Representational Geometries of Perception and Working Memory

Tomoya Nakamura[1,2,3,*], Seng Bum Michael Yoo[2,4,5], Kendrick Kay[6], Hakwan Lau[2,4,5,*], & Ali Moharramipour[1,*]

[1]Center for Brain Science, RIKEN, Wako, Japan
[2]Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, Republic of Korea
[3]Japan Society for the Promotion of Science, Tokyo, Japan
[4]Department of Biomedical Engineering, Sungkyunkwan University, Suwon, Republic of Korea
[5]Department of Intelligent Precision Healthcare Convergence, Sungkyunkwan University, Suwon, Republic of Korea
[6]Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Minneapolis, USA

*Correspondence: tomoya.nakamura@riken.jp; alimoharrami1371@gmail.com; hakwan@gmail.com

Highlights

• Representational geometry rotates when perceptual content enters working memory.

• Such rotation expands the representational dimensionality for flexible coding.

• Perceptual and mnemonic representations are still generalizable across each other.

• Flexibility–generalization tradeoff aligns with the unimodal–transmodal gradient.

Abstract

This study compared the representational geometry of visual perception and visual working memory using human 7T fMRI. Observers viewed a face-scene blended sample, attended to either the face or scene, and judged whether the attended aspect matched a subsequent test image. We found that the stimulus coding dimension distinguishing faces and scenes was rotated when perceptual content entered working memory, in both visual and association areas. In these regions, exclusive combinations (i.e., perceived face and memorized scene vs. perceived scene and memorized face) were linearly decodable, indicating that perception and working memory occupy distinct subspaces. Such high-dimensional, flexible coding may help prevent confusion between perception and memory. In contrast, robust cross-decoding across perception and working memory revealed factorized representational formats, consistent with generalizability of their contents. The balance between flexibility and abstraction varied along the cortical hierarchy: early sensory regions emphasized flexibility, whereas transmodal regions showed greater abstraction.

Keywords

neural state space, rotational dynamics, dimensionality, visual phenomenology, cross-decoding, cognitive flexibility, abstract coding, pre-registered study, 7T fMRI

Introduction

Our visual experiences during perception and working memory differ markedly, such that we rarely confuse the two. At the same time, when perceptual content enters working memory, we readily recognize that both states refer to the same content. How, then, does the human brain maintain the same content, and yet distinguishes between perception and working memory [1–6]? One possibility is that some brain regions represent visual content irrespective of whether it is perceived or maintained in working memory, while other regions represent the content in a format that distinguishes perception and working memory. In the present study, using 7T fMRI, we sought to identify these putative brain areas by examining their neural representational geometry [7–11], when human observers perceived and held in working memory either face or scene stimuli.

Our focus on representational geometry was inspired by recent electrophysiological studies in animal models. For example, Libby and Buschman (2021) found that the stimulus coding axis *rotates* when a perceptual representation transforms into working memory, in the sensory cortex [12]. They proposed that such rotation may serve to protect memory content from interference by newly incoming perceptual information (see also [13, 14]).

Another key property of representational geometry is dimensionality, which has been linked to the capacity or flexibility of neural systems [7, 8, 15–18]. High-dimensional representational structure, in particular, is considered to be a characteristic feature in the association cortices, such as the prefrontal cortex, to support flexible behavior in complex tasks [15, 19]. In the context of distinguishing between perceptual and working memory content [1–3, 20, 21], Mendoza-Halliday and Martinez-Trujillo (2017) found that, in the monkey lateral prefrontal cortex, perceptual and memorized content for the same stimuli are represented by neurons that exhibit similar tuning in both cases as well as by neurons tuned to only one of the two conditions, indicating that there is information enough to distinguish perception and working memory at the neural population level [22]. However, it remains unclear whether such diverse selectivity shapes the high-dimensional structure at the mesoscopic level.

As such, coding axis rotation as well as high-dimensional representations are suited for distinguishing perception and working memory. In contrast, how can the brain determine that the content remains the same? One possibility is that despite the rotation of the coding axis, the two coding schemes do not become fully orthogonal. If this is the case, a shared structure may be preserved that allows generalization across perception and working memory. Indeed, it was empirically demonstrated in monkey neural recordings that high-dimensional structure still exhibits generalization [7].

In the present study, using 7T fMRI in humans, we found that visual content is indeed represented in nonparallel dimensions, between perception and working

memory. Specifically, the rotation implies expansion into an extra effective dimension. However, the rotation does not render the coding dimensions completely orthogonal, and therefore, generalization of content is still preserved across perception and working memory.

Results

Perception and working memory contents were coded in non-parallel axes

Using ultra-high-field (7T) fMRI, we measured cortical activity while observers (n = 8) performed a change detection task in which they were cued to remember either the face or the scene aspect of a sample stimulus and then judged whether the cued aspect had changed after a delay period (Figure 1A). Behavioral performance was high (mean correct response rate = 94.3%), indicating that observers followed the instructions and were well engaged in the task (Figure S1).

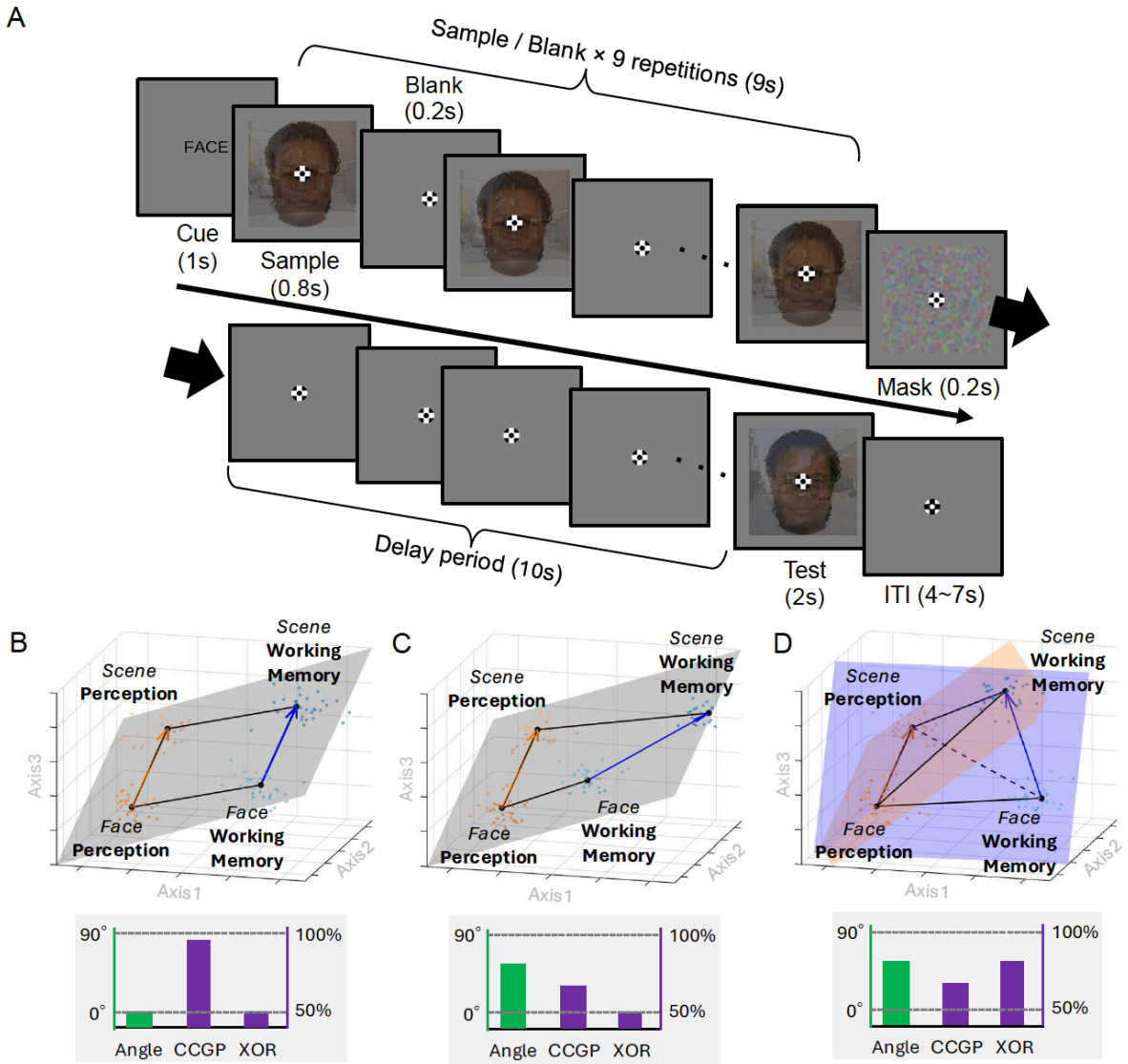

Figure 1. Task design and analytic framework of representational geometry. *(A) In each trial, observers were instructed to memorize only the cued aspect (either the face or the scene) of a face-scene blended sample stimulus, which was repeatedly presented (perception epoch). Following a delay period (working memory epoch), a test stimulus appeared in a randomly selected half of the trials, and observers judged whether the cued aspect of the test stimulus was the same as or different from the sample. In the remaining half of the trials, no test stimulus was presented, and no response was required. In this example of a face-cued trial, the face remains the same while the scene changes; therefore, the correct response is "SAME." The face and scene stimuli are reproduced with permission from Ma et al. (2015) [23] and Konkle et al. (2010) [24]. Abbreviations: ITI, inter-trial interval. (B) A hypothetical representational space where the coding axes distinguishing faces from scenes are parallel across perception (orange arrow) and working memory (blue arrow). In this case, the angle between the coding axes is zero, leading to perfect generalization as indicated by cross-condition generalization performance (CCGP). However, linear, binary* exclusive-or *(XOR) decoding (i.e. binary classification of Scene in Working Memory & Face in Perception vs Scene in Perception & Face in Working Memory)*

would be impossible. (C) A case in which the coding axis rotates within the same subspace as perceptual representations transform into working memory. All four points corresponding to each condition still lie on the same plane (gray). In this case, the angle between coding axes is non-zero, CCGP decreases, and XOR decoding remains impossible. (D) A case where the coding axis rotates to define distinct subspaces for perception (orange plane) and working memory (blue plane). In this case, the angle between coding axes is non-zero, CCGP decreases, and XOR decoding is above chance. Note that here the axes represent three example effective axes within a higher-dimensional space.

We extracted vertex-level whole-brain BOLD activity from 22 brain regions-of-interest (ROIs) defined by the Glasser atlas [25] under four conditions: face in perception, scene in perception, face in working memory, and scene in working memory (Figure S2). These corresponded to activity when observers attended to the face or scene aspect of the sample and when they subsequently maintained the cued aspect in working memory during the delay. In a neural state space in which each dimension corresponds to a vertex, we defined two separate face–scene coding axes for perception and memory epochs, as the lines connecting the multivertex beta patterns for face- and scene-cued trials (face beta pattern − scene beta pattern) (Figure 1B).

Prior to assessing the rotation, we evaluated the robustness of face-vs-scene multivertex activation pattern within each ROI by applying a test-retest reliability approach (e.g., [26]). For each perception and working memory epoch, split-half correlations were computed on face–scene difference patterns. Despite substantial individual variability, we confirmed that nearly all ROIs showed significant reliability for both perception and working memory at the group level (Figure S3).

On this basis, we quantified the geometric relationship between perceptual and mnemonic representations by measuring the angle between the face-vs-scene coding axes (i.e., the axis linking the multivertex beta patterns for face- and scene-cued trials) of perception and working memory. Not only low- and high-level visual cortices but also parietal and frontal association cortices exhibited larger angles compared to the noise floor (the expected angle between parallel axes under noise) (Figure S4).

We calculated the z-values of the angle deviation from the noise floor to perform a formal statistical test. All but three ROIs (somatosensory/motor, paracentral/mid cingulate, and posterior opercular) exhibited significant rotation at the group-level, indicating that rotation from perceptual to mnemonic representation is prevalent among distributed brain regions, irrespective of processing hierarchy (Figure 2A).

Even at the individual level, we found a reliable rotation, especially in the visual cortex. The universality test [27] revealed that the significance of rotation in most

visual areas (early visual, dorsal/ventral stream visual, and MT+), except for primary visual cortex, was universal across observers. In contrast, in addition to the three non-significant ROIs at the group-level, we found the significance of the rotation in early auditory, auditory association, insular/frontoopercular, lateral temporal, anterior cingulate/medial prefrontal, and orbital/polar frontal areas was sporadic. The regions exhibiting reliable rotation were also consistent when tested using fine-grained parcellations (Figure 2B). Note that the lack or sporadicity of significance does not necessarily imply parallel axes; rather, visual tuning in these regions might be too noisy or effectively absent, thereby elevating the noise floor of the angle (Figure S4).

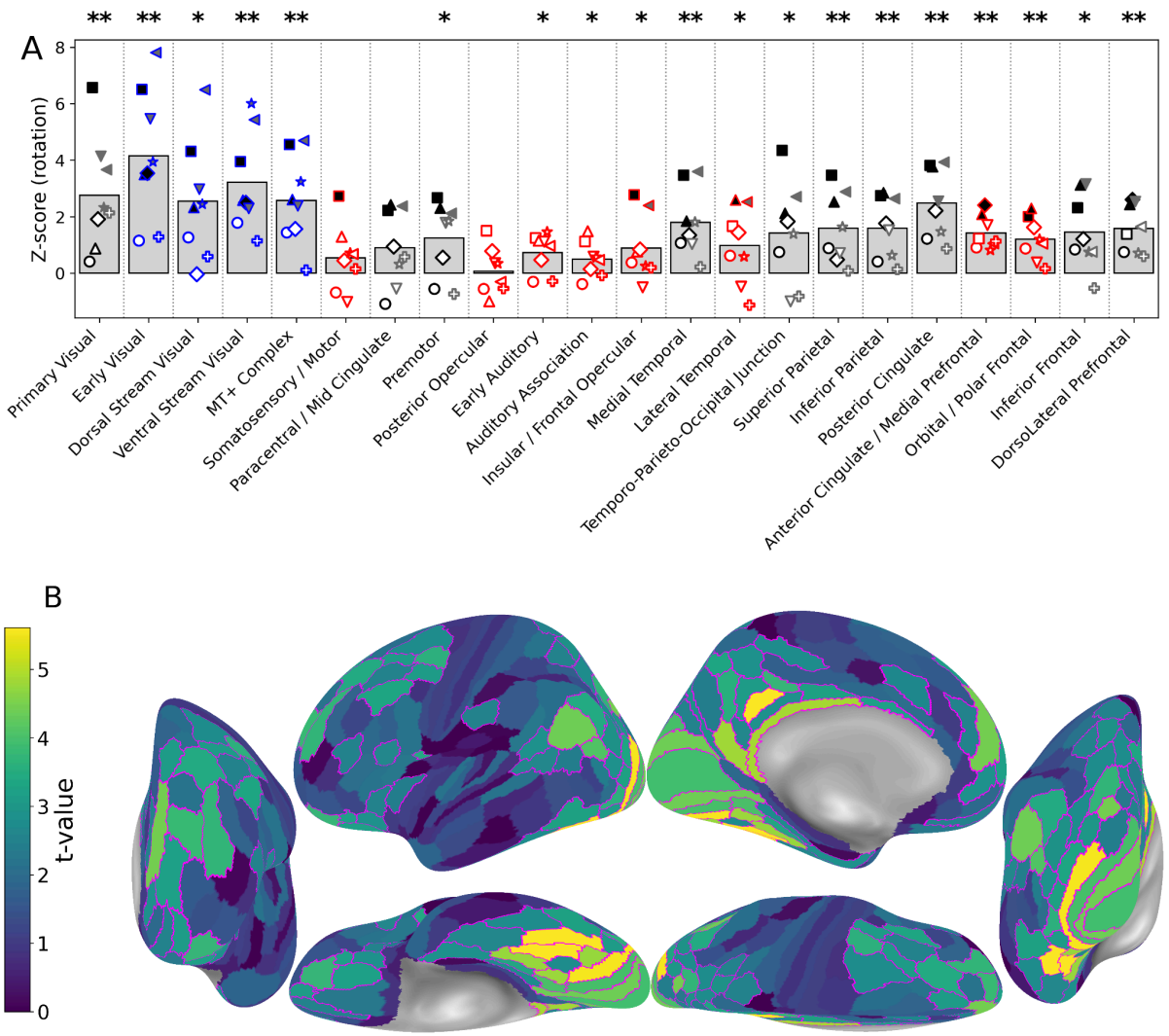

Figure 2. Rotation of the coding axis between perception and working memory. *(A) Z-values reflecting deviation from the noise floor (parallelism under noise; Figure S4) are shown for all observers and ROIs. Bar graphs indicate the inter-individual mean. Asterisks denote FDR-corrected group-level significance (**q < .01, *q < .05). Filled markers indicate observers who exhibit individual-level significance (FDR-corrected), whereas unfilled markers indicate those who do not. Black and gray markers indicate pilot and new observers, respectively. ROIs marked with blue and red indicate universality and sporadicity of significant individuals, respectively. In the majority of observers, the coding axis rotation across perceptual and mnemonic contents was evident in some visual and parietal/frontal association areas. (B) Fine-grained surface maps illustrating significant axis rotation. Group-level t-values are visualized on the Glasser atlas, with negative t-values clipped to zero owing to the one-tailed test. ROIs showing significant rotation (q < .05) are outlined with purple contours.*

Perception and working memory contents were coded in different subspaces

Theoretically, the rotation could occur within the same subspace (Figure 1C), or it could involve distinct subspaces, leading to an increase in the effective dimensionality of the representational space (Figure 1D). In order to distinguish these alternative scenarios, we evaluated the dimensionality by assessing the linear separability of variables. When multiple variables are represented, the greater the number of dichotomies that can be separated by linear hyperplanes, the higher the effective dimensionality [28]. Specifically, when rotation of a coding axis recruits an extra, new dimension (as described in Figure 1D), this expansion can facilitate flexible coding and thereby enable decoding of an exclusive-or (XOR) dichotomy [7]. In our case, where two factorial variables each have two levels (e.g., the combination of content: *face* vs. *scene* and cognitive process: *perception* vs. *working memory*), XOR corresponds to the dichotomy between *face in perception* and *scene in working memory* versus *face in working memory* and *scene in perception*.

The face vs. scene was decodable from all ROIs except the early auditory cortex (Figure 3; even separately within perception and working memory epochs, see Figure S5), indicating that visual content was represented across widespread cortical regions. From these content-sensitive regions, we were also able to decode the XOR dichotomy with the exception of the auditory association cortex. These regions substantially overlapped with those showing rotation of the face-vs-scene coding axes (Figure 2). This pattern indicates that the face-vs-scene coding axes do not merely rotate within a single representational space, but instead define at least partially distinct subspaces for perception and working memory. Such properties were evident across broad cortical regions, except in the auditory areas that showed low sensitivity to visual content in the first place.

For the sake of completeness, we confirmed that perception and working memory, irrespective of face or scene content, was robustly separable by linear decoders across all ROIs and observers (Figure S6), likely reflecting the presence versus absence of the stimulus in addition to the cognitive processes engaged.

Perception and working memory representations were generalizable across each other

Although rotation of the coding axis expanded the dimensionality, it was still possible that two coding schemes are not fully orthogonal and therefore remain generalizable to each other. To test this possibility, we used Cross-Condition Generalization Performance (CCGP) [7] to quantify how well the linear decoder trained in one epoch could generalize to the other epochs (from perception to working memory and vice

versa). Higher CCGP values reflect coding that is more invariant to cognitive processes engaged.

Among the ROIs from which face vs. scene was decodable, we observed consistently above-chance CCGP (except for the primary visual cortex). This indicates that most areas coded the content information in a partially invariant format that could be generalized across perception and working memory (Figure 3).

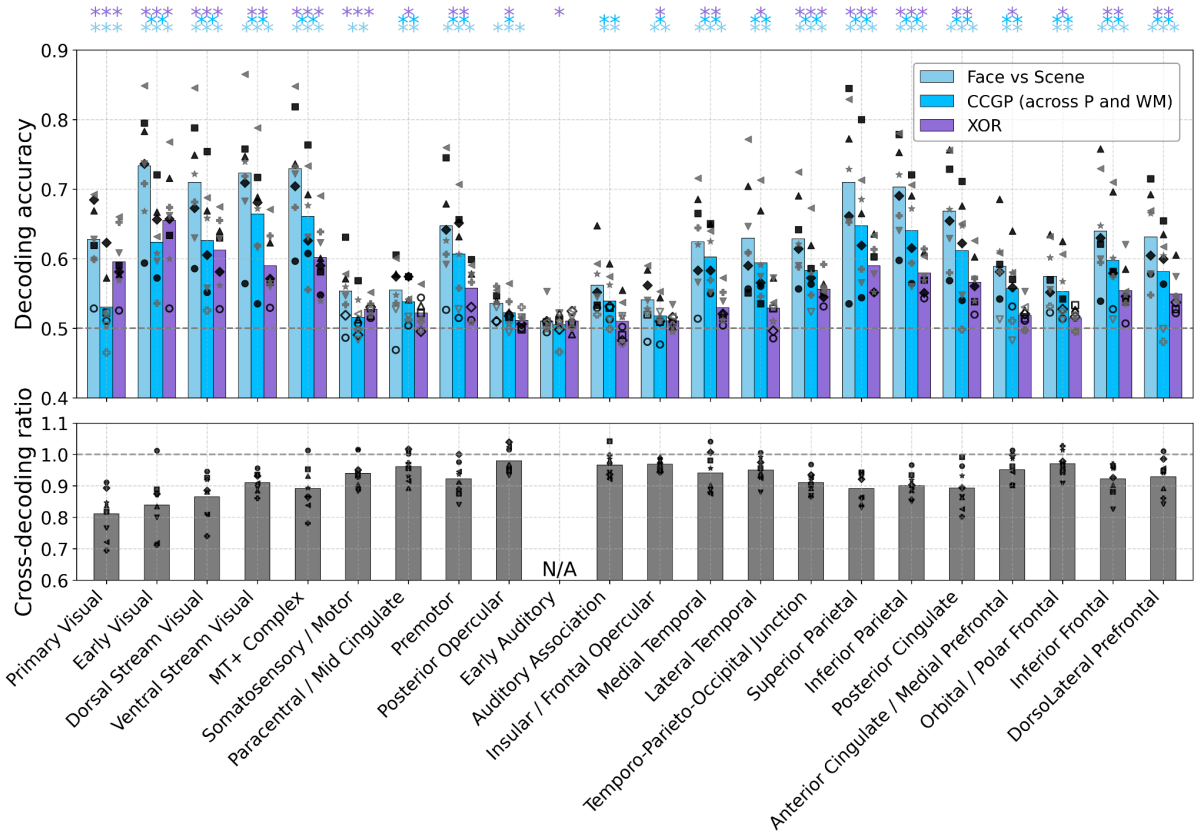

Figure 3. Face-vs-scene decoding, cross-condition generalization performance (CCGP) and exclusive-or (XOR) decoding. *(A) Decoding accuracy between face-cued and scene-cued trials, regardless of epochs (i.e., perception and working-memory), is shown in light blue, CCGP for face–scene dichotomy (trained to decode face vs. scene in one epoch and tested in the other) is shown in deep blue, and XOR (decoding face in perception and scene in working memory vs. scene in perception and face in working memory) is shown in purple. Individual observer data are plotted around the bars (interindividual mean) using markers of different shapes. Filled markers indicate individual-level significance assessed by permutation tests. Black and gray markers indicate pilot and new observers, respectively. Asterisks denote the significance of the group-level one-sample t-test (one-tailed) on z-transformed accuracy, FDR-corrected across all ROIs (**** q < .0001, *** q < .001, ** q < .01, * q < .05). The face-scene dichotomy was decodable except for the early auditory cortex and generalizable across epochs in all regions except for the primary visual cortex, somatosensory/motor cortex, and early auditory cortex. Therefore, visual content is represented in a manner generalizable across perception and*

working memory in all areas except for early sensory-motor areas. XOR dichotomy was decodable from all visual areas and most association areas (e.g., temporal, parietal, and prefrontal regions), indicating partially different subspaces for perceptual and mnemonic content. (B) Cross-decoding ratio for contents quantifies the drop in content-decoding accuracy when training and testing are performed across perception and working memory, relative to within-epoch decoding (see Figure S5). Bars and markers follow the same conventions as in (A); N/A indicates ROIs in which content was not decodable in the first place. When the runs with poor image quality were excluded, XOR and CCGP (across P and WM) in the insular/frontal opercular cortex became non-significant.

Representations were more abstract in the association cortex compared to the early sensory regions

The ROIs with greater XOR decoding tended to exhibit larger cross-decoding performance drops (Figure 3). Such a tradeoff between flexible and abstract coding is theoretically predicted [7]. Importantly, here we provide evidence for the co-occurrence of flexibility and abstraction in human fMRI data, as XOR decodability did not eliminate CCGP in most ROIs.

Interestingly, regions located at the sensory–motor periphery of the cortical hierarchy (i.e., primary/early visual and somatomotor cortices) tended to prioritize flexibility at the expense of abstraction (Figure 4A). In contrast, higher-tier visual regions (especially for the ventral visual stream) and association cortices (especially for the temporal cortex) showed the opposite pattern, exhibiting relatively greater abstraction. This organizational trend was particularly evident within the visual cortex, when the data were visualized on a fine-grained atlas (Figure 4B). Notably, this trend broadly aligned with the principal functional gradient from unimodal to transmodal cortex that has been identified in resting-state functional connectivity [29] and in multitask representational similarity [30]. However this relationship did not survive statistical testing when evaluated against a stringent null model that preserves spatial autocorrelation among parcels [30, 31] (Figure S7).

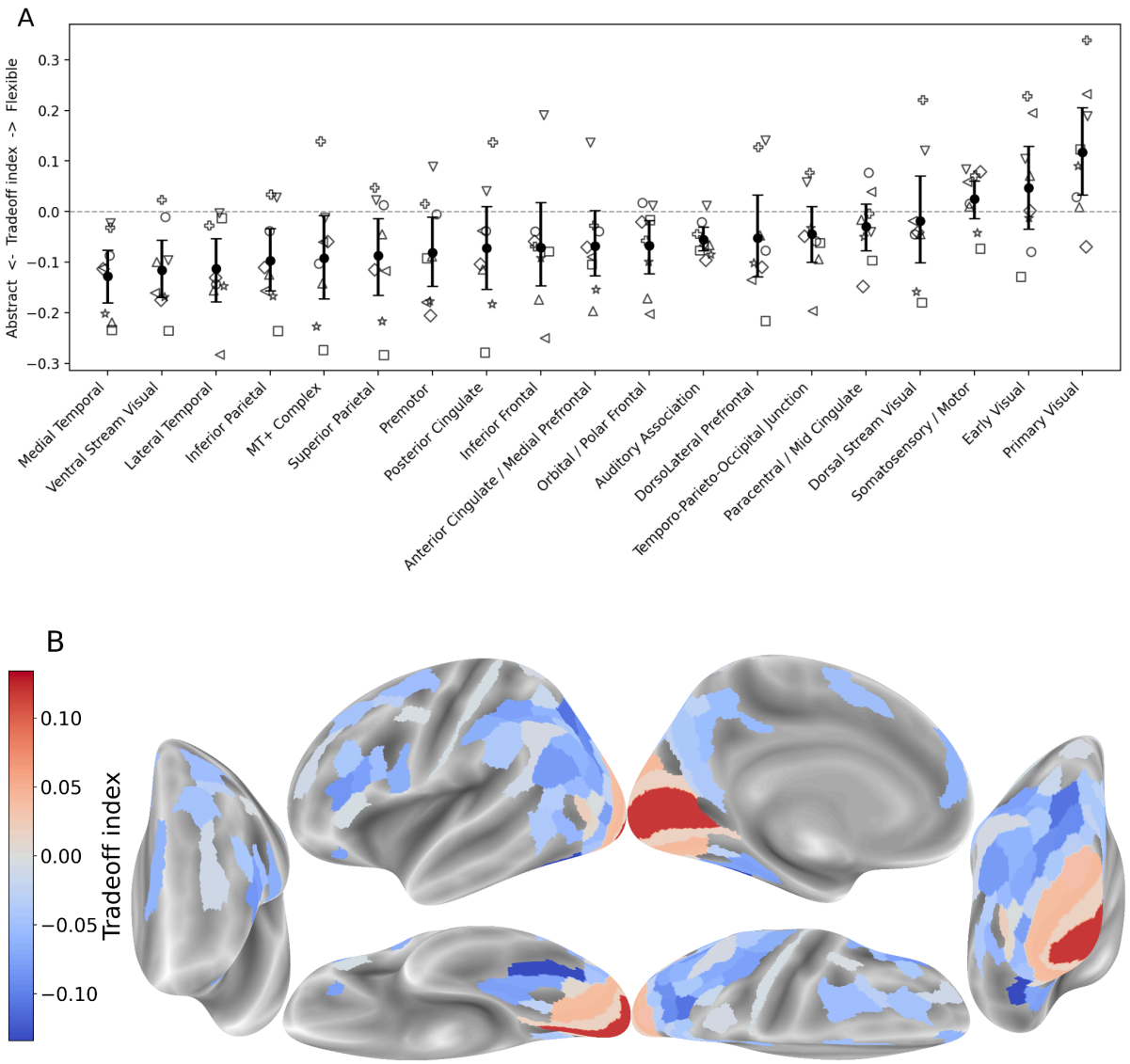

Figure 4. The level of tradeoff between abstract and flexible coding. *(A) Tradeoff index for each of the 22 coarse-grained ROIs (shown in ascending order). Data show the inter-observer mean with 95% bootstrapped confidence intervals, with individual observers plotted as distinct marker shapes. A positive index indicates that XOR decoding exceeds CCGP, reflecting a bias toward flexible coding at the cost of abstraction. (B) The tradeoff index derived for the fine-grained cortical atlas. ROIs with poor face-vs-scene decoding accuracy (q > .01) were excluded because they were not reliably tuned to the face–scene distinction. ROIs with poor CCGP and XOR performance (both q > .01) were also excluded because their tradeoff index could not be reliably estimated. Overall, unimodal regions (primary, early visual, and somatomotor areas) tended to favor flexibility by separating perceptual and mnemonic subspaces, whereas more transmodal regions exhibited a more balanced tradeoff or favored abstraction.*

Discussion

Using 7T fMRI to collect data from multiple sessions for each observer, we identified numerous brain regions that differentially encode faces and scenes during perception and working memory. This supports the notion that working memory contents are represented in many distributed brain regions [4, 32, 33], and allowed us to systematically compare the neural representations of perception and working memory across different levels of the processing hierarchy. Unlike the fusiform face area (FFA) and parahippocampal place area (PPA), which are established as face- and scene-selective regions [35, 36], some of the regions identified in the present study may not explicitly encode face or scene identity *per se*, but rather distinguish them through other visual features (e.g., color distribution in color-selective regions such as V4/VO). However, given that the task was to detect potential change in the stimuli after working memory delay, which could happen for any feature in ways that could not be anticipated *a priori*, it is unlikely that the working memory representations only reflected individual features in our experiment. It is also unlikely that the observed activity solely reflects processes driven by verbal cues or language-related factors, as we were able to decode scene (but not face) subcategories (Figure S8).

In previous studies, comparisons between the representations of perception and memory have primarily focused on the visual cortex [37–40]. Many of these studies were motivated by the sensory recruitment hypothesis, which posits that the early visual cortex plays a critical role in forming “sensory-like” working memory representations [34, 41–44]. Our CCGP results are broadly consistent with those prior findings showing cross-decoding between perception and working memory in the early visual cortex, except for V1, where we did not observe reliable CCGP. For instance, Harrison and Tong (2009) demonstrated that stimulus orientations maintained in working memory could be decoded from the early visual cortex, with partial generalization between perceptual and mnemonic representations [37]. Similarly, Albers et al. (2013) showed that decoders trained on perceptual trials could successfully decode not only working memory but also imagined orientations [38].

However, these earlier studies often diverted observers’ attention to secondary tasks (e.g., letter identification at fixation; see [37]) during perceptual conditions, making it possible that differences between perception and working memory were driven by attentional factors (for the exception, see [45]). Moreover, because only a single stimulus was presented in each perceptual trial, those studies may have captured stimulus-driven activity rather than neural processes directly associated with perceptual phenomenology. In contrast, in the present study, we manipulated perceptual conditions solely through endogenous attention while matching the visual input across conditions, thereby minimizing the difference of attentional state between perception and working memory epochs as well as stimulus-driven activity.

Interestingly, under this design, we observed a chance-level cross-decoding performance in the primary visual cortex (Figure 3).

Except for the primary visual and somatosensory/motor cortices, however, content representation was generalizable (i.e., CCGP was high) across perception and working memory. This suggests that visual contents are largely coded in a format that is somewhat invariant across perception and working memory, despite rotations in representational geometry. Such abstract representation was prevalent across the processing hierarchy with the exception of the primary visual cortex. Importantly, CCGP was assessed using perceptual and mnemonic axes derived from independent groups of runs, making it unlikely that the observed generalization reflects artifactual correlations arising from overlapping slow BOLD responses within the same run.

It is challenging to directly compare decoding performance across ROIs that differ in (raw and functional) SNR, microanatomy, functional topography, and neurovascular coupling (e.g., visual cortex vs. prefrontal cortex; see [46–48]). This may partly explain why many previous investigations using human fMRI have been restricted to the visual cortex [37–40], where regions share relatively similar properties compared to association cortices. For instance, a decrease in cross-decoding performance does not necessarily indicate that working memory relies on a qualitatively different representational format from perception; rather, it may simply reflect noisier visual tuning or region-specific limitations inherent to fMRI-based multi-voxel pattern analysis.

However, quantifying representational geometry can perhaps help address this issue by assessing *how* visual information is represented instead of how much could be read out. According to the “geometry of abstraction” framework [7], assessing CCGP (bidirectional cross-decoding) in conjunction with XOR decodability provides a way to capture differences in coding strategies across brain regions and to infer their underlying computational principles. This framework was initially applied in the context of associative learning [7], where abstract representations of task variables—stimulus, response, and reward—allow the brain to treat the same reward as equivalent even when it is associated with different stimuli or actions. Bernardi et al. (2020) further demonstrated that better task performance was associated with higher CCGP values [7]. In contrast, representing variables in a high-dimensional geometry, which increases the number of linearly separable dichotomies, enhances information capacity and supports episodic memory formation [17] as well as flexible cognitive control [16], and even supports illusory perceptual categories that are not linearly separable in the underlying stimulus space [49]. Thus, abstract and high-dimensional (i.e. flexible) representations may serve distinct yet complementary functional roles.

In the context of perception and working memory, abstracting their representations may facilitate detecting matches or mismatches between internal templates and

external stimuli, as in visual search or change detection tasks (as in our experiment). By contrast, it is not obvious how the decodability of XOR dichotomy (e.g., combining perceptual face and memory scene) would be helpful for human behavior. Rather, representing perceptual and working memory contents in distinct subspaces may protect the information held in working memory from interference by newly incoming perceptual inputs [12]. If this is the case, our findings suggest that such subspace rotation may constitute a default coding strategy of working memory, occurring even in the absence of distractors (i.e., newly incoming perceptual stimuli).

Our XOR decoding results demonstrated that, indeed, most regions distinguishing faces from scenes—including not only visual cortex but also temporal, parietal, and prefrontal association areas—define partially distinct subspaces as perceptual content transforms into working memory. XOR decoding was possible in largely the same regions where the perceptual and mnemonic coding axes deviated from parallel. This suggests that coding axes rotation was generally accompanied by an expansion of representational dimensionality.

Notably, by quantifying the extent to which a given ROI sacrifices generalization for flexibility, we observed a suggestive trend along the unimodal–transmodal axis: unimodal regions tended to prioritize flexibility (manifested as strong subspace rotation) more at the cost of generalization, whereas this tradeoff became progressively favored generalization toward the transmodal end of the gradient. Importantly, this pattern extends prior gradient-based accounts of cortical organization [29, 30, 50] by embedding them within a representational geometry framework. From this perspective, the unimodal–transmodal gradient may be reframed as a macroscopic expression of a more general organizational principle governing how neural systems allocate representational resources to balance context-dependent and context-invariant coding in a graded manner.

Stronger subspace rotation in the early visual cortex suggests that these regions employ qualitatively different formats to represent perception and working memory [3, 41, 51, 52]. Nevertheless, this does not preclude their contribution to working memory functions; rather, these regions may help protect mnemonic information from interference by ongoing perceptual input [39] (but see [53]). Such segregation may be implemented through layer-specific activity, with bottom-up input primarily targeting the middle layers and top-down signals conveyed through deep and superficial layers [54–56]. While earlier studies reported above-chance cross-decoding [37, 38] in the early visual cortex, more recent studies have emphasized the reduction in cross-decoding performance [40, 57] arguing that working memory may be encoded in a distinct format from perception in these regions. Our findings are consistent with this, and further confirmed that working memory is not simply a weaker/noisier version of perception in the early visual cortex, by directly measuring the geometry angle and assessing XOR decoding in addition to cross-decoding.

Compared with the early visual cortex, which favored flexibility, the association cortex favored more abstract representations. Consistent with this result, Xu (2024) reported that working-memory representations in the posterior parietal cortex are more resistant to distraction than those in the visual cortex, maintaining the target coding axis parallel across different distractors [58]. Taken together, abstract representation may be a general coding strategy at least in part of the parietal cortex, holding not only when observers encounter different distractors but also when certain perceptual information enters working memory in the absence of distractors.

To the best of our knowledge, this study provides the first direct comparison of neural representations during visual perception and working memory in the human prefrontal cortex. Particularly in the *lateral* prefrontal regions, beyond their visual tuning to faces and scenes [59–61], we show that the content is represented in partially distinct subspaces yet in a format that generalizes across perception and memory. These mesoscopic characteristics can arise from intermixed populations of neurons—some tuned selectively to perceptual or mnemonic content and others tuned to both—as reported in homologous monkey regions [22].

Closely related to our question, Dijkstra et al. (2025) investigated how the human brain distinguishes or confuses visual perception and visual imagery [62]—a form of internally generated representation that is conceptually and neurally similar to visual working memory [37, 63, 64]. They showed that the behavior and neural responses in fusiform regions were best explained by a model in which congruent imagery and perception summate along a single dimension, such that excessively vivid imagery can be subjectively experienced as *real* even in the absence of sensory input.

By contrast, working memory representations are unlikely to be mistaken for real (i.e., veridical perception) on their own. Rather, they are more plausibly expressed as contaminating or biasing concurrent perception of external stimuli [65, 66], making it more necessary for the brain to keep them separable. Although it remains unclear which brain regions are fundamentally related to subjective experience, we consider that subspace rotation may be sufficient to reduce subjective confusion between perceptual and mnemonic contents [1, 3].

It is noteworthy that one limitation of our study is that we did not directly establish the link between subspace rotation and behavior. Thus, it may be informative for future studies to titrate task difficulty (e.g., through parametric manipulation of synthesized naturalistic stimuli; see [40]) to yield a sufficient number of incorrect trials. This would perhaps enable trial-by-trial prediction of response accuracy based on the degree of subspace rotation in specific brain regions. Alternatively, future work could employ behavioral paradigms that introduce distractors during the delay period and also allow for higher-resolution reports, such as continuous reproduction tasks [66]. Such designs may further allow formalization of models of subjective experience—for example, how working memory representations are contaminated by concurrent

perceptual input—based on the geometric relationship between working memory and perceptual representations.

Methods

Pre-registration

Unless otherwise noted, all analyses and experimental procedures described below were conducted in accordance with a pre-registered plan (pre-registered after collecting data from four pilot observers; see https://doi.org/10.1101/2025.09.07.674590).

Observers

Eight naïve observers and one author (four males and five females; aged 21–28 years) participated in our task. We excluded one observer's data from the analyses, as they were able to complete only one out of four scanning sessions. Except for that observer, all the other eight observers completed four scan sessions, on different days.

All observers reported that they have normal or corrected-to-normal visual acuity, normal color vision and no history of neuropsychiatric disorders. The study was approved by the Institutional Review Board at Sungkyunkwan University. We obtained written informed consent from all observers before the experiment.

As pre-registered, observer recruitment was terminated based on the universality of individual-level significant effects [27]. After collecting data from eight observers (the pre-registered minimum sample size), we met the pre-registered stopping criterion: for each of the three measures of interest (coding axis rotation, CCGP for face–scene dichotomy, and XOR decoding), more than half of the 22 ROIs reached the threshold for sufficient universality (likelihood ratio > 50 following [27]), regardless of whether the specific ROIs overlap across measures. Assuming a reasonable sensitivity of $1 - \beta = .9$, this corresponded to finding at least 5 out of 8 observers showing significant effects or finding at least 6 out of 8 observers showing non-significant effects (with $\alpha = .05$, FDR-corrected across ROIs).

Apparatus

The experimental code was written in MATLAB R2021b (MathWorks, Natick, MA, USA) with the Psychphysics Toolbox 3.0.19 [67, 68] and was run on a Ubuntu-based laptop. Visual stimuli were displayed on a screen behind the scanner and observed by observers through a mirror attached to the head coil. The refresh rate and spatial resolution of the monitor was 60 Hz and 1600 × 1000 pixels, respectively. The effective viewing distance was 104 cm.

Stimuli

In order to focus on the specific cognitive processes (working memory vs. perception) rather than simply the physical stimuli themselves, we presented identical sets of face–scene blended sample images and manipulated only the aspect (face vs. scene) that was attended and remembered. In this way, the contrast between face-cued and scene-cued trials—which is our main focus—reflects primarily the difference in perceptual or mnemonic representation when the same stimulus is presented.

To create the set of face-scene blended sample images, first we selected 32 different stimuli (4 sub-categories × 8 exemplars) each for faces and scenes. These numbers were chosen to balance the diversity of the stimuli with the number of repetitions of the same stimuli. Diversity was expected to help extract the general representation of face/scene and minimize the familiarity effect that can confound the comparison between perception and working memory. In contrast, repetition was expected to improve data reliability.

Face stimuli were drawn from the Chicago Face Database [23], comprising 4 sub-categories defined by the combination of race and sex (Black/Female, Black/Male, White/Female, and White/Male), with all faces displaying neutral expressions. Scene stimuli were drawn from the “Massive Memory” dataset [24], comprising 4 sub-categories defined by crossing natural vs. manmade with open vs. closed (Beach, Cavern, Street, and Livingroom). Within each face/scene sub-category, we selected 8 exemplars as dissimilar as possible to each other. Face dissimilarity was determined using the embeddings from the penultimate layer of the Dlib face model [69, 70], which were further transformed to better reflect how humans perceive similarity between faces [71]. We applied k-means clustering to the embeddings of all faces for each face sub-category in the dataset. We then created eight clusters, and selected one representative exemplar from each cluster, whose embedding was closest to the gravitational center of the cluster, resulting in the selection of eight dissimilar representative exemplars within the sub-category. Scene dissimilarity was determined in a similar way but by using the embeddings from the output layer of GoogleNet pre-trained with Places365 datasets [72] (“imagePretrainedNetwork” function in MATLAB). After selecting all 32 stimuli, we scaled the lightness (Value in the HSV color space) separately for the face and scene stimulus sets so that the mean lightness was identical within each set. The face stimuli were blended and resized so that they fit within the boundaries of the scene stimuli (7.8 × 7.8 degrees of visual angle).

Since it was not feasible to present all 32 × 32 combinations of face and scene exemplars within a single scanning session, we selected 128 balanced subsets of combinations. Each combination was presented only once per session (128 trials in total in a session). Specifically, each face exemplar appeared four times per session, each time paired with a different scene exemplar from a different sub-category, and

vice versa for scene exemplars. Half of the 128 combinations were presented in face-cued trials, and the other half in scene-cued trials. In the second session, the pairing between cues and stimuli was reversed, with all combinations presented again but in a new, randomized order. The third and fourth sessions repeated the same stimulus combinations and cues as the first and second sessions, respectively, but with the order of trials randomized again. Each scanning run showcased 16 out of 128 combinations, including all combinations of face and scene sub-categories (4 face sub-categories × 4 scene sub-categories = 16 trials).

To minimize eye movements in the scanner, we asked observers to focus their gaze on a fixation point [73], displayed at the center of the sample/test (i.e., face–scene blended stimuli).

Task and procedure

Observers performed a change detection task of either face or scene in the MRI scanner. At the beginning of each trial, a 100% predictive word cue (either “FACE” or “SCENE”) was presented in Arial font for 1s to indicate whether to remember a face or a scene in that trial (Figure 1A). Following that cue, a face-scene blended stimulus was presented as a “sample” stimulus, flashing on (0.8 s) and off (0.2 s) for a total of 10 cycles. The final (10th) blank was replaced with a noise mask to signal the end of the perception epoch, followed by 10s delay period (working memory epoch), during which only the fixation point was shown. The working memory epoch was followed by the appearance of another face-scene blended image as a test stimulus, the duration of which was 2s. Observers judged whether the cued aspect (i.e., either a face or scene) of the test stimulus was the same or different from that of the sample stimulus. Thus, the dimension along which the attended aspect was manipulated and the dimension of the behavioral response were orthogonal by design, ensuring that the face-vs-scene coding axis was free from confounds related to decision- or response-related activity. To reduce regressors’ collinearity, the test image was omitted in a random half of the trials, in which no behavioral response was required. The contrast polarity reversal of the fixation point signalled the end of the trial, followed by an inter-trial interval (ITI), which was randomly chosen from 4, 5, 6, and 7s. Each functional run consists of 16 trials, with face-cued and scene-cued trials performed in equal numbers (8 trials each) and randomized order.

Out of eight trials requiring responses in each run, in two trials the face aspect changed, in two the scene aspect changed, in two both changed, and in two neither changed between the sample and test; within each type, one was face-cued and one scene-cued trial. In trials where they were intended to change, the test image was randomly selected from the *same* sub-category but from the entire dataset (rather than from the only eight exemplars, which served as the sample stimuli), to make the task moderately engaging. This, thus, helped to encourage observers to retain the entire image in working memory to do the task rather than relying on labels or verbal

descriptions of its features [45]. This control was important because we were interested in the phenomenological difference between perception and working memory.

MRI data acquisition

Functional images were acquired using a T2*-weighted gradient-echo echo-planar imaging (EPI) sequence on a 7T Siemens MAGNETOM Terra scanner with a 32-channel head coil at the Center for Neuroscience Imaging Research. Acquisition parameters were as follows: repetition time (TR) = 1400 ms; echo time (TE) = 21.0 ms; flip angle = 61°; voxel size = 1.5 × 1.5 × 1.5 mm^3; field of view (FOV) = 210 × 210 mm^2; 92 axial slices; phase encoding direction = posterior to anterior; multiband acceleration factor = 4; GRAPPA acceleration factor = 2; phase partial Fourier = 6/8; bandwidth = 1984 Hz/pixel. Each functional run consisted of 319 volumes (including initial 5 volumes and final 10 volumes without any events), acquired in an interleaved slice acquisition order. Slice orientation was set obliquely to the AC-PC line (approximately -15° to axial). Each observer performed 8 functional runs in each of 4 scanning sessions, a total of 32 runs.

To correct for susceptibility-induced geometric distortions in the functional data, a pair of gradient-echo EPI volumes with reversed phase-encoding directions, one with anterior-to-posterior and one with posterior-to-anterior, was acquired before every functional run. These fieldmap images were acquired with the same pulse sequence as the functional scans. Each had 4 volumes acquired in interleaved slice order.

Structural T1-weighted images were acquired at the beginning of the first session (and additionally at the beginning of the third session for the last four observers). A 3D Magnetization-Prepared Two Rapid Acquisition Gradient Echoes (MP2RAGE) sequence was used with the following parameters: TR = 4500 ms; TE = 2.24 ms; inversion times (TI1/TI2) = 1000/3200 ms; flip angles = 4°/5°; voxel size = 0.7 × 0.7 × 0.7 mm^3; FOV = 224 × 224 mm^2; 320 sagittal slices; phase encoding direction = anterior to posterior; bandwidth = 200 Hz/pixel; acquisition time = 306 s. Automatic background denoising was applied, and the resulting image (UNI-DEN) was used as the T1w anatomical input for subsequent preprocessing in fMRIPrep.

Preprocessing

MRI data were preprocessed using fMRIPrep 23.2.1 [74, 75], which is built on Nipype 1.8.6 [76, 77]. Many internal operations, particularly within the functional processing workflow, rely on Nilearn 0.10.2 [78].

Fieldmaps

B0 nonuniformity maps (fieldmaps) were estimated from EPI references with opposite phase-encoding directions using FSL TOPUP v2203.2 [79] and applied to correct susceptibility distortions for each following functional run. For Obs. 2, due to a technical issue, the posterior-to-anterior reference for run 3 in session 1 was unavailable; in this case, the posterior-to-anterior reference from run 4 was used instead.

Anatomical data

Each T1-weighted (T1w) image was corrected for intensity non-uniformity using N4BiasFieldCorrection (ANTs 2.5.0; [80, 81]) and skull-stripped using a Nipype implementation of the ANTs antsBrainExtraction.sh workflow with the OASIS30ANTs template. Brain tissue segmentation into cerebrospinal fluid (CSF), white matter (WM), and gray matter (GM) was performed on the brain-extracted T1w image using FSL FAST v2111.3 [82]. For observers with two T1w images, an anatomical T1w reference was generated by registering the INU-corrected images using mri_robust_template (FreeSurfer 7.3.2; [83]). Cortical surface reconstruction was performed with recon-all (FreeSurfer 7.3.2; [84]), and the brain mask was refined using a custom procedure based on Mindboggle to reconcile ANTs- and FreeSurfer-derived cortical segmentations [85]. Volume-based spatial normalization to MNI152NLin2009cAsym was performed using nonlinear registration with antsRegistration (ANTs 2.5.0), aligning brain-extracted T1w images to the ICBM 152 nonlinear asymmetric template version 2009c [86] accessed via TemplateFlow (v23.1.0) [87].

For one observer (Obs. 8), skull stripping failed within fMRIPrep; therefore, cortical surfaces were reconstructed using FreeSurfer after manual skull stripping of the T1w image prior to fMRIPrep processing.

Functional data

For each BOLD run, the following preprocessing was performed using fMRIPrep. A reference volume was generated for motion correction, and head-motion parameters were estimated relative to this reference using FSL-MCFLIRT v2111.0 [88]. Susceptibility distortion correction was applied by aligning the estimated fieldmap to the EPI reference, which was subsequently co-registered to the T1w reference using boundary-based registration with six degrees of freedom (FreeSurfer bbregister; [89]). Framewise displacement was calculated for each functional run using both the formulations of Power et al. (2014) [90] and Jenkinson et al. (2002) [88], as implemented in Nipype.

Finally, BOLD time series were resampled onto the fsaverage surface (FreeSurfer format). All resamplings were performed in a single interpolation step by composing

motion correction, susceptibility distortion correction, and corregistration. Volumetric resampling used cubic B-spline interpolation, and surface resampling was performed using FreeSurfer's mri_vol2surf.

Quality control

We flagged two functional runs with poor preprocessed image quality. One deviated from the other runs of the same observer, possibly due to co-registration failure (Obs. 2, Session 1, Run 3), and the other exhibited severe striping artifacts in the inferior brain regions, likely caused by head motion during shimming (Obs. 5, Session 2, Run 4). Importantly, we confirmed that excluding these runs did not alter the subsequent group-level statistical results except where noted.

General linear model

We fitted general linear models (GLMs) to the BOLD time courses per each vertex in the fsaverage space using the run_glm function of Nilearn (version 0.12.0). For each run, the perception and working memory epochs were modeled with four separate regressors corresponding to face in perception (FP), scene in perception (SP), face in working memory (FWM), and scene in working memory (SWM) conditions, each of which was the 10-s box-car function convolved with the canonical SPM hemodynamic response function (HRF) (Figure S9). Note that the perception epoch included cue and mask in addition to the stimulus period (i.e., sample/blank) while the working memory epoch just included the 10s delay period. Responses to test images (appeared only on half of the trials) were modeled with another single regressor, irrespective of condition. Six motion parameters and polynomial drift terms up to the third order were included as nuisance regressors. When framewise displacement exceeded 0.5 mm, the signals for that TR were interpolated using neighboring TRs. This "runwise" GLM produced a beta matrix of size 4 (FP, SP, FWM, and SWM conditions) × N (number of vertices) per run. For each run, the beta matrix was z-standardized within ROI by subtracting its mean and dividing by its standard deviation.

To increase the amount of training data for subsequent decoding analyses, we also estimated "trialwise" beta values using GLMsingle [91]. First, we upsampled the BOLD time series to a TR of 1 s [92], ensuring that event durations were integer multiples of the TR. Each event (FP, SP, FWM, SWM, and the test stimulus) was then modeled with a sequence of consecutive 1-s regressors spanning the entire event duration. GLMsingle selected the best-fit HRF for each vertex from a predefined library, identified nuisance regressors from pools of noise-dominant vertices, and finally applied ridge regression to stabilize single-trial beta estimates, with the degree of regularization determined by cross-validation [91]. Beta estimates were first obtained separately for each 1-s regressor and were then averaged across time points to yield a single beta for each event. This "trialwise" GLM produced a

beta matrix of size 2 (perception and working memory epochs) × N (number of vertices) per trial. Since one run has 16 trials, we obtained a beta matrix of size 32 × N per run, which was z-standardized within ROI prior to decoding analyses.

Definition of ROIs

We examined the entire cerebral cortex, divided into 22 anatomically defined regions of interest (ROIs), which are combined subsets of the Human Connectome Project-MultiModal Parcellation (HCP-MMP) atlas exactly following the grouping strategy in Glasser et al. (2016) [25] (Figure S2). In addition, we also used a more fine-grained definition of ROIs, namely the original 180-region HCP-MMP atlas, for visualization purposes and an exploratory analysis (see *Tradeoff between abstraction and flexibility* section). We analyzed homologous regions in the left and right hemispheres as a single ROI.

Rotation angle of the coding axis

To examine the rotation from perceptual to mnemonic representations, we measured the angle between their corresponding face–scene coding axes (defined by runwise GLM beta patterns for face-cued vs. scene-cued trials). There were a total of 32 runs, each resulting in a 1×N (vertices) face-scene coding axis for the perception and another for the working-memory epoch within each ROI. We averaged the coding axes across runs and computed the angle between the coding axes of the perception and working memory epochs:

$$\theta_{p,\,wm} = arccos\frac{V_p \cdot V_{wm}}{||V_p||\,||V_{wm}||}\ [1],$$

where V_p and V_{wm} denote the run-averaged face-vs-scene coding axes for perception and working memory, respectively. To obtain the distribution of $\theta_{p,\,wm}$, we used bootstrap resampling. For each ROI, starting from the 32 (number of runs) × *N* matrix for runwise coding axes, we generated 10,000 bootstrap replicates by resampling new 32 rows with replacement and averaging across the resampled rows, thereby yielding one bootstrapped estimate of the coding axis each for perception and working memory per replicate. We then calculated the angle between the two axes for every replicate.

The angle between two parallel axes can even appear as non-zero due to noise, with larger noise leading to larger deviations from zero. Thus, to evaluate whether two axes are truly rotated (i.e., non-parallel), we compared their observed angle (i.e., the distribution obtained above) to the angle expected to arise purely from noise, even if the axes are parallel (i.e., null hypothesis of parallelism). This expected rotation caused purely by noise (hereafter referred to as noise floor) follows Equation 2 (see Supplementary Information for the derivation in detail):

$$\theta_{noise-floor} = arccos\left(\frac{1+\frac{\rho_{noise}}{\sqrt{SNRp\ SNRwm}}}{\sqrt{1+\frac{1}{SNRp}}\sqrt{1+\frac{1}{SNRwm}}}\right) [2],$$

$SNRp$ and $SNRwm$ are the signal-to-noise ratios in the coding axis of the perception and working memory epochs, respectively. Lower SNR (i.e., greater noise levels) raises the noise floor of the angle, as indicated by the equation above. This means that even if the axes are parallel, for larger noise levels, the observed angle could deviate further from 0. ρ_{noise} indicates the cross-noise correlation between perception and working memory coding axes. Given that perception and working memory epochs both occurred on the same trials (see Figure 1A), we considered the possibility that their noise components might be correlated. Correlated noise could make the axes appear more parallel, thus it would lower the noise floor, as indicated by the equation above.

The SNR terms are estimated as follows:

$$SNR_p = \frac{||V_p||^2}{\sigma_p^2(noise)} [3],$$

$$SNR_{wm} = \frac{||V_{wm}||^2}{\sigma_{wm}^2(noise)} [4],$$

where $||V_p||^2$ and $||V_{wm}||^2$ are the true squared L2 norms of the perception and working memory axes, respectively, approximated as follows from the observations.

$$||V_p||^2 = ||V_p(observed)||^2 - \sigma_p^2(noise) [5],$$

$$||V_{wm}||^2 = ||V_{wm}(observed)||^2 - \sigma_{wm}^2(noise) [6],$$

$V_p(observed)$ and $V_{wm}(observed)$ indicate the observed perception and working memory coding axes (averaged across runs), respectively. $\sigma_p^2(noise)$ and $\sigma_{wm}^2(noise)$ denote the noise variances. Noise variances were estimated via the bootstrap resampling described above. For each dimension (vertex), we computed the variance across bootstrap replicates and then summed these variances across dimensions to obtain the total noise variance.

Note that when noise is overwhelming, the above estimations of $||V||^2$ could become negative, thus resulting in a negative SNR. To avoid introducing negative SNRs in Equation 3, which would make the cosine undefined, we clipped $||V||^2$ to 0. This

results in a noise floor of 90 degrees, suggesting that in such a noise level, even parallel axes are expected to appear orthogonal on average (in this clipped case, if no true rotation exists, the actual angle would still be distributed around 90 degrees, so the z-value would be around zero).

The cross-noise correlation is defined as follows:

$$\rho_{noise} = \frac{\sigma_{p,\,wm}(noise)}{\sigma_{p}(noise)\,\sigma_{wm}(noise)} \text{ [7]},$$

where $\sigma_{p,\,wm}(noise)$ is the cross-noise covariance estimated as the covariance across bootstrap replicates within each dimension and then summed over dimensions. It is noteworthy that cross-noise correlation was generally small, so this term had little impact on the noise floor.

Finally, z-score of *rotation* was obtained per ROI and per observer by comparing the distribution of $\theta_{p,\,wm}$ with $\theta_{noise-floor}$ which, as discussed above, is the estimated mean of the null hypothesis (i.e., observed angle even if the axes are parallel):

$$Z_{rotation} = \frac{mean\,(\theta_{p,\,wm}) - \theta_{noise-floor}}{sd\,(\theta_{p,\,wm})} \text{ [8]}$$

Multivariate pattern decoding

Decoding perceptual and mnemonic contents

To identify areas encoding information about perceptual and mnemonic content, we first trained linear decoders (logistic regression models in Scikit-learn) to classify contents (face vs. scene) irrespective of epochs (perception vs. working memory). We also evaluated content decoding accuracy by training decoders separately within the perception epoch and within the working memory epoch.

Decoding epochs

Although this was not our primary focus, for the sake of completeness, we also assessed the ability to decode perception vs. working memory. This was examined across all trials regardless of content (face vs. scene).

Exclusive-or (XOR) decoding

More importantly, we also trained decoders to classify exclusive combinations (perceived face and memorized scene vs. perceived scene and memorized face). Successful classification in this so-called exclusive-or (XOR) decoding would indicate that the two factorized variables (epochs and content) are entangled within a stable three-dimensional representational subspace in that ROI [7].

Cross-Condition Generalization Performance (CCGP)

To assess the generalizability of a hyperplane distinguishing face–scene dichotomy across epochs, we also measured cross-condition generalization performance (CCGP; see [7, 17]). In the case of two-by-two factorial design, we defined *CCGP for content* as the average of cross-decoding accuracy in both directions (trained a classifier to distinguish face vs. scene only in perception epoch and tested it in working memory epoch and vice versa). For the sake of completeness, we also measured the *CCGP for epoch* (trained a classifier to distinguish perception vs. working memory only in face-cued trials and tested it in scene-cued trials and vice versa). In addition, we assessed the face-vs-scene cross-decoding ratio, i.e., content CCGP divided by the mean of content decoding accuracies within perception epoch and within working-memory epoch, to show the degree of generalization of content coding [50].

Decoding sub-categories

With regard to content (face vs. scene) decoding in our design, we cannot fully rule out the possibility that the decoders are simply distinguishing the word cue (or associated language processing). Additionally, content decoding might reflect differences in the spatial extent of the attentional window when observers attended to faces versus scenes [93]. Therefore, more compelling evidence that an ROI encodes information about specific sub-categories—beyond just differentiating face versus scene—would strengthen the interpretation that we are capturing differences in visual phenomenology. For this reason, we identified areas in which sub-category information could be decoded. Specifically, we trained classifiers to distinguish among four face sub-categories within face-cued trials and four scene sub-categories within scene-cued trials, regardless of whether they were perceived or held in working memory. Furthermore, we also assessed decoding of specific sub-category dimensions (black vs. white and female vs. male for face sub-categories; open vs. closed and natural vs. manmade for scene sub-categories).

Decoding analysis procedures

All the abovementioned decoding accuracies were assessed using 8-fold cross-validation on trial-wise GLM beta patterns, z-standardized within each run. To prevent information leakage, all data from the same run was assigned to the same fold (using GroupKFold in Scikit-learn), which also helped balance conditions across folds. For example, to examine CCGP, decoders were trained to classify face vs scene using data from one epoch (e.g., perception) in seven folds and tested on data from the other epoch (e.g., working memory) in the held-out fold. For each iteration of cross-validation, PCA was applied to the training data (retaining components that explain 95% of the variance), followed by training a logistic regression model (using Scikit-learn library). Test data was projected onto the same PC space, and the model prediction accuracy was calculated. A null distribution was generated by permuting

condition labels within each run and repeating the entire cross-validation procedure 1,000 times [94].

Tradeoff between abstraction and flexibility

Theoretically, in factorized low-dimensional geometries, CCGP approaches perfect, while XOR performance remains at chance (Figure 1B). Conversely, in entangled high-dimensional geometries, XOR performance exceeds chance, but CCGP decreases (Figure 1D). This pattern reflects a tradeoff between flexibility (indicated by above-chance XOR decodability) and abstraction (indicated by above-chance CCGP) in general information-coding principles. In an exploratory analysis, we quantified this tradeoff per ROI as:

$$Trade\ off\ index = log(XOR\ /\ CCGP)\ [9],$$

where the logarithmic transformation ensures that the metric is symmetric around a balanced tradeoff (index = 0). We used this index as a convenient summary of the flexibility–abstraction tradeoff, while noting that an alternative, potentially fairer measure was explored in simulations and yielded similar results overall (see Supplementary Information and Figures S10, S11, S12, and S13).

Statistical analyses

We performed group-level statistical tests on all eight observers, including pilot data. For the angle rotation, group-level significance was evaluated using one-tailed one-sample t-tests (against zero) on the individual z-values (see *Rotation angle of the coding axis* section). For the decoding analysis, we similarly reported the results of one-tailed one-sample t-tests (against zero) performed on the individual z-values of decoding performances. We also note that the significance patterns were unchanged when logit-transformed decoding accuracies were tested instead. All p-values were corrected for multiple comparisons across all ROIs using the false discovery rate (FDR) [95].

We also report whether significance at the individual level was universal [27]. Based on the pre-registered sensitivity of tests ($1 - \beta = .9$), we considered the significance as universal if at least 5 (out of 8) observers showed significance at the individual level, whereas we considered it as sporadic if at least 6 (out of 8) observers showed non-significance.

References

1. Lau, H. (2019). Consciousness, metacognition, & perceptual reality monitoring. Preprint at PsyArXiv. https://doi.org/10.31234/osf.io/ckbyf
2. Dijkstra, N., Kok, P., & Fleming, S. M. (2022). Perceptual reality monitoring: Neural mechanisms dissociating imagination from reality. *Neuroscience & Biobehavioral Reviews, 135*, 104557. https://doi.org/10.1016/j.neubiorev.2022.104557
3. Kiyonaga, A., & Serences, J. T. (2025). Sensory reformatting for a working visual memory. *Trends in Cognitive Sciences, 29*(12), 1120–1135. https://doi.org/10.1016/j.tics.2025.09.006
4. Buschman, T. J. (2021). Balancing flexibility and interference in working memory. *Annual Review of Vision Science, 7*(1), 367–388. https://doi.org/10.1146/annurev-vision-100419-104831
5. Hogendoorn, H. (2022). Perception in real-time: predicting the present, reconstructing the past. *Trends in Cognitive Sciences, 26*(2), 128–141. https://doi.org/ 10.1016/j.tics.2021.11.003
6. White, P. A. (2020). The perceived present: What is it, and what is it there for?. *Psychonomic Bulletin & Review, 27*, 583–601. https://doi.org/10.3758/s13423-020-01726-7
7. Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell, 183*(4), 954–967. https://doi.org/10.1016/j.cell.2020.09.031
8. Chung, S., & Abbott, L. F. (2021). Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology, 70*, 137–144. https://doi.org/10.1016/j.conb.2021.10.010
9. Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron, 109*(19), 3055–3068. https://doi.org/10.1016/j.neuron.2021.07.011
10. Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience, 37*(1), 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325
11. Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences, 17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007
12. Libby, A., & Buschman, T. J. (2021). Rotational dynamics reduce interference between sensory and memory representations. *Nature Neuroscience, 24*(5), 715–726. https://doi.org/10.1038/s41593-021-00821-9
13. Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature, 592*(7855), 601–605. https://doi.org/10.1038/s41586-021-03390-w
14. Piwek, E. P., Stokes, M. G., & Summerfield, C. (2023). A recurrent neural network model of prefrontal brain activity during a working memory task. *PLoS*

Computational Biology, 19(10), e1011555. https://doi.org/10.1371/journal.pcbi.1011555
15. Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature, 497*(7451), 585–590. https://doi.org/10.1038/nature12160
16. Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2021). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences, 38*, 20–28. https://doi.org/10.1016/j.cobeha.2020.07.002
17. Boyle, L. M., Posani, L., Irfan, S., Siegelbaum, S. A., & Fusi, S. (2024). Tuned geometries of hippocampal representations meet the computational demands of social memory. *Neuron, 112*(8), 1358–1371. https://doi.org/10.1016/j.neuron.2024.01.021
18. Tye, K. M., Miller, E. K., Taschbach, F. H., Benna, M. K., Rigotti, M., & Fusi, S. (2024). Mixed selectivity: Cellular computations for complexity. *Neuron, 112*(14), 2289–2303. https://doi.org/ 10.1016/j.neuron.2024.04.017
19. Parthasarathy, A., Herikstad, R., Bong, J. H., Medina, F. S., Libedinsky, C., & Yen, S. C. (2017). Mixed selectivity morphs population codes in prefrontal cortex. *Nature Neuroscience, 20*(12), 1770–1779. https://doi.org/10.1038/s41593-017-0003-2
20. Roussy, M., Mendoza-Halliday, D., & Martinez-Trujillo, J. C. (2021). Neural substrates of visual perception and working memory: two sides of the same coin or two different coins?. *Frontiers in Neural Circuits, 15*, 764177. https://doi.org/10.3389/fncir.2021.764177
21. Xu, Y. (2017). Reevaluating the sensory account of visual working memory storage. *Trends in Cognitive Sciences, 21*(10), 794–815. https://doi.org/10.1016/j.tics.2017.06.013
22. Mendoza-Halliday, D., & Martinez-Trujillo, J. C. (2017). Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex. *Nature Communications, 8*(1), 15471. https://doi.org/10.1038/ncomms15471
23. Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*, 1122–1135. https://doi.org/10.3758/s13428-014-0532-5
24. Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological Science, 21*(11), 1551–1556. https://doi.org/10.1177/0956797610385359
25. Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., et al. (2016). A multi-modal parcellation of human cerebral cortex. *Nature, 536*, 171–178. https://doi.org/10.1038/nature18933
26. Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in

ventral temporal cortex. *Science, 293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736
27. Schwarzkopf, D. S., & Huang, Z. (2024). A simple statistical framework for small sample studies. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000710
28. Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology, 37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010
29. Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., ... & Smallwood, J. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences, 113*(44), 12574–12579. https://doi.org/10.1073/pnas.1608282113
30. Ito, T., & Murray, J. D. (2023). Multitask representations in the human cortex transform along a sensory-to-motor hierarchy. *Nature Neuroscience, 26*(2), 306–315. https://doi.org/10.1038/s41593-022-01224-0
31. Burt, J. B., Helmer, M., Shinn, M., Anticevic, A., & Murray, J. D. (2020). Generative modeling of brain maps with spatial autocorrelation. *NeuroImage, 220*, 117038. https://doi.org/10.1016/j.neuroimage.2020.117038
32. Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J. D. (2017). The distributed nature of working memory. *Trends in Cognitive Sciences, 21*(2), 111–124. http://dx.doi.org/10.1016/j.tics.2016.12.007
33. Lorenc, E. S., & Sreenivasan, K. K. (2021). Reframing the debate: The distributed systems view of working memory. *Visual Cognition, 29*(7), 416–424. https://doi.org/10.1080/13506285.2021.1899091
34. Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science, 20*(2), 207–214. https://doi.org/10.1111/j.1467-9280.2009.02276.x
35. Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997
36. Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598–601. https://doi.org/10.1038/33402
37. Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature, 458*(7238), 632–635. https://doi.org/10.1038/nature07832
38. Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & De Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology, 23*(15), 1427–1431. https://doi.org/10.1016/j.cub.2013.05.065
39. Rademaker, R. L., Chunharas, C., & Serences, J. T. (2019). Coexisting representations of sensory and mnemonic information in human visual cortex.

Nature Neuroscience, 22(8), 1336–1344. https://doi.org/10.1038/s41593-019-0428-x
40. Pettini, L., Bogler, C., Matic, K., Görgen, K., Doeller, C. F., & Haynes, J. D. (2025). Visual working memory representations of naturalistic images in early visual cortex are not sensory-like. Preprint at bioRxiv. https://doi.org/10.64898/2025.12.11.693774
41. Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience, 6*(2), 97–107. https://doi.org/10.1038/nrn1603
42. Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences, 18*(2), 82–89. https://doi.org/10.1016/j.tics.2013.12.001
43. 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. *Journal of Neuroscience, 33*(15), 6516–6523. https://doi.org/10.1523/JNEUROSCI.5732-12.2013
44. Gayet, S., Guggenmos, M., Christophel, T. B., Haynes, J. D., Paffen, C. L., Van der Stigchel, S., & Sterzer, P. (2017). Visual working memory enhances the neural response to matching visual input. *Journal of Neuroscience, 37*(28), 6638–6647. https://doi.org/10.1523/JNEUROSCI.3418-16.2017
45. Kwak, Y., & Curtis, C. E. (2022). Unveiling the abstract format of mnemonic representations. *Neuron, 110*(11), 1822–1828. https://doi.org/10.1016/j.neuron.2022.03.016
46. Bhandari, A., Gagne, C., & Badre, D. (2018). Just above chance: is it harder to decode information from prefrontal cortex hemodynamic activity patterns?. *Journal of Cognitive Neuroscience, 30*(10), 1473–1498. https://doi.org/10.1162/jocn_a_01291
47. Haynes, J. D. (2015). A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. *Neuron, 87*(2), 257–270. https://doi.org/10.1016/j.neuron.2015.05.025
48. Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*, *70*(2), 581–607. https://doi.org/10.1093/bjps/axx023
49. Ma, H., Jiang, L., Liu, T., & Liu, J. (2025). From sensory to perceptual manifolds: The twist of neural geometry. *Science Advances, 11*(50), eadv0431. https://doi.org/10.1126/sciadv.adv0431
50. Hayden, B. Y., Heilbronner, S. R., & Yoo, S. B. M. (2025). Rethinking the centrality of brain areas in understanding functional organization. *Nature Neuroscience*, 1–12. https://doi.org/10.1038/s41593-025-02166-z
51. Xu, Y. (2025). Transformed visual working memory representations in human occipitotemporal and posterior parietal cortices. *eNeuro, 12*(7). https://doi.org/10.1523/ENEURO.0162-25.2025

52. Chunharas, C., Wolff, M. J., Hettwer, M. D., & Rademaker, R. L. (2025). A gradual transition toward categorical representations along the visual hierarchy during working memory, but not perception. *eLife, 14*. https://doi.org/10.7554/eLife.103347.1
53. Bettencourt, K. C., & Xu, Y. (2016). Decoding the content of visual short-term memory under distraction in occipital and parietal areas. *Nature Neuroscience, 19*(1), 150–157. https://doi.org/10.1038/nn.4174
54. Lawrence, S. J., Formisano, E., Muckli, L., & de Lange, F. P. (2019). Laminar fMRI: Applications for cognitive neuroscience. *NeuroImage, 197*, 785–791. https://doi.org/10.1016/j.neuroimage.2017.07.004
55. Kok, P., Bains, L. J., Van Mourik, T., Norris, D. G., & de Lange, F. P. (2016). Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Current Biology, 26*(3), 371–376. https://doi.org/10.1016/j.cub.2015.12.038
56. Muckli, L., De Martino, F., Vizioli, L., Petro, L. S., Smith, F. W., Ugurbil, K., ... & Yacoub, E. (2015). Contextual feedback to superficial layers of V1. *Current Biology, 25*(20), 2690–2695. https://doi.org/10.1016/j.cub.2015.08.057
57. Duan, Z., & Curtis, C. E. (2024). Visual working memories are abstractions of percepts. *eLife, 13*, RP94191. https://doi.org/10.7554/eLife.94191.3
58. Xu, Y. (2024). The human posterior parietal cortices orthogonalize the representation of different streams of information concurrently coded in visual working memory. *PLoS Biology, 22*(11), e3002915. https://doi.org/10.1371/journal.pbio.3002915
59. Tsao, D. Y., Schweers, N., Moeller, S., & Freiwald, W. A. (2008). Patches of face-selective cortex in the macaque frontal lobe. *Nature Neuroscience, 11*(8), 877–879. https://doi.org/10.1038/nn.2158
60. Haile, T. M., Bohon, K. S., Romero, M. C., & Conway, B. R. (2019). Visual stimulus-driven functional organization of macaque prefrontal cortex. *NeuroImage, 188*, 427–444. https://doi.org/10.1016/j.neuroimage.2018.11.060
61. Rose, O., & Ponce, C. R. (2024). A concentration of visual cortex-like neurons in prefrontal cortex. *Nature Communications, 15*(1), 7002. https://doi.org/10.1038/s41467-024-51441-3
62. Dijkstra, N., von Rein, T., Kok, P., & Fleming, S.M. (2025). A neural basis for distinguishing imagination from reality. *Neuron, 113*(15), 2536–2542. https://doi.org/10.1016/j.neuron.2025.05.015
63. Tong, F. (2013). Imagery and visual working memory: one and the same?. *Trends in Cognitive Sciences, 17*(10), 489–490. https://doi.org/10.1016/j.tics.2013.08.005
64. Keogh, R., & Pearson, J. (2011). Mental imagery and visual working memory. *PLoS One, 6*(12), e29221. https://doi.org/10.1371/journal.pone.0029221
65. Kang, M. S., Hong, S. W., Blake, R., & Woodman, G. F. (2011). Visual working memory contaminates perception. *Psychonomic Bulletin & Review, 18*(5), 860–869. https://doi.org/10.3758/s13423-011-0126-5

66. Teng, C., & Kravitz, D. J. (2019). Visual working memory directly alters perception. *Nature Human Behaviour, 3*(8), 827–836. https://doi.org/10.1038/s41562-019-0640-4
67. Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 443–446. https://doi.org/10.1163/156856897X00357
68. Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. https://doi.org/10.1163/156856897x00366
69. King, D. (2017). High quality face recognition with deep metric learning. Dlib C++ Library, 12.
70. Serengil, S., & Özpınar, A. (2024). A benchmark of facial recognition pipelines and co-usability performances of modules. *Bilişim Teknolojileri Dergisi, 17*(2), 95–107. https://doi.org/10.17671/gazibtd.1399077
71. Hojjati, S., Ghadikolai, D. K., Shakeripour, A., Bahmani, Z., Shibata, K., Lau, H., & Moharramipour, A. (2025). Trade-off between performance and human-like perception in face recognition models. Preprint at PsyArXiv. https://doi.org/10.31234/osf.io/bz3f9_v1
72. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence, 40*(6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009
73. Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research, 76*, 31–42. https://doi.org/10.1016/j.visres.2012.10.012
74. Esteban, O., Markiewicz, C. J., Goncalves, M., Provins, C., Salo, T., Kent, J. D., DuPre, E., Ciric, R., Pinsard, B., Blair, R. W., Poldrack, R. A., & Gorgolewski, K. J. (2024). fMRIPrep: a robust preprocessing pipeline for functional MRI (23.2.1). Zenodo. https://doi.org/10.5281/zenodo.10790684
75. Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., ... & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods, 16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4
76. Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics, 5*, 13. https://doi.org/10.3389/fninf.2011.00013
77. Esteban, O., Markiewicz, C. J., Burns, C., Goncalves, M., Jarecka, D., Ziegler, E., Berleant, S., Ellis, D. G., Pinsard, B., Madison, C., Waskom, M., Notter, M. P., Clark, D., Manhães-Savio, A., Clark, D., Jordan, K., Dayan, M., Halchenko, Y. O., Loney, F., … Ghosh, S. (2025). nipy/nipype: 1.8.6 (1.8.6). Zenodo. https://doi.org/10.5281/zenodo.15054147
78. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., ... & Varoquaux, G. (2014). Machine learning for neuroimaging with

scikit-learn. *Frontiers in Neuroinformatics, 8*, 14. https://doi.org/10.3389/fninf.2014.00014
79. Andersson, J. L., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage, 20*(2), 870–888. https://doi.org/10.1016/S1053-8119(03)00336-7
80. Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging, 29*(6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908
81. Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis, 12*(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004
82. Zhang, Y., Brady, M., & Smith, S. (2002). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging, 20*(1), 45–57. https://doi.org/10.1109/42.906424
83. Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: a robust approach. *Neuroimage, 53*(4), 1181–1196. https://doi.org/10.1016/j.neuroimage.2010.07.020
84. Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage, 9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395
85. Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., ... & Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLoS Computational Biology, 13*(2), e1005350. https://doi.org/10.1371/journal.pcbi.1005350
86. Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., & Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage, 47*, S102. https://doi.org/10.1016/S1053-8119(09)70884-5
87. Ciric, R., Thompson, W. H., Lorenz, R., Goncalves, M., MacNicol, E. E., Markiewicz, C. J., ... & Esteban, O. (2022). TemplateFlow: FAIR-sharing of multi-scale, multi-species brain models. *Nature Methods, 19*(12), 1568–1571. https://doi.org/10.1038/s41592-022-01681-2
88. Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage, 17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132
89. Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage, 48*(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

90. Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage, 84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048
91. Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife, 11*, e77599. https://doi.org/10.7554/eLife.77599
92. Kay, K., Jamison, K. W., Zhang, R. Y., & Uğurbil, K. (2020). A temporal decomposition method for identifying venous effects in task-based fMRI. *Nature Methods, 17*(10), 1033–1039. https://doi.org/10.1038/s41592-020-0941-6
93. Serences, J. T., Schwarzbach, J., Courtney, S. M., Golay, X., & Yantis, S. (2004). Control of object-based attention in human cortex. *Cerebral Cortex, 14*(12), 1346–1357. https://doi.org/10.1093/cercor/bhh095
94. Valente, G., Castellanos, A. L., Hausfeld, L., De Martino, F., & Formisano, E. (2021). Cross-validation and permutations in MVPA: Validity of permutation strategies and power of cross-validation schemes. *NeuroImage, 238*, 118145. https://doi.org/10.1016/j.neuroimage.2021.118145
95. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Data and code availability

• Anonymized fMRI data have been reposited at OpenNeuro and are publicly available at https://openneuro.org/crn/reviewer/eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJzdWIiOiI0Y2MxNTQ4OS1kMTNkLTRmYmUtYTNlMi04ZDUyZjEzNzVmNzciLCJlbWFpbCI6InJldmlld2VyQG9wZW5uZXVyby5vcmciLCJwcm92aWRlciI6Ik9wZW5OZXVybyIsIm5hbWUiOiJBbm9ueW1vdXMgUmV2aWV3ZXIiLCJhZG1pbiI6ZmFsc2UsInNjb3BlcyI6WyJkYXRhc2V0OnJldmlld2VyIl0sImRhdGFzZXQiOiJkczAwNzA4MiIsImlhdCI6MTc2NjAyMjk3OCwiZXhwIjoxNzk3NTU4OTc4fQ.qwMrPdmrP1YVw3oOBa4aGKGC7wA-65TLiNG_lMNRvmg.

• Behavioral data have been deposited at Github and are publicly available at https://github.com/Tomnakam/Representational-geometries-of-perception-and-working-memory/tree/main.

• All original codes to run experiments and analysis have been deposited at Github and are publicly available at https://github.com/Tomnakam/Representational-geometries-of-perception-and-working-memory/tree/main.

• Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Acknowledgements

T.N. is supported by JSPS KAKENHI Grant Numbers 24KJ0233, 25K18957, and 25K00896 in Japan. H.L. is supported by the Institute for Basic Science (IBS-R015-D2) in the Republic of Korea. We thank Joshua Bradley Chan Tan for helpful comments on the manuscript.

Author contributions

Conceptualization, T.N., H.L., and A.M.; methodology, T.N., M.Y., K.K., H.L., and A.M.; software, T.N.; validation, T.N., K.K., and A.M.; formal analysis: T.N., and A.M.; investigation: T.N.; resources: H.L.; data curation: T.N., K.K.; writing - original draft, T.N.; writing - review & editing, T.N., M.Y., K.K., H.L., and A.M.; visualization, T.N., and A.M.; supervision: H.L., M.Y., and A.M.; project administration: T.N.; funding acquisition: T.N., and H.L.

Competing interests

The authors declared no competing interests.

Supplementary Information

Representational Geometries of Perception and Working Memory

Tomoya Nakamura, Seng Bum Michael Yoo, Kendrick Kay, Hakwan Lau, & Ali Moharramipour

Derivation of Equation 2

The full expression for the cosine similarity between the perceptual coding axis V_p and the working-memory coding axis V_{wm}, incorporating noise terms, ϵ_p and ϵ_{wm}, is:

$$cos(V_p + \epsilon_p, V_{wm} + \epsilon_{wm}) = \frac{V_p \bullet V_{wm} + \epsilon_p \bullet V_{wm} + V_p \bullet \epsilon_{wm} + \epsilon_p \bullet \epsilon_{wm}}{\sqrt{||V_p||^2 + 2V_p \bullet \epsilon_p + \epsilon_p \bullet \epsilon_p}\sqrt{||V_{wm}||^2 + 2V_{wm} \bullet \epsilon_{wm} + \epsilon_{wm} \bullet \epsilon_{wm}}}$$

Under the assumption that signal and noise are independent, and the noises have a zero mean, the expression reduces to:

$$cos(V_p + \epsilon_p, V_{wm} + \epsilon_{wm}) = \frac{V_p \bullet V_{wm} + \sigma_{p,wm}(noise)}{\sqrt{||V_p||^2 + \sigma_p^2(noise)}\sqrt{||V_{wm}||^2 + \sigma_{wm}^2(noise)}}$$

$$= \frac{V_p \bullet V_{wm} + \sigma_{p,wm}(noise)}{||V_p||\,||V_{wm}||\sqrt{1 + \frac{\sigma_p^2(noise)}{||V_p||^2}}\sqrt{1 + \frac{\sigma_{wm}^2(noise)}{||V_{wm}||^2}}}$$

$$= \frac{\frac{V_p \bullet V_{wm}}{||V_p||\,||V_{wm}||} + \frac{\sigma_{p,wm}(noise)}{||V_p||\,||V_{wm}||}}{\sqrt{1 + \frac{\sigma_p^2(noise)}{||V_p||^2}}\sqrt{1 + \frac{\sigma_{wm}^2(noise)}{||V_{wm}||^2}}}$$

Here, $\sigma_p^2(noise)$, and $\sigma_{wm}^2(noise)$ represent the noise variances for perceptual and working-memory coding axes, respectively, and $\sigma_{p,wm}(noise)$ indicate their noise covariance. Under the null hypothesis of parallelism, the cosine value between V_p and V_{wm} is 1.

$$cos(V_p + \epsilon_p, V_{wm} + \epsilon_{wm}) = \frac{1 + \frac{\sigma_{p,wm}(noise)}{||V_p||\,||V_{wm}||}}{\sqrt{1 + \frac{\sigma_p^2(noise)}{||V_p||^2}}\sqrt{1 + \frac{\sigma_{wm}^2(noise)}{||V_{wm}||^2}}}$$

Then, for a better intuitive understanding of the equation, we rewrite it using the SNR notion.

$$cos(V_p + \epsilon_p, V_{wm} + \epsilon_{wm}) = \frac{1 + \frac{\frac{\sigma_{p,wm}(noise)}{\sigma_p(noise)\,\sigma_{wm}(noise)}}{\frac{||V_p||}{\sigma_p(noise)}\,\frac{||V_{wm}||}{\sigma_{wm}(noise)}}}{\sqrt{1+\frac{1}{\frac{||V_p||^2}{\sigma_p^2(noise)}}}\sqrt{1+\frac{1}{\frac{||V_{wm}||^2}{\sigma_{wm}^2(noise)}}}}$$

$$= \frac{1+\frac{\rho_{noise}}{\sqrt{SNRp\ SNRwm}}}{\sqrt{1+\frac{1}{SNRp}}\sqrt{1+\frac{1}{SNRwm}}}$$

Supplementary figures

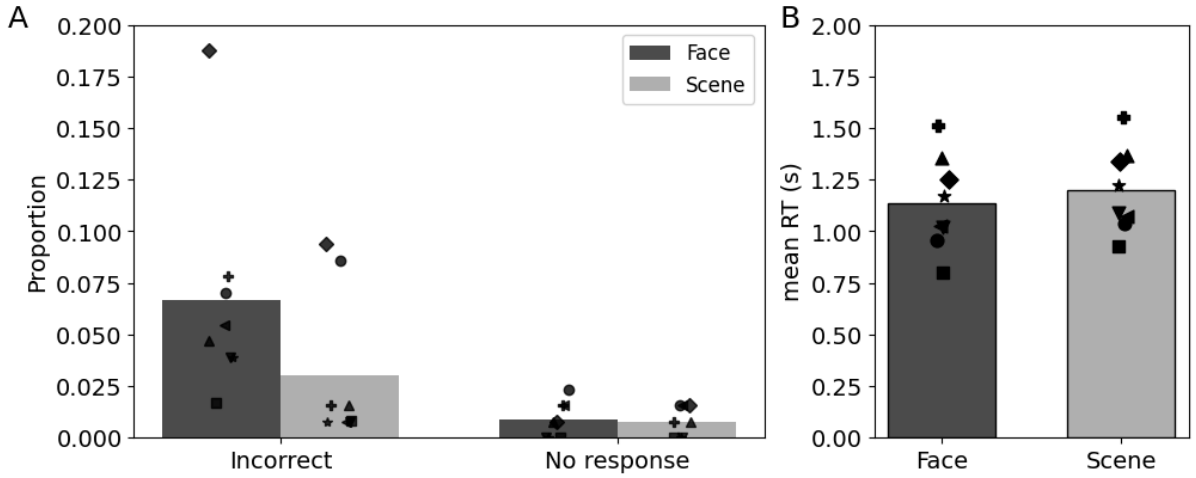

Figure S1. Behavioral data. *(A) Proportion of trials (among those including the Test phase) that resulted in incorrect responses and no responses. The mean correct response rate was 94.3% and the proportion of no-response trials was low (mean = 0.8%). At the group level, correct response rates for the scene-cued trials were significantly higher than for the face-cued trials (paired t-test: t(7) = 3.44, p = .011). (B) Geometric mean reaction time (RT) for correct trials in each observer. In contrast to correct response rates, RTs were significantly shorter for the face-cued trials (paired t-test on mean log-RTs: t(7) = 4.03, p = .005). Bar lengths indicate the inter-observer mean. Note that we excluded two runs (1st run in the 1st and 3rd sessions) for Obs. 2 only from behavioral analyses because the responses were not properly recorded due to technical issues.*

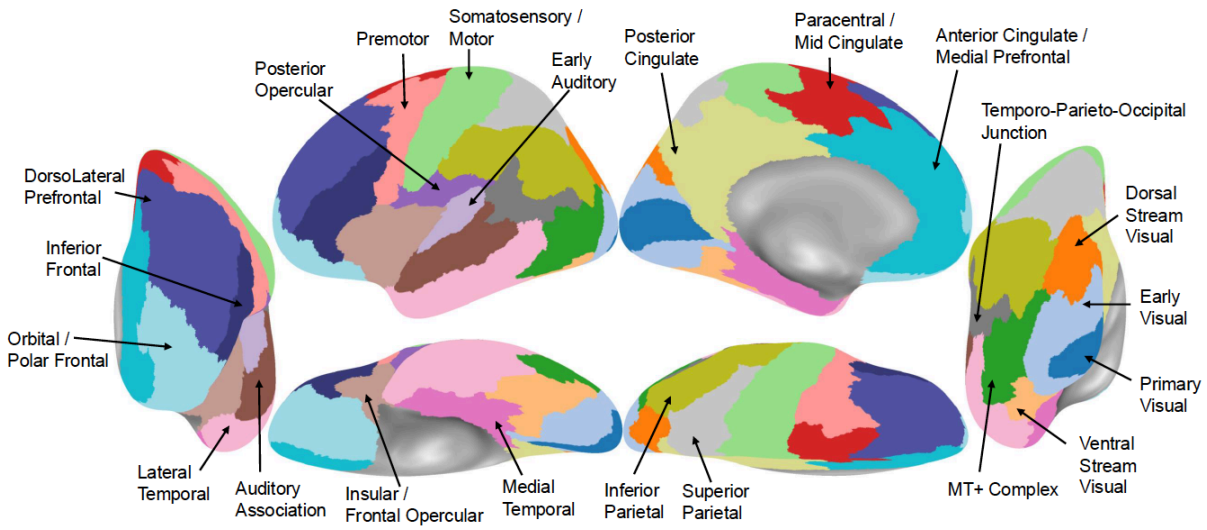

Figure S2. Definition of the regions of interest (ROIs) used for the main analyses. *Different colored regions indicate each of the 22 coarse-grained ROIs defined by Glasser et al. (2016) [s1].*

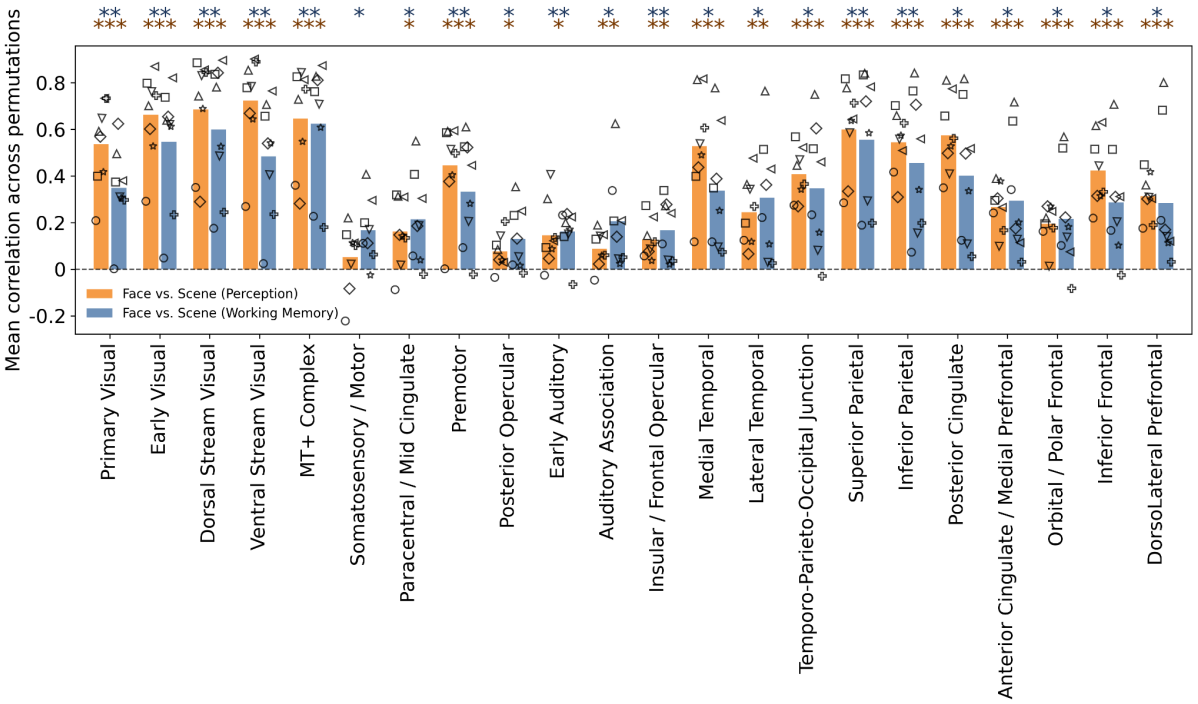

Figure S3. Test-retest reliability of face-vs-scene tuning. *Each marker indicates each observer, showing the mean correlation across 10,000 split-half permutations. For each permutation, runs were randomly split into two halves, and the Pearson correlation was computed between the multivertex activation patterns (face-scene contrast) of the two halves across all vertices within each ROI. Asterisks indicate group-level statistical significance, assessed by one-sample one-tailed t-test (*** q < .001, ** q < .01, * q < .05; FDR-corrected among all ROIs). Face-vs-scene tuning was reliable across different runs in the visual and most association areas (temporal, parietal, prefrontal cortex) both for perception and working memory. Such reliable tuning provides a basis for the subsequent analysis regarding the angle between face-vs-scene coding axes for perception and working memory.*

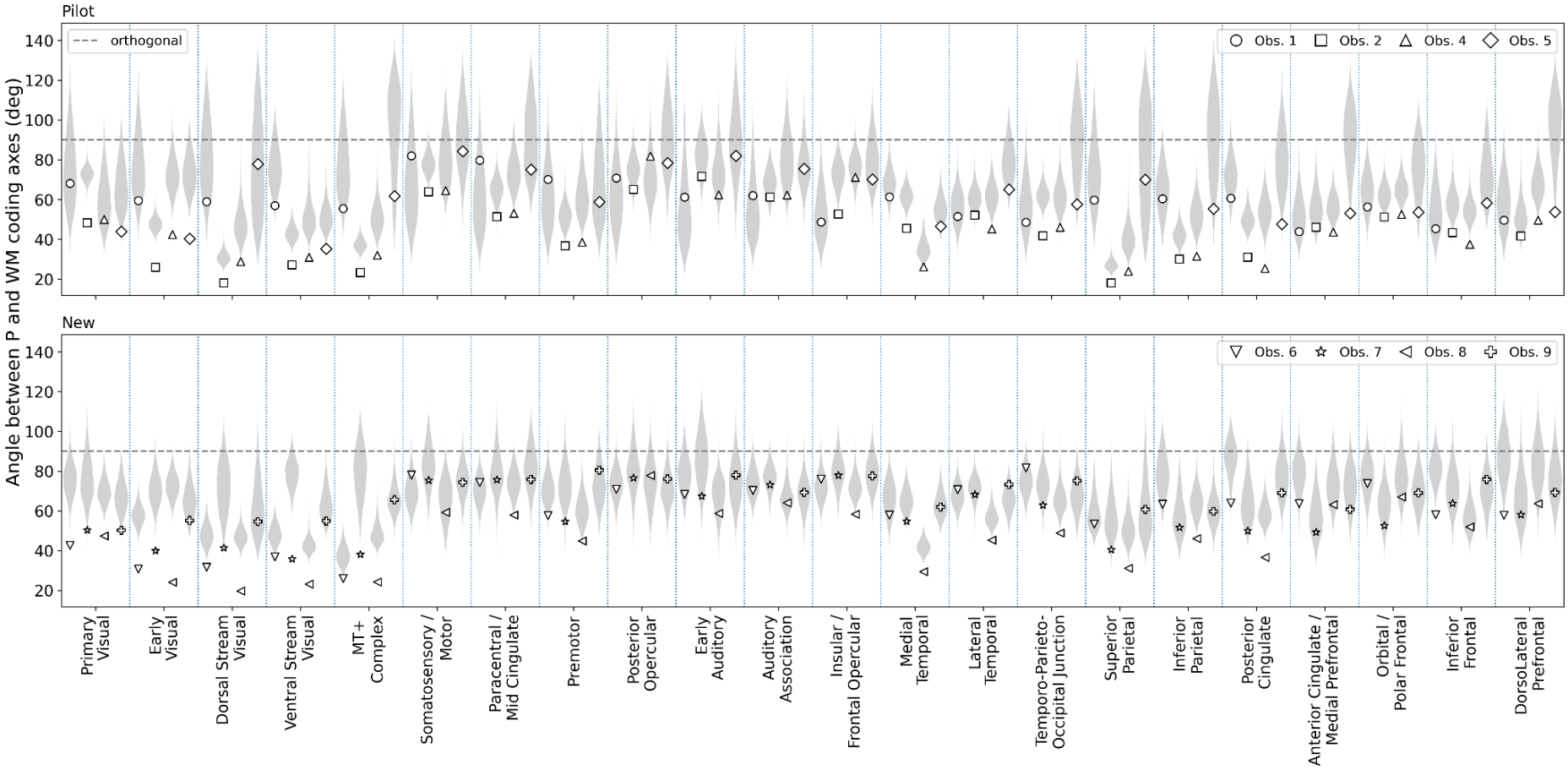

Figure S4. Individual results of angle analysis. *Bootstrapped distribution of angle between face-vs-scene coding axis for perception and working memory was shown along with its noise floor. Different marker shapes indicate the noise floor from different observers.*

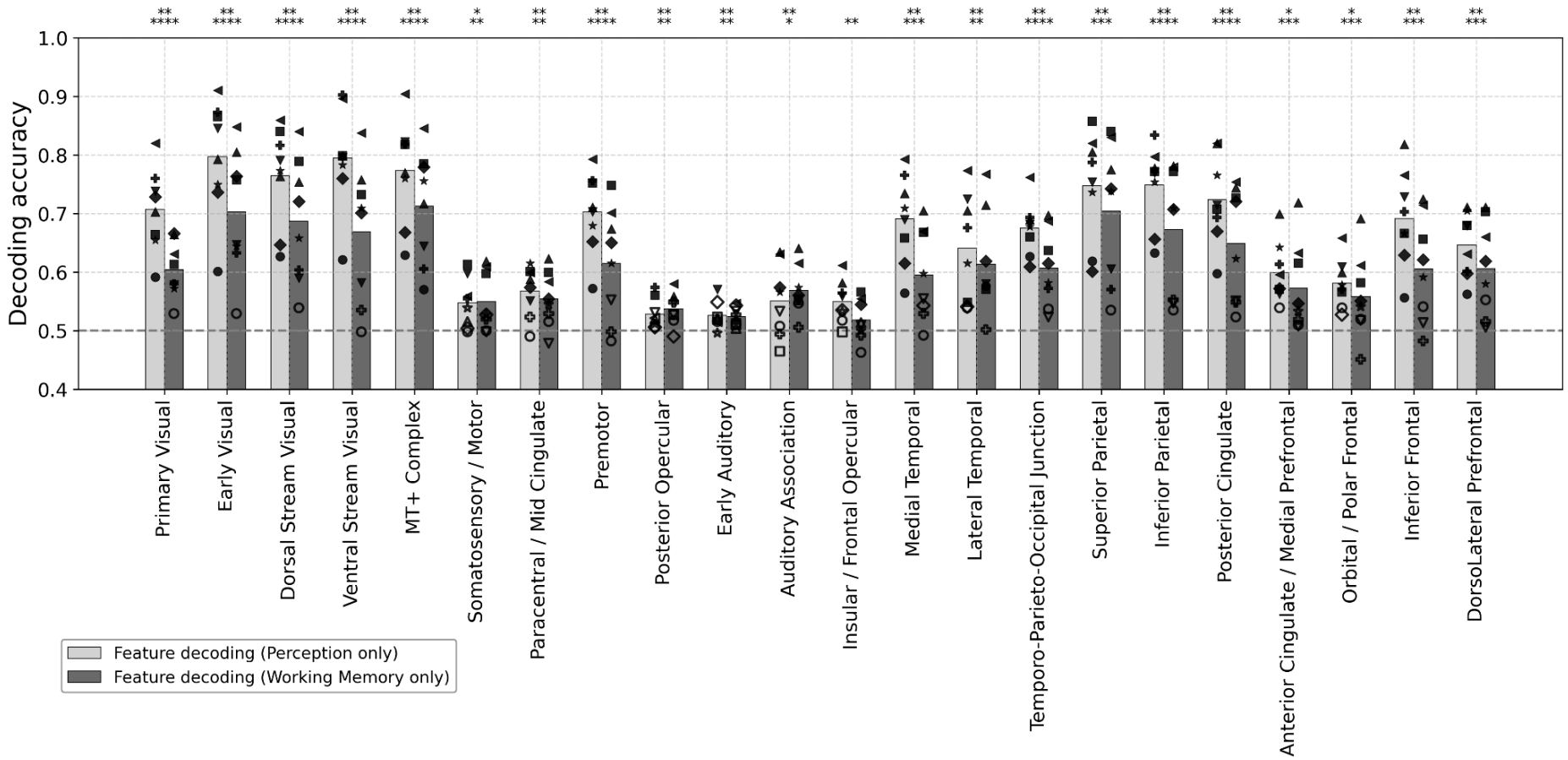

Figure S5. Decoding accuracy for face vs. scene assessed separately within the perception and working-memory epochs. *Asterisks denote the significance of the group-level one-sample t-test (one-tailed) on z-transformed accuracy, FDR-corrected across all ROIs (**** q < .0001, *** q < .001, ** q < .01, * q < .05). The dichotomy between face vs. scene was decodable from both perception and working memory epochs across the entire visual cortex, the paracentral areas, and the temporal, parietal, and frontal association cortices (cf. Figure 3). These results are consistent with face-vs-scene tuning measured via test-retest reliability (Figure S3).*

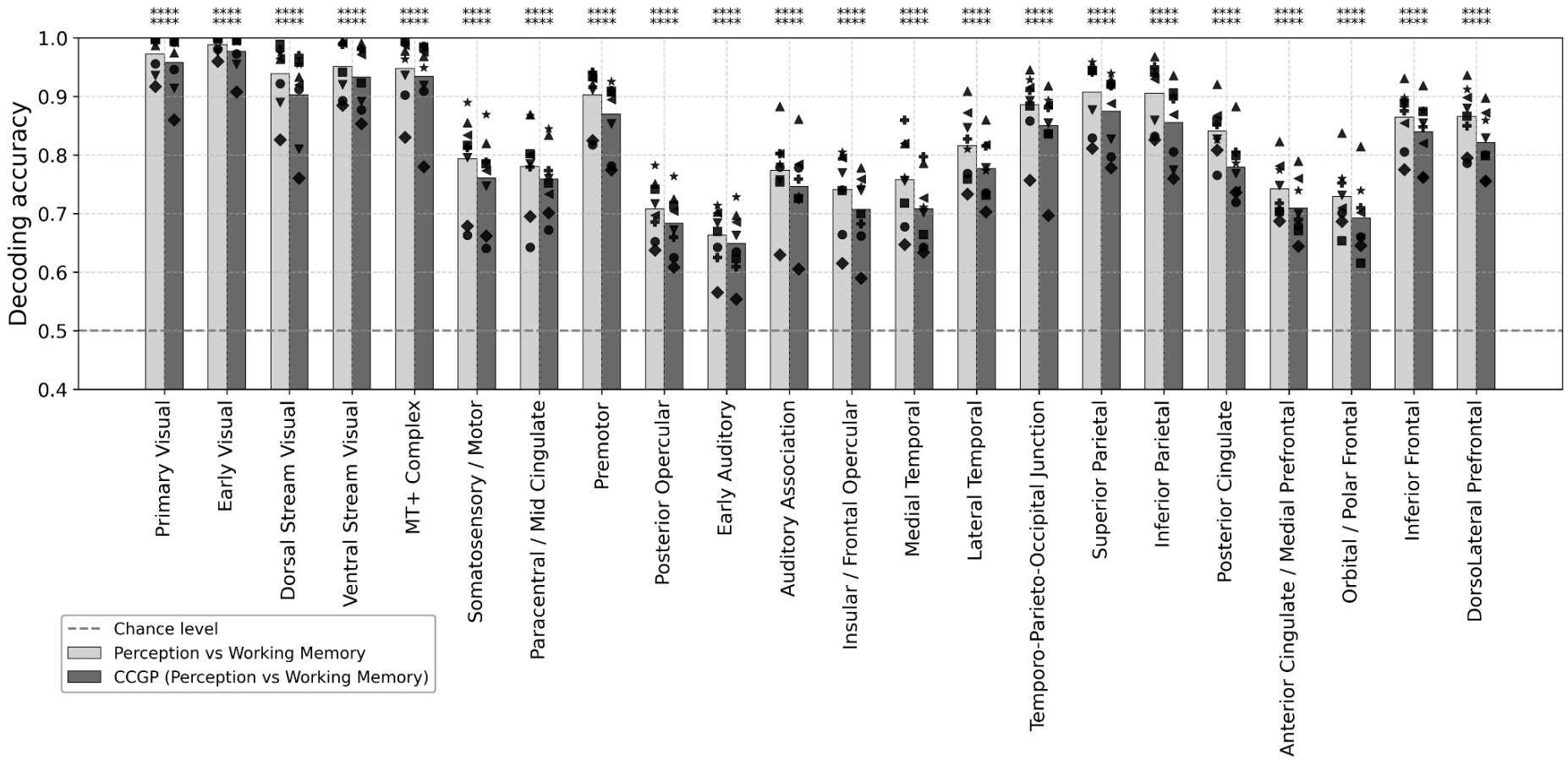

Figure S6. Decoding accuracy and CCGP for perception vs. working memory. *Asterisks denote the significance of the group-level one-sample t-test (one-tailed) on z-transformed accuracy, FDR-corrected across all ROIs (**** q < .0001). The dichotomy between perception vs. working memory was highly decodable and highly generalizable across faces and scenes in all ROIs in all observers.*

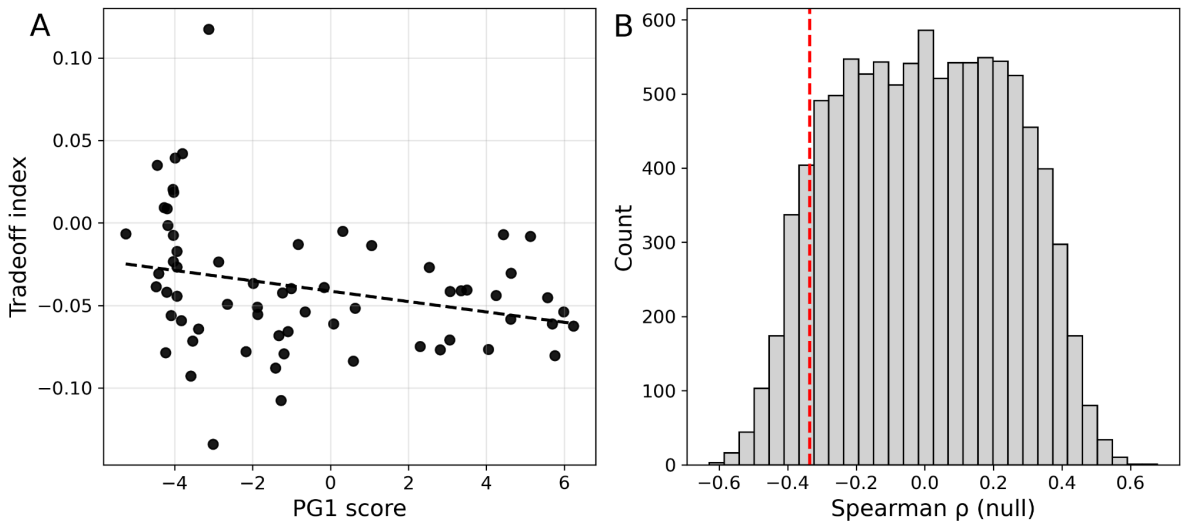

Figure S7. Relationship between the tradeoff index and the functional gradient. *(A) Tradeoff index plotted against the 1st principal gradient scores from Margulies et al. (2016) [s2] for the ROIs selected using the same criteria as in Figure 4. The Spearman correlation was significantly negative when tested against a null hypothesis assuming fully random spatial structure (*$\rho = -.336$*, two-sided* $p = .007$*). (B) Null distribution of Spearman* ρ *obtained while preserving spatial autocorrelation [s3, s4]. Under this more stringent null model, the observed correlation was not significant (two-sided* $p = .188$*). The red vertical line indicates the observed correlation coefficient.*

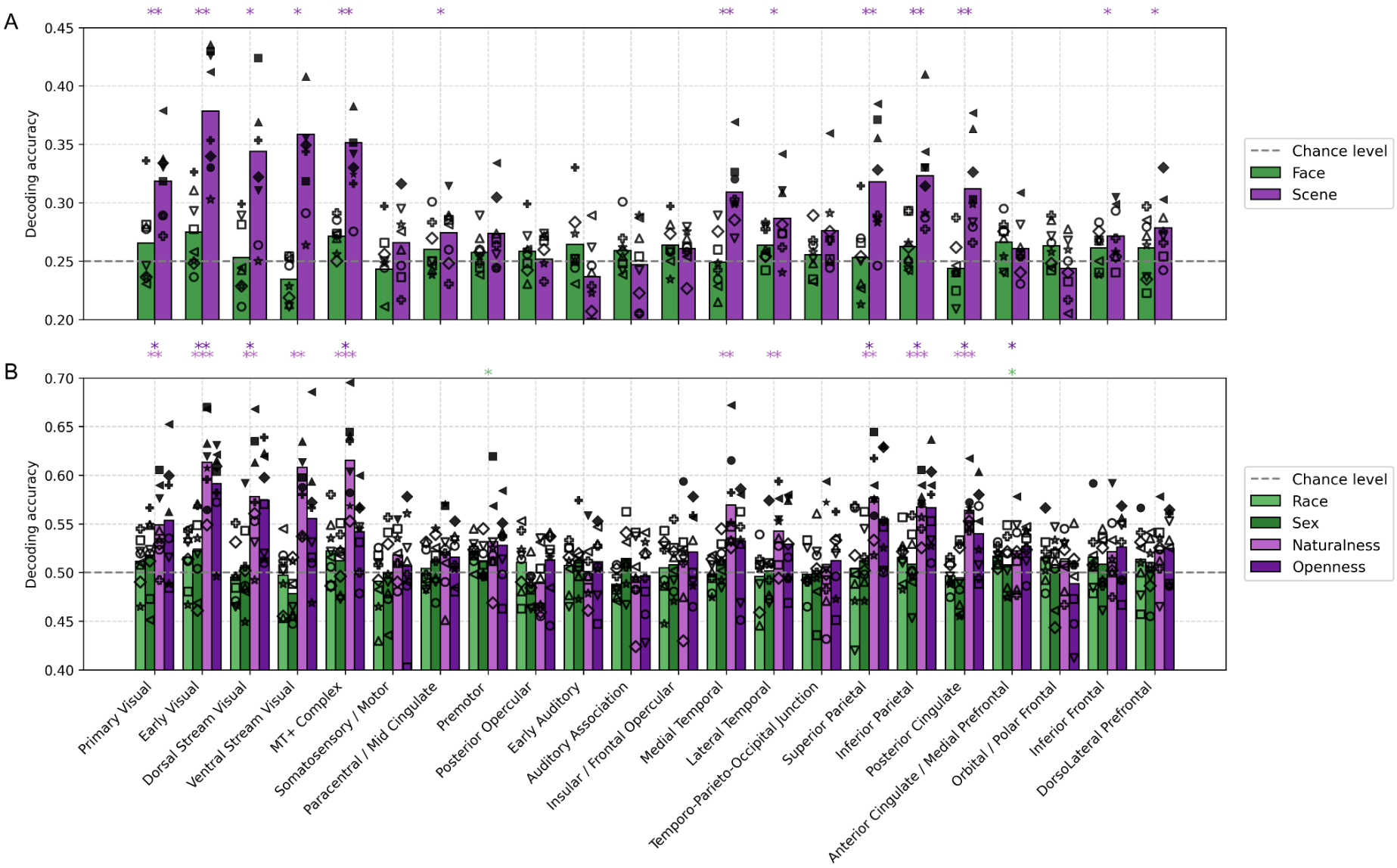
A
B
Decoding accuracy
Chance level
Face
Scene
Race
Sex
Naturalness
Openness
Primary Visual
Early Visual
Dorsal Stream Visual
Ventral Stream Visual
MT+ Complex
Somatosensory / Motor
Paracentral / Mid Cingulate
Premotor
Posterior Opercular
Early Auditory
Auditory Association
Insular / Frontal Opercular
Medial Temporal
Lateral Temporal
Temporo-Parieto-Occipital Junction
Superior Parietal
Inferior Parietal
Posterior Cingulate
Anterior Cingulate / Medial Prefrontal
Orbital / Polar Frontal
Inferior Frontal
DorsoLateral Prefrontal

Figure S8. Decoding results for specific sub-categories. *(A) Decoding accuracy among the four face sub-categories within face-cued trials (green) and among the four scene sub-categories within scene-cued trials (purple). (B) Decoding accuracy for black vs. white (light green) and female vs. male (deep green) within face-cued trials, and for natural vs. manmade (light purple) and open vs. closed (deep purple) within scene-cued trials. All decoders were trained and tested irrespective of whether the content is perceived or held in working memory. Asterisks denote the significance of the group-level one-sample t-test (one-tailed) on z-transformed accuracy, FDR-corrected across all ROIs (*** q < .001, ** q < .01, * q < .05). Overall, beyond face-vs-scene decoding, we successfully decoded scene sub-categories from the entire visual cortex as well as from the entire parietal cortex (superior parietal, inferior parietal, and posterior cingulate cortex), medial/lateral temporal cortex and only lateral parts of prefrontal cortex (inferior frontal and dorsolateral prefrontal cortex). In contrast, we could not reliably decode face sub-categories from any ROI. Furthermore, from similar areas, we were able to decode scene sub-category dichotomies (Figure 7B). The most decodable dichotomy was naturalness (natural vs. manmade scenes), which was decoded from the entire visual cortices, medial/lateral temporal cortex, and from the entire parietal cortices (superior parietal and inferior parietal). The second most decodable dichotomy was openness (open vs. closed scenes), which was decoded from some of the visual, parietal, and prefrontal cortices. Face subcategory dichotomy (only the race) was decoded only from the premotor and anterior cingulate/medial prefrontal cortex. The difference in sub-category decodability between faces and scenes may reflect the greater visual diversity of scenes compared to faces. In any case, these results confirm that, at least in the visual and some association (temporal, parietal, and lateral prefrontal) cortices, each scene sub-category are differentially represented and thus the rotation of the face-vs-scene coding axes of interest cannot be solely explained by language processing regarding the cue or by differences in the spatial extent of the attentional window [s5]. When assessed using logit-transformed accuracy, face decodability in MT+ complex became significant, whereas openness decodability in the primary visual cortex and posterior cingulate cortex became non-significant. Moreover, when the runs with poor image quality were excluded, scene decodability in the insular/frontal opercular cortex and naturalness decodability in the premotor cortex became significant, whereas openness decodability in the primary visual cortex and race decodability in the premotor cortex became non-significant.*

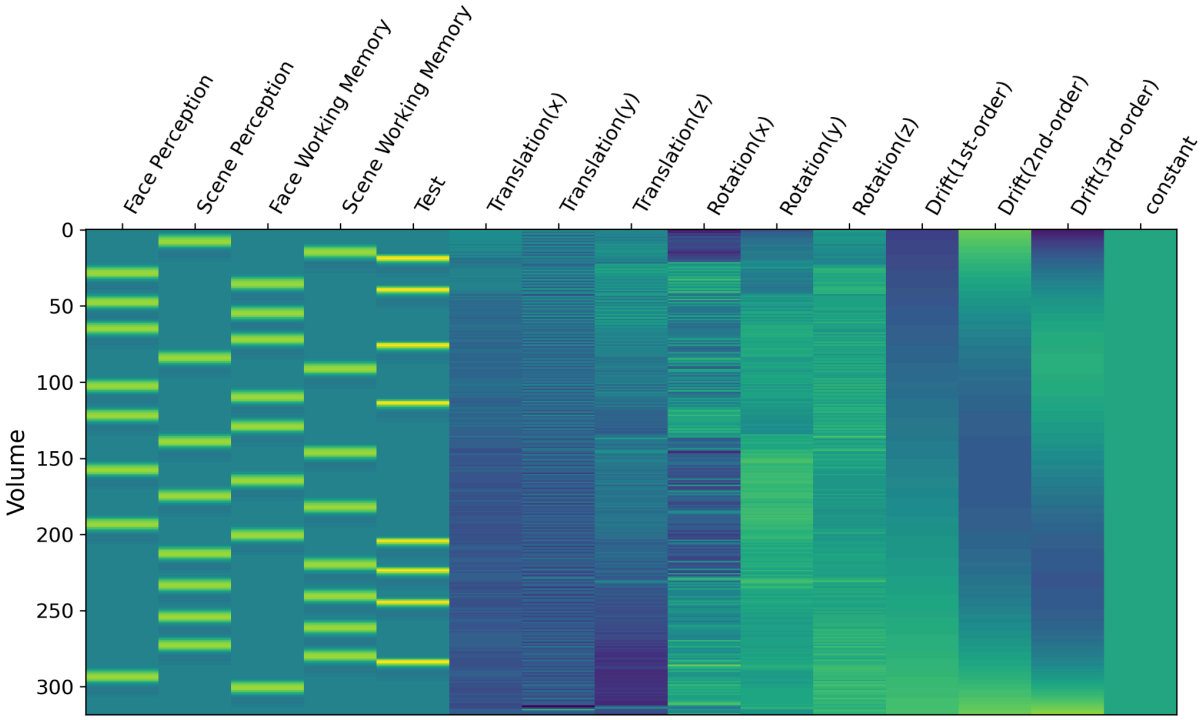

Figure S9. Design matrix for the runwise GLM. *Each column indicates a regressor, and each row indicates a volume. The leftmost four regressors correspond to each condition of interest. The 5th regressor from the left corresponds to the test stimuli. 6th to 11th regressors are motion parameters derived during preprocessing. 12th to 14th regressors model slow polynomial drift (up to 3rd order). The rightmost regressor models the baseline activation for this run. One example run is shown here.*

Simulation of the tradeoff

CCGP is high for factorized, low-dimensional representational geometries, whereas XOR accuracy is high for high-dimensional, entangled geometries. Thus, CCGP and XOR should theoretically be in a tradeoff relationship. However, when we simply plot empirical CCGP accuracy against XOR accuracy across ROIs, they instead show a positive correlation (Figure S10A). This counterintuitive pattern can be explained by differences in functional SNR across regions (high SNR raises decoding accuracy overall). Here, we illustrate how SNR influences the apparent relationship between CCGP and XOR using simulations based on a simplified toy model and test the robustness of the results regarding tradeoff even after SNR effects are better accounted for.

In our simulation, we assumed three dimensions because four conditions can be fully represented in a three-dimensional space, and for simplicity we assumed isotropic Gaussian noise. We began with perfectly factorized ("rectangular") geometry that maximizes CCGP while rendering XOR decoding chance-level (Figure 1B). In this configuration, mean of the activation patterns for each of the four conditions lie on the same two-dimensional plane, and all observed responses can be generated by a model characterized by three parameters (S_{P-WM} : signal strength along the perception-memory axis, S_{F-S} : signal strength along the face-scene axis, σ_{noise}: unit variance Gaussian noise). Starting from this perfectly factorized geometry, we rotated the memory axis stepwise (by 5 degrees) in a direction orthogonal to the initial two-dimensional plane until the perceptual and memory axis became orthogonal to each other. This orthogonal geometry represents the opposite end of the tradeoff continuum, supporting maximal XOR decodability while disabling CCGP (Figure 1D).

We found that under the assumption of independent noise, S_{P-WM} does not change the simulation results as long as we had enough trials. Therefore, here we report results obtained by fixing the S_{P-WM} as 2 (twice the σ_{noise}). When we plot XOR against CCGP, as the memory axis is rotated more, the point moves along the concave. Importantly, S_{F-S} moves the curve toward upright (different colored curves in Figure S11A), which illustrates the fact that even if angle is the same, higher SNR produces higher XOR and higher CCGP simultaneously, which could explain the apparent positive relationship between them across different ROIs (Figure S10A).

To eliminate the effect of SNR, we normalized the measures as follows:

$$normalized\ CCGP = \frac{\Phi^{-1}(ACC_{CCGP})}{\Phi^{-1}(ACC_{FvsS})},\ normalized\ XOR = \frac{\Phi^{-1}(ACC_{XOR})}{\Phi^{-1}(ACC_{FvsS})}\ [S1],$$

where ACC_{FvsS}, ACC_{XOR}, ACC_{CCGP} denote the decoding accuracies for the face-vs-scene dichotomy, XOR dichotomy and CCGP (face-vs-scene), respectively, and Φ^{-1} indicates the probit transformation. Probit normalization is justified under the assumption of Gaussian noise. Our simulations confirmed that the relationship between normalized CCGP and normalized XOR remains stable across different SNR levels (Figure S11B). In the empirical data, normalized CCGP and normalized XOR also showed a negative relationship (Figure S10B), although their values do not exactly follow the simulated trajectory (Figure S11B), because, in the real brain, unlike in the simulation, noise structure is likely more complex, and the vertex space is also higher-dimensional.

Given the improved normalization strategies, we can quantify the tradeoff for each region as the difference between the probit ratios, as follows:

$$Tradeoff\ index = normalized\ XOR - normalized\ CCGP\ [s2],$$

We confirmed that the tradeoff index increased monotonically with rotation angle and, importantly, that this relationship was largely independent of SNR (Figure S11D). It is noteworthy that the simpler definition of the tradeoff index, used in Figure 4, likewise exhibited a monotonic increase with angle (Figure S11C). Although SNR independence was not perfect, deviations were primarily observed in cases of poor SNR, which were largely excluded from the main analyses (see Figure 4), as well as at extreme rotation angles (0° or 90°), which were unlikely for most ROIs in the empirical data (see Figure S4).

Even when the effects of SNR were more precisely accounted for using the index, described in Equation S2, we further confirmed that the overall empirical results remained consistent: early sensory regions favored flexibility whereas association cortex favored abstraction (Figure S12; cf. Figure 4). The correlation between first functional gradient (unimodal-to-transmodal) [s2] and tradeoff index was slightly stronger, but still it was marginally significant ($p = .056$) when considering the spatial autocorrelation in the null distribution (Figure S13). However, given that this alignment with functional gradient was not statistically robust across indices (cf., Figure S7), this finding should be interpreted with caution and remains inconclusive.

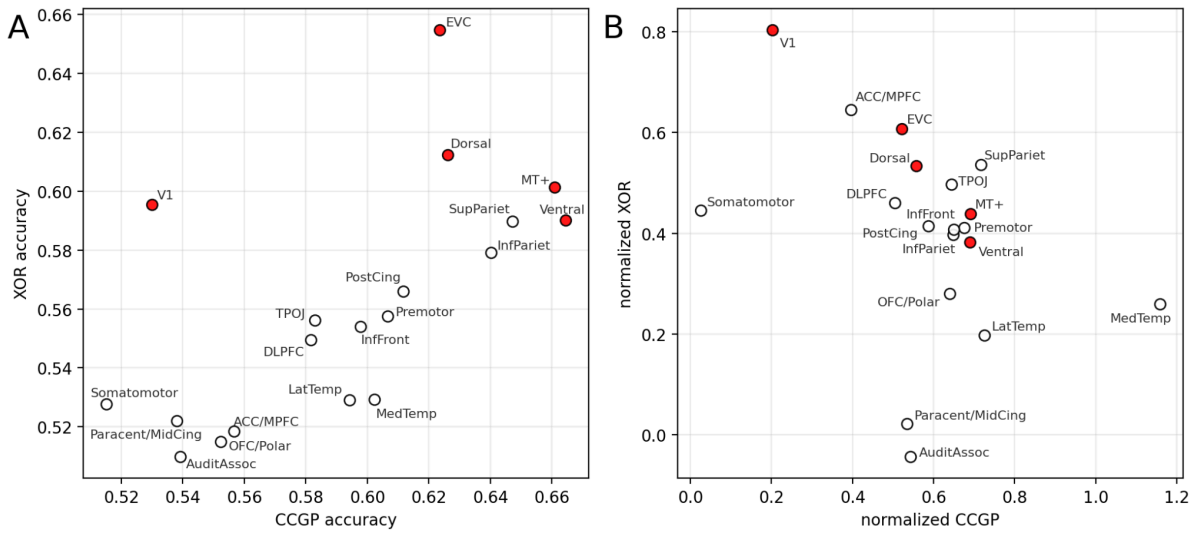

Figure S10. The empirical relationship between CCGP and XOR. *(A) The relationship between interobserver mean of CCGP accuracy and XOR accuracy. (B) The relationship between interobserver mean of normalized CCGP and normalized XOR. Visual areas are highlighted in red.*

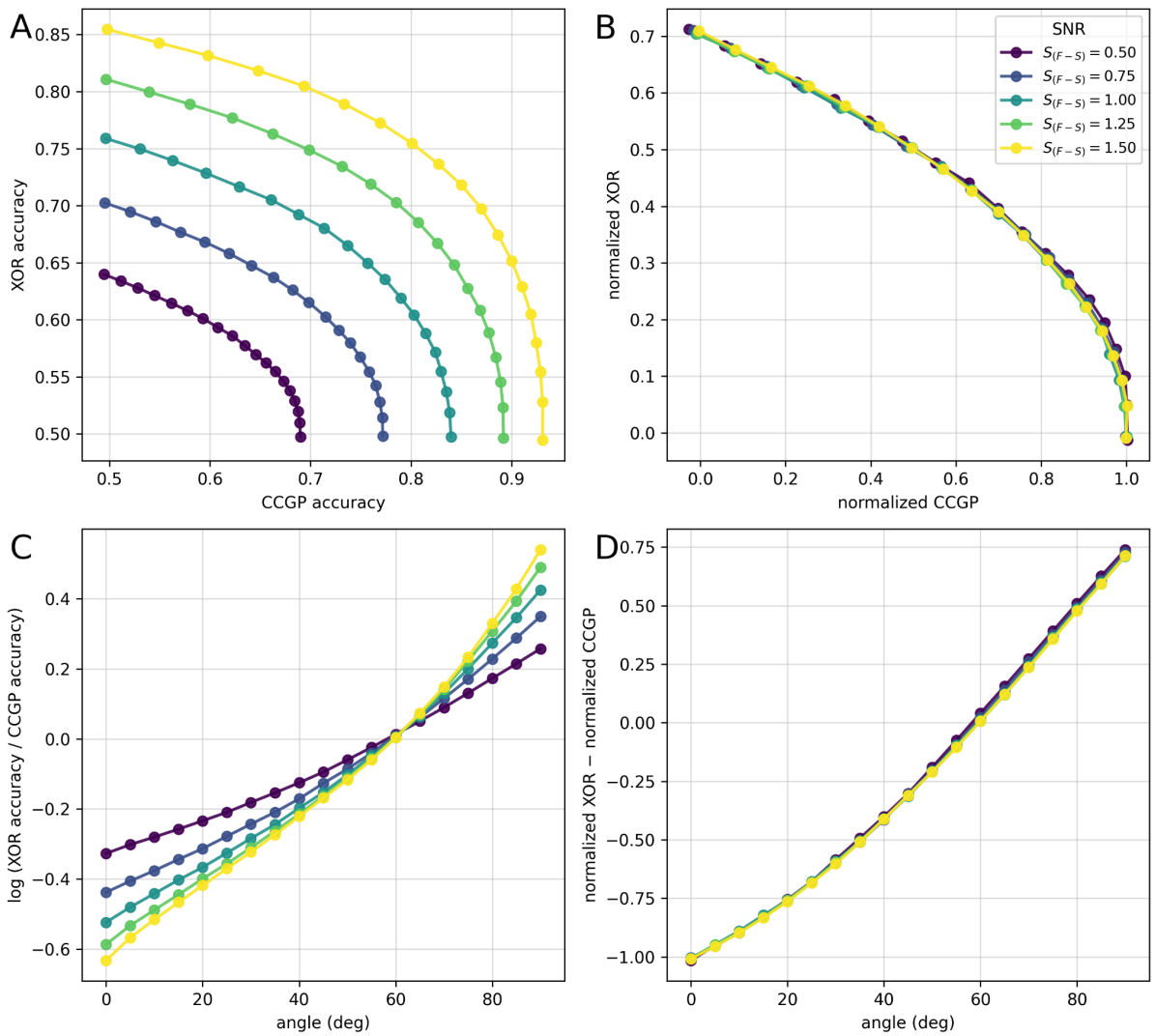

Figure S11. Simulation for SNR-independent index of tradeoff. *(A) The simulated relationship between CCGP accuracy and XOR accuracy. (B) The simulated relationship between normalized CCGP and normalized XOR. Different colors indicate different levels of SNR. As the perceptual and memory axes became increasingly orthogonal, points traversed the curve from the bottom right (maximal CCGP with chance-level XOR) toward the top left (chance-level CCGP with maximal XOR), with points sampled at 5° rotation increments. (C) The simulated relationship between rotation angle and tradeoff index defined in Equation 9. (D) The simulated relationship between rotation angle and tradeoff index defined in Equation S2.*

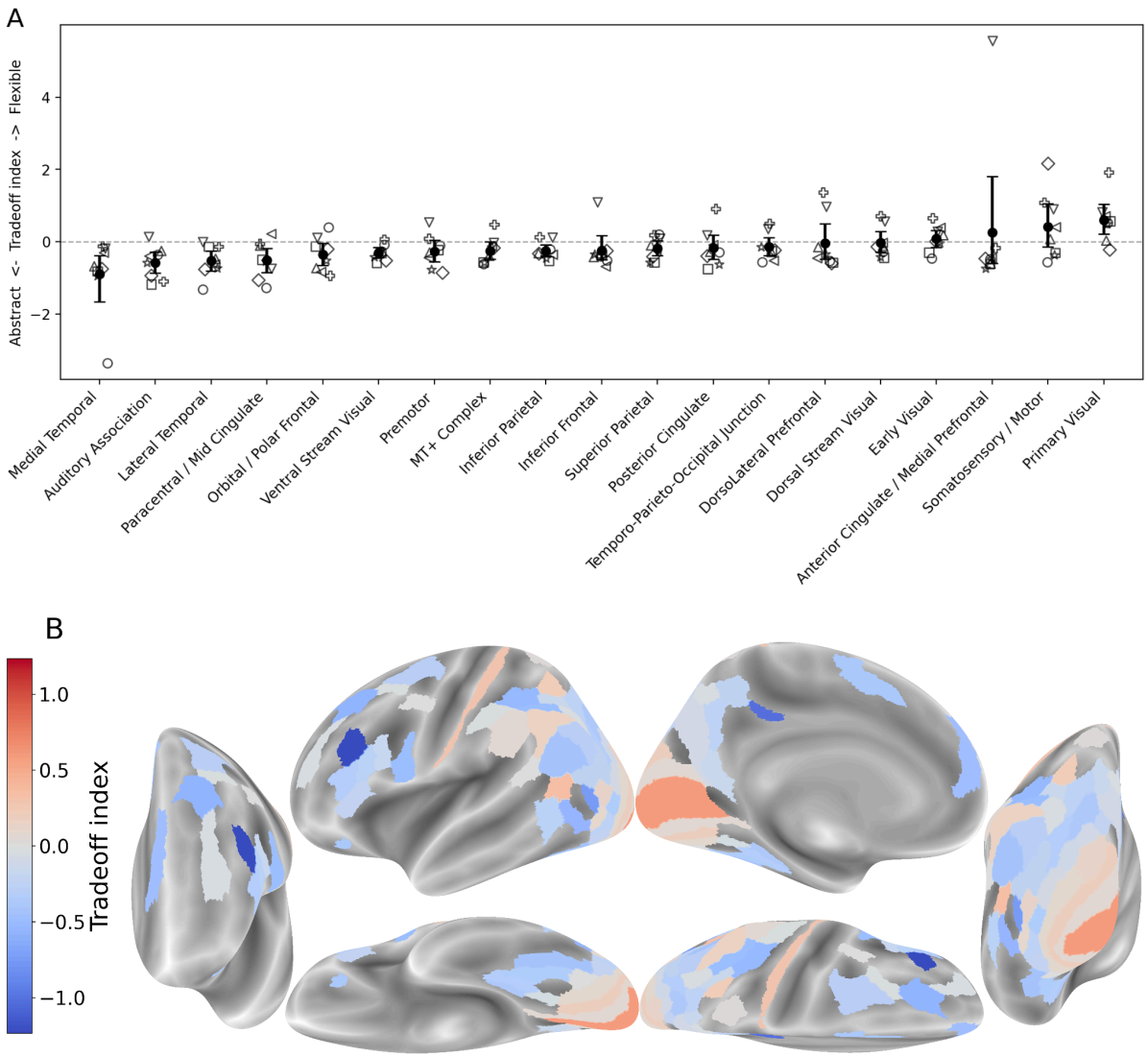

Figure S12.Tradeoff index based on the SNR-independent measure (Equation S2). *The tradeoff index (A) for each of the 22 coarse-grained ROIs, and (B) for the fine-grained cortical atlas. The conventions are the same as in Figure 4. Overall, even when we use a more SNR-independent index of the tradeoff, the general trend remains similar: early sensory regions favored flexibility at the cost of abstraction, whereas association cortex (particularly temporal regions) prioritized abstraction.*

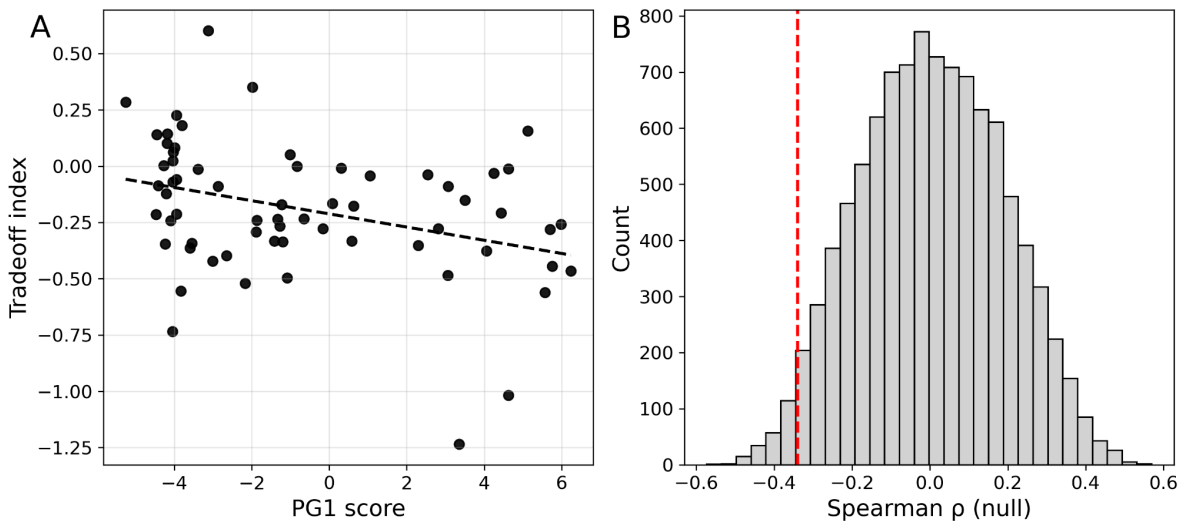

Figure S13. Relationship between the tradeoff index (based on the SNR-independent measure) and the functional gradient. *(A) Tradeoff index plotted against the 1st principal gradient scores from Margulies et al. (2016) [s2] for the ROIs selected using the same criteria as in Figure 4. The Spearman correlation was significantly negative when tested against a null hypothesis assuming fully random spatial structure (*$\rho = -.339$*, two-sided* $p = .006$*). (B) Null distribution of Spearman* ρ *obtained while preserving spatial autocorrelation [s3, s4]. Under this more stringent null model, the observed correlation was barely non-significant (two-sided* $p = .056$*). The red vertical line indicates the observed correlation coefficient.*

Supplementary references

s1. Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., et al. (2016). A multi-modal parcellation of human cerebral cortex. *Nature, 536*, 171–178. https://doi.org/10.1038/nature18933

s2. Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., ... & Smallwood, J. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences, 113*(44), 12574–12579. https://doi.org/10.1073/pnas.1608282113

s3. Ito, T., & Murray, J. D. (2023). Multitask representations in the human cortex transform along a sensory-to-motor hierarchy. *Nature Neuroscience, 26*(2), 306–315. https://doi.org/10.1038/s41593-022-01224-0

s4. Burt, J. B., Helmer, M., Shinn, M., Anticevic, A., & Murray, J. D. (2020). Generative modeling of brain maps with spatial autocorrelation. *NeuroImage, 220*, 117038. https://doi.org/10.1016/j.neuroimage.2020.117038

s5. Serences, J. T., Schwarzbach, J., Courtney, S. M., Golay, X., & Yantis, S. (2004). Control of object-based attention in human cortex. *Cerebral Cortex, 14*(12), 1346–1357. https://doi.org/10.1093/cercor/bhh095