Neuroscience

Reviewed Preprint
v1 • September 3, 2024
Not revised

Stimulus representation in human frontal cortex supports flexible control in working memory

Zhujun Shao, Mengya Zhang, Qing Yu

Institute of Neuroscience, Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China • University of Chinese Academy of Sciences, Beijing, China

Abstract

When holding visual information temporarily in working memory (WM), the neural representation of the memorandum is distributed across various cortical regions, including visual and frontal cortices. However, the role of stimulus representation in visual and frontal cortices during WM has been controversial. Here we tested the hypothesis that stimulus representation persists in the frontal cortex to facilitate flexible control demands in WM. During functional MRI, participants flexibly switched between simple WM maintenance of visual stimulus or more complex rule-based categorization of maintained stimulus on a trial-by-trial basis. Our results demonstrated enhanced stimulus representation in the frontal cortex that tracked demands for active WM control and enhanced stimulus representation in the visual cortex that tracked demands for precise WM maintenance. This differential frontal stimulus representation traded off with the newly-generated category representation with varying control demands. Simulation using multi-module recurrent neural networks replicated human neural patterns when stimulus information was preserved for network readout. Altogether, these findings help reconcile the long-standing debate in WM research, and provide empirical and computational evidence that flexible stimulus representation in the frontal cortex during WM serves as a potential neural coding scheme to accommodate the ever-changing environment.

eLife assessment

This work presents **valuable** findings that the human frontal cortex is involved in a flexible, dual role in both maintaining information in short-term memory, and controlling this memory content to guide adaptive behavior and decisions. The evidence supporting the conclusions is **convincing**, with a well-designed task, best-practice decoding methods, and careful control analyses. The work will be of broad interest to cognitive neuroscience researchers working on working memory and cognitive control.

https://doi.org/10.7554/eLife.100287.1.sa2

Introduction

Real-world flexible behavior relies largely on working memory (WM), which allows the maintenance and manipulation of information in the brain in order to serve diverse behavioral goals (Baddeley, 2003). One central problem in the field of WM is to understand how stimulus information is represented and maintained in WM. Over the past decade, mounting evidence has demonstrated stimulus-specific representation during WM maintenance in a distributed cortical network, including sensory, parietal, and frontal cortices (Christophel et al., 2012; Ester et al., 2015; Gosseries et al., 2018; Harrison & Tong, 2009; Riggall & Postle, 2012; Serences et al., 2009; Sprague & Serences, 2013; Yu & Shim, 2017, 2019). However, the exact nature and functions of stimulus representation in different cortical regions remain controversial. Specifically, while neurophysiological studies in non-human primates have mostly emphasized stimulus representation in the frontal cortex (Funahashi et al., 1989; Fuster & Alexander, 1971; Leavitt et al., 2017), neuroimaging work in humans has reported disparate findings. During maintenance of simple visual features, stimulus representation is robustly encoded in the early visual cortex (EVC), which has been taken as the evidence in support of the sensorimotor recruitment hypothesis of WM (Harrison & Tong, 2009; Riggall & Postle, 2012; Serences et al., 2009). Meanwhile, those in the higher-order frontoparietal cortex are typically weaker and less stable (Emrich et al., 2013; Gosseries et al., 2018; Riggall & Postle, 2012; Yu & Shim, 2019). However, in dynamic environments such as those involving distraction, stimulus representation in EVC could be greatly interrupted or biased (Bettencourt & Xu, 2016; Hallenbeck et al., 2021; Lorenc et al., 2018). In contrast, stimulus representation in the frontal cortex could be robust under certain circumstances including attentional prioritization (Christophel et al., 2018), categorization (Lee et al., 2013) and after extensive training (Miller et al., 2022). To summarize, stimulus representation could vary markedly depending on specific brain regions and memory tasks, complicating the interpretation of potential functions of stimulus representation in WM.

In this study, we consider these apparent discrepancies from the perspective of cognitive flexibility (Badre et al., 2021; Fusi et al., 2016; Musslick & Cohen, 2021). We propose that changes in stimulus representation in different cortical regions might reflect a global reconfiguration in coding strategy and resource allocation in response to varied WM functions (Henderson et al., 2022; Lee et al., 2013). To elaborate, beyond the passive maintenance of incoming sensory information, WM provides an online mental workspace for active manipulation and control of stimulus contents (Baddeley, 2003; Miller & Cohen, 2001). As control functions often result in the generation and maintenance of new information, the brain needs to manage not only the original stimulus information but also the newly generated information in WM. Due to the limited cognitive resources available, it is likely that original stimulus representation in WM could adapt flexibly to various task goals beyond simple maintenance of WM contents, which might also co-vary with changes in the representation of the newly-generated information, leading to a systematic reconfiguration in representations of all levels across various cortices. We make two specific predictions from this account. First, in accordance with the findings of elevated neural activity in the frontal cortex with increasing demand for memory manipulation (D'Esposito et al., 1999; D'Esposito et al., 2000) and cognitive control (Badre, 2008; Badre et al., 2010; Miller & Cohen, 2001), stimulus representation in frontal cortex should be enhanced for active-control-related functions in WM. By contrast, due to the precise nature of stimulus representation in visual cortex, stimulus representation in this region should be enhanced for precise-maintenance-related functions in WM (Henderson et al., 2022; Lee et al., 2013). Second, within the brain regions that encode the newly generated information, a dynamic tradeoff between representations of original and new information should be observed to achieve flexible allocation of limited cognitive resources (Badre et al., 2021; Flesch et al., 2022).

Using functional magnetic resonance imaging (fMRI), we directly tested this account by systematically investigating stimulus representation in visual, parietal, and frontal cortices during WM tasks with varying demands for active control. In particular, we surmised that stimulus representation in the frontal cortex would increase to accommodate complex control demands such as rule-based categorization. To this end, we employed a visual WM paradigm that required flexible switching between maintenance and categorization tasks. Specifically, in the maintenance task, participants maintained one visual orientation throughout a delay period, whereas in the categorization task participants were required to categorize the remembered orientation into one of two categories in accordance with previously learned rules, which could be either switched randomly between two rules on a block-by-block basis (Experiment 1) or fixed with one rule (Experiment 2). Thus, compared with the maintenance task, the categorization task imposed additional control demand of WM information at two different levels.

In line with our prediction, our results showed that stimulus information was more prominently represented in the frontal cortex during categorization than maintenance task, and this differential representation was enhanced with increasing demands for control. Importantly, the strength of stimulus representation in the frontal cortex was predictive of WM behavioral performance in the categorization but not in the maintenance task, implicating a selective involvement of the frontal cortex in control functions. By contrast, stimulus representation in the visual cortex was found to exhibit an opposite pattern, with higher strength for maintenance than categorization task. Moreover, varying control demands across experiments revealed a dynamic tradeoff between stimulus and the newly-generated category representations. To further examine whether the enhanced stimulus representation in the frontal cortex during categorization task could be explained by global coding strategy, we simulated our flexible WM tasks with multi-module recurrent neural networks (RNNs). The results of this computational modeling well replicated our human data when precise stimulus information was preserved at the output during network training. Taken together, our results indicate the importance of the frontal cortex for flexible control in WM and highlight the relative changes of stimulus representation in different cortical regions for varying task demands of WM.

Results

Behavioral learning and performance of WM tasks

In the fMRI session of Experiment 1, human participants (n = 24) completed two tasks, maintenance and categorization, inside an MRI scanner. The maintenance task was a delayed match-to-sample working memory task of orientations. Participants only needed to maintained the cued orientation throughout a memory delay. In categorization task, participants also started with maintaining an orientation. After the task cue, they needed to categorize the remembered orientation into one of two categories using the cued categorization rule. Within an experimental block of nine trials, participants randomly switched between the two tasks. Across blocks, participants randomly switched between two categorization rules acquired during a preceding learning session. We randomized response mapping across trials to avoid potential influence by motor-planning signals (see **Figure 1A** for probe design). Prior to the main session, participants 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 ($t(23) = 0.24$, $p = 0.813$; for Rule A, $M = 0.85$, $SD = 0.050$; for Rule B, $M = 0.85$, $SD = 0.05$; **Figure 1C**) and comparable precision (averaged error in reported boundaries for Rule A were 8.80° ± 6.29° and that for Rule B was 9.02° ± 5.69°; **Figure 1D**).

Overall, participants performed equally well on both tasks in the fMRI session. Accuracy for the maintenance task ($M \pm SD$: 0.81 ± 0.07) and that for the categorization task (0.82 ± 0.05) did not significantly differ ($t(23) = 1.51$, $p = 0.144$; **Figure 1E**), suggesting that the two tasks were

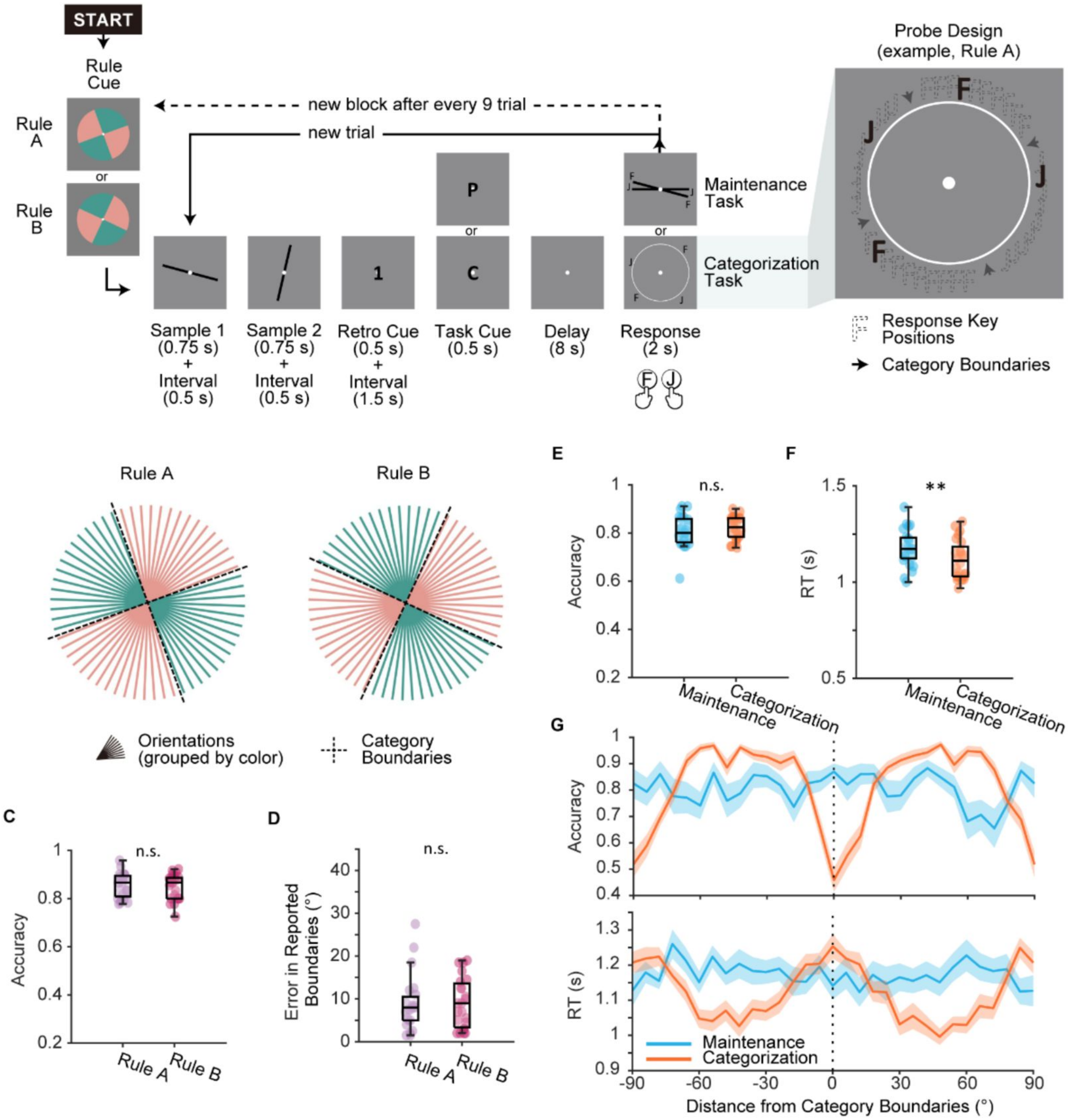


Figure 1.

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 orientations consecutively 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 categorization task (cued by letter ‘C’), participants needed to categorize the remembered orientation following the categorization rule of the current block. maintenance and categorization trials were interleaved within an experimental block of nine trials. Categorization rule (Rule A or Rule B) switched randomly on a block-by-block basis. Response keys (‘F’ and ’J’) for categorization 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) Illustration of the two orthogonal categorization rules (Rule A and Rule B). (C) Rule learning performance during learning session for Rule A (purple) and Rule B (pink). (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 between tasks. 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; **: $p < 0.01$. (F) Reaction time compared between tasks. Same conventions as (E). (G) Upper panel: accuracy in relation to distance from categorization boundaries. Lower panel: reaction time in relation to distance from categorization boundaries. Shaded areas represent ± SEM.

matched in terms of task difficulty. In line with previous categorization studies (Ester et al., 2020; Freedman & Assad, 2006) demonstrating a boundary effect, only in the categorization task but not in the maintenance task did participants perform better for trials distant from category boundaries in terms of both accuracy and reaction time (see **Figure 1G**). These results demonstrated the effect of categorization and confirmed that participants faithfully followed task instructions.

Enhanced stimulus representation in frontal cortex during categorization task

The primary goal of this study was to determine the role of stimulus representation in various cortices in WM. Using conventional multivariate encoding and decoding methods, we tracked stimulus (i.e., orientation) representation in three brain regions of interest (ROIs) that have been implicated in representing WM information, including early visual cortex (EVC), intraparietal sulcus (IPS), and superior precentral sulcus (sPCS) (Christophel et al., 2018; Ester et al., 2015; Hallenbeck et al., 2021; Yu & Shim, 2017).

First, we used multivariate inverted encoding models (IEMs) (Brouwer & Heeger, 2009, 2011; Ester et al., 2015; Rademaker et al., 2019; Yu & Shim, 2017) to reconstruct orientation representation at the population level in each ROI. **Figure 2A** shows example orientation reconstructions from representative time points, and **Figure 2C** demonstrates the time course of orientation reconstruction as quantified by representational fidelity. A larger fidelity value indicates a stronger orientation representation (**Figure 2B**). In EVC, we found significant orientation representation in both maintenance and categorization tasks starting from the sample period (**Figure 2C** left panel; see **Supplemental Table 1** for full statistics), even when the categorization task did not require explicit memory of stimulus information. Additionally, the strength of orientation representation in the maintenance task became significantly higher than that in the categorization task after the task cue during the delay, suggesting the strength of orientation representations in EVC reflected the degrees of task demands for maintaining visual details. In IPS, orientation representation was significant in both tasks, but did not differ from each other at most time points (**Figure 2C** middle panel). In sPCS, a reversed pattern was observed. In the maintenance task, orientation information was maintained during early delay period and then dropped to baseline level during late delay period. By contrast, in the categorization task, orientation representation was persistent throughout the delay and response periods. The strength of orientation representation in the categorization task became statistically higher than that in the maintenance task in late delay period (**Figure 2C** right panel), suggesting that this differential representations of visual stimulus in the frontal cortex reflected the demand for active control of memory contents. To facilitate comparison of the differential stimulus representation across ROIs, we averaged the difference in representational strength across a late task epoch (11 – 16 s), and the difference in stimulus representation between ROIs remained (**Figure 2D**).

We validated the difference in stimulus representations through a series of control analyses. First, we demonstrated that these results cannot be explained by the specific model used to train the data (Liu et al., 2018; Sprague et al., 2018) nor the specific analytical approach used, because similar patterns were observed when we trained the IEM separately for each condition (**Figure S2A**) or adopted a Support Vector Machine (SVM) decoding approach (**Figure S2D**) (Henderson et al., 2022; Rademaker et al., 2019). 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 S2C**). Furthermore, to remove the potential 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.

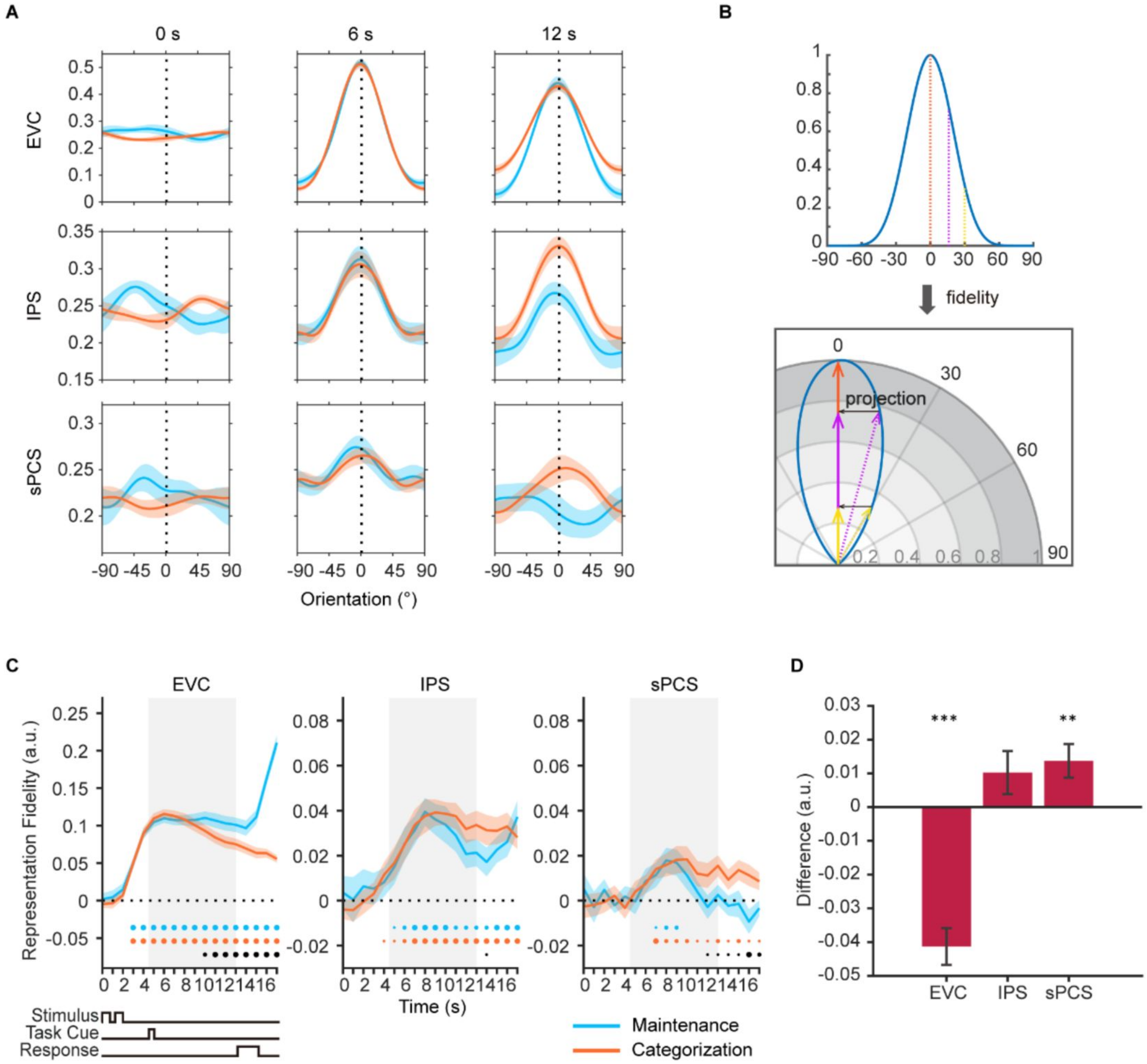


Figure 2.

Orientation reconstructions at the population level using IEMs.

(A) Reconstructed population-level orientation representations from selected time points at EVC, IPS, and sPCS for maintenance (blue) and categorization (orange) tasks, respectively. X axis represents distance from the cued orientation (at 0°), and y axis represents reconstructed channel responses in arbitrary units. Significant orientation representation was observed at 6 s and 12 s but not at 0 s. Shaded areas represent ± SEM. (B) To quantify the strength of orientation reconstructions, we calculated the reconstruction fidelity by first projecting the channel response at each orientation onto a vector at the cued orientation and then averaging the projected vectors. (C) Time course of representational strength of orientations at EVC, IPS and sPCS. Gray shaded areas indicate the entire memory delay following task cue. Blue and orange dots at the bottom indicate the FDR-corrected significance of representational fidelity at each time point of the corresponding task at $p < 0.05$ (small), $p < 0.01$ (medium), and $p < 0.001$ (large). The bottom black dots indicate significant difference in representational fidelity between tasks (uncorrected). Horizontal dashed lines represent a baseline of 0. Shaded areas represent ± SEM. (D) Average difference of representational strength across 11 – 16 s in each ROI (from EVC to sPCS: $p < 0.00001$, $p = 0.063$, $p = 0.007$, respectively). Positive difference indicates higher representational strength for categorization, and vice versa for negative difference. Black asterisks denote FDR-corrected significance, *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

Again, this analysis yielded similar findings (**Figure S2B**). Together, these results demonstrated enhanced stimulus representation in the frontal cortex with increased demand for active control, as well as those in the visual cortex with increased demand for precise WM maintenance.

Prediction of categorization behavior by frontal stimulus representation

Previous WM studies have shown that the strength of stimulus representation in EVC positively correlated with memory performance (Emrich et al., 2013; Ester et al., 2013; Gosseries et al., 2018), suggesting that EVC plays an important role in precise WM maintenance. However, stimulus representation in the frontal cortex rarely predicted behavioral performance in maintenance task (Hallenbeck et al., 2021). Nevertheless, if frontal stimulus representation is involved in WM control, its behavioral relevance should be subject to observation with increased control demands. Therefore, we assessed the behavioral predictability of stimulus representation during the delay period in EVC, IPS, and sPCS (**Figure 3**). Consistent with previous findings, we found the strength of stimulus representation in EVC during the early delay period predicted behavioral accuracies in both maintenance and categorization tasks (see **Supplemental Table 2** for full statistics). Similar predictability was found in IPS, with stimulus representation predicted behavior in the maintenance task during early delay and in the categorization task throughout the entire memory delay. Interestingly, we found that, throughout the entire memory delay, the strength of stimulus representation in sPCS predicted behavioral accuracies only in the categorization task but not in the maintenance task. These results highlighted the functional significance of stimulus representation in sPCS exclusively for the categorization task.

Reduced frontal stimulus representation with lower control demand

In Experiment 1, participants flexibly switched between two categorization rules to prompt the manipulation of WM content on a trial-by-trial basis. The rule switching increased control demand but also complicated the interpretation of our results. To exclude potential impact of rule switching, we conducted Experiment 2, in which participants performed maintenance and categorization tasks with only one fixed rule. Behavioral results of Experiment 2 again demonstrated a classic boundary effect and were comparable to Experiment 1 (**Figure S1**), with no significant difference between experiments in terms of either accuracy or reaction time (*F*s < 1.28, *p*s > 0.26). When using IEMs to reconstruct stimulus representation, we found EVC and IPS both showed patterns similar to those in Experiment 1 (**Figure 4A**), with stimulus representation decreased in EVC in categorization task and remained at the same level in IPS between the two tasks (see **Supplemental Table 3** for full statistics). The frontal region, sPCS, also showed a differential enhancement of stimulus representation in categorization task as in Experiment 1, but in an earlier delay period (**Figure 4A**). To validate such a temporal difference, we defined an additional early task period (5 – 10 s), and confirmed a significant difference in stimulus representation in sPCS during the early (p = 0.015; **Figure 4B**) but not during the late epoch (p = 0.372; **Figure 4C**). In addition, we performed a mixed ANOVA on experiments (Experiment 1 vs. 2) and epochs (early vs. late epoch) and observed a significant interaction effect between the two, $F(1, 46) = 7.43$, p = 0.009, suggesting that the two experiments differed in terms of the temporal emergence of the differential stimulus representation in the frontal cortex. Taken together, these results are consistent with our expectation that, with reduced control demand, the differential enhancement of stimulus representation in frontal cortex was still present but decreased during late memory delay. Nevertheless, stimulus representation in Experiment 2 still predicted behavioral performance as in Experiment 1, although the difference between task was reduced (**Figure S3**).

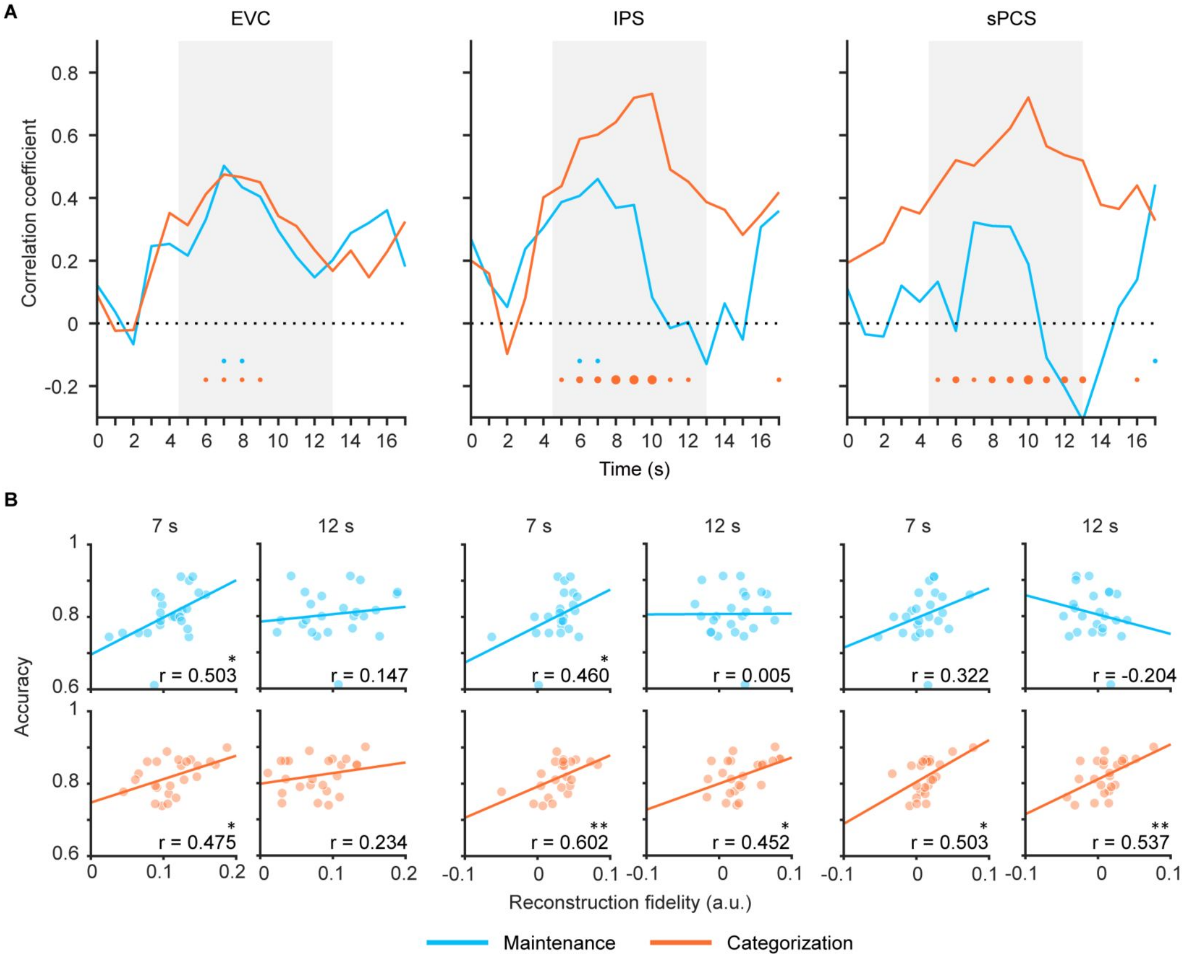


Figure 3.

Behavioral correlation of stimulus representation for maintenance (blue) and categorization (orange) tasks.

(A) Time course of correlation coefficients in EVC, IPS, and sPCS. Gray shaded areas indicate the entire memory delay following task cue. Blue and orange dots at the top indicate significance of correlation (uncorrected) at each time point at $p < 0.05$ (small), $p < 0.01$ (medium), and $p < 0.001$ (large). (B) Correlation scatter plots at representative time points (7 s and at 12 s) in EVC, IPS, and sPCS. R denotes Pearson correlation coefficients. Asterisks denote significant results, *: $p < 0.05$; **: $p < 0.01$.

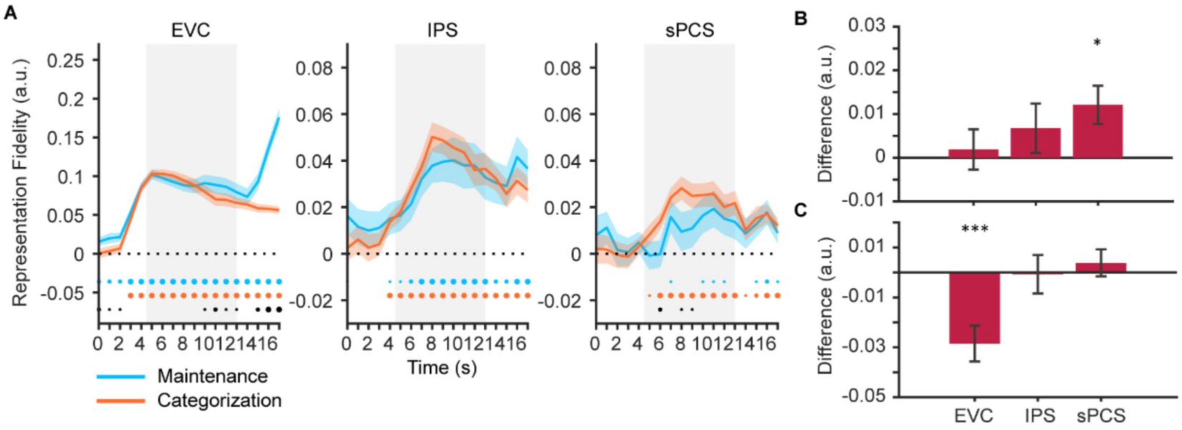


Figure 4.

Orientation reconstructions in Experiment 2 at the population level using IEMs.

(A) Time course of representational strength of orientations at EVC, IPS and sPCS. Gray shaded areas indicate the entire memory delay following task cue. Blue and orange dots at the bottom indicate the FDR-corrected significance of representational fidelity at each time point of the corresponding task at $p < 0.05$ (small), $p < 0.01$ (medium), and $p < 0.001$ (large). The bottom black dots indicate significant difference in representational fidelity between tasks (uncorrected). Horizontal dashed lines represent a baseline of 0. Shaded areas represent ± SEM. (B) Average difference of representational strength across an early task epoch (5 – 10 s) in each ROI. Positive difference indicates higher representational strength for categorization, and vice versa for negative difference (FDR-corrected). n.s.: not significant; *: $p < 0.05$; ***: $p < 0.001$. (C) Average difference of representational strength across a late task epoch (11 – 16 s) in each ROI. Same conventions as (B).

Category representation in WM in various cortices

Having observed a differential representation of stimulus in the frontal cortex, we next asked how newly generated information in WM during the categorization task emerged and sustained in the distributed WM network and how representations of the original stimulus and new information interacted. The categorization task could demand additional generation of category information in WM. We therefore trained SVMs to decode category information during the categorization task. For each rule, the SVM decoder was trained to discriminate between the two categories. In both experiments, we found that during the late epoch, category information could be well decoded across ROIs (*ps* < 0.044, **Figure 5A**; also see **Figure S4A** for full decoding time course), with a marginal difference between experiments in sPCS ($p = 0.055$).

One might argue that the category decoding results could at least be partially attributed to stimulus similarity. 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 orthogonal rule). We then calculated an abstract category index by subtracting decoding accuracy under the opposite rule from that under the true rule (Mok & Love, 2020) (**Figure 5B**). The rationale was that the amount of stimulus similarity would be comparable for the opposite rule, but additional category information, if existed, should result in higher decoding accuracy for the true rule. After removing stimulus-related signals, average decoding performance of abstract category was only evident in Experiment 2 (*ps* < 0.017) but not in Experiment 1 (*ps* > 0.14) for all ROIs. Moreover, decoding performance of abstract category was significantly higher in Experiment 2 than Experiment 1 in sPCS ($p = 0.034$; **Figure 5C**). These results together suggest a potential tradeoff between stimulus difference and category representation in the frontal cortex.

Differential stimulus representation in frontal cortex replicated by RNN modeling

Lastly, we tested how stimulus representation could emerge in frontal cortex at the mechanistic level using recurrent neural network (RNN) models. Our hypothesis is that precise stimulus representation during WM might emerge in frontal cortex in response to complex task demands such as rule-based categorization. In other words, instead of relying (solely) on category representations, the cortical network might have adopted a different strategy to accommodate the flexible task requirements in the current study, for instance, by preserving stimulus information until a later stage of information processing. This different strategy can be implemented by altering the RNN's output structure. Therefore, the logic of this modeling analysis was to examine whether explicitly placing a demand for the model to preserve stimulus representation would recapitulate our fMRI findings in frontal cortex, in comparison to a model that did not specify such a demand.

Two types of modular RNNs were trained on the maintenance and categorization tasks simultaneously (Masse et al., 2019; Zhou et al., 2021). 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 in the hidden layer (80% excitatory + 20% inhibitory units, equally distributed in three separate modules). The only difference was in the structure of the output layer. The first type of RNN (RNN1; n = 20) had only two units in the output layer to indicate networks' choice (**Figure 6A**), whereas the second type of RNN (RNN2; n = 20) had additional units in the output layer corresponding to the original stimulus information. In other words, the second RNN was designed to maintain stimulus information throughout the network modules. For the common hidden layer, we included three hierarchically organized (posterior, middle, and anterior) modules of recurrent units generated according to neurobiological principles of neuronal connections (e.g., denser connectivity within than between modules) to simulate the interconnected brain areas in our ROI-based fMRI analyses above: the

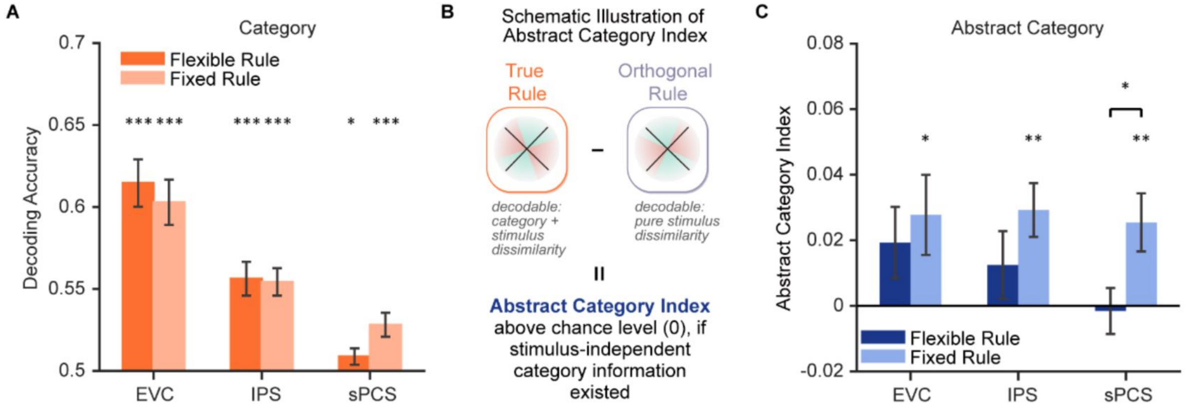


Figure 5.

Decoding performance for category and abstract category information.

(A) Average category decoding accuracy across the delay period (11 – 16 s) in each ROI of both experiments, black asterisks denote FDR-corrected significance, n.s.: not significant; *: p < 0.05; ***: p < 0.001. (B) Schematic illustration of abstract category decoding. In categorization task, category information can be decoded using category labels according to the true categorization rule. On the other hand, category can also be decoded due to stimulus similarity. Thus, to remove stimulus-dependent categorical information, we calculated an abstract category index by removing decoding accuracy using orthogonal category boundaries (assuming comparable stimulus-dependent effect) from that using true rule boundaries. (C) Average abstract category decoding index across the delay period (11 – 16 s) in each ROI of both experiments, black asterisks denote FDR-corrected significance, n.s.: not significant; *: p < 0.05; **: p < 0.01.

posterior module (Module 1, simulating EVC) was directly connected with the input layer, the middle module (Module 2, simulating IPS) received projections from the posterior module and relaying information to the anterior module, and the anterior module (Module 3, simulating sPCS) projected to the output layer. Task events were simulated as numerical inputs to the model, matching the procedures of Experiment 1 (see Methods for details).

After successful training, defined as reaching at least 90% accuracies in all tasks in the same training batch, we applied an SVM decoding approach to investigate population-level stimulus representations in neuronal spiking activities of the RNNs. We found that in RNN1, both the middle and anterior modules showed stronger stimulus representation in the maintenance task than the categorization task during the delay period ($p_{posterior}$ = 0.09, p_{middle} = 0.011, p_{anterior} = 0.007; **Figure 6B** and Figure S7A), opposite to our fMRI observation in IPS and sPCS. In comparison, decoding performance in RNN2, which was explicitly required to maintain stimulus information for the output, yielded results consistent with our human findings, with increased stimulus decoding performance during categorization only in the anterior module ($p_{\mathrm{posterior}}$ = 0.436, p_{middle} = 0.212, p_{anterior} = 0.026; **Figure 6B**).

Besides difference in stimulus representation, we further tested whether RNN2 could also replicate the human results on category representation. For this analysis we focused on abstract category representation to fully remove the impact of stimulus on category decoding. To examine the influence of control demand on category decoding, following our fMRI experiment we trained 20 additional RNNs with the same output structure as RNN2 (preserving stimulus information) to perform the tasks with a fixed categorization rule, mimicking the task structure of Experiment 2. Consistent with our human findings, we observed increased abstract category decoding performance in the fixed-rule RNNs compared to the flexible-rule RNNs, throughout the modules ($p_{posterior}$ = 0.045, p_{middle} = 0.003, p_{anterior} < 0.001; **Figure 6C**). By contrast, abstract category decoding in RNN1 across modules demonstrated a distinct pattern from human data, with numerically increasing decoding accuracy towards later modules (**Figure 6C**).

Altogether, these findings demonstrated that our fMRI results could be simulated by RNN models when stimulus information for readout was preserved, suggesting that the requirement for flexible control of WM content could demand high-fidelity stimulus representation at the output stage of the model. Notably, we found that RNN2 generally took less iterations for training and had fewer failures in learning the task (with a defined maximal number of iterations).

Discussion

In this study, we investigated the emergence and maintenance of stimulus representation with varied control demands of WM. In a distributed human cortical network encompassing visual, parietal, and frontal cortex, we found enhanced stimulus representations in the frontal cortex that tracked increasing demands on active WM control, as well as enhanced stimulus representations in the visual cortex that tracked the demand for the precise maintenance of WM content. The enhanced stimulus representation in frontal cortex was well predicted by RNNs that 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 for distinct aspects of WM, and help to resolve the current controversy in the roles of various cortices in WM.

Role of visual cortex in WM maintenance

The visual cortex has been considered a critical site for maintaining visual WM in the context of sensorimotor recruitment hypothesis (D’Esposito & Postle, 2015; Harrison & Tong, 2009). This idea, however, has been challenged in recent years due to some seemingly contradictory findings from the human neuroimaging studies. For example, compared to the frontoparietal cortex,

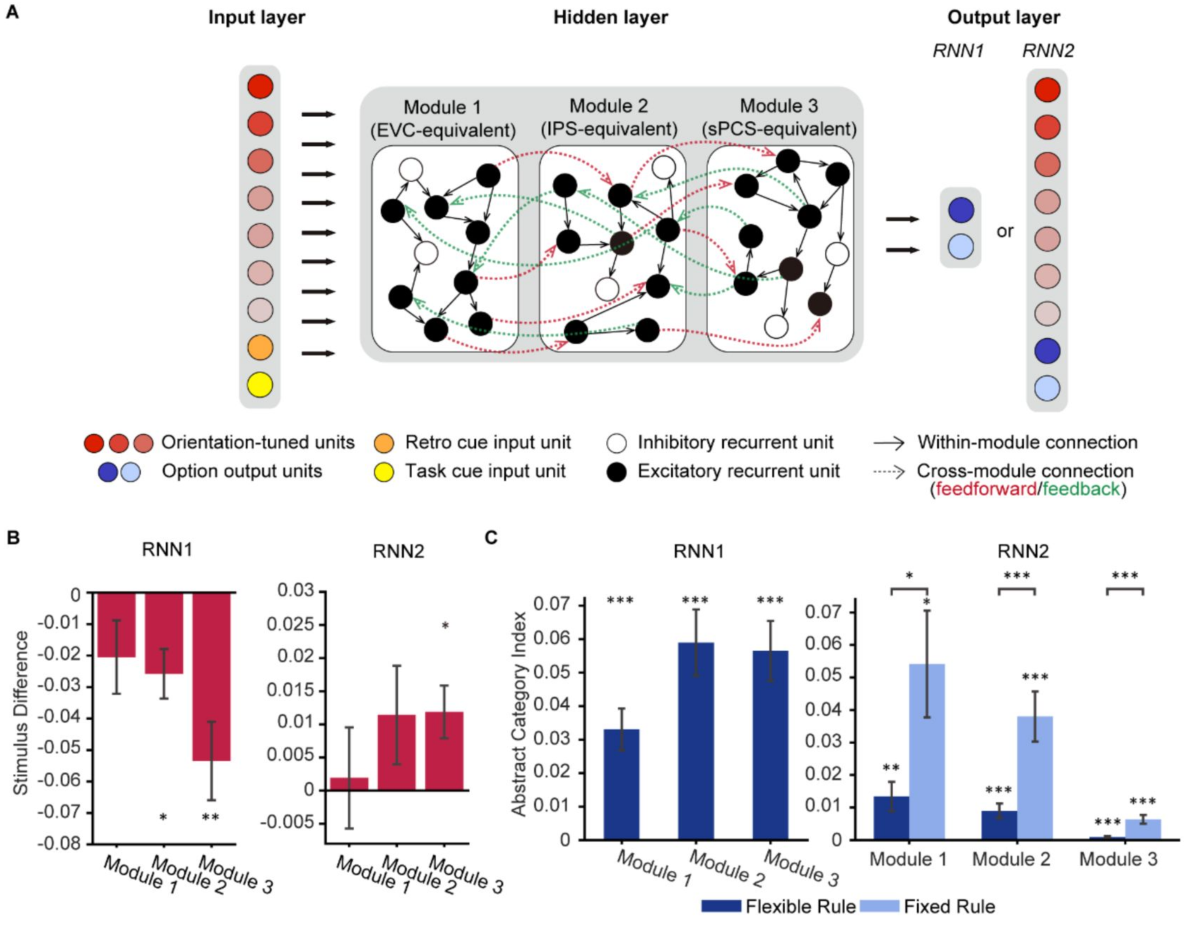


Figure 6.

Architecture of RNNs and simulation results.

(A) All networks consist of 3 layers of artificial units: the input, hidden and output layers. For both RNN1 and RNN2, 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 three modules of 200 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 and green arrows), which only occur between excitatory units. Only excitatory units in module 1 receive projections from the input layer and only excitatory units in module 3 project to the output units. For RNN1, networks output (0,1) or (1,0) through the 2 units in the output layer to indicate responses. For RNN2, the network output (0,1) or (1,0) to report the category to which the cued orientation belonged in the categorization task, or (0,0) in the maintenance task (blue units). Importantly, the models also output the orientation itself through 15 additional orientation-tuned units (red). (B) Difference in orientation decoding between tasks in RNN1 and RNN2. Results were averaged across the delay period. Positive difference indicates higher decoding accuracy for categorization, and negative difference indicates higher decoding accuracy for maintenance. Error bars represent ± SEM. Black asterisks denote FDR-corrected significance, n.s.: not significant; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. (C) Average abstract category information across the delay period for RNN1 and RNN2. Same conventions as (B).

mnemonic representations in EVC were found to be more vulnerable to distractors (Bettencourt & Xu, 2016; Hallenbeck et al., 2021; Lorenc et al., 2018). The decodability of memory contents in visual cortex also depends on the specific task type. A previous study showed that in nonvisual tasks that required judgements on object category instead of visual details, memory contents were no longer decodable in the visual cortex (Lee et al., 2013). In this study, we found that, although the strength of stimulus representation in EVC differed between WM maintenance and categorization tasks, a copy of stimulus representation remained in EVC during the categorization task. Moreover, stimulus representations in both tasks were equally predictive of subsequent memory performance, suggesting the functional significance of EVC representations in WM.

The discrepancy between our results and that of the previous work (Lee et al., 2013) could be attributed to the fact that our categorization task required participants to manipulate remembered information according to arbitrary yet flexible categorization rules, rather than simply paying selective attention to different aspects (visual details vs. category membership) of everyday objects. In our case, maintaining visual details of the memoranda was critical for accurate behavioral responses. Our finding is consistent with the prediction of sensorimotor recruitment hypothesis that representation of memory contents in the visual cortex is necessary for the precise maintenance of visual information. The observation of robust category representation in early visual cortex during the response period further indicated the recruitment of EVC in categorization, 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 (Henderson et al., 2022). In addition, electrophysiological research in non-human primates has also shown robust feature selectivity in the visual cortex during a categorization task (Brincat et al., 2018), and recent computational modeling work has suggested intact maintenance of sensory information during categorical judgements (Luu & Stocker, 2021).

Role of frontal cortex in active WM control

Compared to the prominent role of EVC in memory maintenance, sPCS in the frontal cortex played a dominant role in WM tasks that require active control of memory contents such as categorization. Although stimulus representations in sPCS have been observed during WM in previous studies, the nature of these representations remained debatable. In WM tasks that required mere maintenance of memoranda, stimulus was not always decodable in the frontal cortex (Emrich et al., 2013; Gosseries et al., 2018; Riggall & Postle, 2012), raising the issue of functional significance of stimulus representation in the frontal cortex. On the other hand, stimulus representation in the frontal cortex could become robust in the face of tasks that require attentional prioritization and extensive training (Bettencourt & Xu, 2016; Christophel et al., 2018; Hallenbeck et al., 2021; Lorenc et al., 2018; Miller et al., 2022). Our current study contributes to the resolution of this issue by demonstrating that stimulus representation in sPCS increased with increasing demands for WM control. This finding is in line with recent computational studies proposing that active WM functions may involve neuronal mechanisms different from that for passive maintenance. For example, passive maintenance could rely mainly on synaptic plasticity mechanisms, whereas active control functions such as distractor resistance and information manipulation involve more neuronal spiking activity (Masse et al., 2019; Wang, 2021). In this study, we provided the first empirical evidence that the frontal cortex exhibits enhanced stimulus representation in categorization task requiring active WM control and this representation is predictive of WM performance. In contrast, stimulus representation of WM maintenance failed to predict WM performance at high control demand. It would be of interest to further investigate whether this active control in the frontal cortex could be generalized to tasks that require other types of WM control such as mental rotation.

WM representations in frontal cortex support cognitive flexibility

Our results in the frontal cortex are also in line with recent theoretical proposals in the field of cognitive flexibility. To behave flexibly in complex environments with limited cognitive resources, two mechanisms have been proposed: low-dimensional abstraction of stimulus representation for generalization and efficient learning, and high-dimensional stimulus representation for separability and flexible readout (Badre et al., 2021; Flesch et al., 2022; Fusi et al., 2016). Within this framework, high-dimensional stimulus representations during WM might emerge in the frontal cortex in response to complex control demands such as rule-based categorization. The results of the two fMRI experiments in the current study jointly demonstrate a dynamic tradeoff between high-dimensional stimulus and low-dimensional category representations depending on the control demand. Specifically, when control demand was reduced with a single categorization rule in Experiment 2 compared to Experiment 1, the differential stimulus representation in the frontal cortex was also reduced during the late delay period, accompanied by an increase in category decoding performance especially in the frontal cortex. This result is consistent with neurophysiological findings in non-human primates: while robust category selectivity was observed in frontoparietal cortex during the delay period of categorization tasks when the animal was trained on the categorization task only (Brincat et al., 2018; Freedman & Assad, 2006; Freedman et al., 2001; McKee et al., 2014), category selectivity in the parietal cortex was significantly reduced when the animal had been exposed to a maintenance task prior to categorization training (Latimer & Freedman, 2023). Our RNN simulation further confirmed that this dynamic reconfiguration in information coding at the network level can be well explained by a change in the coding strategy for the network readout. In other words, in flexible environments, and with rich prior experience, the brain might adopt an entirely different strategy for processing information in WM. High-dimensional stimulus information might be preserved in its original identity in the higher-order cortex, potentially reducing processing demands in dealing with each task and thereby facilitating efficiency and flexibility (Badre et al., 2021; Flesch et al., 2022; Fusi et al., 2016). One important future direction would be to further address the meta-control mechanisms that determine the flexible selection of coding strategies for WM (Eppinger et al., 2021).

Differentiating between frontal and parietal cortex in WM functions

While many previous WM studies have focused on the functional distinction between sensory and frontoparietal cortex, it has remained less clear how frontal and parietal cortex might differ in terms of WM functions. Some studies have reported stimulus representations with similar functionality in frontal and parietal cortex (Christophel et al., 2018; Yu & Shim, 2019), while others have observed differential patterns (Hu & Yu, 2023; Lee et al., 2013; Li et al., 2023). We interpret the differential patterns as reflecting a difference in the potential origin of the corresponding cognitive functions. For example, in our study, sPCS demonstrated the most prominent effect for enhanced stimulus representation during categorization as well as the tradeoff between stimulus difference and category representation, suggesting that sPCS might serve as the source region for such effects. On the other hand, IPS did show visually similar patterns to sPCS in some analyses. For instance, stimulus representation in IPS was visually but not statistically higher in the categorization task. These results together support the view that our findings in sPCS do not occur in isolation, but rather reflect a dynamic reconfiguration of functional gradients along the cortical hierarchy from early visual to parietal and then to frontal cortex.

Conclusion

In conclusion, we observed a distributed cortical network, including early visual, parietal, and frontal cortex, in representing stimulus-specific information in WM. These stimulus representations in visual and frontal cortex played distinct functional roles, with those in EVC contributing primarily to precise maintenance and those in frontal cortex contributing primarily to active control in WM. RNN simulations indicated that the stimulus representation in the frontal cortex might have emerged as a result of output selection to facilitate cognitive flexibility. Collectively, these results help to reconcile current debates on the functional roles of different cortical regions in WM, and provide new insights into how a unified WM framework could support varied control demands.

Methods

Participants

A total of 54 participants were recruited at Chinese Academy of Sciences, Shanghai Branch. Twenty-six healthy participants (21 female, all right-handed, mean age = 24.0 ± 1.4 years) were recruited for Experiment 1. Two were excluded due to failure in completing the experiment or low conformity to task instructions, remaining 24 participants who completed the main experiment (19 female, mean age = 23.92 ± 1.41 years). Twenty-eight (22 female, mean age = 24.14 ± 1.51 years) participants were recruited for Experiment 2. Two quitted after behavioral training and two did not finish scanning due to technical problems with the scanner, resulting in 24 participants (20 female, mean age = 24.13 ± 1.60 years) in the final analyses. 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, 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 (Brainard, 1997; Pelli, 1997). 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

In Experiment 1, 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 30 learning trials in each run. They reviewed rule disks after every 10 trials for memory reconsolidation. Each participant completed at least two runs for each rule. After achieving an average accuracy 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 the boundaries of learned rules as a qualification of behavioral training. If the 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.

Behavioral training for Experiment 2 followed the same procedure and used the identical sample set (30 orientations ranging from 4° to 178°) as Experiment 1, except that only one rule was trained. Half of the participants in Experiment 2 learned rule boundaries of 19°/109°; the other half learned rule boundaries of 67°/157°.

Flexible WM Task (fMRI Task)

In Experiment 1, during scanning, participants completed a flexible WM task which implemented levels of control demand with different rules. To be specific, participants randomly switched between a maintenance task and a categorization task. In the maintenance task, participants needed to memorize stimulus information (i.e., orientations). In the categorization task, participants needed to categorize orientations following the rule that was randomly assigned and cued 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 categorization 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 categorization 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 categorization trials and three maintenance trials. Each run contained two blocks (i.e., 18 trials).

In Experiment 2, to isolate potential effect of rule switching, the categorization rule stayed the same throughout the experiment. In Experiment 2, participants randomly switched between the maintenance task and the categorization task. Each trial followed the same procedure as Experiment 1.

In Experiment 1, orientations, tasks, and cued target order (1st or 2nd) were counterbalanced across trials, resulting in an equal trial number of 90 across all three conditions (categorization-Rule A, categorization-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. The same counterbalancing procedure was conducted for Experiment 2 (90 trials for maintenance task and 180 trials for categorization task). In Experiment 2, six participants completed 15 runs of 18 trials; the other seven completed 18 runs of 15 trials each due to scanner limitations. At the end of scanning, participants reported the rule boundaries three times as a final check of their rule memory.

Data Acquisition

MRI data of Experiment 1 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. 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. MRI data of Experiment 2 were collected using identical procedure and settings except that the last eleven participants were scanned using a newly installed 3 Tesla Siemens MRI scanner (Prisma; Siemens Healthineers) at the Functional Brain Imaging Platform.

Preprocessing

All preprocessing of individual MRI data was performed using AFNI (afni.nimh.nih.gov) (Cox, 1996; Cox & Hyde, 1997). 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, WM-related brain areas: early visual cortex (EVC), intraparietal sulcus (IPS) in parietal cortex, and superior precentral sulcus (sPCS) in frontal cortex (Ester et al., 2015; Hallenbeck et al., 2021; Yu & Shim, 2017). We created anatomical ROI masks based on the probabilistic atlas by Wang and colleagues (Wang et al., 2015). 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

Multivariate Inverted Encoding Modeling (IEM)

Neural representations of orientations were reconstructed using inverted encoding modeling (IEM) (Brouwer & Heeger, 2009, 2011; Ester et al., 2015; Rademaker et al., 2019; Yu & Shim, 2017) with custom MATLAB scripts on individuals' BOLD activation patterns in the three 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 to train and to test IEM in order to avoid potential biases from a specific task condition. Results for categorization task were averaged across rules for Experiment 1. 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:

$$B_1 = WC_1$$

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

The least-squared estimates of the weight matrix ($\hat{W}$) 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 (Rademaker et al., 2019; Yu & Postle, 2021). 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° (true orientation of trials). The responses from all trials were averaged to obtain reconstructed orientation representations for the test datasets.

To quantify the strength of each IEM reconstruction, we calculated reconstruction fidelity of channel responses. First, all channel responses were projected to the vector at the true orientation. Then the reconstruction fidelity was calculated as the mean of all projected vectors (Rademaker et al., 2019). A larger fidelity value indicates a stronger positive representation of orientation.

Multivariate Pattern Analysis (MVPA)

Besides IEM, we tracked neural representations of stimulus and of category using linear Support Vector Machine (SVM) decoders. All decoding analyses were performed using the templateSVM and fitcecoc functions in MATLAB.

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 using the same leave-one-run-out cross-validation procedure as in IEM analyses. To avoid biases in model training, we randomly balanced the trial numbers for each bin in the training set. Decoding accuracies were then computed by averaging performance of cardinal and oblique classifiers. For the categorization task, we averaged accuracies across rules.

Decoding of category information for Experiment 1 was performed under each rule (90 trials for each rule) using a leave-one-trial-out cross-validation procedure (see next paragraph for details), and the decoding accuracies were then averaged across rules. Since Experiment 2 adopted a fixed rule with 180 trials in the categorization task, we randomly divided categorization trials into two halves with 90 trials each, and decoded category information for Experiment 2 using identical procedures as for Experiment 1.

Because closer orientations are more similar to each other inherently, orientations per se could contain categorical information by visual similarity. 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 the opposite category decoder. Thus, an abstract category index was calculated by subtracting opposite category decoding accuracy from true category decoding accuracy (i.e., chance level = 0). Since the opposite category decoding used re-assigned labels, to eliminate imbalance in trial number between true and opposite categories, we used a leave-one-trial-out cross-validation procedure for true category and opposite category decoders. Decoding for Experiment 1 was performed separately for each rule and were then averaged. Decoding for Experiment 2 was performed separately for randomly divided halves and averaged.

Recurrent Neural Network Simulations

RNN architecture

The network model was built following the details in previous work (Masse et al., 2019), and implemented in TensorFlow (version: Nvidia-tensorflow 1.15.0) (Abadi et al., 2016). 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. In order to simulate neural activity patterns in hierarchically connected brain regions (EVC, IPS and sPCS), we separated the hidden layer into three modules, each containing 200 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 (Zhou et al., 2021), modularity was achieved by constraining the recurrent connectivity in the hidden layer. Specifically, only posterior module's (module 1) excitatory units received inputs from the input layer and only anterior module's (module 3) excitatory units projected to the output units. Between-module connections were culled so that only 50% of a module's excitatory units were randomly connected to their counterparts in the neighboring module(s), and vice versa (feedforward and feedback connections). Connections among inhibitory units remained strictly within-module in accordance with the observation that inhibitory connections in cortex are largely local. Thus, posterior, middle and anterior modules were intended to simulate the three interconnected ROIs we used in the fMRI analyses that posited differently at the processing hierarchy. We specifically manipulated the output demand to investigate whether it would alter similarity of the results to the fMRI observations. To this end, one type of network architecture (RNN1) implemented a two-unit output layer with each unit corresponding to one of two possible response options, presented through the input units before the test period; In contrast, the other type of RNN architecture (RNN2) had additional units in the output layer, creating a demand for preserving the original stimulus information alongside categorical representations.

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 of its preferred orientation to the input angle. Stimulus values were selected from an array of 20 orientations evenly spanned from [0 to 180) degrees. The sequentially-presented stimuli were presented through the same receptive field, followed by retro-cue and task cue indicated through the separate input units. For RNN1, before the test period when the network was required to make a choice, two response options were presented sequentially through the same input receptive field. The selection of the options varied slightly between the maintenance and categorization tasks: in maintenance, one orientation was always the cued sample while the other was randomly chosen from all other possible angles. In categorization, 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, RNN2 output (0,1) or (1,0) to report the category to which the cued orientation belonged in the categorization task, or (0,0) in the maintenance task. Importantly, the model also needed to report the cued orientation itself through a receptive field consisting of 15 orientation-tuned units in the output layer.

Training parameters and procedure

The hyperparameters and procedure for training the models were consistent with those detailed in previous work (Masse et al., 2019), 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 a similar vein, we also expanded the number of hidden units to 600 and number of training iteration to a maximum of 10000. Additionally, spike penalty was set to 0 for both RNN models to remove constraints on neuronal activity.

We trained 20 models for each type of RNN and results were obtained by averaging over all of them (for single-rule RNN, 10 models for each categorization rule). The goal of the training process was to minimize the mean square error between the model outputs and correct outputs during the test period via back propagation (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 categorization Rule A and 300 categorization Rule B trials, with the task block order randomized (for single rule RNN, each batch consisted of 300 maintenance and 300 categorization trials). Training would automatically stop if the model achieved 90% accuracy in the last training batch on all tasks. For RNN2, the accuracies for category and stimulus outputs were calculated separately to ensure precision of the stimulus 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.

Statistical Testing

Participants' behavioral performance for the main task was assessed using accuracy and reaction time. Paired t-test was conducted for the two task types (maintenance & categorization) to evaluate differences between conditions.

Statistical significance was evaluated via a sign-flip permutation procedure for all other analyses. For example, to characterize the significance of IEM fidelity, we computed the p-value by comparing the true mean fidelity of our sample with a null distribution reflecting no IEM fidelity. The null distribution was created by randomly assigning 1 or -1 to fidelity scores of our sample and then averaging the sign-flipped samples for 10,000 times, resulting in a null distribution of 10,000 fidelity scores. To characterize the difference of IEM fidelity between tasks, we sign-flipped the fidelity sample for each condition and then averaged the difference for 10,000 times. The p-value was calculated by comparing the true mean difference with the generated null distribution of difference. P-values were corrected using FDR across ROIs, time (TRs), and tasks for all analyses unless specified. An early (5-10 s; 6^{th}-11^{th} TR) and late (11-16 s; 12^{th}-17^{th} TR) task epoch was also defined to facilitate comparisons between ROIs and experiments when needed.

For RNN decoding results, we adopted a cluster-based permutation approach (using MNE-Python (Gramfort et al., 2013) function *permutation_cluster_1samp_test* to accommodate the large number of time points to be corrected for) to determine statistical significance of the time course, for stimulus decoding accuracies in maintenance/categorization, difference between stimulus decoding accuracies between tasks, and abstract category decoding accuracy in categorization task. Moreover, we also pooled decoding accuracies across a critical task period (50-75 time points during delay) to produce summary statistics aligning with what was reported in the fMRI results. Average decoding results were corrected using the FDR method.

Acknowledgements

We would like to thank Dr. Mu-ming Poo for valuable comments on an earlier version of the manuscript, and Dr. Tianming Yang for helpful discussions. 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), CAS Project for Young Scientists in Basic Research (YSBR-071), and Shanghai Pujiang Program (22PJ1414400) to Q.Y.

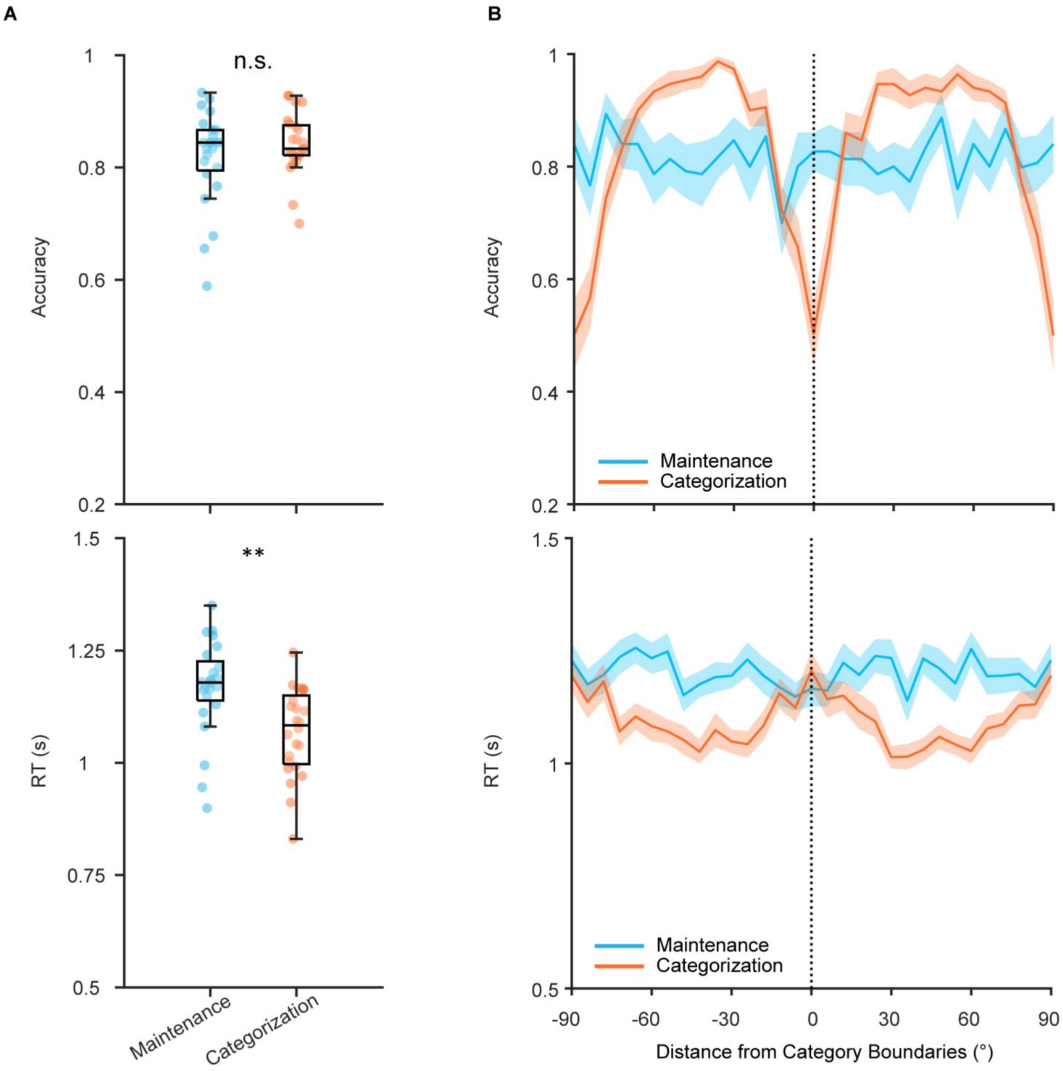


Figure S1.

Behavioral performance of Experiment 2.

(A) Accuracy (upper) and reaction time (lower) of Maintenance (blue) and Categorization (orange) tasks in Experiment 2. Asterisks denote significant results, n.s.: not significant; **: $p < 0.01$. (B) Accuracy (upper) and reaction time (lower) for orientations based on their distances from category center for Categorization task. Shaded areas represent ± SEM. Vertical dashed line represents category center.

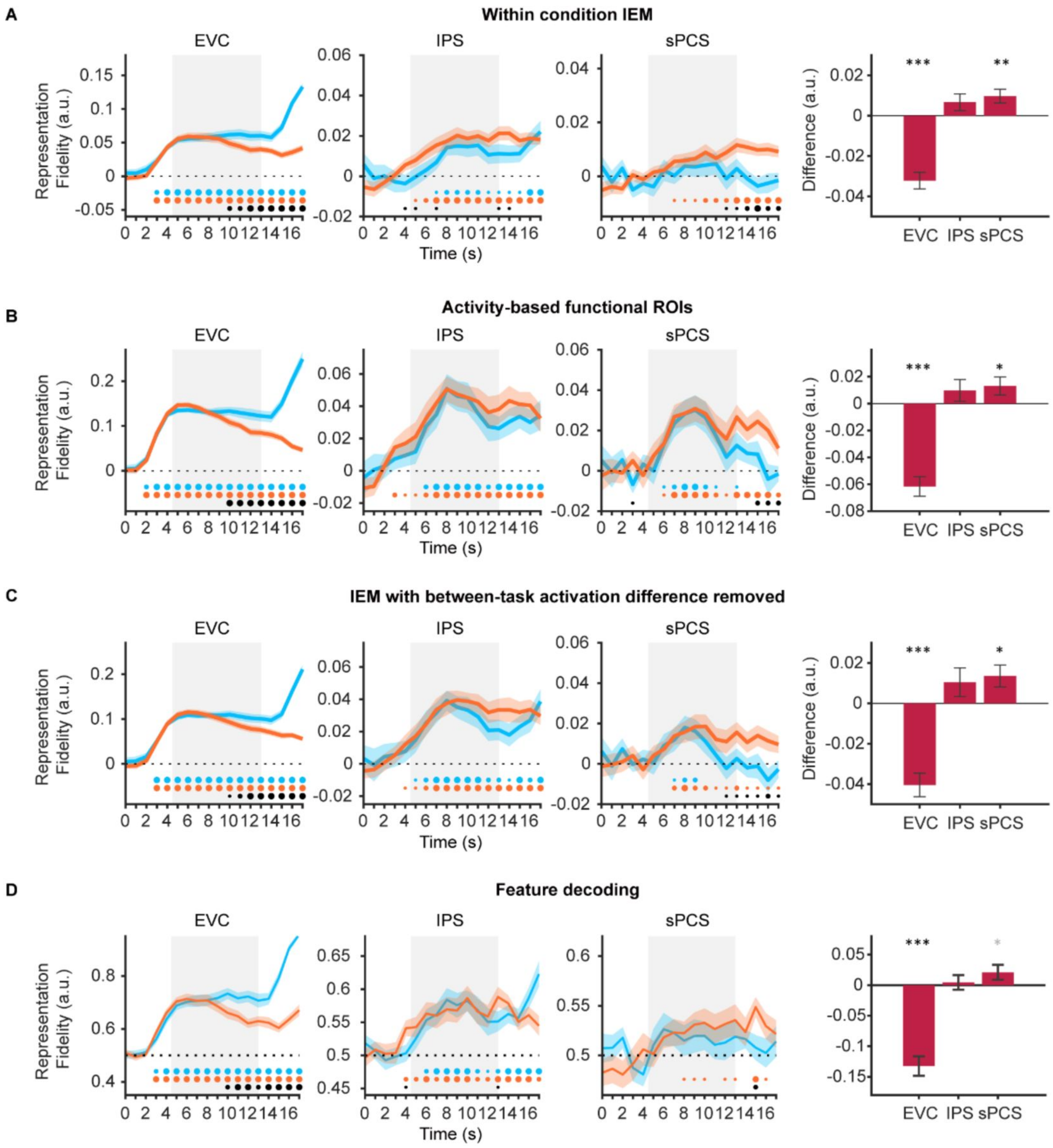


Figure S2.

Control analyses for stimulus representation results.

(A) Time course of representational strength of orientations in EVC, IPS and sPCS using IEMs trained separately for each condition. Bar plot on the right showing corresponding averaged difference between tasks. Average difference of representational strength across later delay period (11 – 16 s) in each ROI. Positive difference indicates higher representational strength for categorization, and vice versa for negative difference. (B) Time course of representational strength of orientations in functional ROIs defined by top 500 most selective voxels during sample or delay period. Bar plot same as (A). (C) Time course of representational strength of orientations after removing voxel-wise mean activation for each condition at each TR. Bar plot same as (A). (D) Time course of stimulus decoding accuracy. Bar plot same as (A). Gray shaded areas indicate the entire memory delay following task cue. Horizontal dashed lines represent a baseline of 0 or 0.5. Blue and orange dots at the bottom indicate the significance of representational fidelity at each time point of the corresponding task at $p < 0.05$ (small), $p < 0.01$ (medium), and $p < 0.001$ (large). The bottom black dots indicate significant difference in representational fidelity between tasks. Shaded areas represent ± SEM. n.s.: not significant; black asterisks denote significance, *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. Gray asterisk denotes marginal significance p < 0.1.

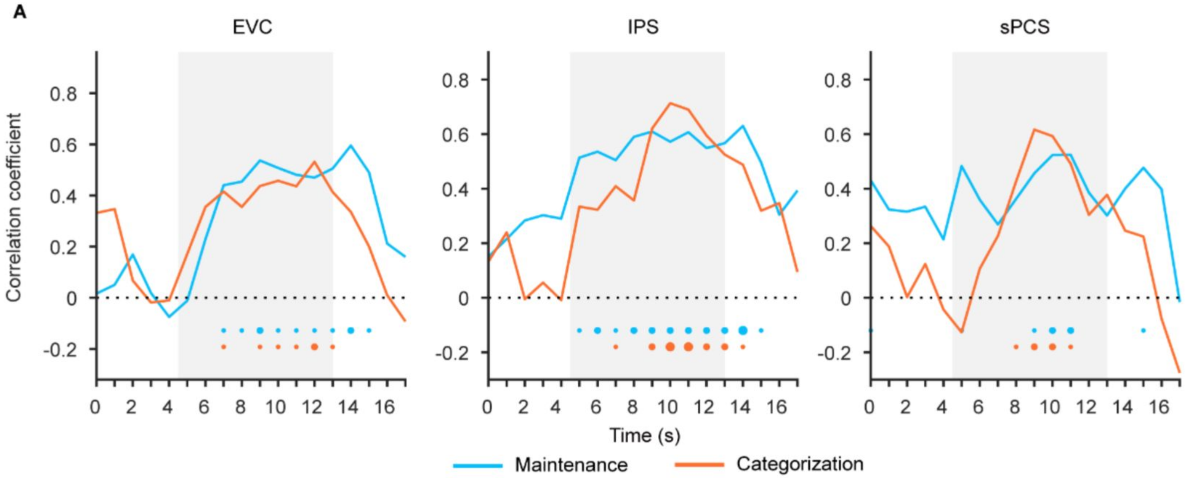


Figure S3.

Behavioral correlation of stimulus representation in Experiment 2.

(A) Time course of correlation coefficients in EVC, IPS, and sPCS. Correlation was performed between strength of stimulus representations and behavioral performance (accuracy) for Maintenance (blue) and Categorization (orange) tasks. Gray shaded areas indicate the entire memory delay following task cue. Horizontal dashed lines represent a baseline of 0. Bottom dots indicate the significance of corresponding analyses at each time point of the corresponding task at $p < 0.05$ (small), $p < 0.01$ (medium), and $p < 0.001$ (large). Shaded areas represent ± SEM.

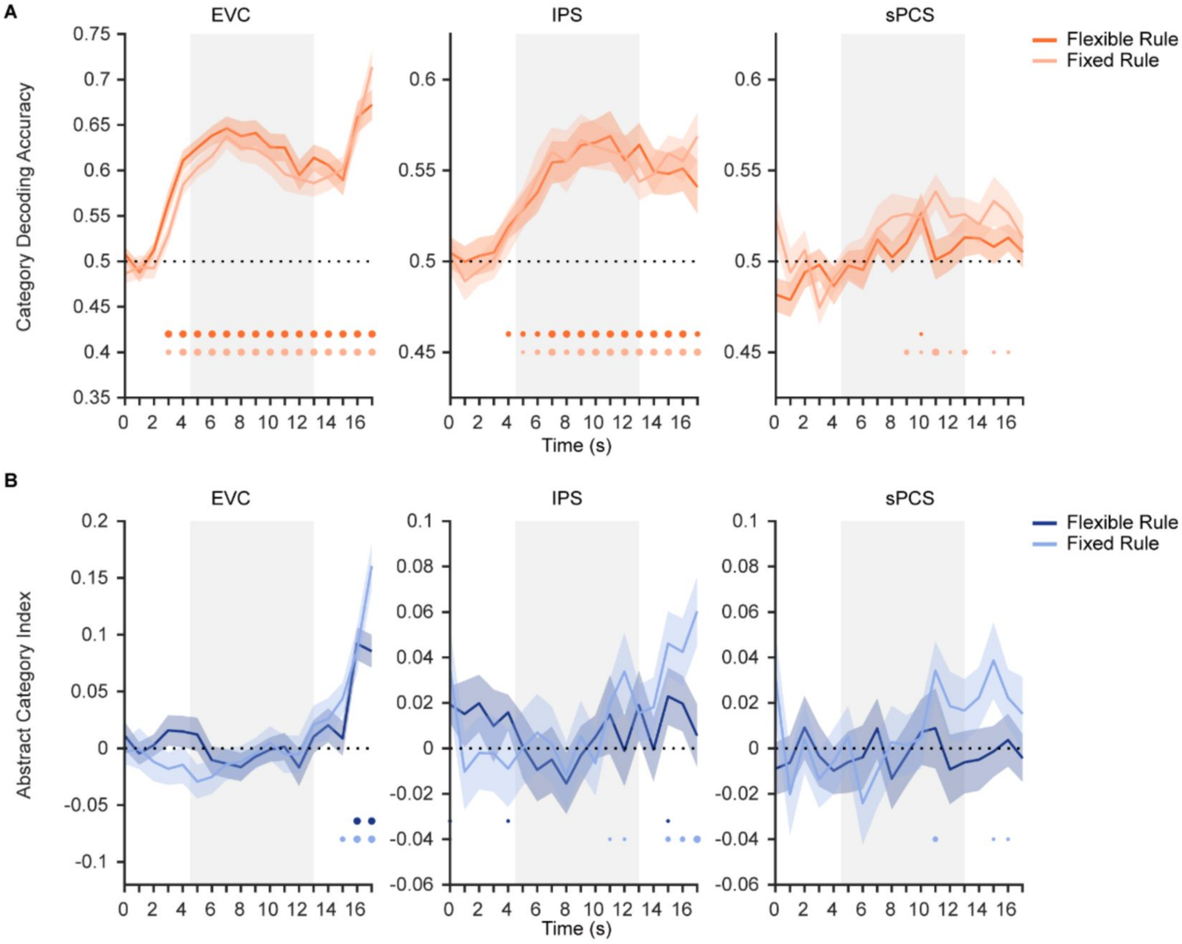


Figure S4.

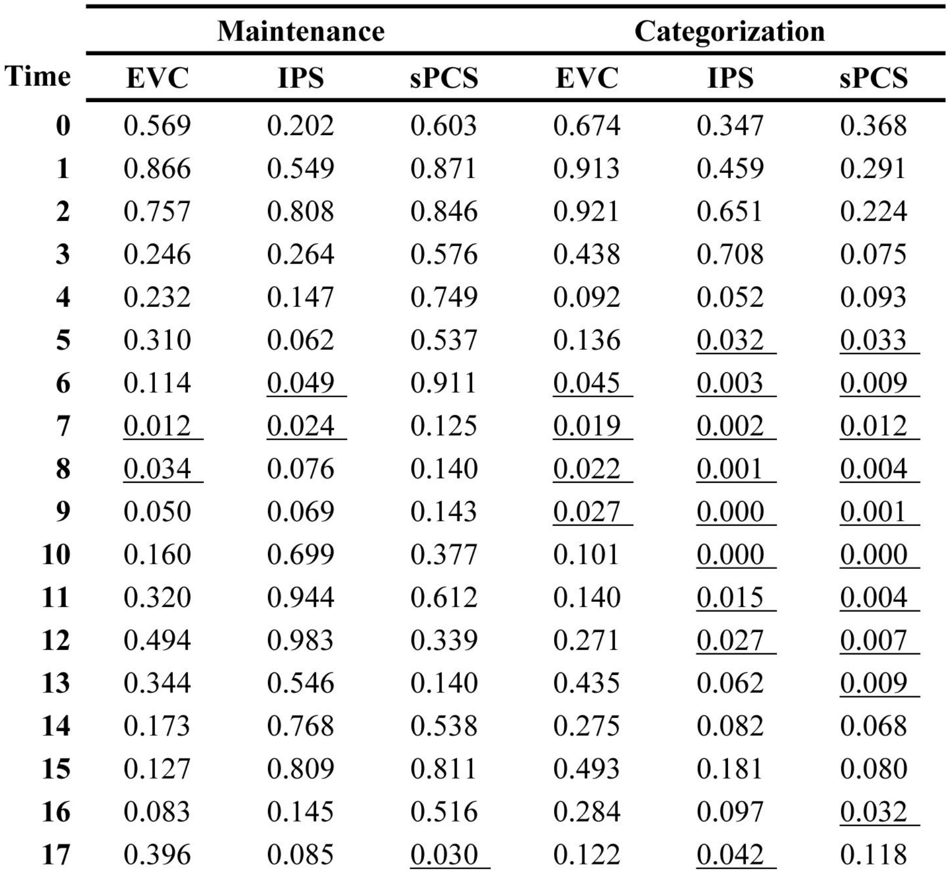
Category, Abstract Category and Rule Information in Experiment 1 and 2.

(A) Time course of category decoding strength in Experiment 1 with flexible rule (orange) and in Experiment 2 with fixed rule (light orange). Horizontal dashed lines represent the chance level of 0.5. Gray shaded areas indicate the entire memory delay following task cue. Bottom dots indicate FDR-corrected significance of decoding accuracy at each time point at $p < 0.05$ (small), $p < 0.01$ (medium), and $p < 0.001$ (large). Error bars represent ± SEM. (B) Time course of abstract category decoding strength in Experiment 1 (dark blue) and in Experiment 2 (light blue). Horizontal dashed lines represent a baseline of 0. Gray shaded areas indicate the entire memory delay following task cue. Bottom dots indicate uncorrected significance of decoding accuracy at each time point at $p < 0.05$ (small), $p < 0.01$ (medium), and $p < 0.001$ (large). Error bars represent ± SEM.

	Maintenance			Categorization			Maintenance vs. Categorization		
Time	**EVC**	**IPS**	**sPCS**	**EVC**	**IPS**	**sPCS**	**EVC**	**IPS**	**sPCS**
0	0.486	0.389	0.286	0.838	0.838	0.747	0.206	0.814	0.827
1	0.385	0.576	0.743	0.747	0.822	0.717	0.205	0.675	0.460
2	0.060	0.286	0.219	0.174	0.626	0.626	0.188	0.788	0.779
3	0.000	0.286	0.764	0.000	0.353	0.389	0.497	0.630	0.232
4	0.000	0.154	0.422	0.000	0.038	0.784	0.534	0.299	0.812
5	0.000	0.042	0.618	0.000	0.027	0.269	0.816	0.389	0.241
6	0.000	0.002	0.094	0.000	0.002	0.056	0.927	0.528	0.594
7	0.000	0.000	0.033	0.000	0.000	0.001	0.914	0.301	0.274
8	0.000	0.000	0.001	0.000	0.000	0.001	0.588	0.651	0.655
9	0.000	0.000	0.004	0.000	0.000	0.007	0.119	0.277	0.399
10	0.000	0.000	0.060	0.000	0.000	0.003	0.007	0.258	0.127
11	0.000	0.001	0.264	0.000	0.000	0.032	0.001	0.162	0.123
12	0.000	0.003	0.759	0.000	0.000	0.039	0.000	0.090	0.018
13	0.000	0.001	0.377	0.000	0.000	0.003	0.000	0.054	0.022
14	0.000	0.007	0.747	0.000	0.000	0.038	0.000	0.046	0.034
15	0.000	0.000	0.667	0.000	0.000	0.003	0.000	0.145	0.012
16	0.000	0.000	0.948	0.000	0.000	0.020	0.000	0.190	0.001
17	0.000	0.000	0.838	0.000	0.000	0.031	0.000	0.889	0.004

Supplemental Table 1.

P-values for the time course of IEM results in Figure 2. Underline denotes significant results (*p* < 0.05).



	Maintenance			Categorization		
Time	**EVC**	**IPS**	**sPCS**	**EVC**	**IPS**	**sPCS**
0	0.569	0.202	0.603	0.674	0.347	0.368
1	0.866	0.549	0.871	0.913	0.459	0.291
2	0.757	0.808	0.846	0.921	0.651	0.224
3	0.246	0.264	0.576	0.438	0.708	0.075
4	0.232	0.147	0.749	0.092	0.052	0.093
5	0.310	0.062	0.537	0.136	0.032	0.033
6	0.114	0.049	0.911	0.045	0.003	0.009
7	0.012	0.024	0.125	0.019	0.002	0.012
8	0.034	0.076	0.140	0.022	0.001	0.004
9	0.050	0.069	0.143	0.027	0.000	0.001
10	0.160	0.699	0.377	0.101	0.000	0.000
11	0.320	0.944	0.612	0.140	0.015	0.004
12	0.494	0.983	0.339	0.271	0.027	0.007
13	0.344	0.546	0.140	0.435	0.062	0.009
14	0.173	0.768	0.538	0.275	0.082	0.068
15	0.127	0.809	0.811	0.493	0.181	0.080
16	0.083	0.145	0.516	0.284	0.097	0.032
17	0.396	0.085	0.030	0.122	0.042	0.118

Supplemental Table 2.

P-values for the time course of correlation results in Figure 3. Underline denotes significant results (p < 0.05).

	Maintenance			Categorization			Maintenance vs. Categorization		
Time	EVC	IPS	sPCS	EVC	IPS	sPCS	EVC	IPS	sPCS
0	0.005	0.028	0.203	0.511	0.359	0.395	0.007	0.976	0.804
1	0.007	0.090	0.079	0.339	0.181	0.412	0.018	0.810	0.979
2	0.001	0.152	0.444	0.154	0.371	0.548	0.021	0.816	0.668
3	0.000	0.079	0.481	0.000	0.237	0.600	0.057	0.851	0.669
4	0.000	0.021	0.170	0.000	0.001	0.224	0.385	0.578	0.675
5	0.000	0.018	0.548	0.000	0.000	0.045	0.578	0.396	0.071
6	0.000	0.006	0.525	0.000	0.000	0.000	0.811	0.197	0.002
7	0.000	0.000	0.018	0.000	0.000	0.000	0.900	0.275	0.072
8	0.000	0.000	0.152	0.000	0.000	0.000	0.884	0.059	0.011
9	0.000	0.000	0.117	0.000	0.000	0.000	0.612	0.090	0.040
10	0.000	0.000	0.028	0.000	0.000	0.000	0.045	0.207	0.066
11	0.000	0.001	0.017	0.000	0.000	0.000	0.005	0.272	0.209
12	0.000	0.000	0.028	0.000	0.000	0.000	0.014	0.574	0.266
13	0.000	0.001	0.052	0.000	0.000	0.000	0.029	0.344	0.139
14	0.000	0.001	0.117	0.000	0.000	0.021	0.120	0.420	0.395
15	0.000	0.002	0.022	0.000	0.000	0.005	0.001	0.644	0.340
16	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.866	0.546
17	0.000	0.000	0.039	0.000	0.000	0.000	0.000	0.881	0.259

Supplemental Table 3.

P-values for the time course of IEM results in Figure 4. Underline denotes significant results ($p < 0.05$).

References

1. Abadi M., Barham P., Chen J., Chen Z., Davis A., Dean J., Zheng X (2016) **TensorFlow: A System for Large-Scale Machine Learning** *12th USENIX Symposium on Operating Systems Design and Implementation* :265–283

2. Baddeley A (2003) **Working memory: looking back and looking forward** *Nat Rev Neurosci* **4**:829–839 https://doi.org/10.1038/nrn1201

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

4. Badre D., Bhandari A., Keglovits H., Kikumoto A (2021) **The dimensionality of neural representations for control** *Curr Opin Behav Sci* **38**:20–28 https://doi.org/10.1016/j.cobeha.2020.07.002

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

6. Bettencourt K. C., Xu Y (2016) **Decoding the content of visual short-term memory under distraction in occipital and parietal areas** *Nat Neurosci* **19**:150–157 https://doi.org/10.1038/nn.4174

7. Brainard D. H (1997) **The Psychophysics Toolbox** *Spat Vis* **10**:433–436

8. Brincat S. L., Siegel M., von Nicolai C., 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 https://doi.org/10.1073/pnas.1717075115

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

10. Brouwer G. J., Heeger D. J (2011) **Cross-orientation suppression in human visual cortex** *J Neurophysiol* **106**:2108–2119 https://doi.org/10.1152/jn.00540.2011

11. Christophel T. B., Hebart M. N., Haynes J. D (2012) **Decoding the contents of visual short-term memory from human visual and parietal cortex** *J Neurosci* **32**:12983–12989 https://doi.org/10.1523/JNEUROSCI.0184-12.2012

12. Christophel T. B., Iamshchinina P., Yan C., Allefeld C., Haynes J. D (2018) **Cortical specialization for attended versus unattended working memory** *Nat Neurosci* **21**:494–496 https://doi.org/10.1038/s41593-018-0094-4

13. Cox R. W (1996) **AFNI: software for analysis and visualization of functional magnetic resonance neuroimages** *Comput Biomed Res* **29**:162–173

14. Cox R. W., Hyde J. S (1997) **Software tools for analysis and visualization of FMRI Data** *NMR in Biomedicine* **10**:171–178

15. D’Esposito M., Postle B. R (2015) **The cognitive neuroscience of working memory** *Annu Rev Psychol* **66**:115–142 https://doi.org/10.1146/annurev-psych-010814-015031

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

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

18. Emrich S. M., Riggall A. C., Larocque J. J., Postle B. R (2013) **Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory** *J Neurosci* **33**:6516–6523 https://doi.org/10.1523/JNEUROSCI.5732-12.2013

19. Eppinger B., Goschke T., Musslick S (2021) **Meta-control: From psychology to computational neuroscience** *Cogn Affect Behav Neurosci* **21**:447–452 https://doi.org/10.3758/s13415-021-00919-4

20. Ester E. F., Anderson D. E., Serences J. T., Awh E (2013) **A neural measure of precision in visual working memory** *J Cogn Neurosci* **25**:754–761 https://doi.org/10.1162/jocn_a_00357

21. Ester E. F., Sprague T. C., Serences J. T (2015) **Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory** *Neuron* **87**:893–905 https://doi.org/10.1016/j.neuron.2015.07.013

22. Ester E. F., Sprague T. C., Serences J. T (2020) **Categorical Biases in Human Occipitoparietal Cortex** *J Neurosci* **40**:917–931 https://doi.org/10.1523/JNEUROSCI.2700-19.2019

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

24. Freedman D. J., Assad J. A (2006) **Experience-dependent representation of visual categories in parietal cortex** *Nature* **443**:85–88 https://doi.org/10.1038/nature05078

25. Freedman D. J., Riesenhuber M., Poggio T., Miller E. K (2001) **Categorical representation of visual stimuli in the primate prefrontal cortex** *Science* **291**:312–316 https://doi.org/10.1126/science.291.5502.312

26. Funahashi S., Bruce C. J., Goldman-Rakic P. S (1989) **Mnemonic coding of visual space in the monkey’s dorsolateral prefrontal cortex** *J Neurophysiol* **61**:331–349 https://doi.org/10.1152/jn.1989.61.2.331

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

28. Fuster J. M., Alexander G. E (1971) **Neuron activity related to short-term memory** *Science* **173**:652–654

29. Gosseries O., Yu Q., LaRocque J. J., Starrett M. J., Rose N. S., Cowan N., 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 https://doi.org/10.1523/JNEUROSCI.2747-17.2018

30. Gramfort A., Luessi M., Larson E., Engemann D. A., Strohmeier D., Brodbeck C., Hamalainen M. (2013) **MEG and EEG data analysis with MNE-Python** *Front Neurosci* **7** https://doi.org/10.3389/fnins.2013.00267

31. Hallenbeck G. E., Sprague T. C., Rahmati M., Sreenivasan K. K., Curtis C. E (2021) **Working memory representations in visual cortex mediate distraction effects** *Nat Commun* **12** https://doi.org/10.1038/s41467-021-24973-1

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

33. Henderson M. M., Rademaker R. L., Serences J. T (2022) **Flexible utilization of spatial- and motor-based codes for the storage of visuo-spatial information** *Elife* **11** https://doi.org/10.7554/eLife.75688

34. Hu Y., Yu Q (2023) **Spatiotemporal dynamics of self-generated imagery reveal a reverse cortical hierarchy from cue-induced imagery** *Cell Rep* **42** https://doi.org/10.1016/j.celrep.2023.113242

35. Latimer K. W., Freedman D. J (2023) **Low-dimensional encoding of decisions in parietal cortex reflects long-term training history** *Nat Commun* **14** https://doi.org/10.1038/s41467-023-36554-5

36. Leavitt M. L., Mendoza-Halliday D., Martinez-Trujillo J. C (2017) **Sustained Activity Encoding Working Memories: Not Fully Distributed** *Trends Neurosci* **40**:328–346 https://doi.org/10.1016/j.tins.2017.04.004

37. Lee S. H., Kravitz D. J., Baker C. I (2013) **Goal-dependent dissociation of visual and prefrontal cortices during working memory** *Nat Neurosci* **16**:997–999 https://doi.org/10.1038/nn.3452

38. Li S., Zeng X., Shao Z., Yu Q (2023) **Neural Representations in Visual and Parietal Cortex Differentiate between Imagined, Perceived, and Illusory Experiences** *J Neurosci* **43**:6508–6524 https://doi.org/10.1523/JNEUROSCI.0592-23.2023

39. Liu T., Cable D., Gardner J. L (2018) **Inverted Encoding Models of Human Population Response Conflate Noise and Neural Tuning Width** *J Neurosci* **38**:398–408 https://doi.org/10.1523/JNEUROSCI.2453-17.2017

40. Lorenc E. S., Sreenivasan K. K., Nee D. E., Vandenbroucke A. R. E., D'Esposito M (2018) **Flexible coding of visual working memory representations during distraction** *J Neurosci* https://doi.org/10.1523/JNEUROSCI.3061-17.2018

41. Luu L., Stocker A. A (2021) **Categorical judgments do not modify sensory representations in working memory** *PLoS Comput Biol* **17** https://doi.org/10.1371/journal.pcbi.1008968

42. Masse N. Y., Yang G. R., Song H. F., Wang X. J., Freedman D. J (2019) **Circuit mechanisms for the maintenance and manipulation of information in working memory** *Nat Neurosci* **22**:1159–1167 https://doi.org/10.1038/s41593-019-0414-3

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

44. Miller E. K., Cohen J. D (2001) **An integrative theory of prefrontal cortex function** *Annu Rev Neurosci* **24**:167–202 https://doi.org/10.1146/annurev.neuro.24.1.167

45. Miller J. A., Tambini A., Kiyonaga A., D'Esposito M (2022) **Long-term learning transforms prefrontal cortex representations during working memory** *Neuron* **110**:3805–3819 https://doi.org/10.1016/j.neuron.2022.09.019

46. Mok R. M., Love B. C (2020) **Abstract Neural Representations of Category Membership beyond Information Coding Stimulus or Response** *J Cogn Neurosci* :1–17 https://doi.org/10.1162/jocn_a_01651

47. Musslick S., Cohen J. D. (2021) **Rationalizing constraints on the capacity for cognitive control** *Trends Cogn Sci* **25**:757–775 https://doi.org/10.1016/j.tics.2021.06.001

48. Pelli D. G (1997) **The VideoToolbox software for visual psychophysics: transforming numbers into movies** *Spat Vis* **10**:437–442

49. Rademaker R. L., Chunharas C., Serences J. T (2019) **Coexisting representations of sensory and mnemonic information in human visual cortex** *Nat Neurosci* **22**:1336–1344 https://doi.org/10.1038/s41593-019-0428-x

50. Riggall A. C., 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 https://doi.org/10.1523/Jneurosci.1892-12.2012

51. Serences J. T., Ester E. F., Vogel E. K., Awh E (2009) **Stimulus-specific delay activity in human primary visual cortex** *Psychol Sci* **20**:207–214 https://doi.org/10.1111/j.1467-9280.2009.02276.x

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

53. Sprague T. C., Serences J. T (2013) **Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices** *Nat Neurosci* **16**:1879–1887 https://doi.org/10.1038/nn.3574

54. Wang L., Mruczek R. E., Arcaro M. J., Kastner S (2015) **Probabilistic Maps of Visual Topography in Human Cortex** *Cereb Cortex* **25**:3911–3931 https://doi.org/10.1093/cercor/bhu277

55. Wang X. J (2021) **50 years of mnemonic persistent activity: quo vadis?** *Trends Neurosci* **44**:888–902 https://doi.org/10.1016/j.tins.2021.09.001

56. Yu Q., Postle B. R (2021) **The Neural Codes Underlying Internally Generated Representations in Visual Working Memory** *J Cogn Neurosci* :1–16 https://doi.org/10.1162/jocn_a_01702

57. Yu Q., 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 https://doi.org/10.1016/j.neuroimage.2017.05.055

58. Yu Q., Shim W. M (2019) **Temporal-Order-Based Attentional Priority Modulates Mnemonic Representations in Parietal and Frontal Cortices** *Cereb Cortex* **29**:3182–3192 https://doi.org/10.1093/cercor/bhy184

59. Zhou Y., Rosen M. C., Swaminathan S. K., Masse N. Y., Zhu O., Freedman D. J (2021) **Distributed functions of prefrontal and parietal cortices during sequential categorical decisions** *Elife* **10** https://doi.org/10.7554/eLife.58782

Editors

Reviewing Editor
Gui Xue
Beijing Normal University, Beijing, China

Senior Editor
Michael Frank
Brown University, Providence, United States of America

Reviewer #1 (Public Review):

Summary:

In this manuscript, Shao et al. investigate the contribution of different cortical areas to working memory maintenance and control processes, an important topic involving different ideas about how the human brain represents and uses information when it is no longer available to sensory systems. In two fMRI experiments, they demonstrate that the human frontal cortex (area sPCS) represents stimulus (orientation) information both during typical maintenance, but even more so when a categorical response demand is present. That is, when participants have to apply an added level of decision control to the WM stimulus, sPCS areas encode stimulus information more than conditions without this added demand. These effects are then expanded upon using multi-area neural network models, recapitulating the empirical gradient of memory vs control effects from visual to parietal and frontal cortices. In general, the experiments and analyses provide solid support for the authors' conclusions, and control experiments and analyses are provided to help interpret and isolate the frontal cortex effect of interest. However, I suggest some alternative explanations and important additional analyses that would help ensure an even stronger level of support for these results and interpretations.

Strengths:

- The authors use an interesting and clever task design across two fMRI experiments that is able to parse out contributions of WM maintenance alone along with categorical, rule-based decisions. Importantly, the second experiment only uses one fixed rule, providing both an internal replication of Experiment 1's effects and extending them to a different situation when rule-switching effects are not involved across mini-blocks.

- The reported analyses using both inverted encoding models (IEM) and decoders (SVM) demonstrate the stimulus reconstruction effects across different methods, which may be sensitive to different aspects of the relationship between patterns of brain activity and the experimental stimuli.

- Linking the multivariate activity patterns to memory behavior is critical in thinking about the potential differential roles of cortical areas in sub-serving successful working memory. Figure 3 nicely shows a similar interaction to that of Figure 2 in the role of sPCS in the categorization vs. maintenance tasks.

- The cross-decoding analysis in Figure 4 is a clever and interesting way to parse out how stimulus and rule/category information may be intertwined, which would have been one of the foremost potential questions or analyses requested by careful readers. However, I think more additional text in the Methods and Results to lay out the exact logic of this abstract category metric will help readers better interpret the potential importance of this analysis and result.

Weaknesses:

- Selection and presentation of regions of interest: I appreciate the authors' care in separating the sPCS region as "frontal cortex", which is not necessarily part of the prefrontal cortex, on which many ideas of working memory maintenance activity are based. However, to help myself and readers interpret these findings, at a minimum the boundaries of each ROI should be provided as part of the main text or extended data figures. Relatedly, the authors use a probabilistic visual atlas to define ROIs in the visual, parietal, and frontal cortices. But other regions of both lateral frontal and parietal cortices show retinotopic responses (Mackey and Curtis, eLife, 2017: https://elifesciences.org/articles/22974) and are perhaps worth considering. Do the inferior PCS regions or inferior frontal sulcus show a similar pattern of effects across tasks? And what about the middle frontal gyrus areas of the prefrontal cortex, which are most analogous to the findings in NHP studies that the authors mention in their discussion, but do not show retinotopic responses? Reporting the effects (or lack thereof) in other areas of the frontal cortex will be critical for readers to interpret the role of the frontal cortex in guiding WM behavior and supporting the strongly worded conclusions of broad frontal cortex functioning in the paper. For example, to what extent can sPCS results be explained by visual retinotopic responses? (Mackey and Curtis, eLife, 2017: https://elifesciences.org/articles/22974).

- When looking at the time course of effects in Figure 2, for example, the sPCS maintenance vs categorization effects occur very late into the WM delay period. More information is needed to help separate this potential effect from that of the response period and potential premotor/motor-related influences. For example, are the timecourses shifted to account for hemodynamic lag, and if so, by how much? Do the sPCS effects blend into the response period? This is critical, too, for a task that does not use a jittered delay period, and potential response timing and planning can be conducted by participants near the end of the WM delay. Regardless, parsing out the timing and relationship to response planning is important, and an ROI for M1 or premotor cortex could also help as a control comparison point, as in reference (24).

- Interpreting effect sizes of IEM and decoding analysis in different ROIs. Here, the authors are interested in the interaction effects across maintenance and categorization tasks (bar plots in Figure 2), but the effect sizes in even the categorization task (y-axes) are always larger in EVC and IPS than in the sPCS region... To what extent do the authors think this representational fidelity result can or cannot be compared across regions? For example, a reader may wonder how much the sPCS representation matters for the task, perhaps, if memory access is always there in EVC and IPS? Or perhaps late sPCS representations are borrowing/accessing these earlier representations? Giving the reader some more intuition for the effect sizes of representational fidelity will be important. Even in Figure 3 for the behavior, all effects are also seen in IPS as well. More detail or context at minimum is needed about the representational fidelity metric, which is cited in ref (35) but not given in detail. These considerations are important given the claims of the frontal cortex serving such an important for flexible control, here.

https://doi.org/10.7554/eLife.100287.1.sa1

Reviewer #2 (Public Review):

Summary:

The authors provide evidence that helps resolve long-standing questions about the differential involvement of the frontal and posterior cortex in working memory. They show that whereas the early visual cortex shows stronger decoding of memory content in a memorization task vs a more complex categorization task, the frontal cortex shows stronger decoding during categorization tasks than memorization tasks. They find that task-optimized RNNs trained to reproduce the memorized orientations show some similarities in neural decoding to people. Together, this paper presents interesting evidence for differential responsibilities of brain areas in working memory.

Strengths:

This paper was strong overall. It had a well-designed task, best-practice decoding methods, and careful control analyses. The neural network modelling adds additional insight into the potential computational roles of different regions.

Weaknesses:

While the RNN model matches some of the properties of the task and decoding, its ability to reproduce the detailed findings of the paper was limited. Overall, the RRN model was not as well-motivated as the fMRI analyses.

https://doi.org/10.7554/eLife.100287.1.sa0

Author response:

(1) Reviewer 1 suggested that we repeat the analyses in additional ROIs in the prefrontal cortex (PFC). We appreciate this suggestion and believe it will contribute to a comprehensive understanding of the current findings. These results will be included in the revision.

(2) Reviewer 1 suggested that we also examine results in motor-related ROIs to rule out influences from response planning. We would like to note that our experimental design makes it unlikely that response planning would have influenced our results, as participants were unable to plan their motor responses in advance due to randomized response mapping on a trial-by-trial basis. Nevertheless, we agree with the reviewer that showing results from motor-related ROIs is important, and will include these results in the revision.

(3) Reviewer 1 raised a question about the effect size of the results across different ROIs. In our manuscript, we tried to avoid direct comparisons of representational strength across ROIs, by focusing on the differences in representational strength between conditions within the same ROI. Nevertheless, we agree that clarifying this issue is important, which we will address in the revision.

(4) Reviewer 2 raised a concern about the similarity between the RNN and fMRI results. We acknowledge that the complexity of our results makes it challenging to replicate all fMRI findings within a single RNN (e.g., simulating three brain regions in a single network with distinct result patterns). Nonetheless, the current RNNs effectively captured our key fMRI findings, including increased stimulus representation in frontal cortex as well as the tradeoff in category representation with varying levels of flexible control. Reviewer 2 also made several suggestions in tweaking the RNN structure and in choosing alternative analysis methods. We are happy to carry out these points as we think they could potentially increase the alignment between the two modalities.

https://doi.org/10.7554/eLife.100287.1.sa3