activity in human primary visual cortex. Proc. Natl. Acad. Sci. USA *96*, 3314–3319.

84. Somers, D.C., Dale, A.M., Seiffert, A.E., and Tootell, R.B. (1999). Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. Proc. Natl. Acad. Sci. USA *96*, 1663–1668.

85. 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.

86. Saber, G.T., Pestilli, F., and Curtis, C.E. (2015). Saccade planning evokes topographically specific activity in the dorsal and ventral streams. J. Neurosci. *35*, 245–252.

87. Moore, T., and Zirnsak, M. (2017). Neural mechanisms of selective visual attention. Annu. Rev. Psychol. *68*, 47–72.

88. Li, H.-H., Hanning, N.M., and Carrasco, M. (2021). To look or not to look: dissociating presaccadic and covert spatial attention. Trends Neurosci. *44*, 669–686. https://doi.org/10.1016/j.tins.2021.05.002.

89. Li, H.-H., Pan, J., and Carrasco, M. (2021). Different computations underlie overt presaccadic and covert spatial attention. Nat. Hum. Behav. *5*, 1418–1431.

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

91. Breedlove, J.L., St-Yves, G., Olman, C.A., and Naselaris, T. (2020). Generative feedback explains distinct brain activity codes for seen and mental images. Curr. Biol. *30*, 2211–2224.e6.

92. Myers, N.E. (2022). Considering readout to understand working memory. J. Cogn. Neurosci. *35*, 11–13.

93. Kay, K.N., Winawer, J., Mezer, A., and Wandell, B.A. (2013). Compressive spatial summation in human visual cortex. J. Neurophysiol. *110*, 481–494.

94. Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods *164*, 177–190.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Data	This paper	https://osf.io/yc86q/
Software and algorithms		
MATLAB	MathWorks	https://www.mathworks.com/products/matlab.html
Python	Python	https://www.python.org/
Custom code and algorithm	This paper	https://osf.io/yc86q/

RESOURCE AVAILABILITY

Lead contact

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

Materials availability

The study did not produce new materials.

Data and code availability

- The processed fMRI and behavioral data generated in this study have been deposited in the Open Science Framework https://osf.io/yc86q/. Processed fMRI data contains extracted voxel activity of each ROI. The data used to plot figures in this paper (participant means) are provided in the Source Data file.
- All code for data analysis has been deposited in the Open Science Framework https://osf.io/yc86q/ and is publicly available as of the date of publication.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Fourteen participants joined the experiment. All participants had normal or corrected-to-normal vision. The experiments were conducted with the written, informed consent of each participant. The experimental protocols were approved by the University Committee on Activities involving Human Subjects at New York University, and participants received monetary compensation ($30/h).

METHOD DETAILS

Procedures

The details of the main experiment, the passive viewing experiment and the retinotopic mapping sessions have been previously reported in Li et al.[46] In the main experiment, participants performed a memory-guided saccade task in the fMRI scanner. Each trial started with the onset of the WM target, a light gray dot (0.65° diameter) with a duration of 500 ms. The target was at 12° eccentricity and the polar angle of the target was pseudo-randomly sampled from 1 of 32 locations evenly tiling a full circle. The target was followed by a 12-s delay period, during which the participants were required to maintain their gaze at the fixation point while remembering the location of the target dot. After the delay, the fixation point changed from a light gray circle into a gray filled dot, serving as the response cue. In addition, an black ring whose radius matches the eccentricity of the target was presented. Upon the onset of the response cue, participants reported the location of the target by making a saccadic eye movement onto the black ring. The reported location was first readout by the eye tracker, and a dot was presented at the reported location. Participants were allowed to further use a manual dial to adjust the reported location, and they pressed a button to finalize the memory report. Upon the button press, an arc centered at the reported location appeared on the ring. The participants were asked to use the manual dial to adjust the length of the arc in a post-estimation wager, in which they should reflect the uncertainty of their WM by the length of the arc, the longer the arc the more uncertain. Participants finalized the uncertainty report by a button press, after which a white dot was presented at the true target location as the feedback. Participants could earn points if the true target location fell within the arc (see details in Li et al.[46]). The results of uncertainty reports are detailed in an earlier study.[46]

A subset of participants (n = 6) joined an additional passive viewing experiment (Figure S1A). The timing of this experiment was similar to the main experiment. Instead of a brief and dim WM target stimulus, we presented a salient high-contrast flickering checkerboard (0.875 deg radius; 1 cycle/deg spatial frequency; 8 Hz flicker) at the same locations as the main experiment. The checkerboard was presented for 12.5 s (throughout the WM target period and the delay in Figure 1A), during which the fixation point, a '+' symbol, changed its width-height ratio, and the participants were asked to attend to the fixation and discriminate the changes of the fixation symbol, widening vs. heightening, by button presses. We adjusted the aspect ratio of the '+' symbol between runs so the participants' performance in the discrimination task maintained at about 75% accuracy. In addition, 6 participants joined a control experiment, where the procedures of the experiment was the same as the WM experiment, except that the response cue only consisted of the onset of the ring at the periphery (Figure S1D).

Each participant was scanned for a 1.5–2 h retinotopic mapping session. The procedures of the retinotopic mapping session followed those used in Mackey et al.[51] In short, during the retinotopic mapping session, participants maintained their gaze at the fixation point while a bar sweeping across the screen 12 times per run in various directions. Participants were required to attentively track the bar and perform a motion discrimination task based on the random dot kinematograms presented within the bar apertures see details in Mackey et al.[51] We fit a pRF model with compressive spatial summation to the BOLD time series of the retinotopic mapping session.[50,93] We projected the voxels' preferred phase angle and eccentricity on the cortical surface and defined ROIs by visual inspection (primarily based on the reversal of voxels' preferred phase angle). We define bilateral dorsal visual ROIs V1, V2, V3, V3AB, IPS0, IPS1, IPS2, IPS3 and two frontal ROIs, iPCS and sPCS.

Setup and eye tracking

Visual stimuli were presented by an LCD (VPixx ProPix) projector located behind the scanner bore. Participants viewed the stimuli through an angled mirror with a field of view of 52° by 31°. We presented a gray circular aperture (30° diameter) on the screen throughout the experiments. Eye position was recorded with a sampling rate at 500 Hz using an EyeLink 1000 Plus infrared video-based eye tracker (SR Research) mounted inside the scanner bore. We monitored gaze data and adjusted pupil/corneal reflection detection parameters as necessary during and/or between each run.

MRI acquisition

We acquired MRI data using a 64-channel head/neck coil on a Siemens Prisma 3T scanner. Functional imaging was collected for the working memory and passive viewing experiments using 44 slices and a voxel size of 2.53 mm (4x simultaneous-multi-slice acceleration; FoV was 200 × 200 mm; no in-plane acceleration; TE/TR: 30/750 ms, flip angle: 50 deg, Bandwidth: 2290 Hz/pixel; 0.56 ms echo spacing; P→ A phase encoding). Spin-echo images were acquired intermittently throughout each scanning session in the forward and reverse phase-encoding direction with identical slice prescription and no simultaneous-multi-slice acceleration (TE/TR: 45.6/3537 ms; 3 volumes per phase encode direction). These pairs are used to estimate a field map used to correct for local spatial distortions. The functional imaging data for retinotopic mapping was acquired in a separate session at a higher resolution, using a slice prescription spanning 56 slices (4x simultaneous-multislice acceleration) with a voxel size of 23 mm (FoV 208 × 208 mm, no in-plane acceleration, TE/TR: 36/1200 ms, flip angle: 66 deg, Bandwidth: 2604 Hz/pixel (0.51 ms echo spacing), P→ A phase encoding).

QUANTIFICATION AND STATISTICAL ANALYSIS

Behavioral data analysis

As participants were allowed to manually adjust the reported location after the saccades, the final dot location after the manual adjustment was used as the participants' memory report. Eye position was analyzed offline. The raw eye position data were first smoothed with a Gaussian kernel, and was converted into eye velocity using the eye positions of the five neighboring time points. Saccades were detected when the eye velocities exceeded the median velocity by 5 SDs with a minimum duration of 8 ms. Trials with ill-defined primary saccade, gaze positions deviated (>2.5°) from the fixation or with the magnitude of memory error larger than 3 standard deviations were excluded from analyses.

Even when participants were asked to maintain fixation, their gaze positions might still exhibit bias toward the WM targets[63] (also see Figure S1E). To investigate whether the neural dynamics we observed was V1 is driven by the gaze bias, we computed the averaged gaze position over the delay for each trial. We binned the trials into two bins based on whether the gaze position deviated toward or away from the target in a trial and visualized V1's neural dynamics for each bin (Figures S1F and S1G).

Temporal generalization

We decoded the location (polar angle) of the target from the BOLD response. We focused on the BOLD response measured from 0 to 13.5 s from the delay onset (18 TRs in total). Here, decoding is a regression problem where we aimed to predict the target location (polar angle) y from single-trial BOLD response X (an $n_{trial} \times n_{voxel}$ matrix) by estimating weights $\mathbf{w}$. As polar angle is a circular variable, we trained two regressions to predict $y_{sin} = sin(y)$ and $y_{cos} = \cos(y)$, and the predicted target location was computed as $\hat{y} = atan\,2(\hat{y}_{sin}, \hat{y}_{cos})$. Note that the decoder is trained to predict the target location (which is the same regardless of whether the testing data was at on- or off-diagonal in Figure 1C), rather than predict the voxel activity pattern of the same time point (the diagonal). Thus, it is not a given that the decoder would perform worse on the off-diagonal elements. We used support vector regression (with linear

kernel) in the scikit-learn Python library to estimate the weights. In a 10-fold cross-validation procedure, all the trials from a subject were separated as the training set (9/10 of the trials) and the testing set (1/10 of the trials). Regression weights were estimated using the training set's BOLD response of a particular time point. The weights were then applied to predict the target location using the testing trials' BOLD response. The testing was not only applied to the time point same as the time point of the training data, but was applied to all the time points of the data of the testing trials to investigate the generalizability of the decoders. Thereby, the training set and the testing set always involved data collected far away in time in terms of data acquisition. This is the case for on-diagonal and off-diagonal decoding (Figure 1C) so the dynamic code we observed can not be explained by neural or measurement noise that might exhibit systematic autocorrelation in time. The performance of the decoder was quantified as (the absolute value of) decoding errors averaged across all the trials for each participant. In Figure 1C, we report the mean decoding error averaged over participants.

Statistical tests for the stable and dynamic code

The stable code was represented by above-chance decoding even when the training and the testing data were from different time points. Thereby we defined the stable code as the elements in the cross-decoding matrix where the decoding error was smaller than 90° (the mean decoding error under the null hypothesis that the decoder generates random predictions uniformly distributed across the entire polar angle space). We first applied t-test to each element in the cross-decoding matrix to evaluate whether the decoding error was smaller than 90°. The neighboring elements that were significant in the t-test (without control for multiple comparisons) were grouped together as a cluster. For each cluster, we then computed the t-score summed across all the elements within. The summed t-scores was used to determine whether a cluster is significant by a permutation test. The permutation test was done by randomly permuting the decoders' predicted target location and computing the t-scores summed over the element in the most-significant cluster. This procedure was repeated 1000 times resulting in a null-hypothesis distribution of the summed t-scores, which was used to decide whether a cluster was statistically significant in the cross-decoding matrix. Details of this method are described in Maris and Oostenveld.[94]

We further investigated whether decoding performance in the off-diagonal elements is distinguishable from the on-diagonal. If at different time points, WM is represented by the same neural code (i.e., the two weight vectors at different time points are the same $\mathbf{w}_{t1} = \mathbf{w}_{t2}$), off-diagonal decoding would be statistically indistinguishable from that of on-diagonal decoding. On the other hand, we considered WM to exhibit dynamics if the off-diagonal elements had statistically poorer performance than their corresponding on-diagonal elements. Note that for an off-diagonal element to be included in a dynamic cluster, its decoding performance had to be statistically lower than both of its corresponding on-diagonal elements, ruling out the possibility that the dynamic clusters are defined simply because the neural or measurement noise was stronger at a specific time point. The cluster-based permutation test was done by randomly permuting the locations of the on- and off-diagonal elements. Overall, the definitions and the statistical tests for the stable and dynamic code were similar to those used in a previous monkey neurophysiological study.[20]

Stable and dynamic subspaces

We used PCA (principal component analysis) to define low-dimensional subspaces that encode target locations. We defined a data matrix $\mathbf{X}$ with a size of $n_{stimulus}$ by n_{voxel}, which represented the voxel activity patterns averaged across all the trials from the same stimulus location. For all the PCA conducted in this study, for the purpose of higher signal-to-noise ratio, we binned four neighboring target locations together resulting in 8 stimulus locations for the data matrix ($n_{stimulus} = 8$). n_{voxel} was the number of voxels of each ROI. $\mathbf{X}$ had column-wise zero mean, as the mean of each column was removed. When visualizing the subspaces (Figures 1D, 1E, 2A, and 2B), for each ROI, we concatenated the voxels across all participants, and applied the PCA on the participant-aggregated ROI. When computing the indices that quantified the stability of the subspaces—the principal angle and the ratio of variance explained—PCA was applied to individuals' ROIs and the indices were computed for each participant.

For the stable subspaces, we disregarded the time-varying information by averaging the data across all the time points that fell within the stable cluster during the delay (pink dashed lines in Figure 1C). Overall, all the time points within the delay were included except the first 2 or 3 TRs depending on the ROI. To obtain the principal components (PCs), we applied eigendecomposition on the covariance matrix $\mathbf{X}^T\mathbf{X} = \mathbf{W}\Lambda\mathbf{W}^T$, where each column of $\mathbf{W}$ was a unit-length vector, with a size of n_{voxel} by 1, representing the weights of each PC and Λ was a diagonal matrix containing the corresponding eigenvalues. Throughout this study, we used the first two PCs with the largest eigenvalues to define the subspaces; thereby we focused on the weight matrix $\mathbf{W}$ with a size of n_{voxel} by 2. To visualize the dynamics of population neural responses in the stable subspace, we projected the data of each time point into the stable subspace by computing $\mathbf{T} = \mathbf{XW}$, where $\mathbf{X}$ was the data matrix of a single time point and $\mathbf{X}$, with a size of 8 by 2, was the projection of the voxel activity pattern of each stimulus location in the subspace defined by the top two PCs, PC1 and PC2. $\mathbf{T}$ is often referred to as PC scores in the context of PCA.

To investigate the dynamical aspect of the neural subspaces, we binned the BOLD response during the delay into three time windows: early, middle and late time windows, each with 5 TRs (Figure 2B). We then estimated the subspace for each time window by applying PCA to the data matrices, $\mathbf{X}_e$, $\mathbf{X}_m$ and $\mathbf{X}_l$, which were the BOLD response averaged over each of the time windows.

To quantify how the early and the late subspaces oriented in the high-dimensional neural space, we computed the principal angle, which measured the alignment between two subspaces.[57–59] We computed the angle for each subject and each ROI, and we applied repeated-measures ANOVA to test the effect of ROIs on the principal angle. The principal angle was computed using the method proposed by Björck and Golub[57]: We applied singular decomposition to the inner-product matrix $\mathbf{W}_e^T\mathbf{W}_l = \mathbf{P}_e\mathbf{C}\mathbf{P}_l^T$, where $\mathbf{W}_e$ and

$\mathbf{W}_l$, both had a size of n_{voxel} by 2, were the weighting matrices of the early and the late subspaces obtained by PCA. The matrix $\mathbf{C}$ was a diagonal matrix whose diagonal elements were the ranked (from small to large) cosines of the principal angles θ_1 and θ_2: $\mathbf{C} = diag(\cos(\theta_1), \cos(\theta_2),)$. The first principal angle was reported in Figure 2E. We also computed the principal angle for two subspaces from the same time window (dashed lines in Figure 2E). This was achieved by a bootstrapping procedure with 1000 iterations. In each iteration, we resampled the trials twice, computed two subspaces using the data of the same time window and calculated the principal angle between them, resulting in a bootstrapped distribution of the principal angle within the early (or the late) time window. We also compared between-time principal angle and within-time principal angle by splitting each subject's data into two-halves. Thereby, the two subspaces used for computing the angles were always from independent datasets (Figure S2C).

In addition to the principal angle, we also computed the ratio of variance explained (RVE), which quantified how much the variance explained decreased when the data of a time window was projected to the subspace of a different time window. For example, for the data of the early time window $\mathbf{X}_e$, the RVE was computed as $\mathrm{Var}(\mathbf{X}_e\mathbf{W}_l)/\mathrm{Var}(\mathbf{X}_e\mathbf{W}_e)$ (Figure S2B).

Visualizations of WM representations

We used voxels' pRF parameters to visualize how neural populations represent remembered locations. To compute the activation maps (Figures 3 and S3), a grid was positioned in the visual field, centered at the fixation. The grid points evenly sampled the visual space with a step of 0.5° and the entire grid covered ±18° in both horizontal and vertical directions from the fixation point. We computed the neural activity a_i for the grid point i, whose coordinate in the visual field was (x_i, y_i), as the weighted sum of voxel responses $a_i = \frac{\sum_{v=1}^{n_{voxel}} \omega_{vi} r_v}{\sum_{v=1}^{n_{voxel}} \omega_{vi}}$. Here, γ_v was the response of voxel v. All the voxels within an ROI were included except the voxels whose responses in the retinotopic mapping session can not be well-fitted by the pRF model (thresholding at variance explained by the pRF model = 10%). ω_{vi} was the weight of voxel v at grid point i, which was determined by the density function of a bivariate circular Gaussian distribution $N(x_i, y_i; \mathbf{u}, \sigma^2\mathbf{I})$, in which $\mathbf{u}$ was the voxel's pRF center and σ was the voxel's pRF size. To compute an activation map for a time window and an ROI, we did the following steps: We computed an activation map for each trial, rotated the map of each trial to align the target at 0° polar angle, averaged map over all trials for each participant, and lastly we subtracted the grand mean from the activation map.

We computed each ROI's polar angle response function from the activation maps. The cartesian coordinates (x_i, y_i) of each grid point were first converted to phase angle θ_i and eccentricity r_i. We then binned the grid points based on their phase angle ranging from -180° to +180° with a step of 8°. We included the grid points with eccentricity smaller than 15°. The polar angle tuning function was computed as the activity averaged over the grid points within each bin. To quantify the dynamics of polar angle response functions, the tuning function of each time point was fitted by von Mises distributions $\frac{e^{k\cos(x-\mu)}}{2\pi I_0(k)}$, where $I_0(k)$ was the modified Bessel function of order 0. We reported the gain of the tuning function defined as the difference between the maximum and the minimum value of the best-fit tuning function, and the tuning width represented by the fitted k value, converted to have a unit in polar angle degree (Figures 4 and S4A).

Simulations

Our temporal generalization (Figure 1C) showed patterns similar to those observed in two previous neurophysiological studies in monkey's prefrontal cortex,[20,21] where the dynamical neural code was mainly driven by the transition between an early phase with a short duration and a late phase with a long duration, presumably reflecting the transition between WM target encoding and WM maintenance. Despite the similarity, there exists major differences on the timescale between theirs and our results. In the neurophysiological studies, the response in the early phase is much shorter, sustaining only about 500 msec (e.g., Figure 2 in Spaak et al.[20]), which seems to be well below the temporal resolution of our measurement. We conducted a simulation to see if the underlying neural dynamics of our results are similar to those previously reported in the neurophysiological studies—at a timescale of a few hundred milliseconds—whether we would observe the dynamical code observed in our temporal generalization.

The procedure and the results of the simulations are illustrated in Figure S4D. We simulated the response of 100 voxels (or neurons) over a trial. For simplicity, for each voxel we simulated responses with two phases, a 500-msec early response followed by a 12-s late response. This is done by generating a normally-distributed random number (with zero mean) twice for each voxel, leading to two multivariate random patterns for the entire population. We let the early response have 50% higher response amplitude than the late response. The duration and the relative amplitude of the early response was comparable to those reported previously (e.g., Figure 2 in Spaak et al.[20]). We then convolved the response of each voxel with a hemodynamic response function (HRF) with a peak between 4 and 5 s. We downsampled the convolved time series to match the temporal resolution of our measurement: 750 msec per TR. To speed up the analysis, instead of decoding, we computed the correlation between voxel activity patterns between different time points (TR), resulting in a representational similarity matrix, which conveys information similar to the cross-decoding matrix.