Title：

Preserved stimulus representations in frontal cortex as a potential neural mechanism for flexible control in working memory

Authors:

Zhujun Shao[1,2,3], Mengya Zhang[1,3], Qing Yu[1,*]

Affiliations:

[1]Institute of Neuroscience, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China

[2]University of Chinese Academy of Sciences, Beijing, China

[3]These authors contributed equally

[*]**Correspondence should be addressed to:**

Qing Yu

Institute of Neuroscience, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences

Shanghai, 200031, China

Email: qingyu@ion.ac.cn

Keywords: visual working memory, flexible control, frontal cortex, visual cortex, fMRI, recurrent neural network

Summary

The nature and functions of stimulus-specific representations in different cortical regions have remained debatable in working memory research. Here we tested the account that high-dimensional stimulus representations emerge and persist in frontal cortex to facilitate flexible control in working memory. During functional magnetic resonance imaging, human participants flexibly switched between memorizing an orientation (maintenance) and categorizing an orientation following different learned rules (control) on a trial-by-trial basis. Our results demonstrated increased stimulus representations in frontal cortex that tracked demands on control, and increased stimulus representations in early visual cortex that tracked demands on precise maintenance. Simulations from two-module recurrent neural network models replicated human neural patterns, when stimulus information was preserved for readout at the output stage. Altogether, these findings provide empirical and computational evidence that preserved stimulus representations in higher-order cortex for output selection might serve as a possible coding scheme for the brain to accommodate the ever-changing environment.

Introduction

Real-world flexible behavior relies largely on working memory, the ability to maintain and manipulate information in mind in order to serve diverse control demands [1]. Over the past decade, converging evidence has suggested that stimulus-specific information in precise maintenance is represented in a distributed

cortical network, including sensory, parietal, and frontal cortex [2-10]. Specifically, during simple memory maintenance, stimulus representations are robustly encoded in early visual cortex (EVC), while those in frontoparietal cortex are typically weaker and less stable [4,5,8,11]. However, in dynamic environments such as during distraction, stimulus representations in EVC can be greatly interrupted [12,13]. In contrast, stimulus representations in frontoparietal cortex can be robust under certain circumstances such as distraction and attentional prioritization [12,14-17]. These mixed findings have led to considerable debates on the exact nature and functions of stimulus representations in different cortical regions in working memory.

From an adaptive perspective, these seemingly discrepancies in fact highlight the interaction between stimulus representations and the multiplex functions of working memory. Beyond the mere, passive maintenance of incoming sensory information, working memory provides an online mental workspace for flexible control over stimulus contents to take place. Hence, it is plausible that stimulus representations in working memory would flexibly adapt to various control demands beyond simple memory maintenance. It has been proposed that, to achieve behavioral flexibility, two seemingly competing accounts need to be reconciled—forming low-dimensional abstract representations over stimulus representations to enable knowledge generalization, and preserving copies of high-dimensional stimulus representations to facilitate flexible transformation [18-20]. Within this framework, stimulus representations in higher-order frontoparietal cortex during working memory can be considered as evidence in support of the view that high-dimensional stimulus representations

emerge in response to control demands exerted on these representations. This would be consistent with the established role of frontoparietal cortex in memory manipulation [21,22] and cognitive control [23,24]. However, there is also ample evidence, particularly from the categorization literature, indicating that frontoparietal cortex can form low-dimensional, abstract category representations over high-dimensional stimulus representations during perceptual categorization [25-29]. With this account, one might predict exactly the opposite: categorization control should result in reduced stimulus representations in frontoparietal cortex, and correspondingly increased category representations in the same brain regions.

Here, to reconcile these two accounts, we tested how these two mechanisms were implemented in working memory with flexible control, by directly comparing the neural representations of high-dimensional stimulus and those of the low-dimensional category in the same cortical network underlying working memory. Using functional magnetic resonance imaging (fMRI), we systematically investigated how information of stimulus and that of category were represented in visual, parietal, and frontal cortex, along with how these representations supported flexible control in visual working memory. To this end, we employed an orientation working memory task with varied levels of control: in the Maintenance task, participants memorized one orientation throughout a memory delay without additional control demands; while in the Control task, participants needed to categorize the remembered orientation into different categories following learned rules.

In brief, we showed that high-dimensional stimulus representations, rather than low-dimensional, control-related category representations, emerged and increased in frontal cortex with increased control demands. Critically, the strength of stimulus representations in frontal cortex was predictive of memory performance in only the Control but not Maintenance task, suggesting a selective involvement of frontal cortex in control functions. By contrast, EVC demonstrated an opposite pattern, with increased stimulus representations in the Maintenance task and significant category representations in the Control task. We further tested whether the increased stimulus representations in frontal cortex could be explained by output dimensionality, using two-module recurrent neural networks (RNNs) to simulate the tasks. We found that our observed neural patterns could be well replicated with an RNN that preserved high-dimensional stimulus information at the output stage. These results together suggest preserved high-dimensional stimulus representations in frontal cortex as a potential neural mechanism for flexible control in working memory, and highlight the adaptive nature of stimulus representations across different cortical regions to meet the diverse demands on working memory.

Results

Behavioral learning and task performance

Participants (n = 24) first completed a behavioral learning session to learn two categorization rules (Rule A and Rule B, see Figure 1B) that were orthogonal to each other. Participants acquired the two rules with equal familiarity as shown in Figure 1C

($t(23) = 0.239$, $p = 0.813$; for Rule A, $M = 0.854$, $SD = 0.050$; for Rule B, $M = 0.851$, $SD = 0.051$).

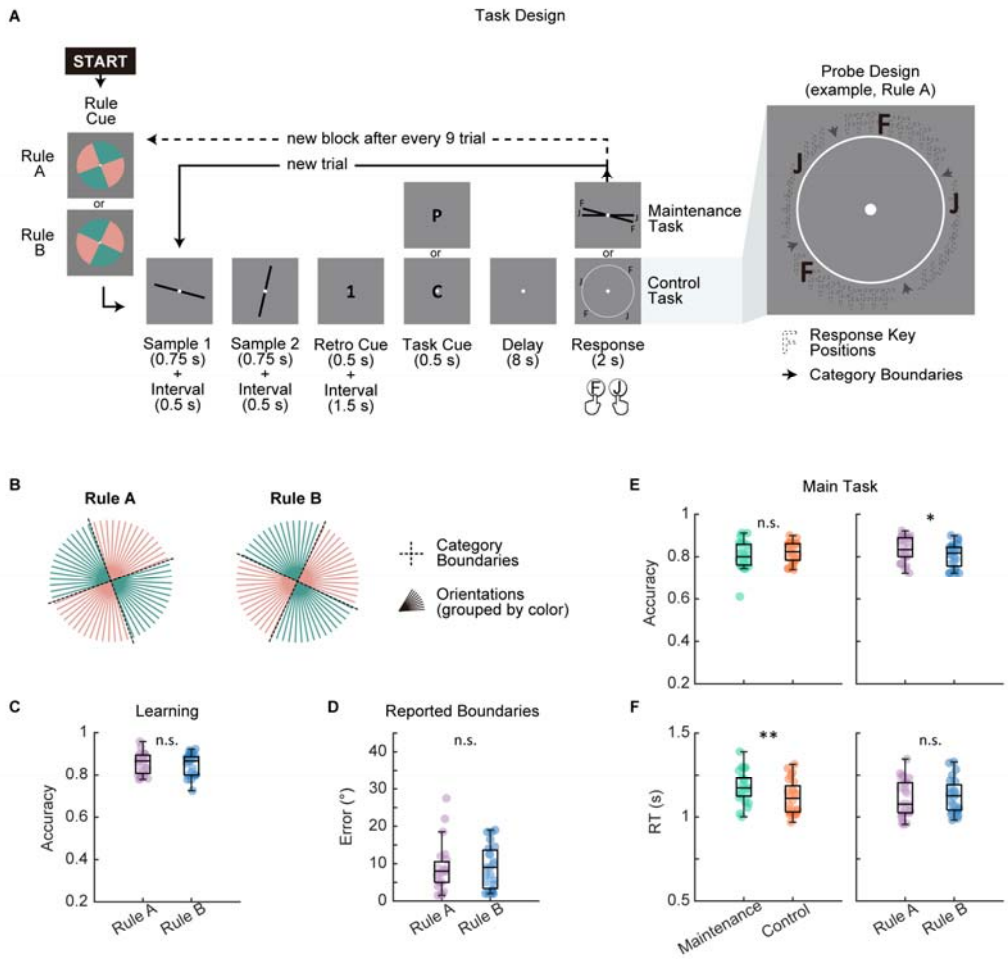

Figure 1. Experimental design and behavioral performance. (A) Main task procedure. Each block started with a rule cue indicating the categorization rule for this block. On each trial, participants saw two sequentially presented orientations and were then cued to remember one of the orientations. In Maintenance task (cued by letter 'P'), participants needed to maintain the remembered orientation as precisely as possible. In Control task (cued by letter 'C'), participants needed to categorize the remembered orientation following the categorization rule of the current block. Maintenance and Control trials were interleaved within an experimental block of 9 trials. Categorization rule (Rule A or Rule B) switched randomly on a block-by-block basis. Response keys ('F' and 'J') for Control task were randomly assigned to the two categories. Each pair of keys displayed at random locations within the category to eliminate information on rule boundaries. (B) The two orthogonal rules (Rule A and Rule B) for categorization. (C) Rule learning performance during learning session for Rule A (purple) and Rule B (blue). (D) Errors in participants' self-reported rule boundaries. Errors were calculated

as the average distance from reported boundaries to ground truth boundaries. (E) Accuracy compared across tasks and rules. (F) Reaction time compared across tasks and rules. Boxplots show the median and the 25th and 75th percentiles. Whiskers extend to 1.5 Inter quartile range (IQR) from the quartiles. Asterisks denote significant results, n.s.: not significant; *: $p < 0.05$; **: $p < 0.01$.

In the main scanning session, participants completed two tasks, Maintenance and Control, inside an MRI scanner. The Maintenance task was a typical delayed match-to-sample task which involved no categorization. Participants only needed to memorize specific visual orientations and maintain the cued orientation throughout the memory delay. In Control task, participants also started with maintaining an orientation. After task cue, they needed to categorize the remembered orientation into one of two categories following one of the two rules acquired in the learning session. Participants randomly switched between the two tasks within an experimental block of nine trials, and randomly switched between the two categorization rules across experimental blocks. We randomized response mapping across trials to avoid influences from motor-planning signals (Figure 1A). Overall, participants performed well on both tasks. Accuracy for the Maintenance task ($M \pm SD$: 0.806 ± 0.065) and that for the Control task (0.821 ± 0.047) did not significantly differ ($t(23) = 1.513$, $p = 0.144$; Figure 1E), suggesting the two tasks were matched in terms of task difficulty. Accuracy for Rule A (0.835 ± 0.056) was slightly higher than that for Rule B (0.807 ± 0.055), $t(23) = 2.306$, $p = 0.031$ (Figure 1E). In terms of reaction time, participants spent slightly more time on Maintenance task (1.177 ± 0.093) than on Control task (1.118 ± 0.098), $t(23) = 3.221$, $p = 0.004$ (Figure 1F). Reaction time for Rule A (1.110 ± 0.108) and that for Rule B (1.126 ± 0.100) were not different, $t(23) = 1.028$, $p =$

0.315 (Figure 1F). Additionally, participants were able to draw accurate boundaries for both rules (errors in Rule A: 8.80° ± 6.29°; errors in Rule B: 9.02° ± 5.69°; Figure 1D).

Instinctively, it is easier to categorize orientations away from boundaries than the ones near boundaries. As shown in Figure S1, participants performed better on trials closer to the category center compared with those closer to the category boundaries, in terms of both accuracy and reaction time. These results demonstrated the effect of categorization and confirmed that participants followed flexible task demands in different conditions.

Frontal cortex showed increased stimulus representations during Control compared to Maintenance

The primary goal of the current study was to evaluate how representations of high-dimensional stimulus and those of low-dimensional category would adapt to varying task control demands in working memory. Previous work on stimulus representations in visual working memory has mostly used multivariate encoding and decoding approaches [2-10]. While these approaches have fructified our understanding on whether stimulus representation is present or absent in specific brain regions, they are often agnostic of the exact representational format of the corresponding representations. Moreover, when dealing with more sophisticated problems, such as when multiplexed representations are involved, dissecting different representations can be non-trivial. To address this problem, here we leveraged two complementary

approaches: first, we developed a factorized representational similarity analysis (RSA) combining linear mixed-effects modeling to construct theoretical representational dissimilarity matrices (RDMs) with different representational structures and tested which model best predicted neural data. This approach allows for simultaneous comparisons between different RDMs, and deals with the collinearity between them. Second, to make our results more comparable to previous work, we also examined representations of stimulus and of category using conventional encoding or decoding methods. To implement these approaches, we focused our primary analyses on three brain regions of interest (ROIs) that have been implicated in representing stimulus information in working memory, including early visual cortex (EVC), intraparietal sulcus (IPS), and superior precentral sulcus (sPCS) [7,9,13].

For the RSA approach, we first constructed three hypothesized RDMs models: a graded stimulus model (assuming equal distance between adjacent orientations), a one-hot stimulus model (assuming equal distance between all orientations), and a category model (assuming equal distance within a specific category). We then tested which model (graded stimulus, one-hot stimulus, or category) predicted population-level neural RDMs in each ROI using linear mixed-effects models, while taking shared explained variance between models into account. In other words, only models contributed significant and unique variance to the neural data would survive the test.

Our results showed that in all ROIs (Figure 2), the graded stimulus model best predicted neural data, suggesting a feature-like representational format was preserved

along the cortical hierarchy. Interestingly, the predictability of the graded stimulus model differed between tasks, as moving more anteriorly into frontal cortex. Specifically, while the graded stimulus model predicted neural RDMs in EVC throughout the trial during both Maintenance and Control tasks, even when Control task did not require explicit memory of stimulus information; its predictability in Maintenance task reduced in IPS and more so in sPCS. On the contrary, the graded stimulus model retained its predictability in Control task in all ROIs, especially during memory delay and response periods. On the other hand, the category model showed little evidence in all ROIs, except for the response period in EVC, suggesting that no category representation was formed in frontal cortex.

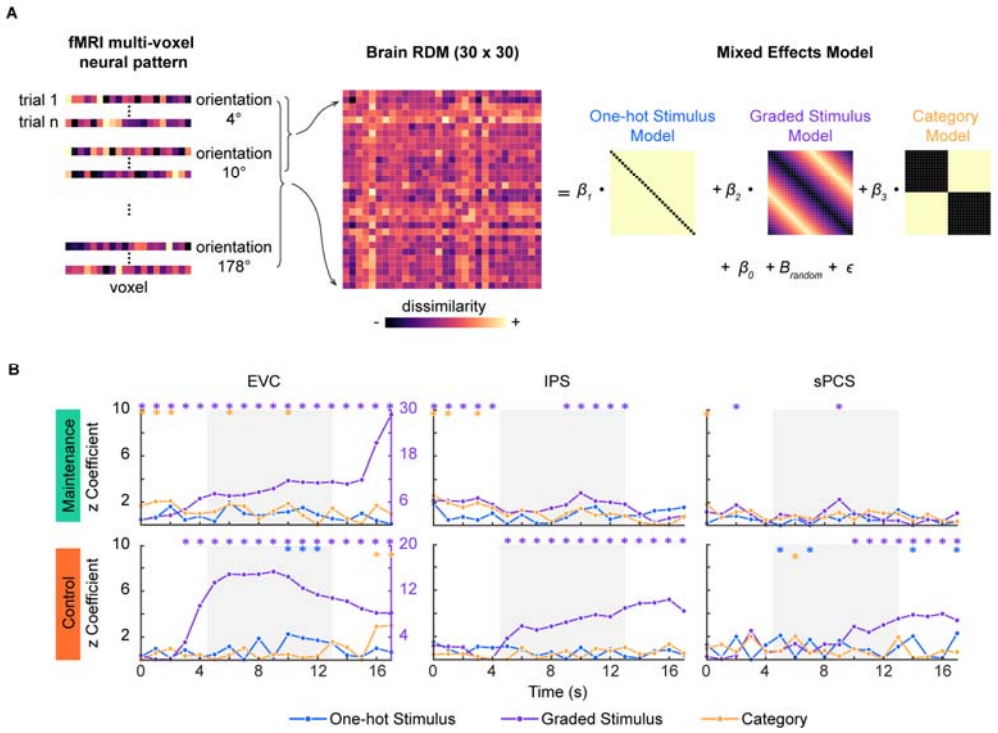

Figure 2. Mixed-effects modeling on neural RDMs using hypothetical stimulus and category models.

(A) To construct individual neural RDMs between orientations, we first averaged all trials for each orientation within the same condition and then calculated the distance

between each pair of orientations. Separately calculated for each condition, we then fit neural RDMs with linear mixed-effects models to estimate which model (the one-hot stimulus model, graded stimulus model, and category model) best estimated the neural RDM at each time point. (B) Predictability of one-hot stimulus model (blue), graded stimulus model (purple), and category model (yellow) at EVC, IPS, and sPCS in the Maintenance task (upper row) and in the Control task (lower row). At EVC, the left y axis refers to one-hot stimulus and category model; the right y axis refers to graded stimulus model. P-values of each predictor in the mixed-effects model was transformed to Z-values for visualization. Gray shaded areas indicate the entire memory delay following task cue. Asterisks denote significant results for each model ($p < 0.05$).

Stimulus representations in visual and frontal cortex reflected differential task demands on maintenance and control

To validate our findings from the RSA approach, we separately examined representations of stimulus and those of category using conventional multivariate encoding or decoding methods. First, we used multivariate inverted encoding models (IEMs) [7,30-32] to reconstruct population-level, orientation representations in each ROI. The orientation reconstructions from IEMs were in line with our RSA results. In EVC, we observed significant orientation reconstructions in both Maintenance and Control tasks (*p*s < 0.00001 during the entire memory delay from 3 to 17 s, Figure 3D). Additionally, the strength of orientation representation in Maintenance task became significantly higher than that in Control task after the task cue (from 10 to 17 s, *p*s < 0.009), suggesting the strength of orientation representations in EVC reflected the degrees of task demand on maintaining visual details. In IPS, orientation reconstructions were significant in both tasks (from 5 to 17 s, *p*s < 0.042), but did not differ from each other (*p*s > 0.352; Figure 3E). In sPCS, a reversed pattern was observed. In Maintenance task, orientation information was maintained during early

delay (7 - 9 s, *p*s < 0.033) and then dropped to baseline level from late delay to response (11 - 17 s, *p*s > 0.303). By contrast, in Control task, orientation representation was persistent throughout the memory delay and response (6 - 17 s, *p*s < 0.046). The strength of orientation representations in Control task became statistically higher than that in Maintenance task in late delay (12 - 17 s, *p*s < 0.05 except at 14 s: p = 0.059; Figure 3F), suggesting that the strength of stimulus representations in frontal cortex reflected the level of control demands.

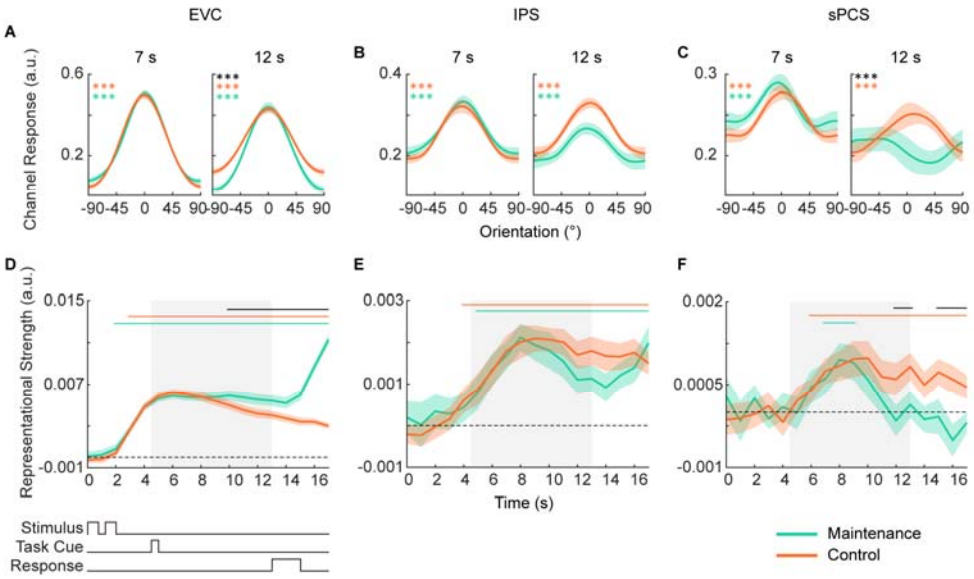

Figure 3. Orientation reconstructions using IEM.
(A) Example orientation reconstructions using BOLD signals at 7 s and 12 s in EVC. for Maintenance (green) and Control (orange) tasks, respectively. X axis represents distance from the cued orientation (at 0º), y axis represents reconstructed channel responses in arbitrary units. Shaded areas represent ± SEM. Asterisks denote significant results for Maintenance task (green), Control task (orange), and task difference (black), *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. (B) Same conventions as (A), but with results from IPS. (C) Same conventions as (A), but with results from sPCS. (D) Time course of representational strength of orientations in EVC. Gray shaded areas indicate the entire memory delay following task cue. Green and orange lines at the top indicate the significance of representational strength of the corresponding task at each time point ($p < 0.05$). The top black line indicates significant difference in representational strength between tasks ($p < 0.05$). Horizontal dashed lines represent a baseline of 0. Shaded areas represent ± SEM. (E) Same

conventions as (D), but with results from IPS. (F) Same conventions as (D), but with results from sPCS.

These results cannot be explained by the specific model used to train the data [33,34] nor the specific analytical approach used, because when we trained the IEM separately for each condition (Figure S2A-C) or adopted a decoding approach (Figure S3) [15,35], similar patterns were observed. Mean activation differences between tasks cannot account for the results either, because when we removed the mean differences in BOLD activity between tasks, the difference in representational strength remained (Figure S2D-F). Furthermore, to remove the impact of voxel number on IEM, we selected the top 500 of most sample- or delay-selective voxels from each ROI and trained IEM using the selected voxels. Again, this analysis yielded similar findings (Figure S2G-I).

Stimulus representations in frontal cortex exclusively predicted behavioral performance in Control task

Previous working memory studies have shown that the strength of stimulus representations in EVC positively correlated with memory performance [8,11,36], suggesting that EVC plays an important role in precise maintenance in working memory. However, although stimulus representations have also been observed in higher-order frontal cortex, they rarely predicted behavioral performance [13]. Here, we separately assessed the behavioral predictability of stimulus representations during memory delay in EVC, IPS, and sPCS (Figure 4). Consistent with previous findings,

in EVC, the strength of stimulus representations predicted behavioral accuracies in both Maintenance (from 7 to 9 s, *rs* > 0.41, *ps* < 0.05) and Control tasks (from 6 to 9 s, *r*s > 0.41, *p*s < 0.049) during early memory delay. Similar predictability was found in IPS, with stimulus representations predicted behavior in Maintenance task during early delay (from 6 to 7 s, *r*s > 0.41, *p*s < 0.049) and in Control task throughout the entire memory delay (*r*s > 0.44, *p*s < 0.034). Interestingly, we found that, throughout the entire memory delay, the strength of stimulus representations in sPCS predicted behavioral accuracies only in Control task (*r*s > 0.44, *p*s < 0.014) but not in Maintenance task (*ps* > 0.1). These results highlighted the exclusive functional significance of stimulus representations in sPCS in Control task.

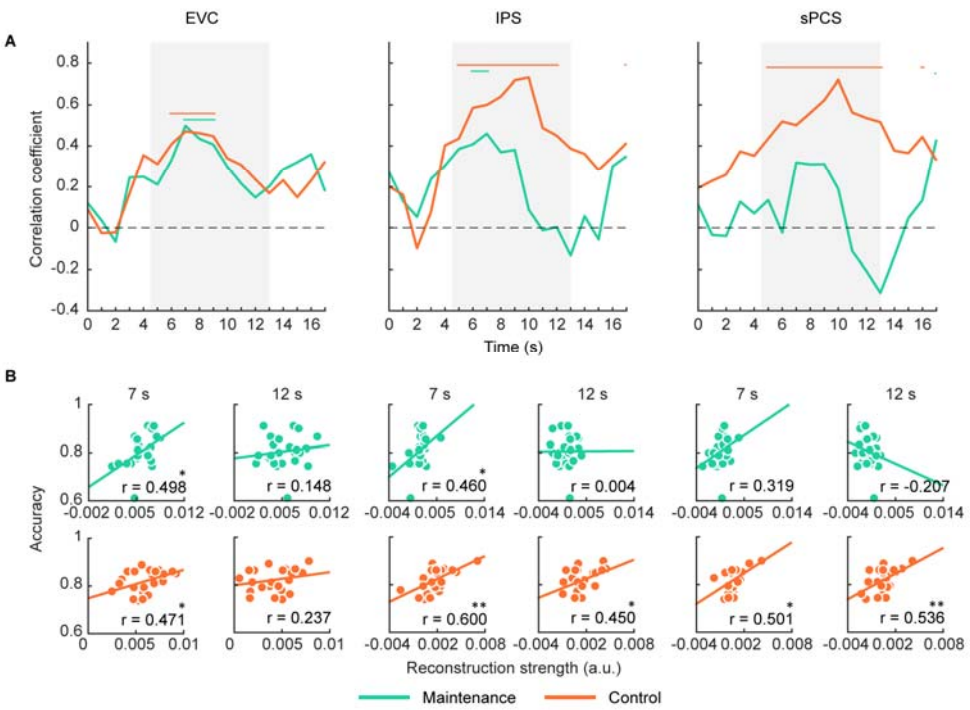

Figure 4. Correlation between strength of stimulus representations and behavioral performance for Maintenance (green) and Control (orange) tasks.
(A) Time course of correlation coefficients in EVC, IPS, and sPCS. Gray shaded areas indicate the entire memory delay following task cue. Green and orange lines at the top indicate significance of correlation at each time point (*p* < 0.05, uncorrected). (B)

Correlation scatter plots at 7 s and at 12 s in EVC, IPS, and sPCS. R denotes Pearson correlation coefficients. Asterisks denote significant results, *: $p < 0.05$; **: $p < 0.01$.

Absent low-dimensional category representations in frontal cortex

Our RSA analysis suggested low-dimensional category information was largely absent, except for the response period in EVC. To test the robustness of this "absence" effect, we conducted the following analyses.

First, we used Support Vector Machines (SVMs) to decode category information in the Control task. For each rule, we trained an SVM decoder to discriminate between the two categories. To minimize the impact of stimulus similarity on category decoding, we additionally trained another decoder using the opposite rule (i.e., using category labels from the other rule) [29]. We then calculated a category decoding score by subtracting decoding accuracy under the opposite rule from that under the true rule (Figure 5C). The rationale was that the amount of stimulus similarity would be comparable for the opposite rule, but additional category information, if existed, would result in higher decoding accuracy for the true rule. Consistent with our RSA results, no significant decoding of category was observed, except in the response period in EVC (16-17 s, ps < 0.00001) and slightly at 9 s at IPS ($p = 0.005$; Figure 5A).

Next, we examined whether categorization led to categorical biases in stimulus reconstructions from IEM [37]. Specifically, if stimulus representation was biased towards the category center, we would expect the orientation reconstruction to be shifted rightward to the target center. Again, we did not observe any significant biases

towards category center, in any of the ROIs, except for the response period in EVC ($p < 0.00001$; Figure S4).

In sum, combining results from RSA, category decoding, and categorical bias, we found little evidence in favor of stimulus-independent, low-dimensional category representation in sPCS, confirming this region contained mostly stimulus representations during memory delay.

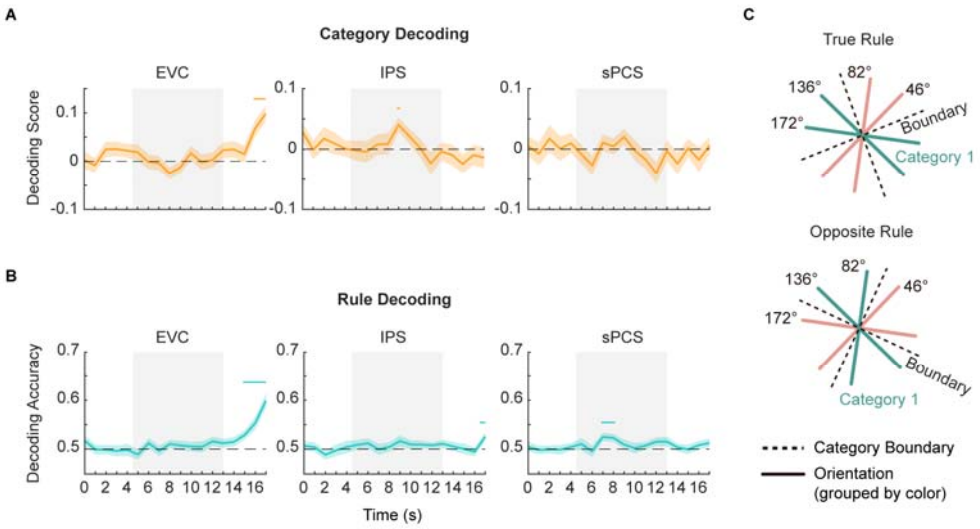

Figure 5. Decoding performance for category and rule information.
(A) Time course of category decoding score, calculated using methods in (C). Gray shaded areas indicate the entire memory delay following task cue. Top horizontal lines indicate significance of decoding accuracy at each time point. Horizontal dashed lines represent chance levels. Error bars represent ± SEM. (B) Time course of rule decoding performance, same conventions as (A). (C) Schematics of category decoding score. Category decoding score was calculated by subtracting category decoding accuracy under the opposite rule from that under the true rule. In this example, A first decoder was trained to differentiate between 46° and 82° vs. 136° and 172° under the true rule. A second decoder was trained to differentiate between 46° and 172° vs. 82° and 136° under the opposite rule. The category decoding score was calculated by subtracting the decoding accuracy of the second decoder from that of the first decoder.

Rule representations in visual, parietal, and frontal cortex with diverse temporal dynamics

Because the current study involved flexible switching between different tasks (Maintenance vs. Control) as well as switching between different categorization rules (Rule A vs. Rule B), it would also be informative to examine how these control-related rule signals evolved in the brain, and how representations of rule differed from those of category. We trained an SVM decoder to discriminate between Rule A and Rule B in Control task. Interestingly, we found that rule representation was evident in sPCS during early period of memory delay (from 7 to 8 s, $ps < 0.024$); while it was evident during response in EVC (from 15 to 17 s, $ps < 0.008$) and IPS (at 17 s, $p = 0.005$; Figure 5B). This result suggested earlier emergence of rule information in sPCS, compared with the lower-level EVC and IPS.

Output dimensionality altered stimulus and category decoding patterns in RNNs

Our fMRI analyses have demonstrated increased high-dimensional stimulus representations in sPCS as well as limited low-dimensional category representations in EVC during the Control task. The absence of category representation in higher-order frontal cortex might seem at odds with the single-neuron recording literature in non-human primates [25-28]. However, one recent electrophysiological study has demonstrated that, when monkeys were trained to perform a discrimination task prior to category learning, the category selectivity in LIP neurons would significantly reduce, suggesting that category selectivity may not necessarily emerge, depending on training history [38]. Here we reasoned that the cortical network might have adopted a different strategy to accommodate the need to flexibly switch between different tasks

and rules in the current study. This different strategy can be implemented via altering the network output dimensionality in response to the tasks.

To test this hypothesis, we utilized the recently advanced RNNs and examined whether modifying the RNN architecture could recapitulate our findings on stimulus and category decoding. Modular RNNs were trained on the Maintenance and Control tasks simultaneously [39,40]. The networks shared common input and hidden layer structures (i.e., orientation-tuned and retro/task cue-related units as the input layer, recurrent units with short-term synaptic plasticity [STSP] in the hidden layer [80% excitatory + 20% inhibitory units, equally distributed in two separate modules]). The only difference was in their output structures. The first type of RNN (RNN 1; n = 20) had only two units in the output layer to indicate networks' choice, making its output structure low-dimensional (Figure 6A). The second type of RNN (RNN 2; n = 20) had additional units in the output layer preserving the original stimulus information, making its output structure high-dimensional (Figure 7A). In other words, the second RNN was designed to preserve stimulus information throughout the network. Additionally, as a proof of concept, we decided to only include two modules of recurrent units, distinguished by denser connectivity within than between modules. The early module was connected with the input layer, imitating EVC which receives visual information first in the processing hierarchy. Meanwhile, the late module received projections from the early module and in turn projected to the output layer, resembling higher-level cortex such as frontal cortex. We simulated the task events as inputs to the models to match the procedures described in our fMRI session.

After successful training, we applied the same decoding approach to investigate population-level stimulus and category representations in spiking activities of the RNNs. We found that decoding performance in the first RNN resembled findings in single-neuron studies that category information emerged significantly in both modules of the network, with higher decoding accuracy for category in the late module than the early module (Figure 6C). Additionally, no increase in stimulus representation was observed in control task in either of the modules (Figure 6B and Figure S5A). By contrast, decoding performance in the second RNN was more comparable to findings in the current study, where category decoding was higher in the early module than the late module (Figure 7C). Critically, increased stimulus decoding during control was also observed in the late module of the second RNN (Figure 7B and Figure S5B), resembling our fMRI results in frontal cortex. These results demonstrated that our fMRI results were replicable when preserving high-dimensional stimulus information for readout at the output stage in the RNN, suggesting that increased stimulus representations in frontal cortex during control might be explained by this high-dimensional output structure.

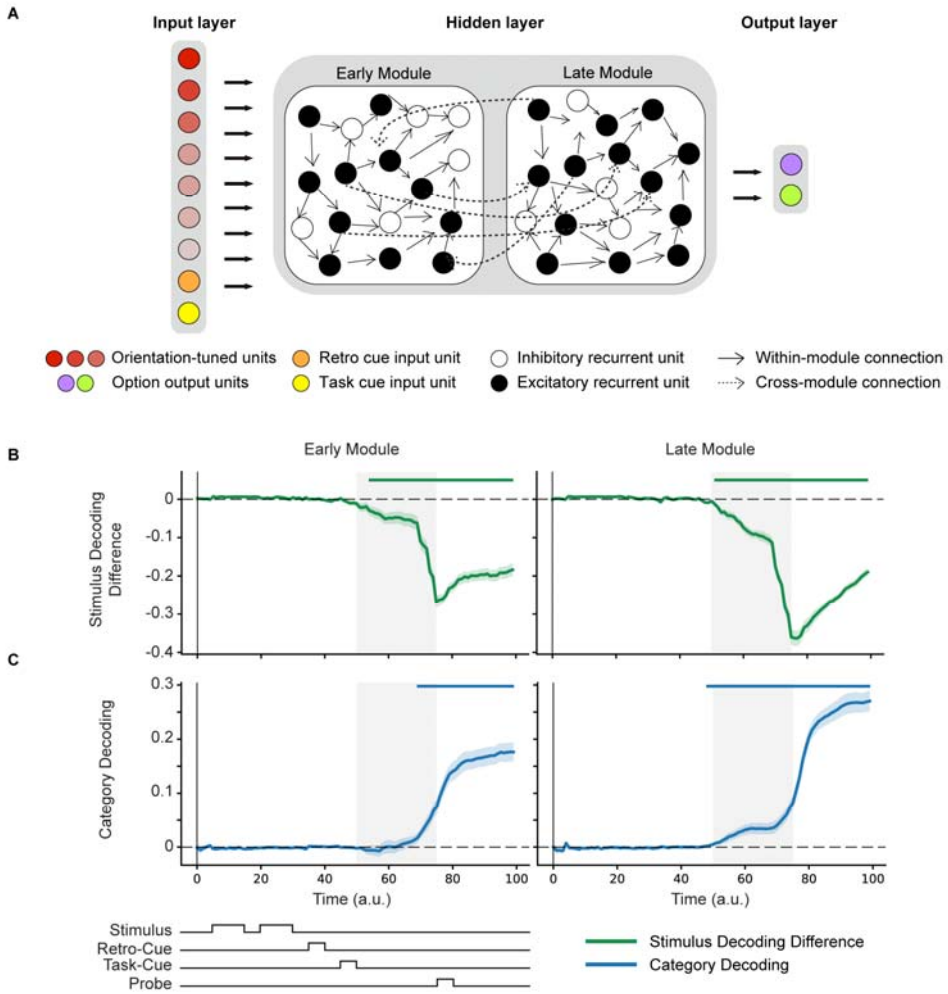

Figure 6. RNN 1 architecture and simulation results.

(A) The network consists of 3 layers of artificial units: the input, hidden and output layers. The input layer contains 50 units including 45 orientation-tuned (15 per receptive field, red) units and 5 cue units (retro-cue and task cue, orange and yellow). The hidden layer consists of two modules of 250 recurrent units with short-term synaptic plasticity (STSP), further divided into 80% excitatory (black) and 20% inhibitory (white). Connectivity within each module (black arrow) is denser compared to between modules (red arrow), which only occur between excitatory units. Only excitatory units in the early module receive inputs from the input layer and only excitatory units in the late module project to the output units. Networks output (0,1) or (1,0) through the 2 units in the output layer to indicate responses. (B) Difference in population-level decoding performance for orientations in the early (left) and late (right) modules of RNN 1. Positive difference indicates higher decoding accuracy for Control, and negative difference indicates higher decoding accuracy for Maintenance. X axis denotes time points. Shaded gray area indicates the delay period. The top colored line indicates significant time points ($p < 0.05$, corrected using a cluster-based

permutation method). Error bars represent ± SEM. (C) Population-level decoding performance for category. Same conventions as (B).

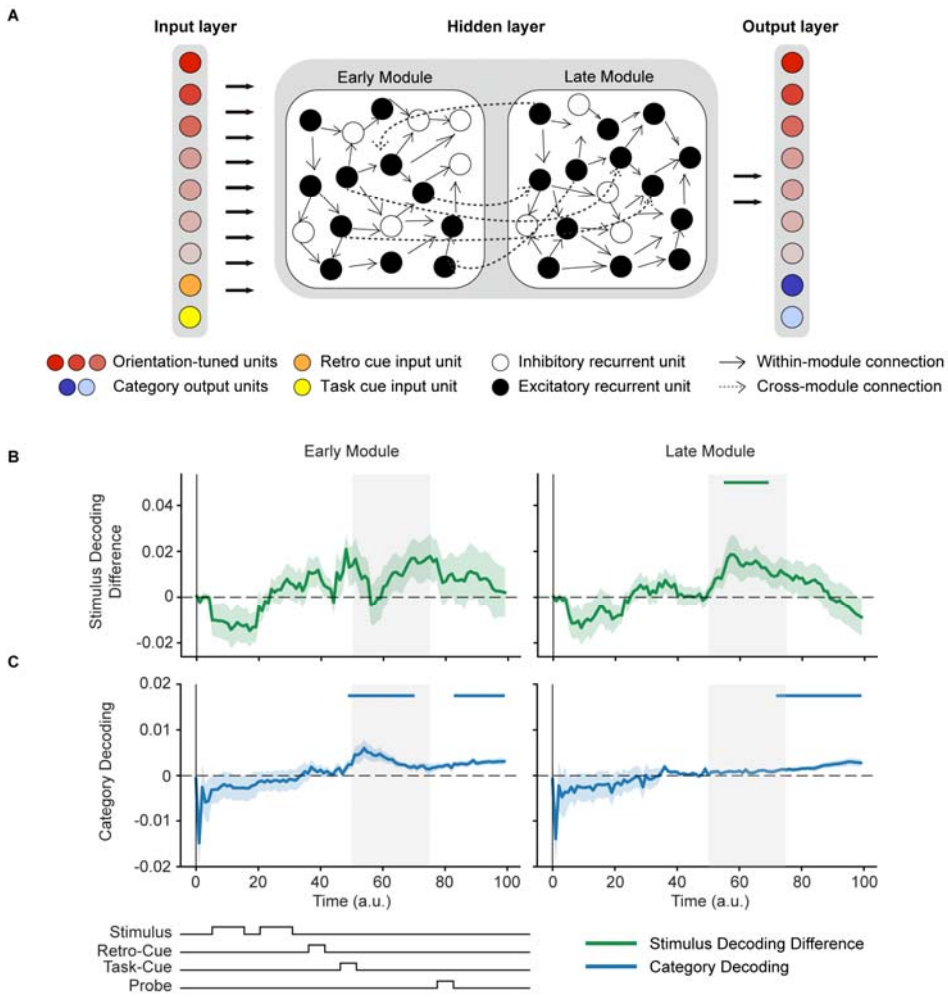

Figure 7. RNN 2 architecture and simulation results.
(A) The network consists of 3 layers of artificial units: the input, hidden and output layers. The input layer contains 20 units including 15 orientation-tuned (red) units and 5 cue units (retro-cue and task cue, orange and yellow). The hidden layer consists of two modules of 250 recurrent units with short-term synaptic plasticity (STSP), further divided into 80% excitatory (black) and 20% inhibitory (white). Connectivity within each module (black arrow) is denser compared to between modules (red arrow), which only occur between excitatory units. Only excitatory units in the early module receive inputs from the input layer and only excitatory units in the late module project to the output units. The network output (0,1) or (1,0) to report the category to which the cued orientation belonged in the control task, or (0,0) in the maintenance task (blue units). Importantly, the model also output the orientation itself through 15 additional orientation-tuned units (red). (B) Difference in population-level decoding

performance for orientations in the early (left) and late (right) modules of RNN 2. Positive difference indicates higher decoding accuracy for Control, and negative difference indicates higher decoding accuracy for Maintenance. X axis denotes time points. Shaded gray area indicates the delay period. The top colored line indicates significant time points ($p < 0.05$, corrected using a cluster-based permutation method). Error bars represent ± SEM. (C) Population-level decoding performance for category. Same conventions as (B).

Discussion

In the current study, we investigated the emergence and maintenance of high-dimensional stimulus and low-dimensional category representations with flexible control demands in working memory. In a distributed cortical network encompassing visual, parietal, and frontal cortex, we found increased stimulus representations in sPCS that tracked demands on control, and increased stimulus representations in EVC that tracked demands on precise maintenance. The increased stimulus representations in higher cortex and the category representations in lower cortex were well predicted by an RNN with preserved stimulus information for readout at the output stage. Together, these results highlight the unique and critical contributions of stimulus representations in different cortical regions to distinct functions in working memory.

Visual cortex has been considered a critical site for maintenance in visual working memory in the context of sensorimotor recruitment hypothesis [2,41]. This idea, however, has been challenged in recent years due to some seemingly contradictory findings from the neuroimaging literature. For example, it has been found that compared to frontoparietal cortex, mnemonic representations in visual cortex were more vulnerable to distractors [12,13,42]. It has also been proposed that the decodability of memory contents in visual cortex depends on the specific task type. In nonvisual

tasks that required judgements on object category instead of visual details, memory contents were no longer decodable in visual cortex [16]. In our experiment, although the strength of stimulus representation in EVC differed between tasks, it remained a copy of stimulus representation even in the Control task. Moreover, stimulus representations in both tasks were equally predictive of subsequent memory performance, suggesting the functional significance of these representations in working memory. We reasoned that the difference between our task and previous work [16] was our Control task required participants to manipulate remembered information according to arbitrary yet flexible categorization rules, rather than selectively paying attention to different aspects (visual details vs. category membership) of everyday objects. In this case, maintaining visual details of the memoranda was still critical for an accurate response to be made. This finding is consistent with predictions from sensorimotor recruitment hypothesis that representations of memory contents in visual cortex are necessary for the precise maintenance of visual information. The observation of robust category representation in early visual cortex during response period further indicated the recruitment of early visual cortex in Control, possibly for boundary comparison and rule implementation. In fact, our results are consistent with a recent study demonstrating significant stimulus representation in EVC even when memoranda had been transformed into a motor format [35]. In addition, monkey research has also shown robust feature selectivity in visual cortex during a categorization task [25]. Recent computational

modeling work has also suggested the possibility of intact maintenance of sensory information during categorical judgements [43].

Compared to the prominent role of EVC in memory maintenance, here we provided converging evidence that sPCS in frontal cortex played a dominant role in Control. Although stimulus representations in sPCS have been observed during working memory in previous studies, the nature of these representations has remained debatable: in working memory tasks that required mere maintenance of memoranda, stimulus was not always decodable in frontal cortex [4,8,11], thereby questioning the functional significance of stimulus representations in frontal cortex. On the other hand, in the face of external distraction or attentional prioritization, stimulus representations in frontoparietal cortex can be resistant [12-14,42]. Our current study further added to this line of research by demonstrating that stimulus representations in sPCS increased with increased levels of control. These seemingly contradictory findings can be potentially reconciled by recent theoretical and computational work which proposes that active working memory functions may involve different neuronal mechanisms from passive maintenance. While passive maintenance relies mainly on synaptic mechanisms, control-related active functions (e.g., distractor resistance and information manipulation) involve more neuronal spiking activity [39,44]. Consistent with this idea, we provided the first empirical evidence that regions in frontal cortex demonstrated enhanced stimulus representations in Control task which required active manipulation of memoranda, and these representations were predictive of memory performance. Meanwhile, stimulus representation in Maintenance failed to predict memory

performance. In other words, stimulus representations in frontal cortex might have emerged to allow implementation of control demands on these representations. It would be interesting for future studies to investigate whether this active manipulation account generalizes to tasks that require other types of working memory manipulation such as mental rotation.

On a separate note, lower-dimensional category representations might emerge as a result of categorization control. However, we found that representations of stimulus and those of category were highly dissociable at the neural level: while stimulus representation was persistent across a wide range of brain regions throughout the trial, robust category representations were only observed during the response period in EVC. We validated this finding by using two complementary approaches, linear mixed-effect modeling and IEM, and by showing an absence of categorical biases in stimulus representations in all ROIs prior to response. Together, these results raised an intriguing possibility that participants accomplished such tasks by utilizing originally-stored stimulus information, rather than creating new representations in working memory. This proposal might seem contradictory to animal findings at first glance, because robust category selectivity has been observed in monkey frontoparietal cortex during the delay period of categorization tasks [25-28]. However, a recent monkey study has demonstrated that, when monkeys had been exposed to a maintenance task prior to categorization training, category selectivity in LIP was significantly reduced, even when behavioral performance of categorization was trained to a high level [38]. We further tested potential mechanisms underlying these

findings through RNN simulations. We showed that by preserving high-dimensional stimulus information at the output stage, similar neural patterns can be observed in a two-module RNN: the late module (corresponding to frontal cortex) demonstrated increased stimulus representations. This suggested that the brain might preserve copies of the original stimulus information in higher-order cortex to enable flexible readout in different tasks. These preserved stimulus representations might in turn facilitate flexible control over stimulus information. In addition, we showed that when only category information was retained for output in the RNN, category representation was more prominent in the late module, suggesting that the previously observed category representations in single-neuron studies might have been primarily driven by this output selection process. In other words, in flexible environments, and with rich prior experience, the brain might adopt an entirely distinct strategy for processing information in working memory. High-dimensional stimulus information might be preserved in its original identity in higher-order cortex, potentially reducing processing demands in dealing with each task and thereby facilitating efficiency and flexibility.

Although we demonstrated high-dimensional stimulus representations in frontal cortex, it does not necessarily mean that high-dimensional representations were the only possible mechanism for flexible control. In fact, real-world flexible decisions need to deal with the tradeoff between low and high dimensionality all the time: while high dimensionality allows for separability and flexible readout, low dimensionality enables generalizability and efficient learning [19,20]. In our study, although we lacked

evidence for low-dimensional category representations in frontal cortex, we did observe such representations in EVC, meaning that low-dimensional representations were also involved in our task, but in a different brain region than commonly assumed. Our RNN work indicated that this dynamic reconfiguration in information coding at the network level can be well explained by a change in the coding strategy for network readout. It would be interesting for future studies to investigate what higher-order neural mechanism determines the selection of these coding strategies. Relatedly, it is noteworthy that any null result should be interpreted with caution. For example, although we were unable to find evidence of category representations in sPCS, we cannot rule out the possibility that an increase in category-induced selectivity might have partially contributed to the observed enhanced stimulus representation, albeit the category selectivity was too weak to be separated from stimulus-specific signals at the decodable level of fMRI [45]. It was also possible that category representations were maintained in other frontal regions beyond sPCS, but the effects might be subtle and would require a finer examination of frontal activities [17]. Exploring these possibilities would require further investigation in future work.

In our study, the inclusion of two independent categorization rules would allow for examination on the representations of contextual control signals [23,46]. With this, we found that rule information was first decodable in sPCS and switched to EVC at a later time point, implicating that sPCS might have served as the source of the control signals that governed implementation of categorization rule later into memory delay. Because the current study adopted a block design of rule presentation to avoid

cue-driven activity at the sensory level, any difference in neural signals observed between the two rules could only be attributed to inherent, abstract rule signals. The observation of rule-based control signals originating from sPCS, together with the increased stimulus representations in the same brain region, jointly support the critical role of sPCS in control processes in working memory.

Our finding of a frontoparietal network, or more broadly speaking, a distributed working memory network (including early visual and frontoparietal regions) in representing a combination of stimulus and rule information, is consistent with the recent proposal that frontoparietal network represents abstract task knowledge, via implementing action selections guided by abstract rules [47]. Working memory plays a key role in this process, as it provides an online sketchpad for the computation of categorization to take place. As such, categorization can be regarded as a mental process that combines building/retrieving abstract categorical knowledge with implementing abstract rule-guided selection in working memory. It would be interesting for future work to examine how the network for representing abstract knowledge (such as the cognitive map) [48] interacts with the working memory network to jointly accomplish this process.

In conclusion, we observed a distributed cortical network, including early visual, parietal, and frontal cortex, in representing stimulus-specific information in working memory. These stimulus representations in visual and frontal cortex played dissociable functional roles, with those in EVC contributing primarily to precise maintenance and those in frontal cortex contributing primarily to control execution in

working memory. RNN simulations indicated that the stimulus representations in higher-order cortex might have emerged as a result of output selection to facilitate behavioral flexibility. Collectively, these results help to reconcile current debates on the functional roles of different cortical regions in working memory, and provide new insights into how a unified working memory framework could support flexible control.

Methods

Participants

Twenty-six healthy participants (21 female, all right-handed, mean age = 24.0 ± 1.4 years) were recruited at Chinese Academy of Sciences, Shanghai Branch. Two participants were excluded due to failure in completing the experiment or low conformity to task instructions, remaining 24 participants who completed the whole study (mean age = 23.92 ± 1.41 years). All participants were neurologically healthy and eligible for MRI, had normal or corrected-to-normal vision, provided written informed consent approved by the Ethics Committee of Institute of Neuroscience, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences (Shanghai, China), and were monetarily compensated for their participation. Sample sizes were not estimated a priori but were comparable and even superior to those in previous studies.

Stimuli and Procedure

All stimuli were generated and presented using MATLAB (The MathWorks) and Psychtoolbox 3 extensions [49,50]. During behavioral training, stimuli were presented on a ThinkVision monitor at a viewing distance of 45 cm. Behavioral responses were acquired with a keyboard. During scanning, stimuli were projected onto a SinoRad monitor (1280 x 1024 pixels, refreshing at 60 Hz) viewed through a coil-mounted mirror in the scanner at a viewing distance of 90.5 cm. Participants' behavioral responses were acquired with a Sinorad MRI-compatible button box.

Behavioral Training

Prior to scanning, participants were trained to learn two novel rules, Rule A and Rule B, for categorizing orientations. Thirty oriented bars were used as sample stimuli, ranging from 5° to 179° (in increments of 6°; two participants used another set of thirty orientations ranging from 4° to 178°). Each abstract rule was constructed by two orthogonal boundaries that divided the thirty orientations into two categories with fifteen orientations each. Rule A and Rule B were orthogonal to each other. Corresponding boundaries were 20°/110° and 65°/155° (15°/105° and 60°/150° for the two participants using different stimuli sets).

Participants learned new rules through a rule learning task. Each run of learning started with a rule disk informing the target rule. To avoid any potential verbal coding, rule specifics were visually illustrated as rule disks containing two distinct colors (colors randomly assigned to categories every time; see Figure 1A for an example). Rule disk was presented on the screen for 2 s followed by 1 s of fixation. On each trial,

an oriented bar (radius=7°) was presented for 1 s followed by a delay of 1 s. Participants were instructed to report the category of the orientation by pressing a response key (‘F’ or ‘J’). To avoid category-response mapping, we randomized the relationship between categories and key buttons across trials. Moreover, to avoid presenting rule boundaries explicitly, we presented key names at random positions within the range defining each category. In other words, participants had to memorize the exact rule boundaries as accurately as possible in order to find the correct key buttons for each trial. Feedback was given at the end of each trial to assist learning.

Participants completed thirty learning trials in each run. They reviewed rule disks after every ten trials for memory reconsolidation. Each participant completed at least two runs for each rule. After achieving an average accuracy of above 86% (26 out of 30 trials) for the first rule, they proceeded to learn the other rule and then to practice the main task for scanning (see next section). Learning order of rules was counterbalanced across subjects. Upon completion of practicing, participants needed to report boundaries of learned rules as a qualification of behavioral training. If total error in reported boundaries had exceeded 20°, participants had one more chance to learn by repeating the learning task. Two participants completed one additional behavioral training to recap rule knowledge prior to scanning.

Flexible Working Memory Task (fMRI Task)

During scanning, participants completed a flexible working memory task which implemented levels of control demands with different rules. To be specific,

participants randomly switched between a Maintenance task and a Control task. In Maintenance task, participants needed to memorize stimulus information (i.e., orientations). In the Control task, participants needed to categorize orientations following the required rule. The categorization rule was randomly assigned on a block basis. Procedure of the main task was visualized in Figure 1A. At the beginning of each block, participants were presented with a rule disk for 3 s, followed by a 2-s interval, instructing the categorization rule of the current block. For each trial, participants saw two oriented bars presented successively. Each bar was presented for 0.75 s, with an inter-stimulus-interval of 0.5 s. Sample sets were the same as those used in behavioral training. After a 0.5-s interval, a retro-cue occurred for 0.5 s, indicating the orientation of which participants should remember. After a 1.5-s delay, a task cue was displayed at fixation for 0.5 s, following by an 8-s memory delay. The task cue was either a letter ‘P’ on Maintenance trials, instructing participants to maintain the cued orientation during memory delay as precisely as possible; or the task cue was a letter ‘C’ on Control trials, asking participants to categorize the cued orientation using the block rule during the delay. Then, participants were probed to respond within 2 s. On Maintenance trials, participants needed to select the memorized orientation from two probe orientations; while on Control trials, participants needed to report the category of the cued orientation. Response mapping followed the same operation as in the learning tasks. Inter-trial-intervals were randomly selected from 3, 5, and 7 s with an equal trial number, resulting in an average trial length of 20 s. Participants switched to the next block after every six

Control trials and three Maintenance trials. Each run contained two blocks (i.e., 18 trials).

Orientations, tasks, and cued target order (1st or 2nd) were counterbalanced across trials, resulting in an equal trial number across all three conditions (Control Rule-A, Control Rule-B, and Maintenance). Nineteen out of the twenty-four participants completed 15 runs of the main task. One participant completed 13 runs due to technical difficulties with the scanner. Another four participants completed 30 runs across two scan sessions. At the end of scanning, participants reported the rule boundaries three times as a final check of their rule memory.

Data Acquisition

MRI data were collected using a 3 Tesla Siemens MRI scanner (Tim Trio; Siemens Healthineers) with a 32-channel head coil at the Functional Brain Imaging Platform at Institute of Neuroscience, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences (CEBSIT, CAS). Functional scanning was performed using a gradient-echo echo-planar sequence with the following parameters: repetition time (TR) = 1000 ms; echo time (TE) = 30 ms; flip angle (FA) = 40°; voxel size = 3 x 3 x 3 mm; multi-band accelerate factor = 4; matrix size = 74 x 74; slice number = 60. A high-resolution anatomical T1 image was collected before functional scanning (TR = 2300 ms; TE = 2.98 ms; FOV = 256 x 240 x 192 mm; voxel size = 1 x 1 x 1 mm). During scanning, participants' head positions were restricted with surrounding paddings to prevent head movements.

Preprocessing

All preprocessing of individual MRI data was performed using AFNI (afni.nimh.nih.gov) [51,52]. Functional data of all runs were registered to the last volume of the final run with the first eight volumes of each run removed. Then, individuals' aligned functional data were registered to their corresponding T1 volume. Alignment of registration was manually checked for each subject to ensure quality. The registered data were further motion corrected and detrended.

ROI Definition

Our primary ROI-based analyses focused on three most commonly studied working-memory-related brain areas: early visual cortex (EVC), intraparietal sulcus (IPS) in parietal cortex, and superior precentral sulcus (sPCS) in frontal cortex [7,9,13]. We created anatomical ROI masks based on the probabilistic atlas by Wang and colleagues [53]. EVC (merging bilateral V1, V2, and V3), IPS (merging bilateral IPS0-5), and sPCS masks were generated by warping masks from the probabilistic atlas to individuals' anatomical image in their native space. In order to generate functional ROI masks, we then performed general linear models (GLMs) to quantify task-related univariate activity changes in each voxel. Task events were modeled using boxcar functions convolved with a canonical hemodynamic response function (durations of event epochs for sample, post retro-cue delay, memory delay, and response were 2.5 s, 2 s, 8.5 s, and 2 s, respectively). Six nuisance regressors were

also included to account for head motion artifacts in the six dimensions of rigid body motion. Functional EVC mask was defined by the 500 most active voxels during sample display. Functional IPS and sPCS masks were defined by the 500 most active voxels during memory delay.

MRI Data Analyses

Representational Similarity Analysis and Mixed-Effects Modeling

We combined Representational Similarity Analysis (RSA) with linear mixed-effects models to investigate neural representations of stimulus and of category. First, three hypothesized representational dissimilarity matrix (RDM) models were constructed: a graded stimulus model (assuming equal distance between adjacent orientations), a one-hot stimulus model (assuming equal distance between all orientations), and a category model (assuming equal distance within a specific category; Figure 2A). Second, neural RDMs were computed by calculating the Spearman correlation distance (1 - r) on multivariate neural patterns between trials; distances were then averaged within each orientation and task, resulting in a 30 x 30 matrix for every task. To align trials of Rule A and Rule B, we shifted the neural RDMs correspondingly in a way that the first half of orientations were counterclockwise to the lower category boundary and the second half were counterclockwise to the higher category boundary. For instance, if boundaries of Rule A were 20°/110°, respectively, then the first fifteen orientations in Rule A RDMs would be 22°-106° and the other fifteen orientations being 112°-16°.

We then tested which model (graded stimulus, one-hot stimulus or category) predicted population-level neural RDMs in each ROI using linear mixed-effects models. We fitted our neural RDMs with the three models being fixed effects using maximum likelihood estimation. Random effect terms were determined based on model fits. Specifically, we tested different models with one or more random intercepts and random slopes for each fixed effect as well as random slopes for participants, and chose the model with best model fit as our final model [54]. Significance of fixed effects was evaluated using p-values of the fixed effect coefficients.

Multivariate Inverted Encoding Modeling (IEM)

Neural representations of orientations were reconstructed using inverted encoding modeling (IEM) [7,30-32] with custom MATLAB scripts on individuals' BOLD activation patterns in the three functional ROIs. IEM provides an estimate of population-level reconstructions of stimulus-specific information. The general procedure for IEM includes using training data to train model weights and then applying weights to testing data to obtain reconstructed channel responses. For the main analyses, we used trials from all conditions (Maintenance, Control-Rule A, Control-Rule B) to train and to test IEM in order to avoid potential biases from a specific task condition. Results for Control task were averaged across Control-Rule A and Control-Rule B conditions. Training and testing were performed for each TR separately. As a control, IEMs were also estimated for each condition separately

(within-condition IEM). Training and testing underwent a leave-one-run-out cross-validation procedure, in which each run was taken out as the testing run, and the rest of the data served as the training run. This procedure was iterated until all runs had served as training and testing runs. Results from all cross-validated folds were averaged. Detailed computations for each fold were elucidated below:

We first modeled responses of voxels into nine equidistant orientation channels (initial channels were 1°, 21°, 41°, 61°, 81°, 101°, 121°, 141°, 161°), characterizing voxel selectivity for orientations. At each channel, the modeled orientation tuning curve was a half-wave-rectified and squared sinusoid raised to eighth power, defined as the function below (c was the center of the channel):

$$\int(\theta) = \cos(\theta - c)^8$$

Population-level tuning responses of voxels was described using the function:

$$B_1 = W C_1$$

B_1 was the training dataset from our fMRI data (v voxels $\times n$ trials). C_1 represented the hypothesized channel responses (k channels $\times n$ trials) which were modulated by W, a weight matrix (v voxels $\times k$ channels).

The least-squared estimates of the weight matrix ($\widehat{W}$) was computed using linear regression:

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

The weight matrix was then applied on the test dataset to reconstruct estimated channel responses ($\hat{C}_2$):

$$\hat{C}_2 = (\widehat{W}^T \widehat{W})^{-1} \widehat{W}^T B_2$$

The analysis above were repeated for 20 times in step of 1° using leave-one-run-out cross-validation so that the nine channel centers covered all 180 orientations [15,32]. All channel responses were combined to create responses for all 180 orientation channels. For statistical comparisons and for visualization, all channel responses were shifted to a common center of 0°. The responses from all trials were averaged to obtain reconstructed orientation representations for the test datasets.

To characterize the strength of each IEM reconstruction, we collapsed over the channel responses on both sides of the common center, averaged them and then calculated the slope of each collapsed reconstruction using linear regression [32,55,56]. A larger positive slope indicates stronger positive representation.

Categorical biases were quantified using following procedures. First, channel responses of each orientation were shifted to its corresponding category center. To conjugate categorical biases of the two flexible rules, channel responses of orientations to the right of category centers were flipped. As a result, any categorical bias towards the category center would manifest as a rightward bias in the averaged channel reconstruction for both Maintenance and Control tasks. To quantify categorical bias, we fit the averaged reconstruction with a multidimensional nonlinear minimization algorithm (Nelder-Mead) to the following exponentiated cosine function [37]:

$$f(x) = \alpha(e^{k(cos(\mu - x) - 1)}) + \beta$$

Where α, β, k, and μ represent amplitude, baseline, concentration, and center, respectively, and x represents vector of channel responses. μ was taken as the categorical bias for quantification.

Multivariate Pattern Analysis (MVPA)

Besides RSA and IEM, we tracked neural representations of stimulus, of category, and of categorization rules using linear Support Vector Machine (SVM) decoders. All decoding analyses were performed using PyMVPA toolbox [57], following a leave-one-run-out cross-validation procedure.

Decoding of stimulus was performed for every TR. We divided the thirty orientations into four bins of 45° each, two cardinal bins centered at 90° or 180° and two oblique bins centered at 45° or 135°. We then performed two two-way classifications, one trained and tested on cardinal bins, and the other trained and tested on oblique bins. We trained and tested decoders separately for each condition. Decoding accuracies were then computed by averaging performance of cardinal and oblique classifiers. For the Control task, we averaged accuracies of Control-Rule A and Control-Rule B conditions.

Decoding of category and of rule information for Control task was performed for every TR following a similar two-way classification procedure. Moreover, orientations could contain categorical information by visual similarity, as closer orientations were more similar to each other inherently. Thus, to isolate the influence of stimulus on category, in addition to the decoder using true category labels, we

trained an opposite category decoder using category labels based on the opposite rule. If the two-way classification on categories only captured stimulus information, then true category and opposite category decoding should have had comparable performance. If abstract category information existed beyond stimulus information, then true category decoder should have outperformed opposite category decoder. Thus, a category decoding score was calculated by subtracting opposite category decoding accuracy from true category decoding accuracy (i.e., chance level = 0). Category decoding was performed separately for each rule and were then averaged.

Statistical Testing

Participants' behavioral performance for the main task was assessed using accuracy and reaction time. Paired t-tests were conducted for the two task types (Maintenance & Control) and two rules (Rule A & Rule B) to evaluate differences between conditions.

Statistical significance was evaluated via a bootstrapping procedure for IEM reconstructions, categorical bias, and MVPA classification accuracy [7,32,58]. To characterize the significance of IEM slopes, for each reconstruction, twenty-four orientation reconstructions were randomly sampled with replacement from the pool of twenty-four participants and averaged. This procedure was repeated 10,000 times, resulting in 10,000 average orientation reconstructions, and correspondingly 10,000 slopes. The probability of obtaining a negative slope among the 10,000 slopes was counted as the one-tailed p-value of the slope. To characterize the difference between

the slopes of two IEM reconstructions, we first calculated the difference between two bootstrapped slopes 10,000 times, generating 10,000 slope differences. The significance of the slope difference was then calculated using the same one-tailed method as above. Statistical testing of categorical bias was performed using similar procedures that bias reconstructions were randomly selected (with replacement) and averaged from all participants for 1,000 times, resulting in a distribution of 1,000 estimated bias centers (μ). Statistical significance was then obtained by calculating the proportion of permuted centers smaller than the unbiased curve center (0°). A similar procedure was conducted to characterize the significance of classification accuracy and classification accuracy difference between tasks. Comparison between stimulus reconstruction of Maintenance and Control tasks in Figure 3D-F were corrected for multiple comparisons using the False Discovery Rate (FDR) method across time. For other analyses, p-values were corrected using FDR across ROIs, time (TRs), and tasks.

Recurrent Neural Network Simulations

RNN architecture

The network model was built following the details in previous work [39], and implemented in TensorFlow (version: nvidia-tensorflow 1.15.0) [59]. The general network architecture consisted of three layers of artificial units: the input, hidden, and output layers. The input layer contains units served to present various task-related signals corresponding to those in the fMRI paradigm, including orientations,

retro-cues, task cues (and response options in RNN 1 as detained below). In order to simulate responses in hierarchically connected brain regions (EVC and sPCS), we separated the hidden layer into two modules, each containing 250 recurrent units with short-term synaptic plasticity (STSP). Within each module, units were further divided into 80% excitatory and 20% inhibitory following Dale's principle. Similar to previous work [40], modularity was achieved by constraining the recurrent connectivity in the hidden layer. Specifically, only early module (module 1) excitatory units received inputs from the input layer and only late module (module 2) excitatory units projected to the output units. Between-module connectivity were culled so that 50% of early module excitatory units randomly projected to their counterparts in late module, and vice versa. Connections among inhibitory units remained strictly within-module in accordance with the observation that inhibitory connections in cortex are largely local. Thus, early module and late module simulated the lower, information-receiving and the higher, output-projecting brain areas of the processing hierarchy, respectively. We specifically manipulated the dimensionality of the outputs to investigate whether it would alter the results in terms of similarities to the fMRI observations. To this end, one type of network (RNN 1) implemented a low-dimensional, two-unit output layer with each unit corresponding to one of two possible options, presented through the input units before response; In contrast, the other type of RNN (RNN 2) had additional units in the output layer preserving the original stimulus information, making the output high-dimensional.

Task simulation

Orientations were simulated as Gaussian signals from 15 orientation-tuned units in the input layer distributed equally across 0-180 degrees, forming a ring of receptive field. The magnitude of an orientation-tuned input unit represented the closeness to the input angle. Sample values were selected from an array of 20 orientations evenly spanned from [0 to 180) degrees. The sequentially-presented stimuli were presented in the same receptive field, followed by retro-cue and task cue indicated through the separate cue units. For RNN 1, before the response period when the network was required to make a choice, two options were presented simultaneously through two other input receptive fields (with 15 orientation-tuned units each). The selection of the options varied slightly between the Maintenance and Control tasks: in Maintenance, one orientation was always the cued sample while the other was randomly chosen from all other possible angles. In Control, one option was taken from the same category as the cued sample but not necessarily the exact angle, while the other option was randomly chosen from orientations belonging to the other category. The network output (0,1) or (1,0) in the output units to report its choice. In comparison, RNN 2 output (0,1) or (1,0) to report the category to which the cued orientation belonged in the control task, or (0,0) in the maintenance task. Importantly, the model also output the orientation itself through 15 additional orientation-tuned units.

Training parameters and procedure

The hyperparameters and procedure for training the models were consistent with those detailed in previous work [39], with the following exceptions: standard deviations of input and recurrent noise were set to 0.01 as our tasks were much harder to train compared to those used in the reference study (especially networks were trained on both tasks simultaneously). Lowering the noise level may provide an edge for the models to successfully learn to perform the tasks. In similar vein, we also expanded the number of hidden units to 500 and training batch to a maximum of 5000. Additionally, spike penalty was set to 0 for RNN 2 to remove constraints on high neuronal activity in the original models.

We trained 20 models for each type of RNN and results were obtained by averaging over all of them. The goal of the training process was to minimize the mean square error between the model outputs and correct outputs during the response period (with a 50 ms grace period at onset when model output was not taken into account in calculating error). Training was conducted in a block-interleaved fashion in which each gradient batch consisted of 300 Maintenance, 300 Control rule A and 300 Control rule B trials, with the task block order randomized. Training would automatically stop if the model achieved 90% accuracy in the last training batch on all tasks. For RNN 2, the accuracies for category and orientation were calculated separately to ensure precision of the orientation outputs.

Population decoding

We measured the strength of stimulus and category representations through training SVMs on time-resolved neuronal activity. Activities were obtained by feeding new batches of tasks into the already-successfully trained networks after freezing all connection weights to prevent further changes to the models' behaviors. The intrinsic noise for the recurrent layer was also set to 0 for decoding analyses. To ensure accurate decoding results, we sampled large number of trials (900 trials for each condition) and implemented a 5-fold cross-validation procedure in which 80% of trials were used as training set and the remaining 20% as testing set in each fold. Decoders were trained separately for each module and time point. We adopted a cluster-based permutation approach (using MNE-Python [60] function *permutation_cluster_1samp_test* to accommodate the high number of time points to be corrected for) to determine statistical significance of the time course, for stimulus decoding accuracies in Maintenance/Control, difference between stimulus decoding accuracies between tasks, or stimulus-independent category decoding accuracy in Control.

Acknowledgements

This work was supported by the Ministry of Science and Technology of China (STI2030-Major Projects 2021ZD0203701, 2021ZD0204202), the National Natural Science Foundation of China (32271089), Shanghai Pujiang Program (22PJ1414400), CAS Project for Young Scientists in Basic Research (YSBR-071), and Shanghai Municipal Science and Technology Major Project (2018SHZDZX05).

Reference

1. Baddeley, A. (2003). Working memory: looking back and looking forward. Nat Rev Neurosci *4*, 829-839. 10.1038/nrn1201.
2. Harrison, S.A., and Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. Nature *458*, 632-635. 10.1038/nature07832.
3. 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. 10.1111/j.1467-9280.2009.02276.x.
4. Riggall, A.C., and Postle, B.R. (2012). The Relationship between Working Memory Storage and Elevated Activity as Measured with Functional Magnetic Resonance Imaging. Journal of Neuroscience *32*, 12990-12998. 10.1523/Jneurosci.1892-12.2012.
5. Yu, Q., and Shim, W.M. (2019). Temporal-Order-Based Attentional Priority Modulates Mnemonic Representations in Parietal and Frontal Cortices. Cereb Cortex *29*, 3182-3192. 10.1093/cercor/bhy184.
6. Christophel, T.B., Hebart, M.N., and Haynes, J.D. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. J Neurosci *32*, 12983-12989. 10.1523/JNEUROSCI.0184-12.2012.
7. 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. 10.1016/j.neuron.2015.07.013.
8. Gosseries, O., Yu, Q., LaRocque, J.J., Starrett, M.J., Rose, N.S., Cowan, N., and Postle, B.R. (2018). Parietal-Occipital Interactions Underlying Control- and Representation-Related Processes in Working Memory for Nonspatial Visual Features. J Neurosci *38*, 4357-4366. 10.1523/JNEUROSCI.2747-17.2018.
9. Yu, Q., and Shim, W.M. (2017). Occipital, parietal, and frontal cortices selectively maintain task-relevant features of multi-feature objects in visual working memory. Neuroimage *157*, 97-107. 10.1016/j.neuroimage.2017.05.055.
10. Sprague, T.C., and Serences, J.T. (2013). Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nat Neurosci *16*, 1879-1887. 10.1038/nn.3574.
11. 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. 10.1523/JNEUROSCI.5732-12.2013.
12. 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. 10.1038/nn.4174.
13. 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. 10.1038/s41467-021-24973-1.
14. 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. 10.1038/s41593-018-0094-4.
15. 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. 10.1038/s41593-019-0428-x.
16. 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. 10.1038/nn.3452.

17. Miller, J.A., Tambini, A., Kiyonaga, A., and D'Esposito, M. (2022). Long-term learning transforms prefrontal cortex representations during working memory. Neuron *110*, 3805-3819 e3806. 10.1016/j.neuron.2022.09.019.

18. Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., and Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. Neuron *110*, 1258-1270 e1211. 10.1016/j.neuron.2022.01.005.

19. Badre, D., Bhandari, A., Keglovits, H., and Kikumoto, A. (2021). The dimensionality of neural representations for control. Curr Opin Behav Sci *38*, 20-28. 10.1016/j.cobeha.2020.07.002.

20. Fusi, S., Miller, E.K., and Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. Curr Opin Neurobiol *37*, 66-74. 10.1016/j.conb.2016.01.010.

21. D'Esposito, M., Postle, B.R., Ballard, D., and Lease, J. (1999). Maintenance versus manipulation of information held in working memory: an event-related fMRI study. Brain Cogn *41*, 66-86. 10.1006/brcg.1999.1096.

22. D'Esposito, M., Postle, B.R., and Rypma, B. (2000). Prefrontal cortical contributions to working memory: evidence from event-related fMRI studies. Exp Brain Res *133*, 3-11. 10.1007/s002210000395.

23. Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. Trends Cogn Sci *12*, 193-200. 10.1016/j.tics.2008.02.004.

24. Badre, D., Kayser, A.S., and D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. Neuron *66*, 315-326. 10.1016/j.neuron.2010.03.025.

25. Brincat, S.L., Siegel, M., von Nicolai, C., and Miller, E.K. (2018). Gradual progression from sensory to task-related processing in cerebral cortex. Proc Natl Acad Sci U S A *115*, E7202-E7211. 10.1073/pnas.1717075115.

26. McKee, J.L., Riesenhuber, M., Miller, E.K., and Freedman, D.J. (2014). Task dependence of visual and category representations in prefrontal and inferior temporal cortices. J Neurosci *34*, 16065-16075. 10.1523/JNEUROSCI.1660-14.2014.

27. Freedman, D.J., and Assad, J.A. (2006). Experience-dependent representation of visual categories in parietal cortex. Nature *443*, 85-88. 10.1038/nature05078.

28. Freedman, D.J., Riesenhuber, M., Poggio, T., and Miller, E.K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. Science *291*, 312-316. 10.1126/science.291.5502.312.

29. Mok, R.M., and Love, B.C. (2020). Abstract Neural Representations of Category Membership beyond Information Coding Stimulus or Response. J Cogn Neurosci, 1-17. 10.1162/jocn_a_01651.

30. Brouwer, G.J., and Heeger, D.J. (2009). Decoding and reconstructing color from responses in human visual cortex. J Neurosci *29*, 13992-14003. 10.1523/JNEUROSCI.3577-09.2009.

31. Brouwer, G.J., and Heeger, D.J. (2011). Cross-orientation suppression in human visual cortex. J Neurophysiol *106*, 2108-2119. 10.1152/jn.00540.2011.

32. Yu, Q., and Postle, B.R. (2021). The Neural Codes Underlying Internally Generated Representations in Visual Working Memory. J Cogn Neurosci, 1-16. 10.1162/jocn_a_01702.

33. Sprague, T.C., Adam, K.C.S., Foster, J.J., Rahmati, M., Sutterer, D.W., and Vo, V.A. (2018). Inverted Encoding Models Assay Population-Level Stimulus Representations, Not Single-Unit Neural Tuning. eNeuro *5*. 10.1523/ENEURO.0098-18.2018.

34. Liu, T., Cable, D., and Gardner, J.L. (2018). Inverted Encoding Models of Human Population Response Conflate Noise and Neural Tuning Width. J Neurosci *38*, 398-408. 10.1523/JNEUROSCI.2453-17.2017.
35. Henderson, M.M., Rademaker, R.L., and Serences, J.T. (2022). Flexible utilization of spatial- and motor-based codes for the storage of visuo-spatial information. Elife *11*. 10.7554/eLife.75688.
36. 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. 10.1162/jocn_a_00357.
37. Ester, E.F., Sprague, T.C., and Serences, J.T. (2020). Categorical Biases in Human Occipitoparietal Cortex. J Neurosci *40*, 917-931. 10.1523/JNEUROSCI.2700-19.2019.
38. Latimer, K.W., and Freedman, D.J. (2023). Low-dimensional encoding of decisions in parietal cortex reflects long-term training history. Nat Commun *14*, 1010. 10.1038/s41467-023-36554-5.
39. Masse, N.Y., Yang, G.R., Song, H.F., Wang, X.J., and Freedman, D.J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. Nat Neurosci *22*, 1159-1167. 10.1038/s41593-019-0414-3.
40. Zhou, Y., Rosen, M.C., Swaminathan, S.K., Masse, N.Y., Zhu, O., and Freedman, D.J. (2021). Distributed functions of prefrontal and parietal cortices during sequential categorical decisions. Elife *10*. 10.7554/eLife.58782.
41. D'Esposito, M., and Postle, B.R. (2015). The cognitive neuroscience of working memory. Annu Rev Psychol *66*, 115-142. 10.1146/annurev-psych-010814-015031.
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. 10.1523/JNEUROSCI.3061-17.2018.
43. Luu, L., and Stocker, A.A. (2021). Categorical judgments do not modify sensory representations in working memory. PLoS Comput Biol *17*, e1008968. 10.1371/journal.pcbi.1008968.
44. Wang, X.J. (2021). 50 years of mnemonic persistent activity: quo vadis? Trends Neurosci *44*, 888-902. 10.1016/j.tins.2021.09.001.
45. Riley, M.R., and Constantinidis, C. (2015). Role of Prefrontal Persistent Activity in Working Memory. Front Syst Neurosci *9*, 181. 10.3389/fnsys.2015.00181.
46. Badre, D., and Nee, D.E. (2018). Frontal Cortex and the Hierarchical Control of Behavior. Trends Cogn Sci *22*, 170-188. 10.1016/j.tics.2017.11.005.
47. Vaidya, A.R., and Badre, D. (2022). Abstract task representations for inference and control. Trends Cogn Sci *26*, 484-498. 10.1016/j.tics.2022.03.009.
48. Behrens, T.E.J., Muller, T.H., Whittington, J.C.R., Mark, S., Baram, A.B., Stachenfeld, K.L., and Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. Neuron *100*, 490-509. 10.1016/j.neuron.2018.10.002.
49. Brainard, D.H. (1997). The Psychophysics Toolbox. Spat Vis *10*, 433-436.
50. Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis *10*, 437-442.
51. Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res *29*, 162-173.
52. Cox, R.W., and Hyde, J.S. (1997). Software tools for analysis and visualization of FMRI Data.

NMR in Biomedicine *10*, 171-178.

53. Wang, L., Mruczek, R.E., Arcaro, M.J., and Kastner, S. (2015). Probabilistic Maps of Visual Topography in Human Cortex. Cereb Cortex *25*, 3911-3931. 10.1093/cercor/bhu277.

54. Bellmund, J.L.S., Deuker, L., Montijn, N.D., and Doeller, C.F. (2022). Mnemonic construction and representation of temporal structure in the hippocampal formation. Nat Commun *13*, 3395. 10.1038/s41467-022-30984-3.

55. Samaha, J., Sprague, T.C., and Postle, B.R. (2016). Decoding and Reconstructing the Focus of Spatial Attention from the Topography of Alpha-band Oscillations. J Cogn Neurosci *28*, 1090-1097. 10.1162/jocn_a_00955.

56. Foster, J.J., Sutterer, D.W., Serences, J.T., Vogel, E.K., and Awh, E. (2016). The topography of alpha-band activity tracks the content of spatial working memory. J Neurophysiol *115*, 168-177. 10.1152/jn.00860.2015.

57. Hanke, M., Halchenko, Y.O., Sederberg, P.B., Hanson, S.J., Haxby, J.V., and Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. Neuroinformatics *7*, 37-53. 10.1007/s12021-008-9041-y.

58. Yu, Q., Teng, C., and Postle, B.R. (2020). Different states of priority recruit different neural representations in visual working memory. PLoS Biol *18*, e3000769. 10.1371/journal.pbio.3000769.

59. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). TensorFlow: A System for Large-Scale Machine Learning. 12th USENIX Symposium on Operating Systems Design and Implementation, 265–283.

60. Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hamalainen, M. (2013). MEG and EEG data analysis with MNE-Python. Front Neurosci *7*, 267. 10.3389/fnins.2013.00267.

Supplemental Figures

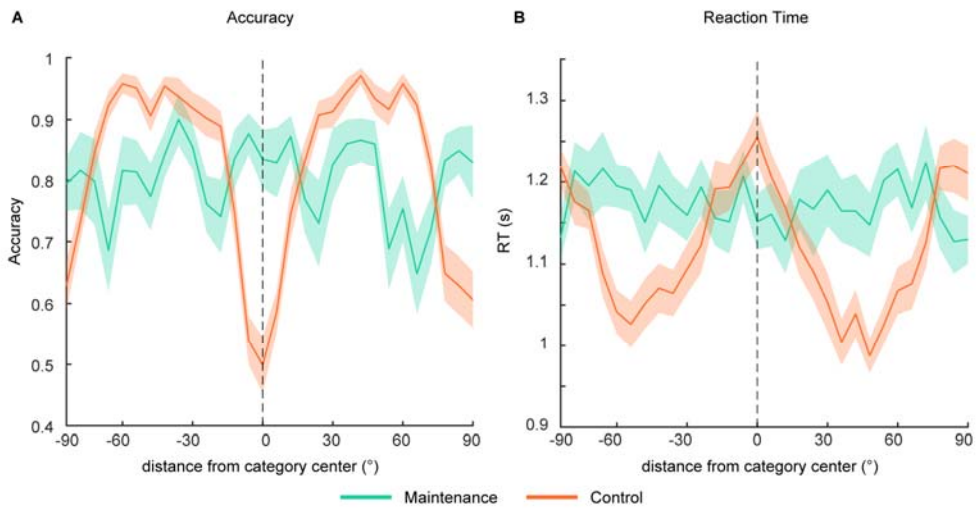

Figure S1. Main task behavioral performance as a function of distance from category center.

(A) Accuracy for orientations based on their distances from category center. (B) Reaction time for orientations based on their distances from category center. Shaded areas represent ± SEM. Vertical dashed line represents category center.

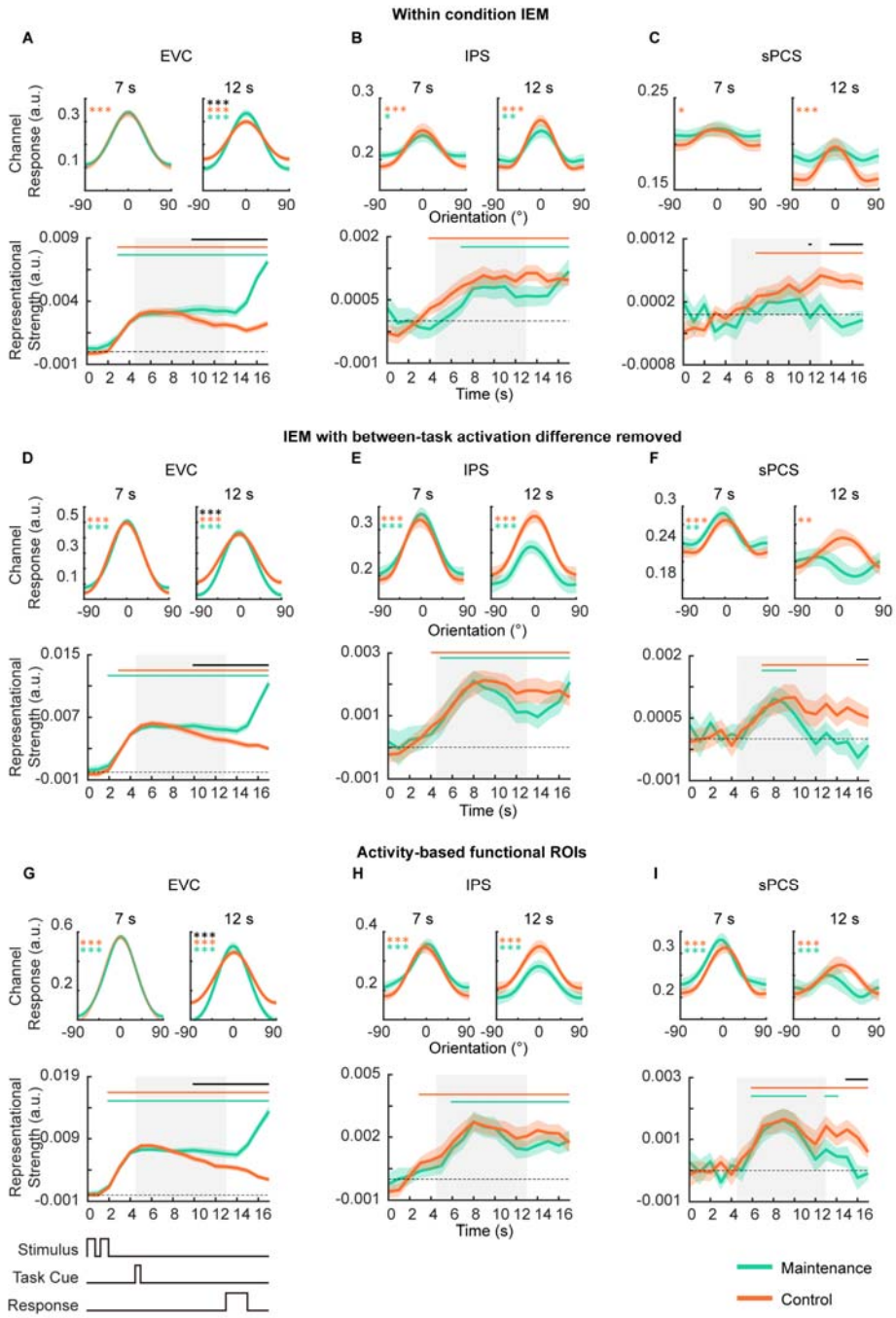

Figure S2. Control analyses for orientation reconstructions using IEM.
Upper panel (A to C) shows stimulus reconstruction using IEMs trained separately for each condition; middle panel (D to F) shows reconstructions from IEM after removing voxel-wise mean activation for each condition at each TR; lower panel (G to I) shows results in functional ROIs defined by top 500 most selective voxels during sample or delay period. Within each panel, the first row shows stimulus reconstructions at 7 s and at 12 s for Maintenance (green) and Control (orange) tasks. X axis represents distance from the cued orientation (at 0º), y axis represents reconstructed channel

responses in arbitrary units. Asterisks denote significant results for Maintenance task (green), Control task (orange), and task difference (black), *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. The second row shows the time course of representational strength of orientation. Gray areas indicate the entire memory delay following task cue. Green and orange lines at the top indicate significance of representational strength at each time point ($p < 0.05$). The top black line indicates significant difference in representational strength between tasks ($p < 0.05$). Shaded areas represent ± SEM.

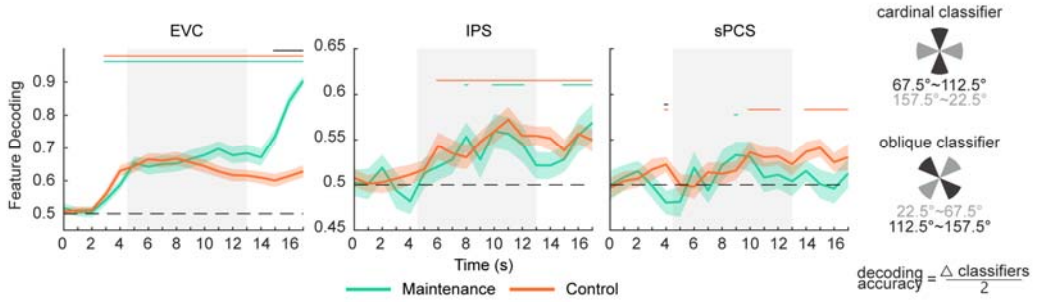

Figure S3. Decoding performance for stimulus information.

Thirty orientations used in the task were divided into four bins of which the centers were cardinal and oblique orientations (90° and 180°; 45° and 135°). Each bin covered 45°. Two two-way classifications were performed between the cardinal bins and between the oblique bins, respectively, and the resulting decoding accuracies were averaged. Classification was performed separately for Maintenance (green) and Control (orange) tasks. Similar to IEM reconstructions, as Control demands increased, the strength of stimulus representations in EVC decreased and that in sPCS increased. Gray shaded areas indicate the entire memory delay following task cue. Green and orange lines at the top indicate significance of decoding accuracies at each time point ($p < 0.05$). The top black line indicates significant difference in decoding between the two tasks ($p < 0.05$). Horizontal dashed lines represent chance level of 0.5. Shaded areas represent ± SEM.

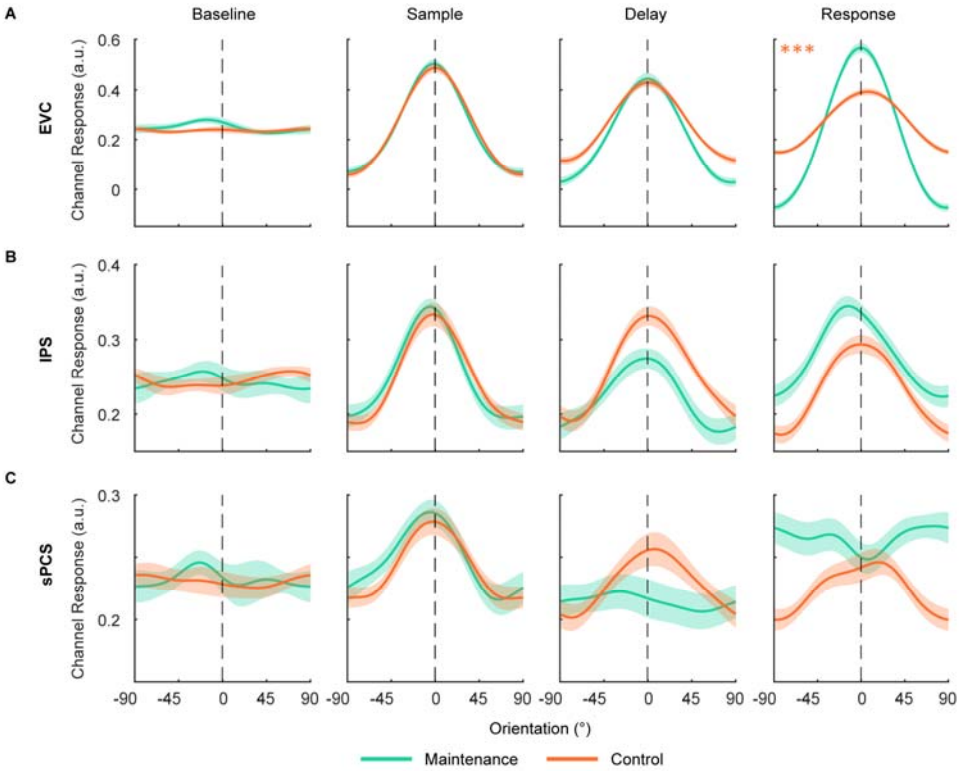

Figure S4. Categorical biases in stimulus reconstructions.

Categorical biases in stimulus reconstructions in EVC (A), IPS (B), and sPCS (C) during baseline (0 - 2 s), sample (7 - 9 s), delay (11 - 13 s) and response (15 - 17 s) epochs. X axis represents orientations, y axis represents reconstructed channel responses in arbitrary units. Categorical biases towards category centers would manifest as rightward to the vertical dashed line, if existed. No significant categorical bias was found except at EVC during response for Control (orange) task. Shaded areas represent ± SEM. Colored asterisks denote significance for the corresponding task, ***: $p < 0.001$.

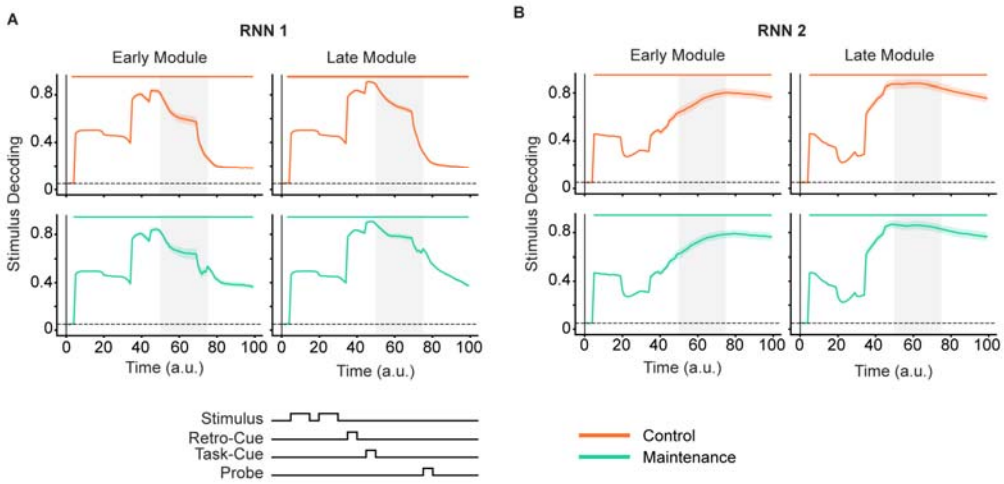

Figure S5. Supplemental RNN simulation results.
(A) RNN 1 population-level orientation decoding performance for both Control (upper row, orange) and Maintenance (lower row, green) tasks in the early (left) and late (right) modules. X axis denotes time points. Shaded gray area indicates the delay period. The top colored line indicates significant time points ($p < 0.05$, corrected using a cluster-based permutation method). Error bars represent ± SEM. (B) RNN 2 population-level orientation decoding performance. Same conventions as (A).