Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Decomposition of Conscious Access: Working Memory for Masked Sequences in Monkey Prefrontal Cortex

Liping Wang

liping.wang@ion.ac.cn

Institute of Neuroscience, Chinese Academy of Sciences https://orcid.org/0000-0003-2038-0234

Hao Zhou

Institute of Neuroscience, Key Laboratory of Primate Neurobiology, CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences

Wen Fang

Institute of Neuroscience, Key Laboratory of Primate Neurobiology, CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031

Guobin Fu

Institute of Neuroscience, Key Laboratory of Primate Neurobiology, CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences

Yiteng Zhang

Lingang Laboratory

Bin Min

Lingang Laboratory https://orcid.org/0000-0003-1006-9629

Stanislas Dehaene

Chair of Experimental Cognitive Psychology, College de France, Paris

Article

Keywords:

Posted Date: July 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10130477/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

The Decomposition of Conscious Access:

Working Memory for Masked Sequences in Monkey Prefrontal Cortex

Hao Zhou[1], Wen Fang[1], Guobin Fu[1], Yiteng Zhang[3], Bin Min[3], Stanislas Dehaene[4], Liping Wang[1,2,#]

Affiliations:

[1]Institute of Neuroscience, Key Laboratory of Primate Neurobiology, CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China.

[2]University of Chinese Academy of Sciences, Beijing 100049, China.

[3]Lingang Laboratory, Shanghai, China.

[4]Cognitive Neuroimaging Unit, INSERM, CEA, CNRS, Université Paris-Saclay, NeuroSpin center, Gif-sur-Yvette, France.

[#]Corresponding author: Liping Wang

Address: Institute of Neuroscience, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China.

Email: liping.wang@ion.ac.cn

Abstract

Conscious access is thought to involve two main stages: transient linear encoding of the objective sensory stimulus, followed by nonlinear bifurcation towards a sustained state of activity (ignition) encoding the subjective percept. To test this hypothesis, we recorded thousands of neurons in the prefrontal cortex (PFC) of monkeys trained to perform a sequence working memory (WM) task with masked visual stimuli of variable contrast, thus modulating their subjective visibility. PFC responses revealed the predicted sequence of linear and nonlinear processing stages as a function of stimulus contrast. Crucially, both stages were instantiated by orthogonal neural subspaces within the same PFC neurons. A shared entry subspace transiently encoded objective sensory inputs, while rank-ordered WM subspaces exhibited all-or-none bifurcations towards sustained states. On error and stimulus-absent trials, endogenous neural signals competed with objective inputs within the entry subspace, and only rank subspace ignition predicted upcoming responses. Routing from the entry to the WM subspace was not automatic but involved an active gating process which vanished when the animal was distracted. Thus, within local PFC, multiple neural subspaces implement the successive neural processes underlying conscious access.

Introduction

Conscious perception is often associated with the ability to hold a representation in mind. Such representations are not merely passive, faithful copies of perceived events; they actively integrate with internal models of the world that constrain, complete, and prioritize conscious information according to task goals and prior experiences [1-3]. Examples include interpreting an ambiguous visual scene (such as a Necker cube) as one of two possible percepts; detecting the number of passes in a basketball game, while failing to aware a passing gorilla; or understanding the specific meaning of a word within its context, but not its other possible meanings. The prefrontal cortex (PFC) has long been proposed as a central brain region for this controlled conscious processing, including the selection, maintenance, and manipulation of cognitive representations [4,5]. A critical ingredient in conscious information control is the use of internal models to integrate and predict subsequent inputs [6], which is sometimes equated with the concept of a "cognitive map" [7,8]. Previous studies in PFC have shown various types of representations of internal cognitive maps that bind transient sensory inputs into coherent and contextually grounded percepts or memories for later choices [9,10].

The global neuronal workspace (GNW) hypothesis complements this view by proposing that conscious access operates via controlled neuronal ignition and global broadcasting, which, in turn, serve as a gateway to a working memory (WM) system capable of holding online, manipulating, and sharing internal representations [11-14]. Crucially, GNW decomposes conscious access into two distinct steps: 1) the propagation of objective bottom-up sensory signals, proportional to stimulus intensity, across the sensory hierarchy to the prefrontal cortex; 2) the crossing of a threshold level of reverberating activity, which leads to non-linear ignition and WM maintenance via persistent firing loops. The latter stage is modulated by attention and task demands [15,16], yielding a subjective mental model of the world that, for weak, ambiguous, or illusory stimuli, may depart from objective sensory information.

In Bayesian terms, the first stage reflects the continuous integration of the objective evidence provided by the stimulus with top-down priors, while the second reflects the outcome of an inference process that selects and stabilizes a single *a-posteriori* conclusion, the most probable subjective percept [17-19].

So far, direct evidence for the role of internal models in shaping conscious perception is limited even in human neuroimaging and electrophysiological recordings [20-22]. Recently, using monkey electrophysiology, van Vugt et al. [23] demonstrated a prefrontal neural correlate of threshold-crossing and ignition, revealing an all-or-none switch whereby weak stimuli failed to elicit sustained prefrontal activity. However, the detailed neural population mechanisms by which the PFC integrates objective sensory inputs into internal models of the world remain unknown.

In the present work, we capitalize on a sequence working memory task (SWM) in which monkeys memorize a sequence of spatial locations, hold it in WM, and later reproduce it with their eyes. Our team recently dissected the representational geometry of SWM in the macaque PFC, showing that the spatial sequence is encoded in WM using near-orthogonal low-dimensional neural subspaces, each storing the spatial location at a given ordinal rank [24,25]. The rank-WM subspaces may thus serve as internal temporal structures (e.g., WM slots) that constrain the conscious perception of sequences [26]. Indeed, in this sequence context, the very same input (a spatial location) can be dispatched at a different position in the internal WM model depending on its ordinal rank. During memorization, we showed that access to WM occurs in two successive steps: sensory information about stimulus location first activates a common entry subspace and is then routed to the appropriate rank-WM subspace [27].

Therefore, to render this paradigm informative to consciousness research, we modulated the visibility of each sequence item by masking it and by varying its contrast from 100% (full visibility) to 0% (stimulus absence). This design allowed us

to dissociate the objective content of sensory stimulation (6 possible locations, each varying from 0 to 100% contrast) from its subjective registration in WM (as reflected by the monkey's subsequent behavioral report, which could differ from the actual sensory stimuli). As in van Vugt et al., we relied on performance (here, sequence reproduction) as an approximate proxy for subjective report, classifying trials as consciously accessed if both locations were correctly recalled in proper order (chance level = 1/30). Using this design, we aimed to answer two central questions about the neural mechanisms of conscious access: 1) Can PFC activity be decomposed into a sequence of neural states reflecting, respectively, objective sensory processing versus integration into a model of the sequence in rank-dependent WM subspaces? 2) Given the monkeys' intensive training on the sequence structure, can these rank-specific WM-subspaces in the PFC serve as learned internal models capable of generating conscious percepts, even in the absence of ascending sensory signals (zero-contrast trials)?

Results

Paradigm and behavior

Three macaque monkeys (M1, M2, and M3) were trained to perform a delayed-masked spatial sequence reproduction task (Fig. 1A). Target stimuli were presented at six possible spatial locations on the screen. The intensity of the target stimuli varied across five contrast levels (0%, 4%, 8%, 15%, and 100%), including a null-contrast condition where no visible target was presented. A fixed backward mask immediately followed each target stimulus, obscuring its visibility. On each trial, two spatial locations (or only masks in the null-contrast condition) were sequentially presented, while the monkey maintained fixation on a white cross at the center of the screen. The monkeys were required to maintain fixation and memorize the detected spatial sequence during a delay of 2.5-3.5 seconds. After the delay, the monkeys reproduced the sequence to obtain a reward (see Methods for details).

Overall, the three monkeys performed the task well. At 100% contrast, the sequence reproduction accuracy was 70% (±5.59) for M1, 85% (±8.55) for M2, and 72% (±4.58) for M3 (Fig. 1C). There were no significant biases in spatial item and ordinal rank (Fig. S1A). As stimulus contrast levels decreased, accuracy declined linearly (all $R^2 > 0.99$, $P < 0.001$), approaching 50/50 performance at 8% contrast (38-44% for sequence reproduction, 53-69% for individual location accuracy) and dropping to chance levels at 0% contrast (3.25-3.81% for sequences and 15.27-17.02% for locations; Fig. 1C and S1B). The proportions of error trials were equally distributed across response offsets from the target locations at all contrast levels (Fig. 1D).

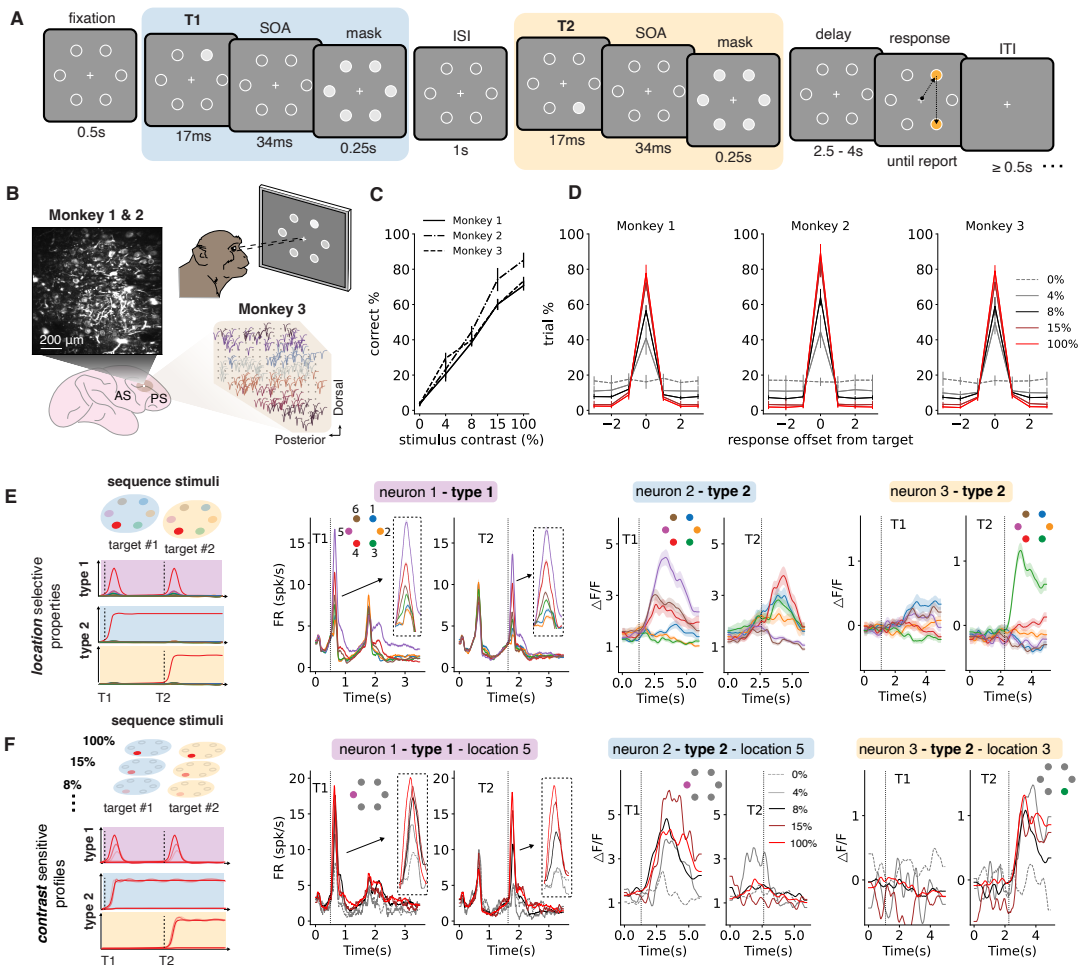


Figure 1. Large-scale neuronal recording of the macaque frontal lobe during a masked delayed spatial sequence reproduction task

(A) Schematic of the task paradigm, illustrating the sequence of spatial locations presented with backward masking and the monkeys' task to memorize and reproduce the spatial sequence after a delay. Stimulus contrast varied (0%, 4%, 8%, 15%, 100%).

(B) A "perceptual snapshot" of target stimuli followed by the backward mask during the task, while two-photon calcium imaging (upper left) was recorded simutaneously in the dorsolateral prefrontal cortex (dlPFC) for Monkey 1 and 2, showing an example field of view (FOV), and a 157-channel microdrive electrophysiology array (lower right) was used for Monkey 3 to record spiking activity in the frontal lobe with example spiking waves in the frontal channel map (details in Methods).

(C) Behavioral performance showing the correct rates at the sequence level, averaged across sessions. Sequence reproduction accuracy decreased linearly with reduced stimulus contrast, reaching chance level (1/30) at 0% contrast. Error bars represent standard error (SD).

(D) Distribution of response offsets for individual location in error trials (distance = ±1, ±2, 3), uniform across contrasts (no spatial bias). Error bars represent standard error (SE).

(E) Example neurons exhibiting spatial location selectivity at different temporal stages: Type 1 neuron (neuron 1) with shared transient tuning during the onsets of S1 and S2 (independent of rank). Type 2 neuron (neuron 2 and 3) showing sustained, rank-specific delay activity. Neuronal activity traces are aligned to trial onset and pooled across trials according to spatial location. Shaded areas indicate SE.

(F) Contrast sensitivity: Type 1 neuron (neuron 1) activity scaled with contrast (gradient); Type 2 neuron (neuron 2 and 3) showed all-or-none pattern (4%-100% vs. 0%). Correct trials only; shuffled labels for 0% contrast.

Recordings and single-neuron responses

To measure the neural state, we recorded single neurons in the PFC using two methods across three monkeys (Fig. 1B). In Monkey 1 and 2, GCaMP6f virus was injected into the lateral PFC, enabling calcium imaging of large-scale neuronal activity (Fig. S1C, n = 2653 for Monkey 1, n = 6790 for Monkey 2). In Monkey 3, a 157-channel microdrive electrophysiology array was deployed over the frontal lobe to capture high-resolution spiking activity (Fig. S1D; n = 1876). We focused on neuronal activities from the sample to the delay period.

Neural responses showed diverse, mixed task-related activity. Two types of single-neuron activity were immediately apparent (Fig. 1E). One type of neuron

transiently encoded locations at each stimulus onset for both stimulus-1 (S-1) and stimulus-2 (S-2) during the sample periods (example neuron 1), demonstrating transient sensory responses independent of ordinal rank. A second type of neuron exhibited sustained rank- and location-selective memory activity during the delay (example neurons 2 and 3). These two types of activities were deeply entwined at the single-neuron level, with many neurons showing mixed selectivity (see the proportion of mixed coding neurons in each monkey in Fig. S2). Other mixed task-related neurons are shown in Figure S3. Importantly, the transient sensory and sustained delay activities displayed distinct contrast sensitivity profiles (Fig. 1F): the activity of neuron 1 showed a contrast gradient, with firing rates increasing with stimulus contrast, whereas the activities of neurons 2 and 3 displayed no such gradient dependency, instead with an all-or-none pattern separating only the trials with a stimulus (4% to 100% contrast) and without a stimulus (0% contrast). Note that only correct trials were included and plotted in this analysis (with shuffled labels for 0% contrast).

Dynamics in PFC neural subspaces during conscious access

We next examined how PFC neuronal populations reflect the dynamics of conscious access to spatial sequences. Using the method developed in our previous studies [24,27], we first tracked neural codes for item locations at ranks 1 and 2 over time and across ranks, on correct trials only. The results revealed distinct temporal patterns in the neuronal dynamics (Fig. 2A and Fig. S4A). In the within-rank decoding, the dynamic patterns for both rank-1 and rank-2 showed stable, persistent coding of items. In the cross-rank decoding, a shared transient signal was immediately apparent during the sample periods, followed by persistent rank-specific item codes during delay periods, during which no significant cross-rank generalization was observed.

To uncover the unique contributions of neural processing stages to spatial sequence encoding, we then searched for low-dimensional subspaces that represent sensory and rank-WM information by analyzing neural activity during the sample and delay periods, using only correct trials from the full-contrast condition (100%). Basically, we modeled population neuronal responses to visuospatial locations (items) as multivariate Gaussian distributions: $x_i \sim \mathcal{N}(\mu_i, \sum_i)$, where x_i is n-dimensional neuronal population response vector for location i, μ_i is the mean response vector and $\sum_i$ is the covariance matrix. We employed linear regression models to assess how the six spatial locations influenced neural responses within different neuronal subspaces. Using regression coefficients for the six locations across neurons during correct trials, we computed high-dimensional vector representations for each time window of interest (see Methods for details). In line with our previous findings [24,27], the analysis revealed three distinct neural subspaces: one shared entry subspace (sensory) and two rank-WM subspaces. These subspaces were visualized in a 2-dimensional Euclidean plane by projecting the data onto the two major principal component (PC) axes.

This analysis revealed a highly reliable state-space portrait: 1) each subspace contained a common geometric ring structure (Fig. 2B & Fig. S4B), reminiscent of the ring shape of the spatial items presented to the monkeys; and 2) these subspaces were near-orthogonal in neural state space, as evidenced by the large principal angles between them (Fig. 2C & Fig. S4C). To further quantify the relationship between subspaces, we calculated the variance accounted for (VAF) ratio by projecting data from one subspace onto another subspace and measuring the residual variance. If the two subspaces are nearly orthogonal, the projection will capture little variance, resulting in a low VAF ratio. The results showed low VAF ratios across all cross-subspace pairs (Fig. S4D).

The identified subspaces thus provided putative neural substrates for the two-step process of conscious access to spatial sequences. We hypothesized that 1)

sensory information is first continuously propagated into a common entry subspace in the PFC; this encoding stage is necessary but not sufficient for conscious access to sequences; and 2) activity in the entry subspace then gets integrated, in an all-or-none manner, in the appropriate rank-WM subspace (serving as the internal temporal structure), at which state the information is consciously accessed and becomes available to control behavior (global broadcasting); 3) the transfer from entry subspaces to all-or-none ignited WM subspaces depends on active control signals.

To test those hypotheses, we next investigated the neural dynamics of the entry and rank subspaces under different contrasts (100%, 15%, 8%, and 4%). We applied a Bayesian probabilistic decoding method over time to every correct trial across conditions, using the decomposed subspaces from the correct trials of the full-contrast condition (see Methods for details). We discovered that transient activity in the entry subspace varied monotonically with the strength of input stimulus contrast level, which varied from 4% to 100% (All Pearson's R = 0.55–0.91; $P < 0.0001$; Fig. 2D-E & Fig. S4B). Instead, the long-lasting activity in the rank subspaces remained invariant to changes in stimulus intensity over the same range and exhibited a single all-or-none metastable pattern of high-level neural activity in each rank subspace ($-0.14 < R < 0.05$; $P > 0.29$ for all tests; note that the 0% contrast was excluded from these tests and is further analyzed below). Across monkeys, the entry subspace displayed significantly higher contrast-dependency than rank-WM subspaces (all $P <$ 0.0001 with paired t-test and Wilcoxon signed-rank test; Fig. 2F).

If the entry subspace indeed served as a gateway controlled by rank-specific mechanisms, the functional relationship between these subspaces should manifest as directional information flow, with entry activity preceding and driving rank activity. To examine this, we analyzed the neural activity at the single-trial level in each subspace, and found that entry subspace activity was significantly earlier than rank subspace activity (e.g., electrophysiology data: mean entry onset = 35ms, mean

256 rank onset = 97 ms; all $P < 0.0001$, two-sided paired *t*-tests; Fig. 2G) and positively
257 correlated with it across trials (all Pearson's $R > 0.12$, $P < 0.001$; Fig. 2H, see the
258 controls in Fig. S4E). Furthermore, the strength of entry activity was significantly
259 correlated with the onset slope of rank activity (all $R > 0.11$, $P < 0.001$; Fig. 2I, and
260 controls in Fig. S4F; note that *rank slope* refers to the rate of change or, equivalently,
261 the steepness of neural activity amplitude during the initial phase in the rank-WM
262 subspaces). Those findings are compatible with a process of evidence accumulation
263 [18,19] and causal flow, with stronger initial entry signals accelerating the non-linear
264 buildup of sustained rank representations [27].
265 Thus, the decomposed neural subspaces and their activity dynamics within the
266 same PFC neurons supported the two-step process of conscious access to spatial
267 sequences: an initial continuous sensory input is followed by sustained all-or-none
268 activity, during which the sensory signal is integrated into the internal temporal
269 structure of ordinal ranks.
270

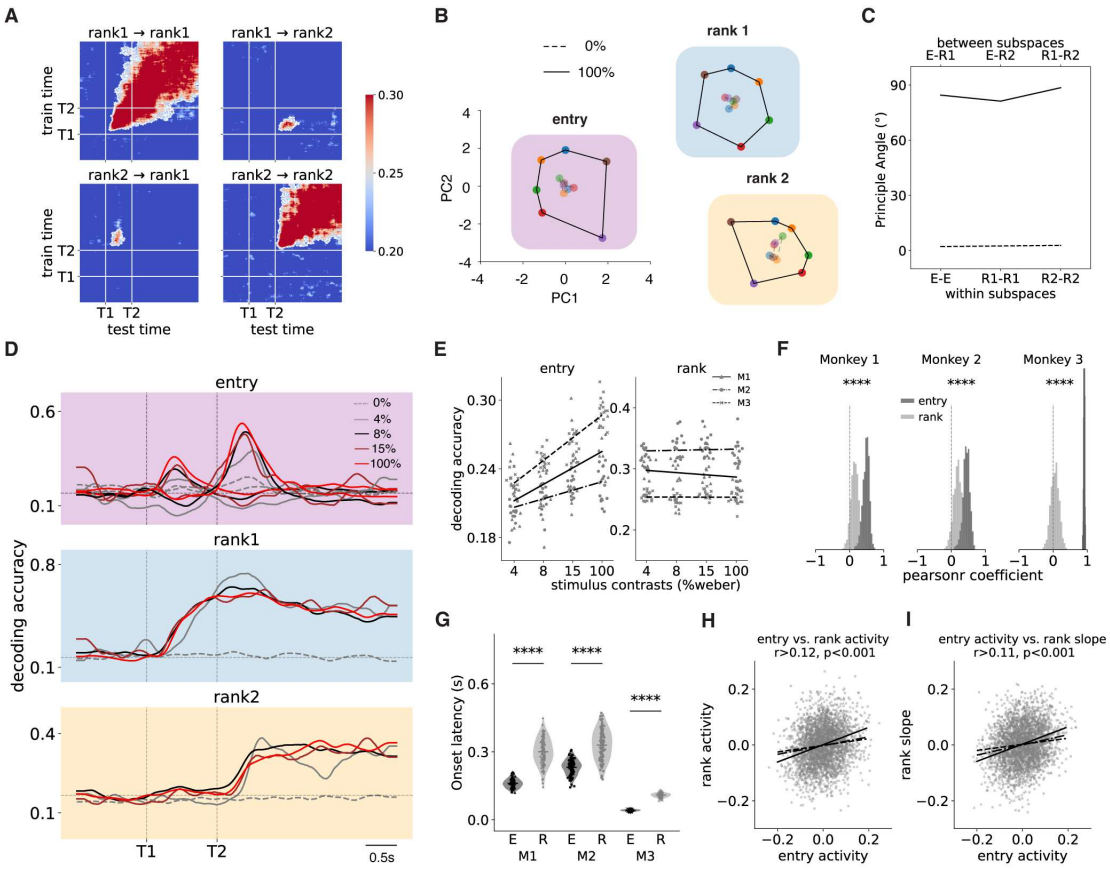


Figure 2. PFC neural subspaces and two-step dynamics of conscious access.

(A) Temporal generalization decoding for item locations at rank 1 and rank 2, showing cross-rank generalization during sample but not delay periods. Decoding of spatial locations at rank 1 (T1) and rank 2 (T2) for correct trials at 100% stimulus contrast in Monkey 1, using 20-fold cross-validation. White contours indicate significant decoding ($P < 0.001$, cluster-based permutation test).

(B–C) Geometry of neural subspaces on 100% contrast trials. **(B)** Decomposition of entry (transient sensory) and rank-WM (sustained) subspaces, projected onto two principal components, forming ring structures mirroring the spatial layout of stimuli. Entry subspace captures averaged activity during the sample period across ranks, while rank subspaces isolate delay-period activity for rank-specific components. Data includes 100% and 0% (shuffled label control) contrast conditions. **(C)** Orthogonality of subspaces, quantified by principal angles. Dark line shows large angles between rank subspaces (e.g., rank-1 vs. rank-2), indicating near-orthogonality. Dotted line (control) represents angles within the same subspace for baseline comparison.

(D) Temporal dynamics of decomposed entry and rank subspaces across contrast levels. Entry subspace activity is transient during the sample period and scales monotonically with stimulus contrast (0%, 4%, 8%, 15%, 100%) on correct trials (note that for 0% contrast, a shuffled label was used since the correct response is undefined). Rank-WM subspaces show an all-or-none pattern of persistent activity across sample and delay periods, invariant to contrast changes. Time courses are aligned to T1/T2 onsets.

(E-F) Average bayesian decoding accuracy across contrasts (100%, 15%, 8%, 4%) during sample and delay periods. **(E)** Entry subspace shows contrast-dependent activity (All Pearson's $R > 0.55$; $P < 0.0001$), whereas rank-WM subspaces exhibit invariant, all-or-none sustained activity ($-0.14 < R < 0.05$; $P > 0.29$ for all tests). All decoding accuracies were computed using 15 resampled pseudo-trials per contrast level. **(F)** Across monkeys, entry subspace displayed significant higher contrast-dependence than rank-WM subspace (mean Pearson's R: entry = 0.44-0.91, rank = 0.05-0.18; **** $P < 0.0001$, two-sided t-test, n = 2000). **(G-I)** Functional relationship between entry and rank subspace. **(G)** Single-trial activity onset latencies. Entry subspace precedes rank subspaces. **** $P < 0.0001$, two-sided paired t-tests. (H) Trial-by-trial correlations between entry and rank subspace activities (z-scored decoding accuracy; Pearson's $R > 0.12$; $P < 0.001$). **(I)** Trial-by-trial correlations between entry activity (z-scored decoding accuracy) and the onset slope of rank activity (also z-scored; Pearson's $R > 0.11$; $P < 0.001$). All analyses were performed separately in each of the three monkeys (n = 2000 resampled trials per monkey).

Shared neural processing stages for 100%- and 0%-contrast conditions

Conscious perception can also be decoupled from objective incoming stimuli and be generated internally in the absence of external inputs [28,29]. We next analyzed trials with 0% contrast, in which no external stimulus was presented. On the vast majority of such trials (mean response rate: 71% across three monkeys, comparable to 68% in the 100% contrast condition), monkeys continued to produce sequences. Based on monkeys' behavioral reports, we asked whether conscious access under the 0%-contrast condition also follows the same two-step process with a similar neural geometry and dynamics. In behavior, we observed that monkeys produced sequences without location bias (Fig. 3A), and the locations they chose were unaffected by the locations of previous trials (Fig. 3B).

At the single-neuron level, during both the sample and the delay period, we first found that a significant proportion of neurons (30.6% for M1, 34.4% for M2, 56.3% for M3; Fig. 3C) showed the same location preference on 100% contrast trials (where they fired to the actual stimulus location) and on 0% contrast trials (where they fired according to the upcoming response location) (see example neurons in Fig. 3D). In the state space analysis, we projected the activity of 0%-contrast single trials onto

the entry and rank-WM subspaces identified by the 100%-contrast trials. We found similar geometric ring structures in the three subspaces, albeit with a slightly smaller ring in the entry subspace (Fig. 3E and Fig. S5A). The temporal dynamics of item information in the corresponding subspaces was further quantified by decoding accuracy (Methods), again showing similar patterns to those observed in the 100%-contrast condition, i.e., transient activity in the entry subspace and sustained activity in the rank-WM subspaces (Fig. 3F and Fig. S5D). Thus, activity predictive of the upcoming response could already be found in those neural subspaces, several seconds earlier, as if it came from an actual stimulus.

Furthermore, although the trial number from the 0%-contrast condition was low, we employed the same decomposition methods to search for low-dimensional subspaces using the activity of 0%-contrast trials. We could still identify the separate entry and rank subspaces, closely resembling the decomposition from 100%-contrast trials, as evidenced by small principal angles and a high variance accounted for (VAF) ratio (Fig. S5B and S5C). Therefore, the results suggest that conscious access under the 0%-contrast condition shares the same neural states and dynamics as those under the 100%-contrast condition, involving transient encoding and integration with rank-WM subspaces within the same population of PFC neurons.

Interestingly, we also found differences between activity in the 100%- and 0%-contrast conditions. The example-1 neuron (Fig. 3D) shows that the peak of the neural response preferred location 3 after the onset of S-1 in the 0%-contrast trials was later than in the 100%-contrast trials. At the population level, we found significantly weaker signal strength and longer latencies (Fig. 3G-H) in both entry and rank-WM subspaces in the 0%-contrast condition. Furthermore, the correlation analysis revealed that the reduced and delayed neural activity in the entry subspace led to slower (i.e., less steep) gating into the rank-WM subspaces (Fig. 3I and Fig. S5E). Thus, on stimulus-absent trials, both subspaces fired as if a weak actual stimulus had been presented.

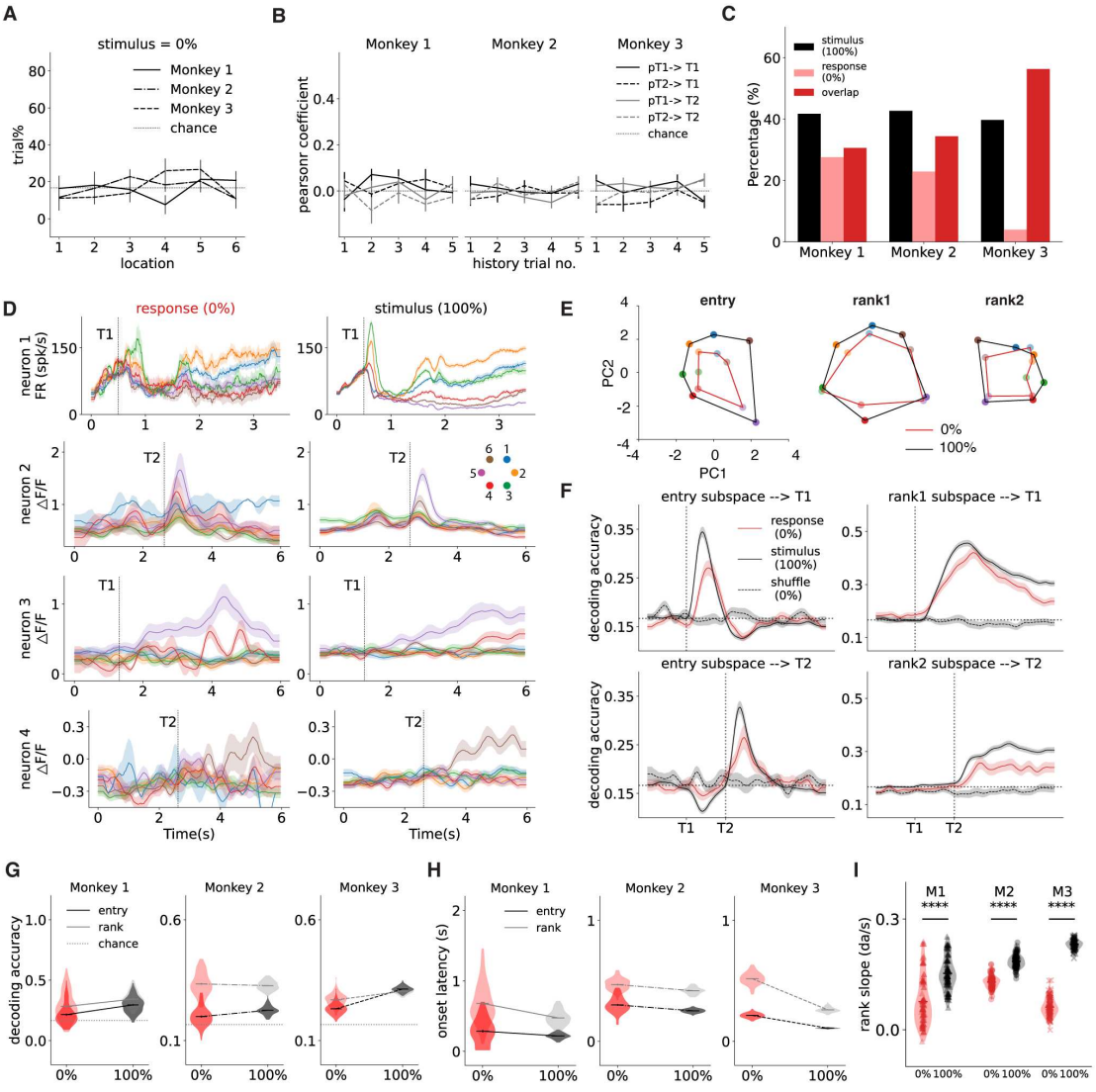


Figure 3. Shared access processes for 100%- and 0%-contrast conditions.

(A-B) Behavioral sequence production in 0% contrast condition. **(A)** Monkeys produced sequences without location bias, with trial percentages uniformly distributed across six spatial locations (chance = 1/6). **(B)** Independence analysis shows no influence of previous trials on chosen locations (Pearson's $R < 0.08$).

(C-D) Shared neuronal item preference between 100% and 0% contrast condition. **(C)** A significant proportion of neurons exhibited consistent item preference between 100% contrast (stimulus locations as item labels) and 0% contrast (response locations as item labels) during sample or delay periods (30.6% for Monkey 1, 34.4% for Monkey 2, 56.3% for Monkey 3). **(D)** Example neurons illustrating shared item preference between 100% and 0% contrast condition. Shaded areas indicate SE.

(E-F) Decomposition of mental sequence in 0% contrast. **(E)** State-space projections reveal similar ring structures in entry and rank-WM subspaces for 0% contrast trials, projected onto subspaces defined by 100% contrast trials, with a slightly smaller ring in the entry subspace.

(F) Decoding dynamics show transient activity in the entry subspace and sustained activity in rank-WM subspaces, resembling patterns observed in the 100% contrast condition. Shaded areas represent SE.
(G-I) Relationship between 0% and 100% conditions. Two-way ANOVA comparison of response-driven (0% contrast) and stimulus-driven (100% contrast) activities in entry and rank-WM subspaces for each monkey. Response-driven activities exhibit: **(G)** weaker activity strength (main effects of subspace [$F(1, 396) > 46.46$, $P < 0.001$] and contrast [$F(1, 396) > 8.63$, $P < 0.003$] across monkeys; Post-hoc: 0% < 100% in entry [meandiff = 0.026 to 0.082, all $P < 0.001$]; in rank, significant for two monkeys [meandiff = 0.007 to 0.061, $P < 0.001$] but not one [$P = 0.43$]), **(H)** longer latency in both entry and rank-WM subspaces (main effects of subspace [$F(1, 396) > 333.25$, $P < 0.001$] and contrast [$F(1, 396) > 58.79$, $P < 0.001$]; Post-hoc: 0% > 100% in entry [meandiff = 49 to 105ms, all $P < 0.038$] and rank [meandiff = 51 to 258ms, all $P < 0.001$]), and **(I)** less steep onset slope in rank-WM subspaces compared to stimulus-driven activities due to reduced/delayed entry activity (**** $P < 0.0001$, two-sided t-test; Pearson's $R > 0.10$, all $p < 0.0001$, Fig. S5E).

Competition between exogenous and endogenous activity for conscious access

We next examined what happened on near-threshold trials, when stimulus location was sometimes but not always correctly reported. To this aim, we focused on the near-threshold contrast condition (8%) and compared neural activity across correct and error trials. Thus, the sensory inputs were identical, but the conscious reports differed. We found that, on correct trials, both the entry and rank subspaces exhibited strong activity. By contrast, on error trials, the entry subspace showed much weaker activity, and the rank subspaces displayed no activity selective to the actual stimulus location, but only to the location of the upcoming response (Fig.4A). The result indicated that high activity in the entry subspace and persistent activity in the rank subspaces are crucial for conscious access to spatial sequences.

We further analyzed the neural activity on error trials only. Remember that, for high contrast, the entry subspace reflects actual stimulus location, while for 0% contrast, purely endogenous activity encoded the location that would ultimately be reported. We reasoned that, on intermediate low-contrast trials, a competition may occur between those exogenous and endogenous signals, as indeed found in attractor-based neural network simulations [30,31]. If this was the case, we should

observe neural activities representing both stimulus and response items in the entry subspace. Indeed, on error trials at threshold contrast (8%), we found significant signal strength for both stimulus and response locations in the entry subspace (both $P < 0.001$, two-sided *t*-tests; Fig. 4B), with no significant difference between them ($P = 0.92$, two-sided *t*-tests). However, in this case, only the response location from the entry subspaces was gated into the rank-WM subspaces. This gating and information flow were further confirmed by the trial-by-trial correlation in the response signal between entry and rank subspaces (Pearson's $R = 0.36$, $P < 0.0001$), but not in the stimulus signal (Pearson's $R = 0.02$, $P = 0.62$; Fig. 4C). The results revealed an internal competition between stimulus- and endogenous-driven activity in the entry subspace, particularly when stimulus uncertainty increases under low-contrast conditions. That is, stimulus-driven signals predominated on correct trials, leading to conscious access to the actual stimulus sequence, whereas endogenous-driven signals predominated in error trials, resulting in access to an internally generated sequence (Fig. S6A). Thus, conscious perception of spatial sequence arises from a selection process that resolves the competition between stimulus-driven and endogenous signals.

How did the competition change in other contrast conditions on the error trials? As shown in Fig. 4D and Fig. S6B, we observed that neural activity representing the actual stimulus location in the entry subspace gradually decreased as contrast decreased. Conversely, the endogenous activity displayed the opposite trend, with strength increasing as the contrast decreased. The competition was most prominent under the near-threshold condition, where the two signals showed similar strength, but was always resolved in an all-or-none manner by the rank subspaces.

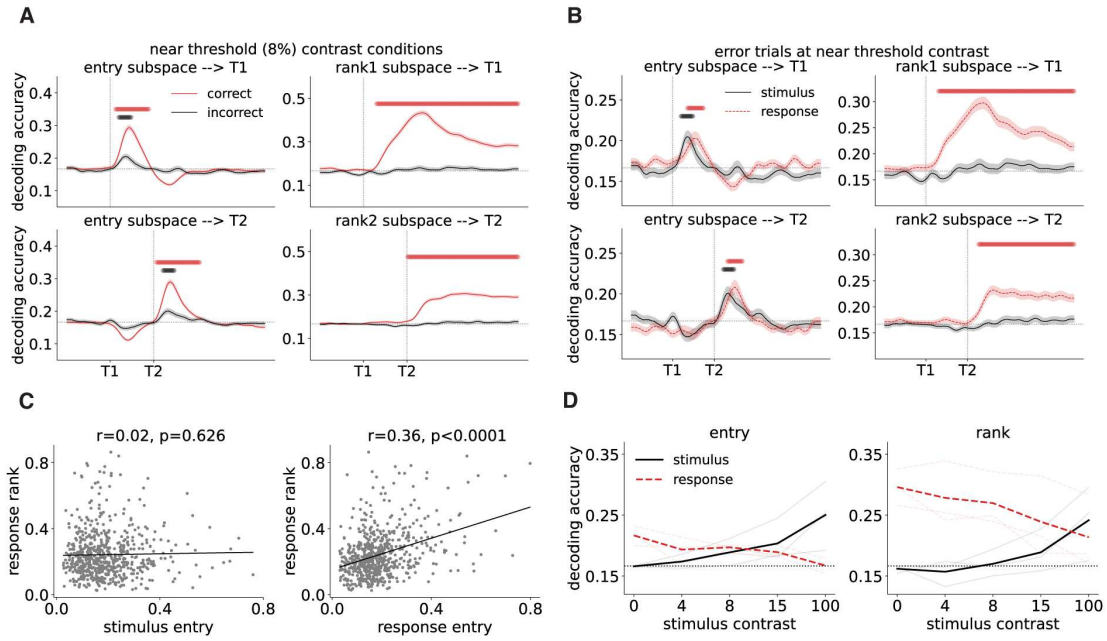


Figure 4. Competition between exogenous and endogenous activity

(A) Subspace activities on near threshold (8%) contrast trials. From an example session of Monkey 1, correct trials show strong activity in both entry and rank-WM subspaces (all $P <$ 0.0001, two-sided *t*-test, n = 1136), while error trials exhibit weak entry subspace activity ($P <$0.0001, n = 635) and absent rank-WM subspace activity (P = 0.21, n = 635). Shaded areas represent SE. Colored horizontal lines denote above-chance decoding ($P < 0.01$, two-sided *t*-test, corrected for multiple comparisons using cluster-based permutation test, n = 5000). Stronger entry subspace activity in correct trials ($P < 0.0001$, two-sided *t*-test, n = 1771) correlates with both all-or-none activity in the corresponding rank subspace and behavioral report of the stimulated sequence.

(B) Coexistence of stimulus- and response-location codes in the entry subspace on error trials. In error trials at 8% contrast, both entry and rank-WM subspaces show strong activity based on subjective reports ($P < 0.0001$, two-sided *t*-test, n = 635), similar to the 0% contrast condition. Signal strengths for stimulus and response items in the entry subspace are comparable (P = 0.92, two-sided *t*-test). Shaded areas indicate SE. Colored horizontal lines denote above-chance decoding ($P < 0.01$, corrected, n = 5000).

(C) Trial-by-trial correlations between subspaces. Response-driven activity in the rank-WM subspace is highly correlated with response-driven activity in the entry subspace (Pearson's R = 0.36, $P < 0.0001$, n = 635), but not with stimulus-driven activity (R = 0.02, P = 0.62).

(D) Stimulus-response interaction across contrast levels in error trials. Across monkeys, stimulus-driven activity in the entry subspace decreases as contrast lowers, while response-driven activity increases. Opaque lines indicate the average across three monkeys; faint lines show data from individual monkeys. The competition is most pronounced at near-threshold (8%) contrast, where stimulus and response signal strengths are similar, and is resolved in an all-or-none manner by the rank-WM subspaces.

Active control of conscious access

The results indicated that conscious access involves a competition process mediating the transfer from entry to WM subspaces, which did not always reflect the actual stimulus, even when it was present. We considered two possibilities for the functional relationship between the entry and rank subspaces: 1) Persistent neural activity in WM subspaces is entirely determined by non-linear thresholding of activities in the entry subspace. Access occurs whenever activity in the entry subspace (possibly driven by noise, alone or in combination with the stimulus) reaches a critical threshold; 2) Access does not depend solely on signal strength within the item subspace itself, but also relies on active, attention- and task-dependent control states that gate information from entry to the rank-WM subspace.

If a control signal gates access to WM, such access should depend on top-down attentional allocation, not just feedforward signals. To test this, we introduced a color-shape distraction task, in which monkeys still had to report the spatial sequences while ignoring the novel colors and shapes at each location (Fig. 5A). Two monkeys were trained on this task: Monkey 2 exhibited few-shot learning with rapid adaptation (Fig. S7A-B), while Monkey 3 displayed a behavioral gradient with gradual improvement (Fig. 5B). On the first day of Monkey 3 performing the distraction task, we found that, during the first ~200 trials, the monkey failed to report locations, exhibiting poor performance (around 20% correct). Between ~200 and ~600 trials, the monkey began to report the correct locations, but most of the time in the wrong order. Only after ~600 trials did sequence WM performance return to normal. Those behavioral results suggested that, when the monkey's attention was diverted by irrelevant features (here novel colors and shapes), conscious access to spatial sequence was severely affected, resulting in either complete loss or instability of the mental representation.

We examined the neural geometry and dynamics of conscious access during the period when the monkey was distracted. Figure 5C shows an example entry neuron that, remarkably, exhibited constant activity across the 3 distraction stages. We further confirmed that, at the population level, all the sensory information, including spatial locations, colors, and shapes, could be decoded equally well across all stages (yet in different entry subspaces; Fig. S7B). Those results indicated that all sensory signals continued to be successfully propagated to the PFC, as also previously observed by others [9].

In the early stage (first ~100 trials), robust item information was observed in the entry subspace, whereas only weak, discontinuous signals were detected in the rank subspaces (Fig. 5E; also see the early stage responses of two example WM neurons in Fig. 5D). Notably, the strength of location information in the entry subspace was stronger than that observed under the near-threshold (8% contrast) condition on correct trials ($P < 0.001$, Welch's *t*-test and Mann-Whitney U test; early error trials: n = 150, mean DA = 0.45, SD = 0.16; correct 8% contrast trials: n = 658, mean DA = 0.40, SD = 0.14; DA refers to decoding accuracy). Nevertheless, this strong entry subspace activity did not lead to any gating into the rank-WM subspaces, where activity remained at a very low, unignited level (Fig. 5E), indicating that bottom-up sensory propagation alone is insufficient for access to WM. Consequently, the robust signal in the entry subspace was insufficient for accurate location reporting (monkey 3 made almost 80% of location errors during the early stage), suggesting that additional selective control of task-relevant information is required for successful conscious access to spatial information.

This necessity became evident as adaptation progressed. During the interim stage, location errors decreased, while swap errors (in which the item order was incorrect) gradually increased. Besides the strong activity in the entry subspace, we also observed transient activities representing S-1 in the rank1-WM subspace; however, it did not persist sufficiently throughout the whole delay period (Fig. 5F),

but rather, was replaced by the rank2-WM subspace at the stage of S-2 onset. Fig. 5D shows the evolution of delay activity during learning for single example neurons. Finally, in the third stage, once the monkey's performance reached 60% correct or higher, the neural activities representing item information in the rank-WM subspaces became more coherent and persistent (Fig. 5G), indicating restored top-down control and stable conscious access.

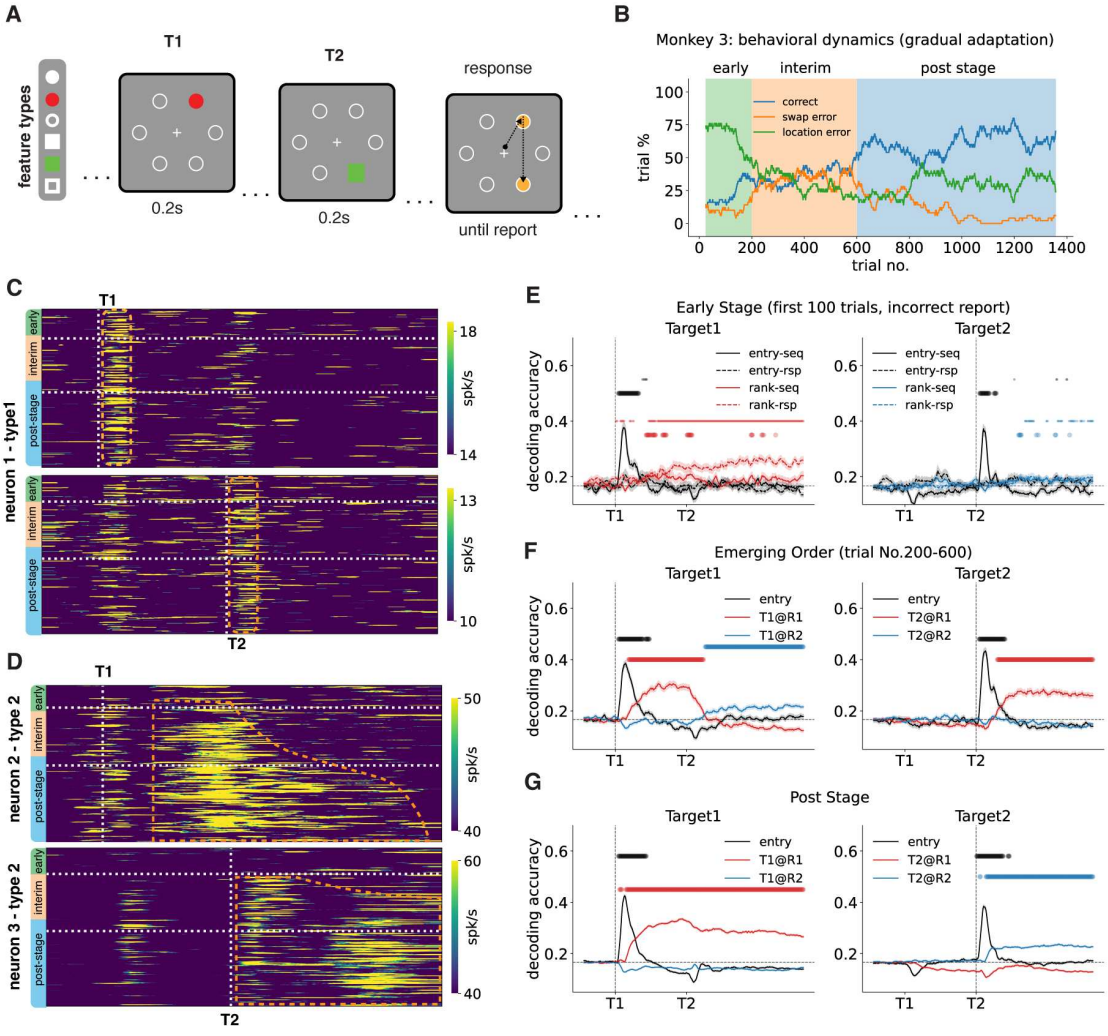


Figure 5. Active control of conscious access

(A) Diagram of the color-shape distraction task. After the completion of the mask experiment, the monkeys were tested on a new sequence reproduction task in which the stimulus cues were changed to a novel combination of color-shape features. This paradigm assessed whether the monkeys could generalize previously learned sequence knowledge to new sensory contexts. Note that monkey 2 and 3 performed this task, while monkey 2

showed fast learning during the first 10 trials (Fig. S7A), while Monkey 3 displayed a behavioral gradient with gradual improvement as follows.
(B) Behavioral dynamics on the first day of Monkey 3. The first day of training revealed three distinct stages of sequence learning based on correct and error response patterns: 1. early exploring stage (first ~200 trials): The monkey predominantly made location errors, suggesting that previously learned sequence knowledge did not directly transfer to the new color-shape stimuli. 2. Middle transitional stage (trials ~200-600): The correct rate rose to about 40%, and location errors dropped from 80% to 40%. Notably, swap errors (wrong order but correct location) increased in this phase, indicating partial learning of item information but incomplete mastery of stable sequence order. 3. Post-adaptation stage (after ~600 trials): The monkey achieved a further increase in correct rate while swap errors declined and location errors increased, suggesting successful integration of the new color-shape cues into the learned sequence structure.
(C-D) Example neurons. **(C)** Type 1 entry neuron (neuron 1) exhibiting constant transient activity across the learning stages. **(D)** Type 2 rank-WM neuron (neuron 2 and 3) displaying evolving activity: no delay activity in the early stage, transient unsustained in the interim stage, and persistent activity in the post-adaptation stage.
(E-G) Population neural dynamics during the learning of the distraction task. **(E)** During the early phase (first ~100 trials), entry subspace activity remained robust, indicating successful encoding of the new sensory features. However, only weak and fragmented signals appeared in the rank-WM subspaces (slightly stronger and more persistent for response labels than for stimulus labels), reflecting an inability to organize the location inputs into a stable temporal sequence (Colored horizontal lines indicate $P < 0.01$, one-sided *t*-test, uncorrected; n = 75 per rank; thick line: stimulus label, thin line: response label. **(F)** In this interim phase (~200-600 trials), transient rank subspace activity began to emerge in rank1 subspace, consistent with the formation of basic ordinal encoding. However, this activity remained unstable and did not persist throughout the delay period, resulting in frequent swap errors ($P < 0.01$, one-sided *t*-test, uncorrected, n = 200 per rank). **(G)** Once the monkey had adapted to the new stimuli and improved performance (after ~600 trials), rank subspace activity became more coherent and persistent, supporting accurate sequence production ($P < 0.01$, one-sided *t*-test, uncorrected, n = 661 per rank). Time courses aligned to T1/T2. Shaded areas indicate SE.

Modeling the controlled ignition process

The key mechanism for ignition, therefore, lies in the delicate control of how information flows from the sensory subspace to each of the two structured WM subspaces. Can a computational model capture this process and shed light on our experimental findings? We used a recently developed interpretable recurrent neural

network (RNN) model with a low-dimensional neural control state space implemented in various cognitive tasks, including perceptual decision-making and WM control [32,33]. When trained on sequences of two or three spatial locations, the model spontaneously develops gating and WM mechanisms capable of storing and reproducing the locations in correct order after a delay. Analyzing how this model operates provides a precise hypothesis about the gating mechanism's role in controlling the ignition process. Inside the RNN, we found a low-dimensional vector space, hereafter called the *control state*, whose activity dynamically modulates the effective connectivity between subspaces by aligning its gain with fixed structural coupling vectors. The theory that we developed to explain this organization [33] posits that 1) flexible routing requires multiple neural subpopulations, and 2) the minimal dimensionality of the control state equals the number of distinct neural subpopulations needed, reflecting task complexity. In the sequence WM task, the control complexity involved in storing a length-K sequence in WM is equal to K, requiring distinct neural subpopulations to carry out information routing from the entry to each of the K WM subspaces.

When an item is gated from the entry subspace to a rank-k WM subspace, we denote the associated information pathway as $input \rightarrow WM_k$ (Fig. 6A). The effective connectivity strength $E_{input \rightarrow WM_k}$ of this pathway depends on two multiplicative factors: one is the structural connectivity of the $input \rightarrow WM_k$, denoted as $\boldsymbol{c}_k$, a K-dimensional vector representing the rank-specific structural coupling from fixed long-term connectivity, which is given by the product of the input and output connectivity weights of the pathway; and the other one is the control state, denoted as g, a K-dimensional vector representing a dynamic modulator of the excitability of the network, such that $E_{input \rightarrow WM_k} = \boldsymbol{g} \cdot \boldsymbol{c}_k$. Crucially, the control state g can be influenced by both external inputs and endogenous signals.

This formulation highlights three critical aspects that unify the model's predictions with empirical observations across experimental conditions. First, the

589 total input comprises two parts: 1) stimulus-driven activity, modulated by contrast,
590 where high contrast (e.g., 100%) amplifies total input via strong bottom-up drive,
591 leading to robust entry subspace activity (Fig. 6B); 2) endogenous activity which
592 elevate total input when external inputs are weak, compensating via internal
593 generation and sustaining rank-WM activity despite absent stimuli (Fig. 6C). Second,
594 $\boldsymbol{c}_k$ embodies the fixed long-term connectivity that anchors the internal model and
595 temporal structure, ensuring pathway fidelity independent of transient inputs. As
596 the product of input and output weights along the $input \rightarrow WM_k$ pathway, $\boldsymbol{c}_k$
597 quantifies the inherent capacity for rank-specific information flow, reflecting learned
598 structural priors stored in PFC circuits, e.g., ordinal rank subspaces in long-term
599 memory (Fig. 6D). Third, the control state $\boldsymbol{g}$ dynamically modulates the strength of
600 the $input \rightarrow WM_k$ pathway. The alignment between $\boldsymbol{g}$ (dynamic) and $\boldsymbol{c}_k$ (static)
601 determines the overall pathway strength $\boldsymbol{E}_{input \rightarrow WM_k}$. In the current sequence WM
602 task, the network must route inputs from the entry subspace to different rank-
603 specific WM subspaces depending on ordinal progression (e.g., rank-1, rank-2, rank-
604 3; Fig. 6E). During the rank-k sensory period, the control state $\boldsymbol{g}$ aligns with the
605 corresponding structural vector c_k while remaining orthogonal to others, effectively
606 gating on the pathway to the rank-k WM subspace while gating off the irrelevant
607 ones (Fig. 6F).

608 Simulations presented in Figure 6 show that the model therefore reproduces the
609 data so far, including the monotonic modulation of entry subspace activity as a
610 function of input strength, followed by its all-or-none gating into the appropriate
611 rank subspaces (Figure 6B). Moreover, the control state is assumed to depend on all
612 kinds of inputs, such as the color-shape distractor input. Therefore, the global top-
613 down input modulation of conscious access (for instance by the color-shape
614 distractor) can be conveniently modelled through changing the control state $\boldsymbol{g}$ (Fig.
615 6D & F). In the simulation, modifying the control state ($\boldsymbol{g}$) to reduce its alignment
616 with the relevant structural coupling vector cancels the WM subspace activation

while leaving the entry subspace activation intact (Fig. 6B), as experimentally observed. This interaction between $\boldsymbol{g}$ and $\boldsymbol{c}_k$ highlights how internal models constrain perception, resolving competition and stabilizing attractors even under uncertainty (Fig. 4D), and provides a mechanistic bridge between sensory processing and conscious ignition, where $\boldsymbol{c}_k$ enforces prior expectations to shape conscious perception.

Crucially, this control state logic generates two non-obvious predictions. First, for the length-2 sequence WM gating problem, the control state theory predicts that entry neurons, although their activity generalizes across ranks (Fig. 2A), should nevertheless exhibit rank-dependent changes in gain, forming at least two distinct neural subpopulations: one that prefers the first item to the second, and another that prefers the second to the first (Fig. 6G, upper panels). This mechanism arises in the model because the same entry population must be differentially gated depending on the current ordinal context. Separating the entry into distinct rank-modulated subpopulations is the mechanism by which the RNN solves the problem of receiving identical sensory inputs at rank 1 and rank 2, and yet routing them to distinct WM subspaces. When we tested whether the monkey brain uses a similar mechanism, we indeed found two clusters of entry neurons with stable preferences for WM1 and WM2 ($|\Delta BIC| = 86$, $P < 0.0001$, Shapiro-Wilk and Kolmogorov-Smirnov normality tests; Fig. 6G, middle panels). No such clusters were observed in the control data before learning ($P = 0.46$ with Kolmogorov-Smirnov test; Fig. 6G, lower panels).

A second prediction is that this separation of the entry subspace into two clusters should disappear in the monkey who was distracted, and should progressively re-emerge as it overcomes the distraction; furthermore, the gain component ($\boldsymbol{g}$) of the control state, which can be measured by the difference in the activity of the two entry neuronal subpopulations, should track behavioral sequence reproduction performance. This prediction was also confirmed: 1) the two-cluster

structure was weak in the first ~400 trials and became progressively clearer in the last 400 trials (Fig. 6H, upper panel); 2) the trial-by-trial control gain estimates derived from the differential activity of those two neuronal subpopulations correlated strongly with the progressive recovery of good performance in the distraction task (Spearman's $R = 0.47$, $P < 0.0001$; Fig. 6H, lower panel). Thus, the model proposes that conscious access emerges from dynamic effective connectivity within local PFC circuits: structural connectivity remains constant, but selective gating via the control state acts as a "switch" to ignite the appropriate PFC subspace.

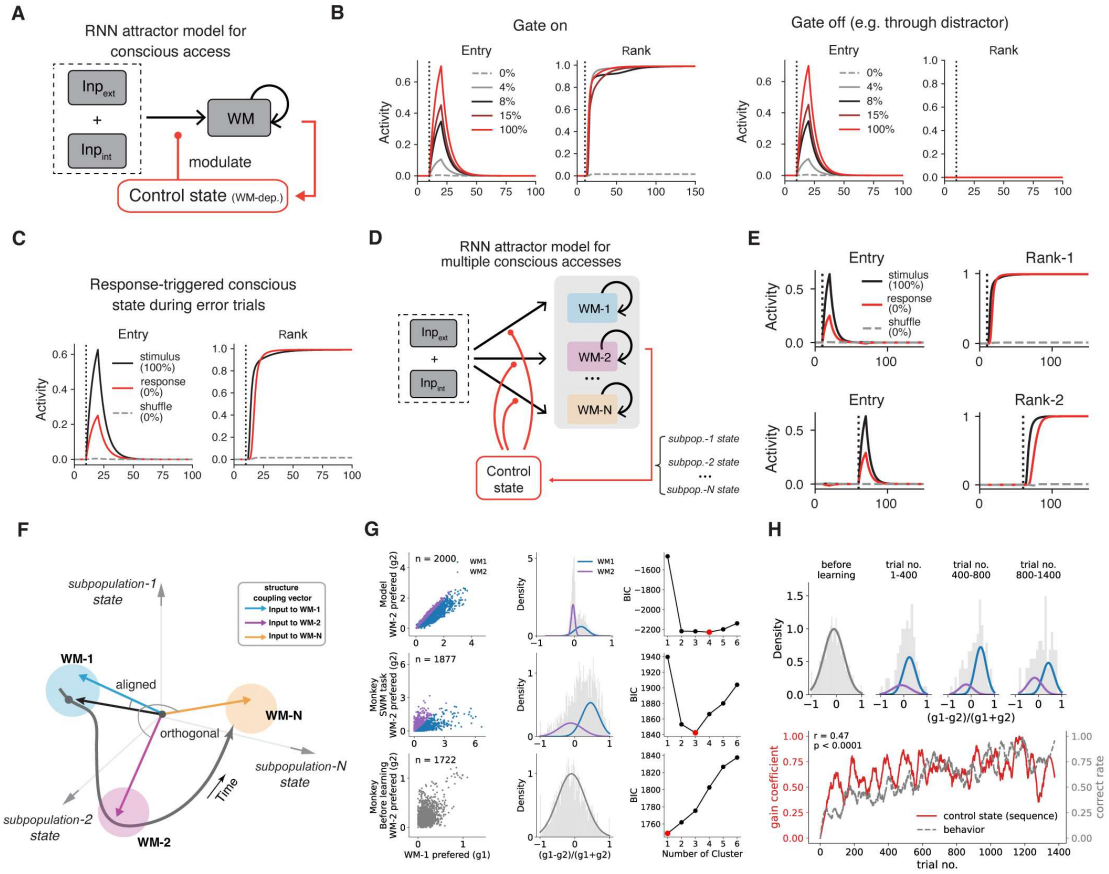


Figure 6. Control state-dependent gating in a compositional RNN model of conscious access.

(A) Schematic of the RNN model for two-step conscious access. Sensory input is transiently encoded in the entry subspace and routed into a working-memory subspace under the control of a population-gain-based control state (red box). The pathway strength is $\boldsymbol{E}_{input \to WM} = \boldsymbol{g} \cdot \boldsymbol{c}$, where $\boldsymbol{g}$ is the subpopulation gain and $\boldsymbol{c}$ the structural coupling.

(B) Model dynamics in the single-stimulus case. Entry subspace activity increases monotonically with stimulus contrast, while WM subspace activity exhibits all-or-none ignition, consistent with empirical recordings of sequence consciousness (mimicking Fig.2D-F).
(C) Endogenous response-related activity. In the absence of external input (error trials or 0% contrast trials), internally generated activity in the entry subspace can drive ignition in the WM subspace (red line), producing conscious-like responses without stimulus (mimicking Figs. 3F, 4B).
(D) Schematic of the RNN attractor model for multiple conscious accesses. Inputs can be routed into multiple WM subspaces (WM-1, WM-2, …, WM-N) through distinct pathways. The vector-valued control state regulates which subspace is selected, with alignment between $\boldsymbol{g}$ and rank-specific $\boldsymbol{c}$ determning effective connectivity strength.
(E) Sequential access of two stimuli. Entry activity encodes both stimuli transiently, but WM subspaces show persistent, rank-specific activity: WM1 maintains the first stimulus after its onset and WM2 maintains the second. Error trials or 0% contrast trials (red line) show response-triggered activity in the absence of stimulus inputs for both sequence rank.
(F) Mechanism of dynamic gating. The control state is represented as a vector of population gains ($\boldsymbol{g}$). Effective connectivity into each WM subspace is determined by alignment between $\boldsymbol{g}$ and the corresponding structural coupling vector (e.g. blue, green or yellow arrow for different conscious states, respectively). When aligned, the gate is open; when orthogonal, it is closed. Sequential conscious access arises from time-dependent re-alignment of the control state vector across ordinal stages.
(G-H) Model predictions and verifications. The control state mechanism generates a key, model-derived prediction beyond a homogenous entry population: even though these neurons receive identical sensory input, they must be differentially gated via discrete subpopulations exhibiting rank-specific gain preferences tuned to WM1 versus WM2. **(G)** During the input period of the sequence WM task, item-selective neurons exhibit stable tuning phases across different ordinal ranks (WM-1 and WM-2). Left column: Scatter plot of preferred amplitude values for WM-1 and WM-2 across neurons (each point represents one neuron). Middle column: Histogram of normalized difference ratio with 2-component Gaussian Mixture Model (GMM) fit (blue: WM-1 component; purple: WM-2 component), revealing a clear multimodal structure. Gray bars represent the empirical density. Right column: Bayesian Information Criterion (BIC) as a function of the number of clusters (subpopulations). The substantial drop in BIC from k=1 to k=2 supports the presence of at least two distinct clusters. Upper row show results from the model simulations, $|\Delta BIC| = 751$, $P < 0.0001$ with Shapiro-Wilk and Kolmogorov-Smirnov tests. Middle row: empirical data during monkey performing masked SWM task ($|\Delta BIC| = 86$, $P < 0.0001$ with Shapiro-Wilk and Kolmogorov-Smirnov tests), confirming rank-specific gain preferences in entry neurons as predicted. Lower row: control analysis showing no clear clusters before the monkey learned the task (sequence reproduction performance at near 20%; $P = 0.46$ with Kolmogorov-Smirnov test). **(H)** Upper: Evolution of the two-cluster structure over the course

of the distraction task (control distribution before learning shown on the leftmost side). Lower: Trial-by-trial analysis reveals that the gain component ($\boldsymbol{g}$) of the control state correlates with behavioral sequence reproduction performance (gradual improvement in the distraction task; Spearman's $R = 0.47$, $P < 0.0001$), with higher $\boldsymbol{g}$ associated with correct sequence reproduction and lower $\boldsymbol{g}$ linked to distraction by task irrelevant features.

Discussion

We recorded thousands of neurons in the lateral PFC of macaque monkeys while they performed a masked sequence WM task with variable stimulus contrast (0-100%). Our findings revealed that conscious access to spatial sequences follows a two-step process: transient monotonic sensory encoding in an entry subspace, followed by sustained, all-or-none activity in rank-WM subspaces. Furthermore, such access was not solely determined by stimulus strength but also involved competition between stimulus-driven and endogenous signals, and was controlled by top-down attentional gating and internal temporal structures. A recurrent neural network model, formalizing the gating mechanism as the outcome of a dynamic control state capable of modulating structural couplings, replicated the main empirical observations (ignition, competition, and distraction effects).

Sensory signals reach the PFC, but do not guarantee access to WM

Our study demonstrates the existence of the two postulated stages of sensory processing (linear followed by non-linear), shows that they coexist as vector states over the same PFC neurons, and provides direct content-specific evidence that neural signals propagating to the PFC do not invariably lead to a second stage of ignition and access to WM. We observed an early transient sensory component in PFC activity, monotonically increasing with input strength, that could, in turn, on different trials, ignite or fail to ignite sustained all-or-none activity in rank subspaces. Previous research has implicated the PFC in conscious perception, but its sustained signals were largely attributed to long-distance reciprocal connections with sensory

areas [11,34,35], suggesting that the fate of sensory inputs is controlled and broadcast beyond the prefrontal regions via reciprocal loops. Van Vugt et al. [23] demonstrated that the activity evoked by weak visual stimuli in areas V1 and V4 elicited subthreshold PFC activity insufficient for recurrent processing, yet their methods (a single point-of-light stimulus, present or absent) did not allow them to separate the neural subspaces that we report here. Our decomposition into entry (transient, contrast-dependent) and rank-WM (sustained, invariant) subspaces resolves this by showing that PFC activity encodes both sensory strength and the all-or-none active gating of information for structured access to WM. The transient sensory inputs, sustained WMs, and the ignition process seem to be instantiated by the same prefrontal circuitry [9].

Moreover, we also found that the entry subspace is the seat of a competition between inputs from diverse sources, with endogenous-driven signals being processed in parallel to weak external inputs, or replacing them in full on stimulus-absent trials (0% contrast). This competition between exogenous and endogenous signals is fully resolved by the time the information reaches the rank subspaces, allowing only a single location to enter WM. This aspect of our finding fits with classical bottleneck models of cognitive architecture, where multiple parallel processors compete for access to a single capacity-limited central stage, which is capable of processing only a single item at a time [36]. Such models have been related to the prefrontal cortex in humans [16,37], where the PFC is proposed to act as a limited-capacity workspace that resolves competition through selective attention and gating. This competition can lead to misperceptions under sensory uncertainty, as internal signals dominate on error trials.

According to a recent proposal, signal strength may be one of the key factors that allows the viewer to distinguish sensory from endogenous signals [29,38]. Since we found that endogenous entry-space signal strength can be equal or even stronger than sensory signals, this model would imply that, on low-contrast stimuli, from the

760 monkey's viewpoint, the endogenous signal may have been indistinguishable from
761 reality. It would readily explain that the entire processing chain of WM subspace
762 activation and behavior unfolded normally even on 0% trials.
763
764 **Report and no-report paradigms**
765 Like much previous work in animals [23,39-42], the present work assesses conscious
766 perception using an explicit-report paradigm, in which monkeys had to actively
767 reproduce the two locations they had seen in proper order. It rests on the
768 assumption that trials with correct task performance must have been accompanied
769 by conscious access to the stimulated location. This seems a safe assumption:
770 although above-chance delayed performance has been reported in humans on trials
771 that were subjectively reported as unseen [17,43-45], such non-conscious task
772 performance is generally very low and short-lived, and therefore unlikely to concern
773 more than a very small number of correct trials, given the low chance level of the
774 present task (1/30) and the delay of several seconds between stimulus and
775 response. Conversely, on error trials, the objective stimuli were not consciously
776 reportable after a short delay. Our neurophysiological recordings indicate that, on
777 such trials, following an initial transient, there was a complete replacement of the
778 stimulus-induced activity by identical neural activity coding for another location,
779 suggesting that the monkey may have consciously perceived an illusory stimulus at
780 the wrong location.
781 Nevertheless, we readily acknowledge that our use of an explicit report task
782 leaves some uncertainty about which of the observed processing stages
783 implemented the "neural correlates of consciousness proper", and which reflected
784 additional post-processing [46]. The present results add complexity to this issue by
785 showing that the very same prefrontal neurons may contribute to distinct
786 representational subspaces, some of which exhibit early continuous signals and
787 others late all-or-none signals [47]. The data, therefore, leave open two distinct

interpretations. First, early activity in the entry subspace may index non-conscious sensory processing, perhaps merely reflecting contemporaneous activity in posterior inferotemporal and occipito-parietal brain areas. Alternatively, it may already reflect some partial or fleeting conscious perception of the stimulated location – a minimal degree of phenomenal awareness [48] which is quickly lost and replaced by a different content on error trials, thus rendering it non-reportable.

Several arguments nevertheless seem to us to favor the first possibility. Although the present work relied on overt conscious reports, a similar late all-or-none bifurcation of neural activity was found to be a signature of conscious access even on no-report task-irrelevant trials, where participants merely observe the stimuli without performing any subsequent cognitive task [49,50]. In spite of initial doubts [51,52], several studies have found that neural activity on visible trials clearly reaches the PFC [53,54], that PFC plays a causal role in conscious perception [55-58], and that the neural activity that separates physically identical trials into seen and unseen corresponds to the beginning of sustained WM activity [59]. Thus, all-or-none bifurcated neuronal activity in WM-related neurons appears as the most likely correlate of consciousness in the present study.

The role of internal structures in shaping conscious content

Conscious experience is not a passive readout of sensory inputs but an active construction shaped by prior knowledge and expectations [1-3,60]. Most traditional consciousness research has overlooked the role of internal structures by focusing on unstructured paradigms with impoverished stimuli, limiting insights into how long-term memory priors constrain perception. Our sequence task reveals conscious access to structured content, e.g., temporal sequences, that rely on rank subspaces as an internal scaffold, acting as a geometric prior in the neural state space that guides the transformation of transient sensory signals into stable, ordered conscious representations. The rank-specific subspaces we identified function as temporal

templates, acquired through learning and stored as structural connectivity patterns ($\boldsymbol{c}$ in our model). Each subspace encodes not just spatial locations but their ordinal position within a sequence, effectively representing the abstract concept of "first" versus "second" item. This geometric architecture may reflect a general principle of conscious organization: complex conscious experiences are composed of simpler elements represented in separable neural dimensions. Such compositionality allows for the vast combinatorial possibilities of conscious contents while maintaining the stability of individual elements. We tentatively propose that similar principles of neural geometry likely operate for other structured contents, such as objects composed of features, scenes composed of objects, sentences composed of words, or narratives composed of events. The contents of human consciousness may be particularly rich in that respect, owing to the human faculty to compose more complex, tree-based symbolic structures [61,62].

Finally, how the ignition process is implemented and shaped within the neural states of the prefrontal circuit is still not fully answered. Nevertheless, based on our neural data and computational model, we can make following postulations: 1) The multiple mixed-selectivity neural populations (particularly those entry neurons that also carry rank-related information) in frontal circuits may serve as the mechanistic underpinning for conscious access; such compositional codes are expected to be generalized to the access of other contents, e.g., visual features in temporal cortex, spatial relations in parietal cortex, or somatosensory information in frontal areas, etc. [63,64]. 2) Multiple subspaces could interact with each other through the proposed gain modulation mechanism [9,65]. The WM control process implemented in PFC seems directly linked to conscious access, thus offering a universal framework for conscious information sharing across subspaces or even across brain regions. 3) According to the model, conscious access is channeled by the anatomical connectivity among the identified neural populations, likely formed during the task structure (ordinal ranks) learning. In the present study, we focused solely on the neural state of entry

subspaces responsive to varying stimulus visibility. However, the control or access strength between subspaces may also depend on the dynamic excitability (gain) of multiple subpopulations and their structural coupling (as shown in our model). In particular, the alignment between gain vectors and structural coupling across subspaces provides a mechanistic explanation for how a learned internal structure can shape conscious perception. Future optogenetic or pharmacological perturbations of PFC connections could directly test our current framework by selectively modulating the strength of effective connectivity, thereby causally preventing or controlling entry-to-rank transfer and validating the predicted role of PFC dynamics in conscious access.

References

1 Bar, M. The proactive brain: using analogies and associations to generate predictions. *Trends Cogn Sci* **11**, 280-289, doi:10.1016/j.tics.2007.05.005 (2007).

2 Girshick, A. R., Landy, M. S. & Simoncelli, E. P. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat Neurosci* **14**, 926-932, doi:10.1038/nn.2831 (2011).

3 Bays, P. M., Schneegans, S., Ma, W. J. & Brady, T. F. Representation and computation in visual working memory. *Nat Hum Behav* **8**, 1016-1034, doi:10.1038/s41562-024-01871-2 (2024).

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

5 Miller, E. K., Lundqvist, M. & Bastos, A. M. Working Memory 2.0. *Neuron* **100**, 463-475, doi:10.1016/j.neuron.2018.09.023 (2018).

6 Seth, A. *Being you: A new science of consciousness*. (Penguin, 2021).

7 Tolman, E. C. Cognitive maps in rats and men. *Psychol Rev* **55**, 189-208, doi:10.1037/h0061626 (1948).

8 Behrens, T. E. J. *et al.* What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron* **100**, 490-509, doi:10.1016/j.neuron.2018.10.002 (2018).

9 Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78-84, doi:10.1038/nature12742 (2013).

10 El-Gaby, M. *et al.* A cellular basis for mapping behavioural structure. *Nature* **636**, 671-680, doi:10.1038/s41586-024-08145-x (2024).

11 Mashour, G. A., Roelfsema, P., Changeux, J. P. & Dehaene, S. Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron* **105**, 776-798, doi:10.1016/j.neuron.2020.01.026 (2020).

12 Baars, B. J. The conscious access hypothesis: origins and recent evidence. *Trends Cogn Sci* **6**,

47-52, doi:10.1016/s1364-6613(00)01819-2 (2002).
13 Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J. & Sergent, C. Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn Sci* **10**, 204-211, doi:10.1016/j.tics.2006.03.007 (2006).
14 Dehaene, S., Sergent, C. & Changeux, J. P. A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proc Natl Acad Sci U S A* **100**, 8520-8525, doi:10.1073/pnas.1332574100 (2003).
15 Sergent, C., Baillet, S. & Dehaene, S. Timing of the brain events underlying access to consciousness during the attentional blink. *Nat Neurosci* **8**, 1391-1400, doi:10.1038/nn1549 (2005).
16 Marti, S., Sigman, M. & Dehaene, S. A shared cortical bottleneck underlying Attentional Blink and Psychological Refractory Period. *Neuroimage* **59**, 2883-2898, doi:10.1016/j.neuroimage.2011.09.063 (2012).
17 King, J. R. & Dehaene, S. A model of subjective report and objective discrimination as categorical decisions in a vast representational space. *Philos Trans R Soc Lond B Biol Sci* **369**, 20130204, doi:10.1098/rstb.2013.0204 (2014).
18 Dehaene, S. in *Biological Physics: Poincaré Seminar 2009.* 141-168 (Springer).
19 Shadlen, M. N. & Kiani, R. in *Characterizing consciousness: from cognition to the clinic?* 27-46 (Springer, 2011).
20 Fang, Z. *et al.* Human high-order thalamic nuclei gate conscious perception through the thalamofrontal loop. *Science* **388**, eadr3675, doi:10.1126/science.adr3675 (2025).
21 Quiroga, R. Q., Mukamel, R., Isham, E. A., Malach, R. & Fried, I. Human single-neuron responses at the threshold of conscious recognition. *Proc Natl Acad Sci U S A* **105**, 3599-3604, doi:10.1073/pnas.0707043105 (2008).
22 Levinson, M., Podvalny, E., Baete, S. H. & He, B. J. Cortical and subcortical signatures of conscious object recognition. *Nat Commun* **12**, 2930, doi:10.1038/s41467-021-23266-x (2021).
23 van Vugt, B. *et al.* The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science* **360**, 537-542, doi:10.1126/science.aar7186 (2018).
24 Xie, Y. *et al.* Geometry of sequence working memory in macaque prefrontal cortex. *Science* **375**, 632-639, doi:10.1126/science.abm0204 (2022).
25 Tian, Z. *et al.* Mental programming of spatial sequences in working memory in the macaque frontal cortex. *Science* **385**, eadp6091, doi:10.1126/science.adp6091 (2024).
26 Li, S. *et al.* Flexible Use of Limited Resources for Sequence Working Memory in Macaque Prefrontal Cortex. *Nat Commun* **16**, 10386, doi:10.1038/s41467-025-65380-0 (2025).
27 Chen, J., Zhang, C., Hu, P., Min, B. & Wang, L. Flexible control of sequence working memory in the macaque frontal cortex. *Neuron* **112**, 3502-3514 e3506, doi:10.1016/j.neuron.2024.07.024 (2024).
28 Dijkstra, N. & Fleming, S. M. Subjective signal strength distinguishes reality from imagination. *Nat Commun* **14**, 1627, doi:10.1038/s41467-023-37322-1 (2023).
29 Wadia, V. S. *et al.* A shared code for perceiving and imagining objects in human ventral temporal cortex. *Science* **392**, 207-215, doi:10.1126/science.adt8343 (2026).

30 Wong, K. F. & Wang, X. J. A recurrent network mechanism of time integration in perceptual decisions. *J Neurosci* **26**, 1314-1328, doi:10.1523/JNEUROSCI.3733-05.2006 (2006).
31 Dehaene, S. & Changeux, J. P. Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentional blindness. *PLoS Biol* **3**, e141, doi:10.1371/journal.pbio.0030141 (2005).
32 Zhang, Y., Feng, J. & Min, B. Elucidating the selection mechanisms in context-dependent computation through low-rank neural network modeling. *Elife* **13**, doi:10.7554/eLife.103636 (2025).
33 Zhang, Y. *et al.* Discovering Novel Circuit Mechanisms in Higher Cognition through Factor-Centric Recurrent Neural Network Modeling. *bioRxiv*, 2026.2004. 2016.718933 (2026).
34 Amit, D. J. & Brunel, N. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb Cortex* **7**, 237-252, doi:10.1093/cercor/7.3.237 (1997).
35 Bouchacourt, F. & Buschman, T. J. A Flexible Model of Working Memory. *Neuron* **103**, 147-160 e148, doi:10.1016/j.neuron.2019.04.020 (2019).
36 Pashler, H. Dual-task interference in simple tasks: data and theory. *Psychol Bull* **116**, 220-244, doi:10.1037/0033-2909.116.2.220 (1994).
37 Tombu, M. N. *et al.* A Unified attentional bottleneck in the human brain. *Proc Natl Acad Sci U S A* **108**, 13426-13431, doi:10.1073/pnas.1103583108 (2011).
38 Dijkstra, N., Bosch, S. E. & van Gerven, M. A. J. Shared Neural Mechanisms of Visual Perception and Imagery. *Trends Cogn Sci* **23**, 423-434, doi:10.1016/j.tics.2019.02.004 (2019).
39 Nieder, A., Wagener, L. & Rinnert, P. A neural correlate of sensory consciousness in a corvid bird. *Science* **369**, 1626-1629, doi:10.1126/science.abb1447 (2020).
40 Cowey, A. & Stoerig, P. Blindsight in monkeys. *Nature* **373**, 247-249, doi:10.1038/373247a0 (1995).
41 Leopold, D. A. & Logothetis, N. K. Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* **379**, 549-553, doi:10.1038/379549a0 (1996).
42 de Lafuente, V. & Romo, R. Neuronal correlates of subjective sensory experience. *Nat Neurosci* **8**, 1698-1703, doi:10.1038/nn1587 (2005).
43 Soto, D., Mantyla, T. & Silvanto, J. Working memory without consciousness. *Curr Biol* **21**, R912-913, doi:10.1016/j.cub.2011.09.049 (2011).
44 Bergstrom, F. & Eriksson, J. Neural Evidence for Non-conscious Working Memory. *Cereb Cortex* **28**, 3217-3228, doi:10.1093/cercor/bhx193 (2018).
45 Lau, H. C. & Passingham, R. E. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proc Natl Acad Sci U S A* **103**, 18763-18768, doi:10.1073/pnas.0607716103 (2006).
46 Aru, J., Bachmann, T., Singer, W. & Melloni, L. Distilling the neural correlates of consciousness. *Neurosci Biobehav Rev* **36**, 737-746, doi:10.1016/j.neubiorev.2011.12.003 (2012).
47 Del Cul, A., Baillet, S. & Dehaene, S. Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biol* **5**, e260, doi:10.1371/journal.pbio.0050260 (2007).
48 Block, N. What Is Wrong with the No-Report Paradigm and How to Fix It. *Trends Cogn Sci* **23**,

1003-1013, doi:10.1016/j.tics.2019.10.001 (2019).

49 Sergent, C. *et al.* Bifurcation in brain dynamics reveals a signature of conscious processing independent of report. *Nat Commun* **12**, 1149, doi:10.1038/s41467-021-21393-z (2021).

50 Cohen, M. A., Dembski, C., Ortego, K., Steinhibler, C. & Pitts, M. Neural signatures of visual awareness independent of postperceptual processing. *Cereb Cortex* **34**, doi:10.1093/cercor/bhae415 (2024).

51 Tsuchiya, N., Wilke, M., Frassle, S. & Lamme, V. A. F. No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. *Trends Cogn Sci* **19**, 757-770, doi:10.1016/j.tics.2015.10.002 (2015).

52 Frassle, S., Sommer, J., Jansen, A., Naber, M. & Einhauser, W. Binocular rivalry: frontal activity relates to introspection and action but not to perception. *J Neurosci* **34**, 1738-1747, doi:10.1523/JNEUROSCI.4403-13.2014 (2014).

53 Panagiotaropoulos, T. I., Deco, G., Kapoor, V. & Logothetis, N. K. Neuronal discharges and gamma oscillations explicitly reflect visual consciousness in the lateral prefrontal cortex. *Neuron* **74**, 924-935, doi:10.1016/j.neuron.2012.04.013 (2012).

54 Kapoor, V. *et al.* Decoding internally generated transitions of conscious contents in the prefrontal cortex without subjective reports. *Nat Commun* **13**, 1535, doi:10.1038/s41467-022-28897-2 (2022).

55 Weilnhammer, V. *et al.* An active role of inferior frontal cortex in conscious experience. *Curr Biol* **31**, 2868-2880 e2868, doi:10.1016/j.cub.2021.04.043 (2021).

56 Bellet, J. *et al.* Decoding rapidly presented visual stimuli from prefrontal ensembles without report nor post-perceptual processing. *Neurosci Conscious* **2022**, niac005, doi:10.1093/nc/niac005 (2022).

57 Bellet, M. E. *et al.* Spontaneously emerging internal models of visual sequences combine abstract and event-specific information in the prefrontal cortex. *Cell Rep* **43**, 113952, doi:10.1016/j.celrep.2024.113952 (2024).

58 Dellert, T. *et al.* Dissociating the Neural Correlates of Consciousness and Task Relevance in Face Perception Using Simultaneous EEG-fMRI. *J Neurosci* **41**, 7864-7875, doi:10.1523/JNEUROSCI.2799-20.2021 (2021).

59 Trubutschek, D. *et al.* A theory of working memory without consciousness or sustained activity. *Elife* **6**, doi:10.7554/eLife.23871 (2017).

60 Seth, A. K. & Bayne, T. Theories of consciousness. *Nat Rev Neurosci* **23**, 439-452, doi:10.1038/s41583-022-00587-4 (2022).

61 Penn, D. C., Holyoak, K. J. & Povinelli, D. J. Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behav Brain Sci* **31**, 109-130; discussion 130-178, doi:10.1017/S0140525X08003543 (2008).

62 Dehaene, S., Meyniel, F., Wacongne, C., Wang, L. & Pallier, C. The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. *Neuron* **88**, 2-19, doi:10.1016/j.neuron.2015.09.019 (2015).

63 de Lafuente, V. & Romo, R. Neural correlate of subjective sensory experience gradually builds up across cortical areas. *Proc Natl Acad Sci U S A* **103**, 14266-14271,

doi:10.1073/pnas.0605826103 (2006).

64 He, B. J. Towards a pluralistic neurobiological understanding of consciousness. *Trends Cogn Sci* **27**, 420-432, doi:10.1016/j.tics.2023.02.001 (2023).

65 Barbosa, J. *et al.* Early selection of task-relevant features through population gating. *Nat Commun* **14**, 6837, doi:10.1038/s41467-023-42519-5 (2023).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- suppmaterials20260623.docx