Representational Geometries of Perception and Working Memory: A Pilot Study

Tomoya Nakamura[1,2,3,*], Seng Bum Michael Yoo[2, 5, 6], Kendrick Kay[4], Hakwan Lau[2,5,6], & 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]Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Minneapolis, USA
[5]Department of Biomedical Engineering, Sungkyunkwan University, Suwon, Republic of Korea
[6]Department of Intelligent Precision Healthcare Convergence, Sungkyunkwan University, Suwon, Republic of Korea

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

Abstract

This study aims to compare the neural representational geometry of visual perception and visual working memory using human fMRI. In our pilot experiment, observers viewed a face-scene blended image (sample) and selectively attended to either the face or scene aspects of the sample. They then judged whether the attended aspects were the same as or different from that of the subsequent test image. In three out of four observers, we found that the coding axis, which distinguishes faces and scenes, was rotated when perceptual content transforms into working memory, not only in the visual cortex but also in some higher association cortical areas. In these regions, exclusive combinations (i.e., perceived face and memorized scene vs. perceived scene and memorized face) were successfully classified by linear decoders, indicating that perception and working memory are represented in partially distinct subspaces. Such specific coding may help prevent confusion between perception and memory. Meanwhile, however, cross-decoding of visual contents across perception and working memory was also robust, indicating common coding components between them. Such common coding may help to explain why even though we generally have no difficulty in distinguishing working memory content from perceptual content, there is yet a sense in which their contents are qualitatively similar. Hereby, we pre-register the hypotheses, experimental design, analysis plan, and stopping rule for data collection before examining whether these preliminary results can be replicated in a larger sample.

Introduction

Our visual experiences have at least two phenomenologically distinct modes: perception and working memory. Perception refers to the representation of external stimuli that are present in the environment, whereas working memory refers to the actively maintained representation of stimuli that were present in the recent past but are no longer physically available. Typically, when needed, working memory follows perception, in order for us to maintain the content of what we have just perceived. The working memory content is thus similar to the perceptual content, and yet, we rarely confuse the two.

How does the human brain maintain the same (or similar) content, and yet distinguishes between perception and working memory (the issue of perceptual reality monitoring; Lau, 2019; Dijkstra et al., 2025)? One possibility is that certain brain regions represent visual content irrespective of whether it is perceived or maintained in working memory. In contrast, some other regions possibly represent the content in a format that distinguishes perception and working memory. In the present study, using fMRI, we sought to identify such brain areas by examining the neural representational geometry (e.g., Bernardi et al., 2020; Chung & Abbott, 2021; Ebitz & Hayden, 2021; Haxby et al., 2014; Kriegeskorte & Kievit, 2013) when observers perceived and held in working memory either face or scene stimuli (see Figure 1).

Our interest in representational geometry was inspired by several previous neurophysiological studies. For example, Libby and Buschman (2021) found that the stimulus coding axis in the subspace defined by the neuronal population in the sensory cortex *rotates* when a perceptual representation transforms into working memory. They proposed that such rotation may serve to protect memory content from interference by newly incoming perceptual information (see also Panichello & Buschman, 2021). Accordingly, we tested if some areas of the brain exhibit a similar kind of representational rotation, as captured by multivariate fMRI data (Aim 1).

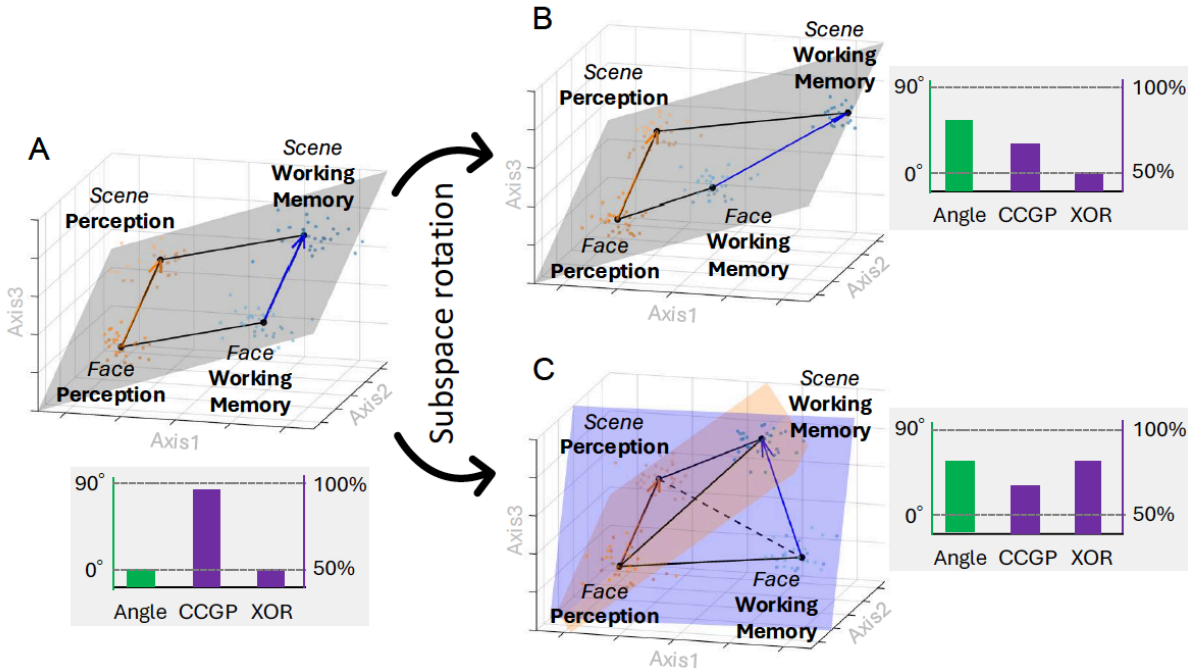

Figure 1. Schematics of hypothetical representational geometry in the present study. *A) Case where the coding axes distinguishing faces from scenes are parallel across perception (orange arrow) and working memory (blue arrow). In this case, the angle between the coding axes is zero, cross-condition generalization performance (CCGP) is perfect, and* exclusive-or (*XOR) decoding is at chance. B) Case where 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 is at chance. C) 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 higher dimensional vertex space and all cases (A, B, and C) presuppose that the ROI has reliable face-vs-scene tuning both for perception and working memory (see Figures A3 and A5).*

Importantly, such rotation of the coding axis can occur within the same subspace (Figure 1B), or it can result in the definition of distinct subspaces before and after rotation (Figure 1C). Only the latter type of rotation would increase the effective dimensionality of the entire representational space. Therefore, it is also intriguing to examine the dimensionality of subspace jointly defined by perceptual and mnemonic representations.

To evaluate the dimensionality in noisy neural data, it is useful to assess 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 (Fusi et al., 2016). Especially in our case having two variables, each with two levels (e.g., the combination of content: *face* vs. *scene* and

cognitive process: *perception* vs. *working memory*), three balanced (2-to-2) dichotomies can be defined: the content dichotomy (*face perception* and *face working memory* vs. *scene perception* and *scene working memory*), the cognitive-process dichotomy (*face perception* and *scene perception* vs. *face working memory* and *scene working memory*), and the exclusive-or (XOR) dichotomy (*face perception* and *scene working memory* vs. *face working memory* and *scene perception*). While successful decoding of the first and second dichotomies indicates the presence of independent content and cognitive-process codings, successful XOR decoding is a hallmark of a three-dimensional representational structure, which is the maximum attainable for this two-by-two variable combination (Bernardi et al., 2020).

Such high-dimensional representational structure is considered to be a characteristic feature in the association cortices, such as the prefrontal cortex, to support flexible behavior in complex tasks (Rigotti et al., 2013; Parthasarathy et al., 2017). Moreover, this high-dimensionality usually arise from the neuron's nonlinear mixed selectivity to multiple variables (Rigotti et al., 2013). In the context of perceptual reality monitoring (Lau, 2019; Dijkstra et al., 2025), indeed, 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 partially distinct neuron populations, while some neurons display similar tuning for both, indicating there is information enough to distinguish perception and working memory. However, it remains unclear whether such mixed selectivity to content and cognitive process is linear or nonlinear and how it shapes the representational geometry at the neural population level. Therefore, in this study we also examined XOR decoding accuracy, as a sign of nonlinear mixed selectivity (Aim 2).

If the coding axis for face–scene dichotomy is rotated and/or representation is high-dimensional (in our case three-dimensional), how can the brain determine that the perceptual and mnemonic representations reflect the same content, even though their phenomenology is distinct? Here, geometry-based decoding approaches may be helpful. Bernardi et al. (2020) introduced the Cross-Condition Generalization Performance (CCGP) to quantify how well a linear separation of a variable generalizes across different contexts. In this approach, a linear decoder is trained in one context and tested in another, with the average accuracy across all context pairs representing the CCGP. Higher CCGP values reflect coding that is more invariant to context. Therefore, in our study we also attempted to identify brain areas with successful CCGP regarding visual content, i.e., successful generalization of face-vs-scene decoder across two cognitive processes, which can be a basis of the continuity of perceptual and mnemonic representations (Aim 3).

In our pilot data reported here, we employed the three complementary measures mentioned above, to compare the representational geometry of perception and working memory in the human cerebral cortex.

Methods

Observers

Four naïve observers and one author (two males and three 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 four 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.

Apparatus

The experimental code was written in MATLAB R2021b (MathWorks, Natick, MA, USA) with the Psychphysics Toolbox 3.0.19 (Brainard,1997; Pelli, 1997) and was run on a Ubuntu 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 (Ma et al., 2015), comprising 4 sub-categories defined by the combination of race and sex

(Black/Female, Black/Male, White/Female, and White/Male). Scene stimuli were drawn from the “Massive Memory” dataset (Konkle et al., 2010), 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 (Serengli & Özpınar, 2024), which were further transformed to better reflect how humans perceive similarity between faces (unpublished work from our group). 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 (Zhou et al., 2017; “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 (Thaler et al., 2013), 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 2). 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 phase, followed by 10s delay period (working memory phase), during which only the fixation point was shown. The working memory phase 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. To reduce regressor 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 such trials. Face-cued and scene-cued trials are performed in equal numbers and randomized order within each run.

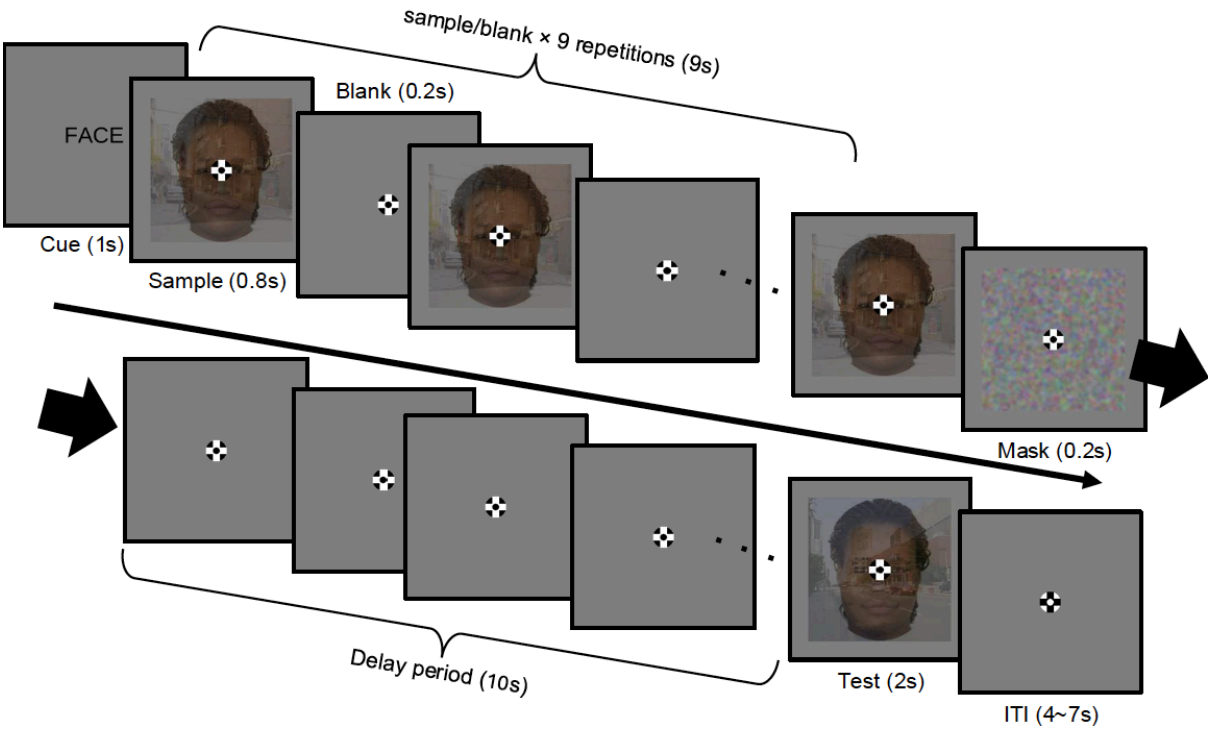

Figure 2. Time course of a single trial. *In each trial, observers were instructed to memorize only the cued aspect (either a face or a scene) of the sample stimulus, which was repeatedly presented during the perception phase. Following a delay period (working memory phase), a test stimulus appeared 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) and Konkle et al. (2010).*
Abbreviations: ITI, inter-trial interval

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. 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 (Kwak & Curtis, 2022). This control was important because we were interested in the phenomenological difference between perception and working memory.

MRI data aquisition

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 for each of 4 scanning sessions, in total 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.

At the beginning of the first session, structural T1-weighted images were acquired. 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

Preprocessing of MRI data was performed using fMRIPrep 23.2.1 (Esteban et al., 2018; Esteban et al., 2019; RRID:SCR_016216), which is based on Nipype 1.8.6 (Gorgolewski et al., 2011; Esteban et al., 2025; RRID:SCR_002502).

Preprocessing of B0 inhomogeneity mappings

A B0-nonuniformity map (or fieldmap) was estimated based on two echo-planar imaging (EPI) references having opposite phase-encoding directions with topup (Andersson, Skare, & Ashburner, 2003; fsl-topup v2203.2). In total, 32 field maps were obtained for each observer, with each applied to the corresponding subsequent functional run to correct for susceptibility distortions: due to MR technical issues, the posterio-to-anterior image for the 3rd run in the 1st session of Obs. 2 could not be corrected; in this case, we instead use the posterior-to-anterior image from the 4th run to estimate the fieldmap for the 3rd run.

Anatomical data preprocessing

One T1-weighted (T1w) image was acquired from each observer. The T1w image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al., 2010), distributed with ANTs 2.5.0 (Avants et al., 2008; RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM), and gray-matter (GM) was performed on the brain-extracted T1w using fast (fsl-fast4 v2111.3, RRID:SCR_002823, Zhang, Brady, & Smith 2001). Brain surfaces were reconstructed using recon-all (FreeSurfer 7.3.2, RRID:SCR_001847, Dale, Fischl, & Sereno 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (RRID:SCR_002438, Klein et al., 2017). Volume-based spatial normalization to two standard spaces (MNI152NLin6Asym, MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.5.0), using brain-extracted versions of both T1w reference and the T1w template. The following templates were selected for spatial normalization and accessed with TemplateFlow (23.1.0, Ciric et al., 2022): FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model (Evans et al., 2012; RRID:SCR_002823; TemplateFlow ID: MNI152NLin6Asym), ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009; RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym).

Functional data preprocessing

For each of the 32 BOLD runs per observer (across all sessions), the following preprocessing was performed. First, a reference volume was generated, using a

custom methodology of fMRIPrep, for use in head motion correction. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (fsl-mcflirt v2111.0, Jenkinson et al., 2002). The estimated fieldmap was then aligned with rigid-registration to the target EPI (echo-planar imaging) reference run. The field coefficients were mapped on to the reference EPI using the transform. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve & Fischl, 2009). Co-registration was configured with six degrees of freedom. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions, Power et al., 2014) and Jenkinson (relative root mean square displacement between affines, Jenkinson et al., 2002). FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, Behzadi et al., 2007). Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM, and combined CSF+WM) are generated in anatomical space. The implementation differs from that of Behzadi et al. (2007), in that instead of eroding the masks by 2 pixels on BOLD space, a mask of pixels that likely contain a volume fraction of GM is subtracted from the aCompCor masks. This mask is obtained by dilating a GM mask extracted from the FreeSurfer's aseg segmentation, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. Additional nuisance timeseries are calculated by means of principal components analysis of the signal found within a thin band (crown) of voxels around the edge of the brain, as proposed by (Patriat, Reynolds, & Birn 2017). The BOLD

time-series were resampled onto the following surfaces (FreeSurfer reconstruction nomenclature): fsaverage. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e., head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using nitransforms, configured with cubic B-spline interpolation. Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Many internal operations of fMRIPrep use Nilearn 0.10.2 (Abraham et al., 2014, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in fMRIPrep's documentation.

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 statistical results except where noted.

General linear model

We fitted general linear models (GLMs) to the BOLD time courses per each vertex in the average space (FreeSurfer format called "fsaverage") using the run_glm function of Nilearn (version 0.12.0). For each run, the perception and working memory phases were modeled with four separate regressors corresponding to Face-cued Perception (FP), Scene-cued Perception (SP), Face-cued Memory (FM), and Scene-cued Memory (SM) conditions, each of which were the 10-s box-car function convolved with the standard SPM hemodynamic response function (Figure A1). Note that the perception phase included cue and mask in addition to the sample/blank periods while the working memory phase just included the entire 10s delay period. Responses to test images (appeared only on the half of trials) were regressed 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 (FD) exceeds 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, FM, and SM conditions) × N (number of vertices) per run. For each run, the beta matrix was 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 (Prince et al., 2022). First, we upsampled the BOLD time series to a TR of 1 s (Kay et al., 2020), ensuring that

event durations were integer multiples of the TR. Each event (FP, SP, FM, SM, and the test stimulus) was then modeled as a sequence of consecutive 1 s events. Finally, the estimated betas were averaged within each event. This “trialwise” GLM produced a beta matrix of size 2 (perception and working memory phases) × 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 standardized within ROI prior to decoding analyses.

ROI definition

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) (Figure 3). In addition, as an exploratory analysis and for visualization purposes only, we also use a more fine-grained definition of ROIs, namely the original 180-region HCP-MMP atlas. We analyzed homologous regions in the left and right hemispheres as a single ROI.

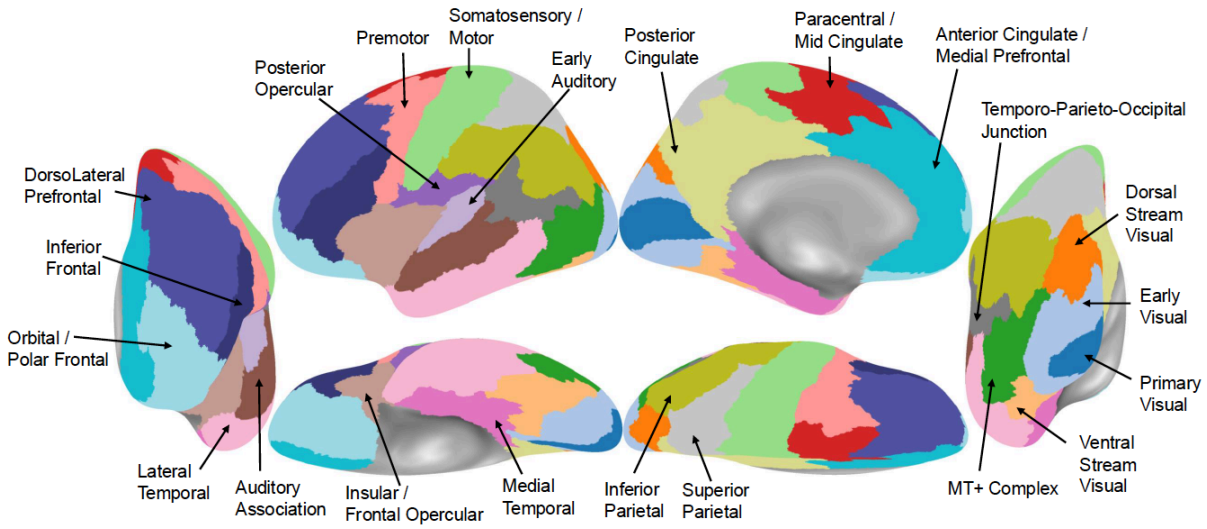

Figure 3. Definition of the region-of-interests (ROIs) used for the main analyses. *Different colored regions indicate each of 22 coarse-grained ROIs defined in Glasser et al. (2016).*

Rotation angle of the coding axis

To examine the rotation between coding axes for perceptual and mnemonic representations, we measured the angle between the face–scene coding axes (defined by runwise GLM beta patterns for face-cued vs. scene-cued trials) across the two cognitive processes. 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 phase within each ROI. We averaged the coding axes across runs

and computed the angle between the coding axes of the perception and working memory phase:

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

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 Appendix 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 phases, 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 phases both occurred on the same trials (see Figure 2), 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.

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

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

$V_p(observed)$ 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) to classify contents (face vs. scene) irrespective of cognitive process (perception vs. working memory). We also evaluated content decoding accuracy separately within the perception phase and within the working memory phase.

Decoding cognitive process

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.

Exclusive-or (XOR) decoding

More importantly, we also trained the 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 (cognitive process and content) are entangled within a stable three-dimensional representational subspace in that ROI (Bernardi et al., 2020).

Cross-Condition Generalization Performance (CCGP)

To assess the generalizability of a hyperplane distinguishing face–scene dichotomy across cognitive processes, we also measured the cross-condition generalization performance (CCGP; Bernardi et al., 2020; Boyle et al., 2024). In the case of two-by-two factorial design, we can define *CCGP for content* as the average of cross-decoding accuracy in both directions (trained a classifier to distinguish contents only in perception phase and tested it in working memory phase and vice versa). For the sake of completeness, we also measured the *CCGP for cognitive process* (trained a classifier to distinguish cognitive processes 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 phase and within working-memory phase, to show the degree of generalization of content coding (Xu, 2025).

Decoding sub-categories

With regard to content 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 (Serences et al., 2004). 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 accuracy was assessed using 8-fold cross-validation on trial-wise GLM beta patterns, 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 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. 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 (Valente et al., 2021).

Results

Behavior

Across observers, the task was performed with high accuracy, indicating that they understood and followed the instructions correctly (Figure A2). The mean correct response rate was 93% and the proportion of no-response trials was low (mean = 1.0%). There was no significant difference in the correct response rate between face-cued and scene-cued trials at an individual level, Obs. 1: $\chi^2(1) = 0.05$, $p = .830$; Obs. 2: $\chi^2(1) = 0.00$, $p = 1$; Obs. 4: $\chi^2(1) = 1.16$, $p = .281$; Obs. 5: $\chi^2(1) = 3.82$, $p = .051$. However, reaction times (RTs) tended to be faster for scene-cued trials. Welch's t-test on the log-transformed RT revealed that three out of four observers showed significantly shorter RTs for scene-cued than for face-cued trials, Obs. 1: $t(229) = 2.50$, $p = .013$; Obs. 2: $t(235) = 9.80$, $p < .001$; Obs. 4: $t(244) = 0.33$, $p = .011$; Obs. 5: $t(211) = 2.58$, $p = .013$. 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.

Rotation angle of the coding axis

Prior to assessing the rotation, we assessed the robustness of face-vs-scene multivertex activation pattern within each ROI (Figure A3) by applying a test-retest reliability approach (e.g., Haxby et al., 2001). Separately for the perception and working memory phase, we calculated the difference of multivertex beta patterns between face and scene and then obtained the Pearson's correlation coefficients of these differential patterns between different halves of runs. A distribution of correlation coefficients was obtained from 10,000 different splits and tested against zero. Although one observer did not show reliable face-vs-scene tuning (particularly

for working memory) in the majority of the ROIs, the other three observers demonstrated significant reliability both for perception and working memory in most of the visual cortex (primary, early, dorsal, and ventral stream) and the parietal association cortex (superior parietal, inferior parietal, and posterior cingulate) as well as in the premotor, medial temporal, temporo-parieto-occipital junction, and inferior frontal cortex.

On this basis, we proceeded to measure 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. As shown in Figure 4A, not only low- and high-level visual cortices but also parietal and frontal association cortices exhibited larger across-cognitive-process angles compared to the noise floor.

As a formal statistical test, we calculated the z-values of the angle deviation from the noise floor. One observer consistently showed no significant rotation of face-vs-scene coding axes across all areas (except for the ventral stream visual cortex), due to a higher noise floor relative to the others. Therefore, when we focused on ROIs in which at least three out of four observers showed significant rotation, we found the early and ventral stream visual cortex within the visual areas. Moreover, in the association areas, temporo-parieto-occipital junction, inferior parietal, posterior cingulate cortex and orbital/polar frontal cortex also showed evidence of rotation (Figure 4B).

Additionally, we also examined the angle rotation in a low-dimensional subspace using principal component analysis (PCA). In this additional analysis, instead of using Equation 2, we estimated the noise floor by splitting runs into two halves and calculating the angle between coding axes from the same cognitive process (perception or working memory phase) but based on different halves of runs (see Figure A4 for details). However, since we found this split-half approach less efficient compared to our main approach (using one half of the data to find PCs forced us to rely only on the other half of the data for the angle rotation analysis, thus it reduced the reliability of angle estimation), we presented it in Appendix. Nevertheless, the overall trends (while a bit weaker) were similar to the results from our main analysis.

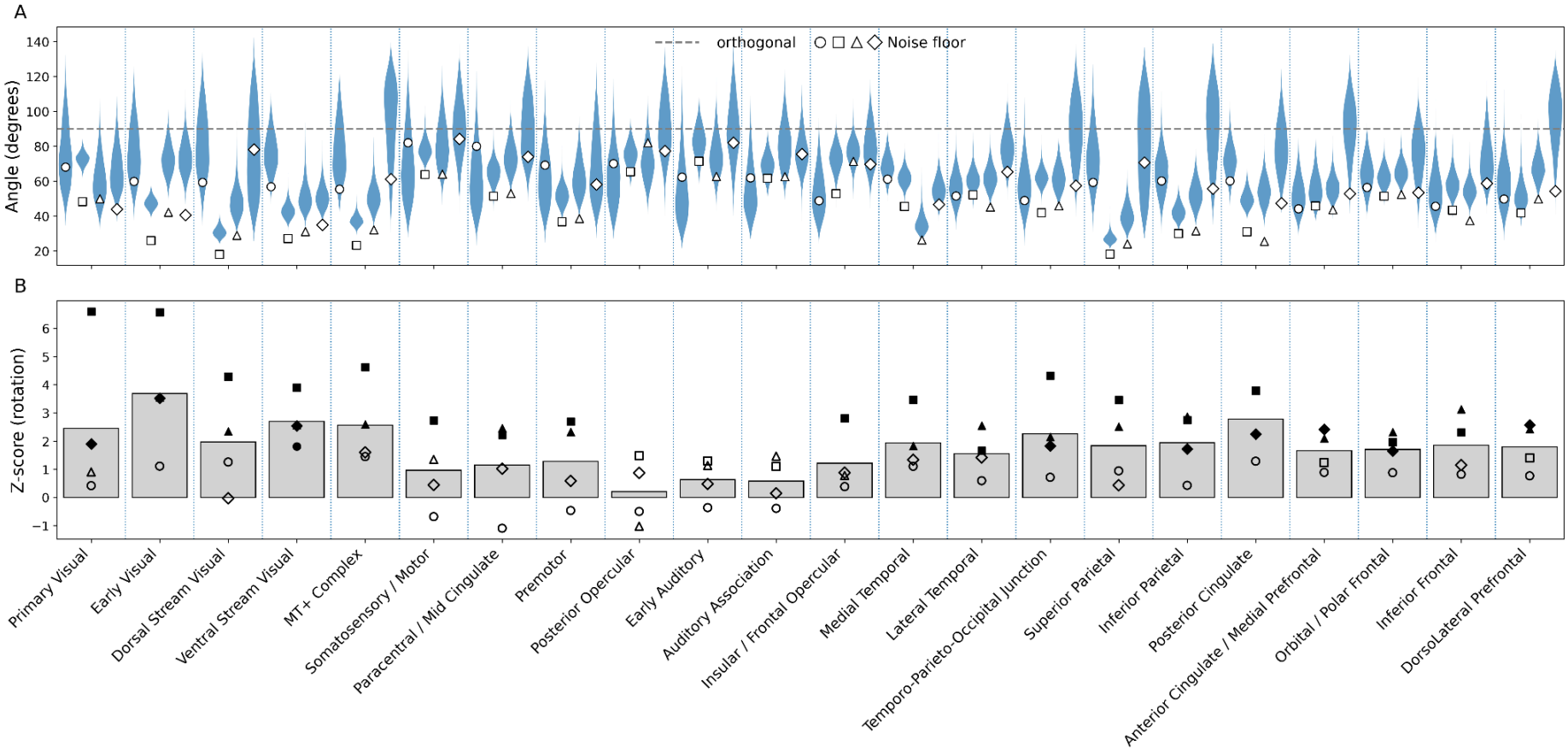

Figure 4. Results of angle analysis. *A) Bootstrapped distribution of angle between face-vs-scene coding axis for perception and working memory was shown with the noise floor. Different marker shapes indicate the noise floor from different observers. B) Z-values (derived using distribution and noise floor in A) for all observers and ROIs. Bar graphs indicate the inter-individual mean. Filled markers indicate observers who exhibit individual-level significance (FDR-corrected), whereas unfilled markers indicate those who do not. In the majority of observers, the coding axis rotation across perceptual and mnemonic contents was evident in some visual and parietal / frontal association areas. For Obs. 2 (square marker), it should be noted that the lateral temporal cortex did not exhibit a significant rotation when the runs with poor image quality were excluded.*

Geometry assessed by decoding

In the following decoding results, we also report the ROIs in which at least three out of four observers exhibited statistical significance at an individual level (Figure 5). First, the cognitive process (perception vs. working memory) was highly decodable from all ROIs in all observers. This may be because the dichotomy between perception and working memory reflects the mere presence or absence of the stimulus as well as the cognitive process operating on it.

Second, the content (face vs. scene) was decodable from the entire visual cortex, the paracentral areas, and the entire temporal, parietal, and frontal association cortices, but not reliably from the early auditory, somatosensory-motor, insular or opercular regions.

Third, we were able to decode the XOR dichotomy not only from the visual cortex (primary, early, dorsal stream, ventral stream, and MT+ complex) but also from the association cortex (temporo-parieto-occipital junction, superior parietal, inferior parietal, posterior cingulate, and inferior frontal cortices). These areas showed substantial overlap with the regions exhibiting rotation of the face-vs-scene coding axes (Figure 4). Therefore, in these areas, the face-vs-scene coding axes do not simply rotate within the same subspace but instead define partially separate subspaces for perception and working memory.

Fourth, among the ROIs from which both variables (cognitive process and content) were decodable, we also observed consistently high CCGP. This indicates that most areas coded the content (and cognitive process) information in a partially cognitive-process-invariant (and content-invariant) format that could be generalized across different cognitive processes (and contents). The content cross-decoding ratio indeed exceeded 90% in all areas except primary visual cortex. Although the current sample size is not yet sufficient to assess group-level significance for the cross-decoding performance drop, the numerical trend was intriguing: the ROIs with reliable XOR decoding tended to exhibit greater cross-decoding performance drop. Such a tradeoff between flexibility and abstraction is theoretically predicted (Bernardi et al., 2020), and here we provide the first evidence of their co-occurrence in human fMRI data. However, flexibility never completely eliminated abstraction in any ROI, which is consistent with electrophysiology data from non-human primates using context-dependent associative tasks (Bernardi et al., 2020).

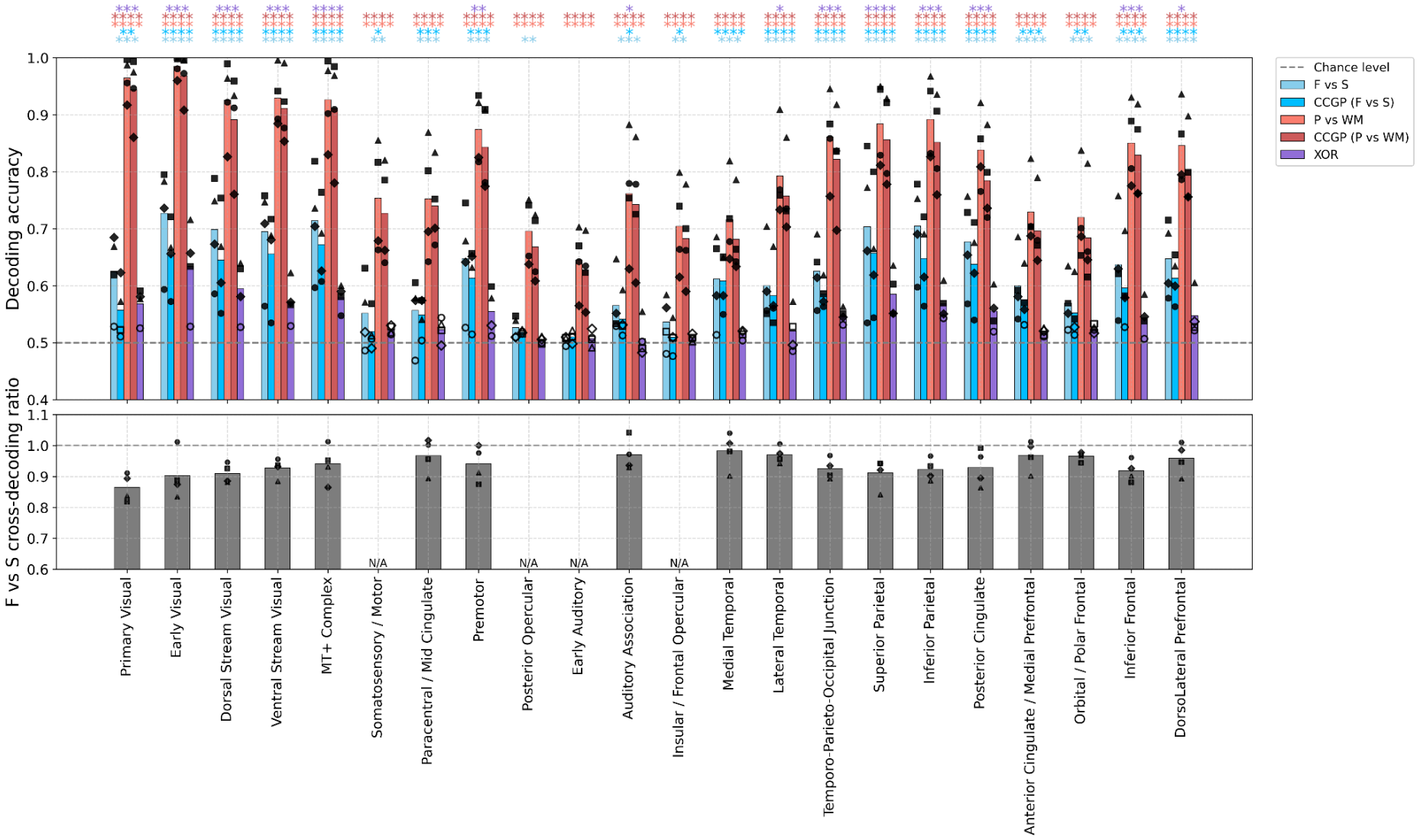
Decoding accuracy
F vs S cross-decoding ratio
Chance level
F vs S
CCGP (F vs S)
P vs WM
CCGP (P vs WM)
XOR
N/A
Primary Visual
Early Visual
Dorsal Stream Visual
Ventral Stream Visual
MT+ Complex
Somatosensory / Motor
Paracentral / Mid Cingulate
Premotor
Posterior Opercular
Early Auditory
Auditory Association
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 5. Results of decoding, cross-condition generalization performance (CCGP) and exclusive-or (XOR) decoding. *A) Decoding accuracy between face-cued and scene-cued trials, regardless of cognitive process, is shown in light blue, and decoding accuracy between perception and working memory phases is shown in light red. CCGP for face–scene dichotomy (trained to decode face vs. scene in one cognitive process and tested in the other) is shown in deep blue, and CCGP for perception-memory dichotomy (trained to decode perception vs. working memory in one content and tested in the other) is shown in deep red. Individual observer data are plotted around the bars (which show the interindividual mean) using markers of different shapes. Filled markers indicate that decoding accuracy for that observer in the given ROI was significantly higher than the permuted null distribution. The number of asterisks above each bar indicates the number of observers showing significance at the individual level (FDR-corrected among all ROIs). We found that the dichotomy between perception vs. working memory was decodable and generalizable across faces and scenes in all ROIs of all observers. The dichotomy between face vs. scene was decodable and generalizable across perception and working memory in the entire visual cortex, the paracentral areas, and the temporal, parietal, and frontal association cortices. XOR dichotomy was decodable from most of the visual areas and some of the association areas (especially in the parietal cortex), indicating that these areas represent perceptual and mnemonic contents in partially different subspaces. 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 when they are performed within the same cognitive process (see Figure A5). Individual observer data are plotted around the bars (which show the interindividual mean) using markers of different shapes. N/A indicates the ROIs wherein content was not reliably decoded in the first place. Overall, cross-decoding performance drop was minimal in all areas from which face vs. scene was decodable (less than 10% except for primary visual cortex), indicating that visual content is represented in a manner generalizable across perception and working memory in those areas. When the runs with poor image quality were excluded, the following statistical results changed: for Obs. 2 (square), XOR in the orbital / polar frontal cortex and CCGP (F vs. S) in the posterior opercular cortex became significant, whereas XOR in the posterior cingulate cortex became insignificant; for Obs. 5 (diamond), XOR in the premotor cortex and CCGP (F vs. S) in the auditory association and orbital / polar frontal cortex became significant.*

Beyond face-vs-scene decoding, we successfully decoded scene sub-categories from visual areas, including primary, early visual cortex, the dorsal and ventral visual streams, and the MT+ complex, as well as from the entire parietal cortex (superior parietal, inferior parietal, and posterior cingulate cortex) and medial temporal cortex (Figure 6A). 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 6B). The most decodable dichotomy was naturalness (natural vs. manmade scenes), which was decoded from some visual cortices (early visual, ventral visual stream and MT+ complex) and from parietal association cortices (superior parietal and inferior parietal). The second most decodable dichotomy was openness (open vs. closed scenes), which was decoded only from early visual areas.

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 parietal cortices, 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 (Serences et al., 2004).

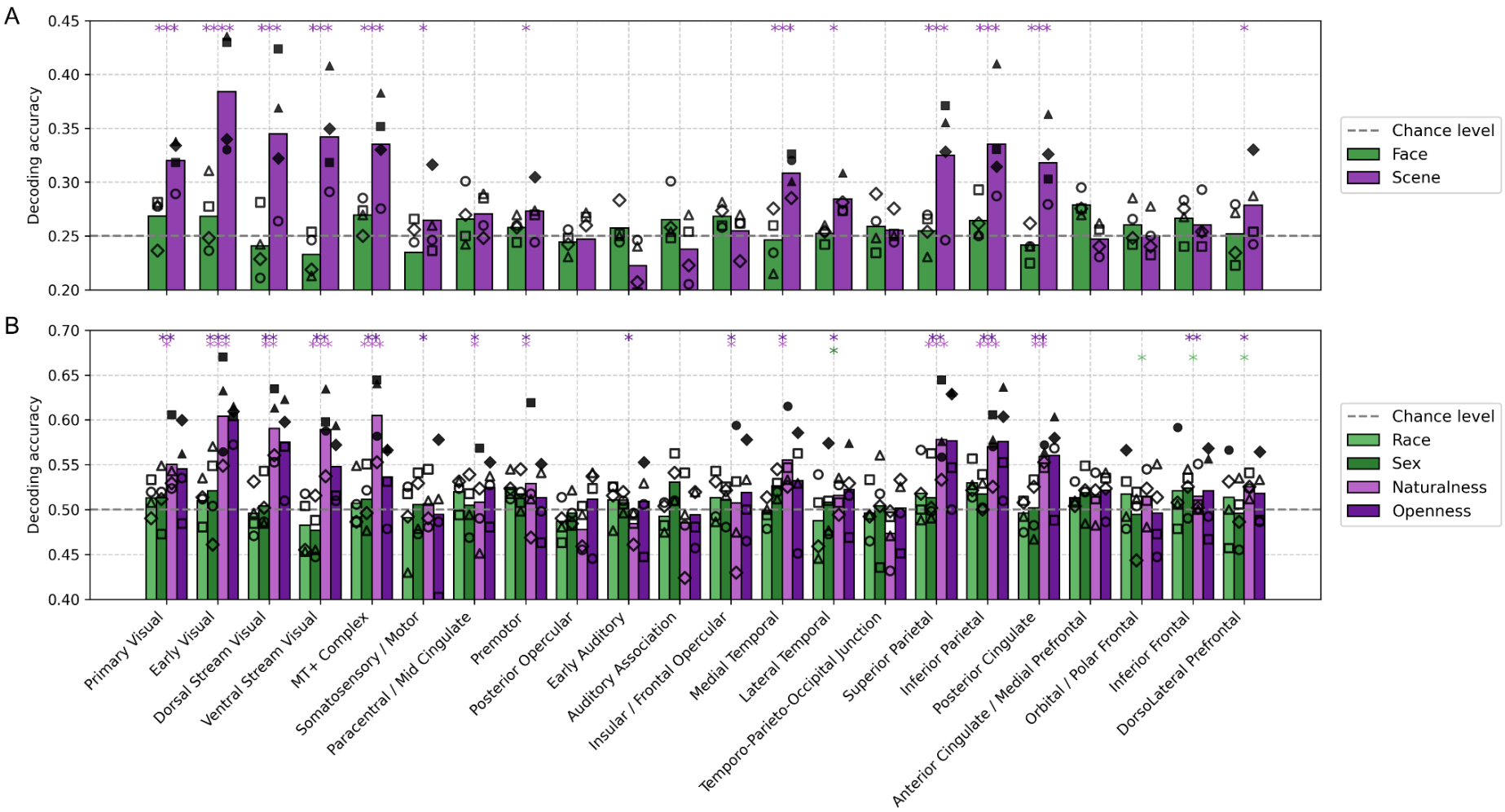
A
B
Decoding accuracy
0.45
0.40
0.35
0.30
0.25
0.20
0.70
0.65
0.60
0.55
0.50
0.45
0.40
Chance level
Face
Scene
Race
Sex
Naturalness
Openness
Primary Visual
Early Visual
Dorsal Stream Visual
Ventral Stream Visual
MT+ Complex
Somatosensory / Motor
Paracentral / Mid Cingulate
Premotor
Posterior Opercular
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 6. 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. The conventions for the markers and asterisks are the same as Figure 5. Overall, for the majority of observers, scene sub-categories were decodable from visual areas and from some association areas, particularly in the parietal cortex. In these regions, face–scene dichotomy cannot be explained solely by differences in attentional window size or by language processing driven by the word cue; rather it reflects at least some difference in visual phenomenology. When the runs with poor image quality were excluded, the following statistical results changed: for Obs. 2 (square), openness in the dorsal stream visual cortex and sex in the auditory association cortex became significant, whereas scene in the posterior cingulate cortex and naturalness in the dorsolateral prefrontal cortex became insignificant; for Obs. 5 (diamond), scene in the medial temporal cortex, naturalness in the early, dorsal stream visual cortex, and MT+ complex became significant, whereas scene in the posterior cingulate and dorsolateral prefrontal cortex, openness in the paracentral / mid cingulate, early auditory, and insular / frontal opercular cortex, sex in the lateral temporal cortex, race in the orbital / polar frontal cortex became insignificant.*

Discussion

Although the results are still preliminary due to the small sample size, we found rotation of the content-coding axis from perception to working memory in three out of four observers, not only in the visual cortex but also in the association cortex. Furthermore, this rotation does not appear to occur within the same subspace but rather entails the creation of new subspaces, as indicated by successful XOR decoding in similar areas.

In most areas from which we were able to decode contents (face vs. scene), however, CCGP was also high, and the cross-decoding performance drop was minimal (~10% except for the primary visual cortex). This suggests that visual contents are largely coded in a format that is invariant across cognitive processes, regardless of their position in the processing hierarchy. At least in early visual areas, our results align with previous studies that have assessed cross-decoding between perception and working memory. For example, Harrison and Tong (2009) showed that orientations maintained in working memory were decodable from the early visual cortex, with partial generalization across actual perception and working memory. Furthermore, Albers et al. (2013) demonstrated that decoders trained on perceptual trials could successfully decode not only working memory but also imagined orientations. However, these earlier studies were limited to the early visual cortex, did not rigorously match attentional conditions between perception and working memory, and typically evaluated only one decoding direction (unlike bidirectional CCGP). Therefore, our study hopefully represents a small step towards a more stringent assessment of representational generalization across the entire human cortex, in the context of perception and working memory.

Notably, the cross-decoding performance drop tended to be greater in areas from which XOR was reliably decoded. This captures a tradeoff between flexibility (indicated by above-chance XOR decodability) and abstraction (indicated by above-chance CCGP), which may help prevent confusion between perceptual and mnemonic content, while simultaneously supporting the experience of continuity of them.

We hereby treat this preprint as a form of pre-registration for the following replication study in which we invite another 4–8 observers with the same experimental procedures, except that each observer will undergo two structural scans (cf. one scan in the pilot) to improve co-registration quality. As is the case with those who have already participated, any new observer whose average correct response rate falls below 75% in a session will not be invited back for further sessions, and their data will be excluded from all analyses unless they finish at least 3 sessions.

We plan to perform group-level statistical analyses on all observers including pilot data. For the angle rotation, group-level significance will be evaluated using one-tailed one-sample t-tests on the individual z-values (against zero). For the

decoding analysis, one-tailed one-sample t-tests will be performed on both the z-values (against zero) and the log-transformed decoding accuracies (against chance levels). All p-values will be corrected for multiple comparisons across all ROIs using the false discovery rate (FDR; Benjamin & Hochberg, 1995), as was done for the pilot results.

In addition, because some of the effects are expected to be detectable even at the individual level, we will also test the universality of individual-level effects across the population using the method proposed by Schwartzkopf and Huang (2024). And our stopping rule of data collection will be based on this universality. We will stop collecting data once more than half of the 22 ROIs reach the criterion of sufficient universality (likelihood ratio > 50 following Schwartzkopf & Huang, 2024) in any of the three measures of interest (coding axis rotation, CCGP for face–scene dichotomy, and XOR decoding), regardless of whether the specific ROIs overlap across measures. Assuming a reasonable sensitivity of $1 - \beta = .9$, this will correspond to finding at least 5 out of 8 (6 out of 10 or 7 out of 12) observers showing significant effects or finding at least 6 out of 8 (7 out of 10 or 8 out of 12) observers showing insignificant effects (with $\alpha = .05$ after correcting for multiple ROIs). However, we will require a minimum of 4 additional observers to assess replicability. Thereafter, if the stopping criterion is not met, we will recruit two additional observers at a time and re-evaluate the criterion after each addition, up to a maximum of 8 additional observers due to resource limitations.

References

Lau, H. (2019). Consciousness, metacognition, & perceptual reality monitoring. Preprint at PsyArXiv https://doi.org/10.31234/osf.io/ckbyf

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

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

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

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

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

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

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

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

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology, 37*, 66–74.

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

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

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

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 443–446. https://doi.org/10.1163/156856897X00357

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

Ma, Correll, & Wittenbrink (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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Evans, A. C., Janke, A. L., Collins, D. L., & Baillet, S. (2012). Brain templates and atlases. *NeuroImage, 62*(2), 911–922. https://doi.org/10.1016/j.neuroimage.2012.01.024

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

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

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

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

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage, 37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., ... & Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage, 64*, 240–256. https://doi.org/10.1016/j.neuroimage.2012.08.052

Patriat, R., Reynolds, R. C., & Birn, R. M. (2017). An improved model of motion-related signal changes in fMRI. *NeuroImage, 144*, 74–82. https://doi.org/10.1016/j.neuroimage.2016.08.051

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

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

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

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

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

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

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

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

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

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

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

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

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

Johnston, W. J., Fine, J. M., Yoo, S. B. M., Ebitz, R. B., & Hayden, B. Y. (2024). Semi-orthogonal subspaces for value mediate a binding and generalization trade-off. *Nature Neuroscience, 27*(11), 2218–2230. https://doi.org/10.1038/s41593-024-01758-5

Acknowledgements

T.N. is supported by JSPS KAKENHI Grant Numbers 24KJ0233 and 25K18957 in Japan. H.L. is supported by the Institute for Basic Science (IBS-R015-D2) in the Republic of Korea.

Competing interests

The authors declared no competing interests.

Appendix

Derivation of Equation 2

The full expression for the cosine similarity between the perceptual coding axes V_p and the working-memory coding axes 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:

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

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.

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

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

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

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

Figures

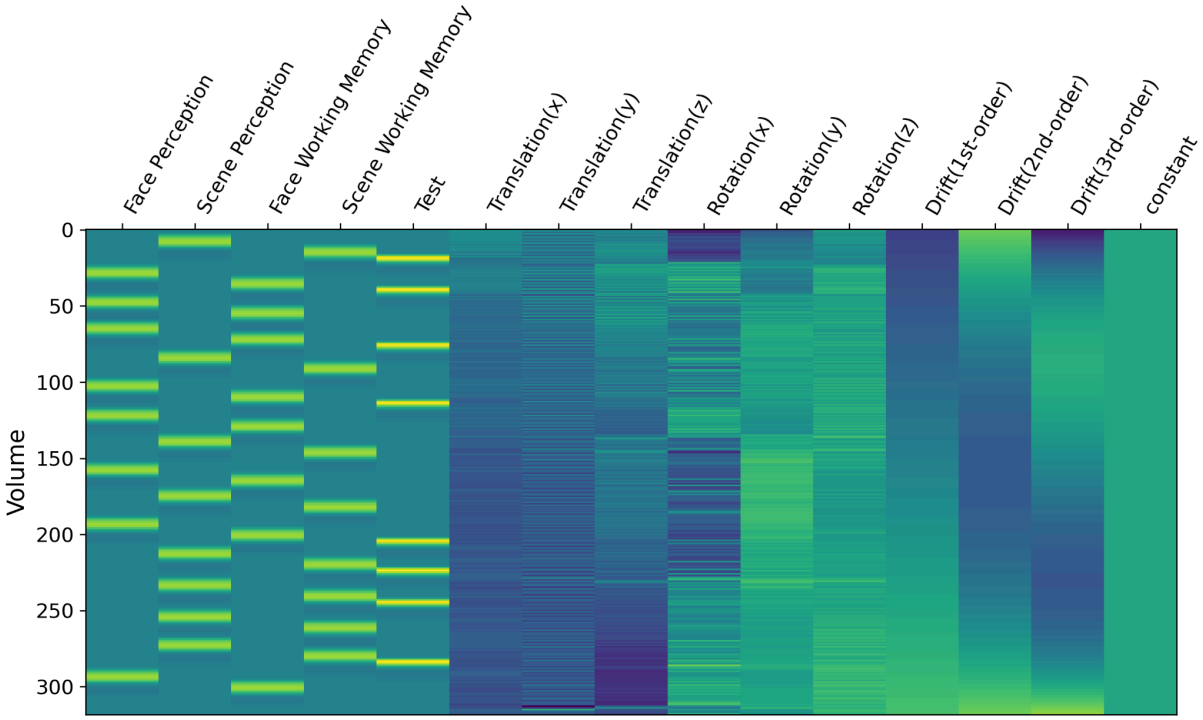

Figure A1. Design matrix for the runwise GLM. *A) Each column indicates a regressor, and each row indicates a volume. The leftmost four regressors correspond to each condition of interest. The 5th regressor from the left corresponds to test stimuli. 6th to 11th regressors are motion parameters derived 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.*

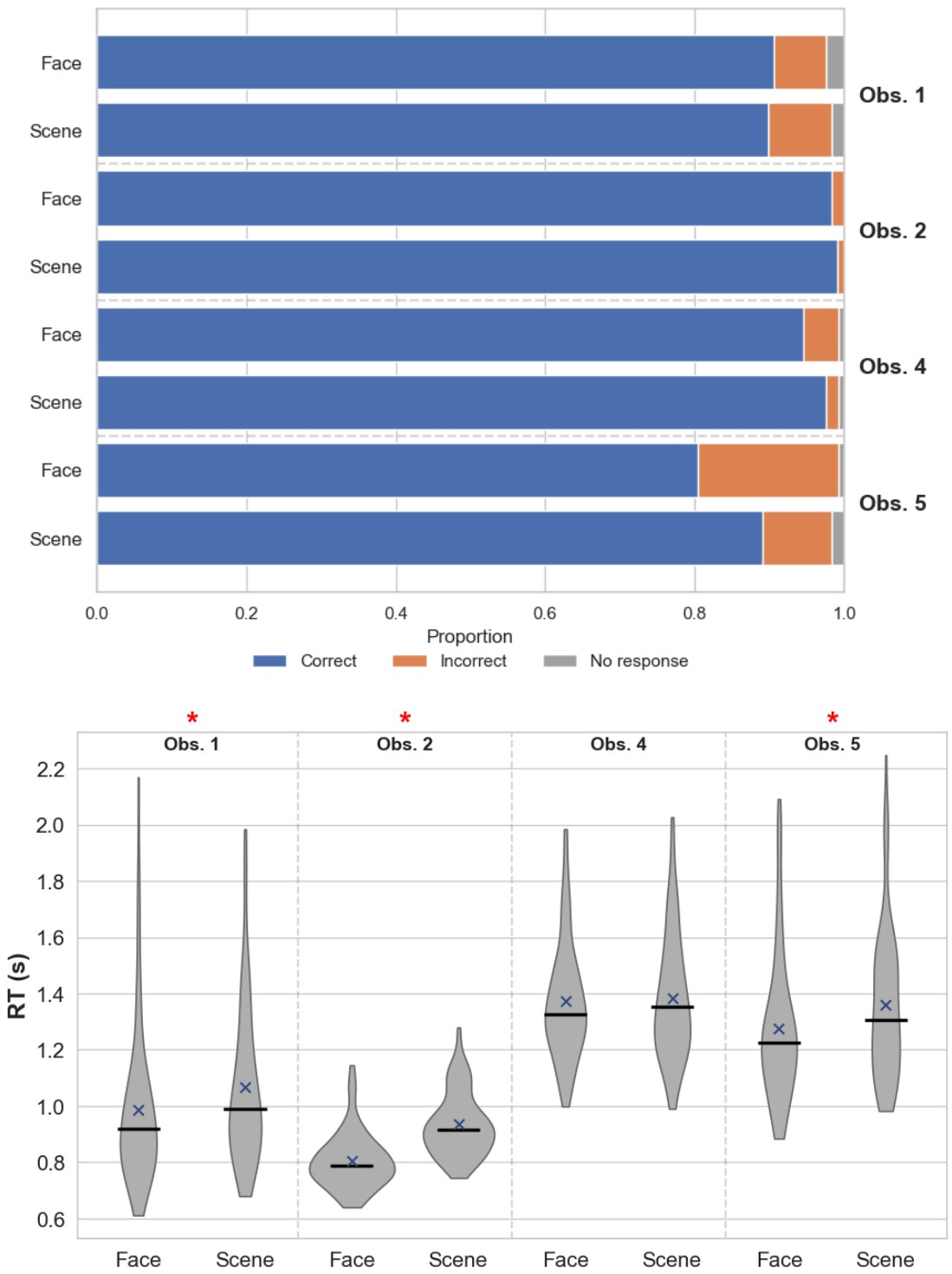

Figure A2. Behavioral data. *A) Proportion of trials resulting in correct, incorrect, and no responses for each observer. No observers exhibited significant difference of correct response rate between face-cued and scene-cued trials. B) RT distribution for each observer. x and black horizontal line indicate the mean and median, respectively. Asterisks indicate statistical significance at an individual level (p < .05).*

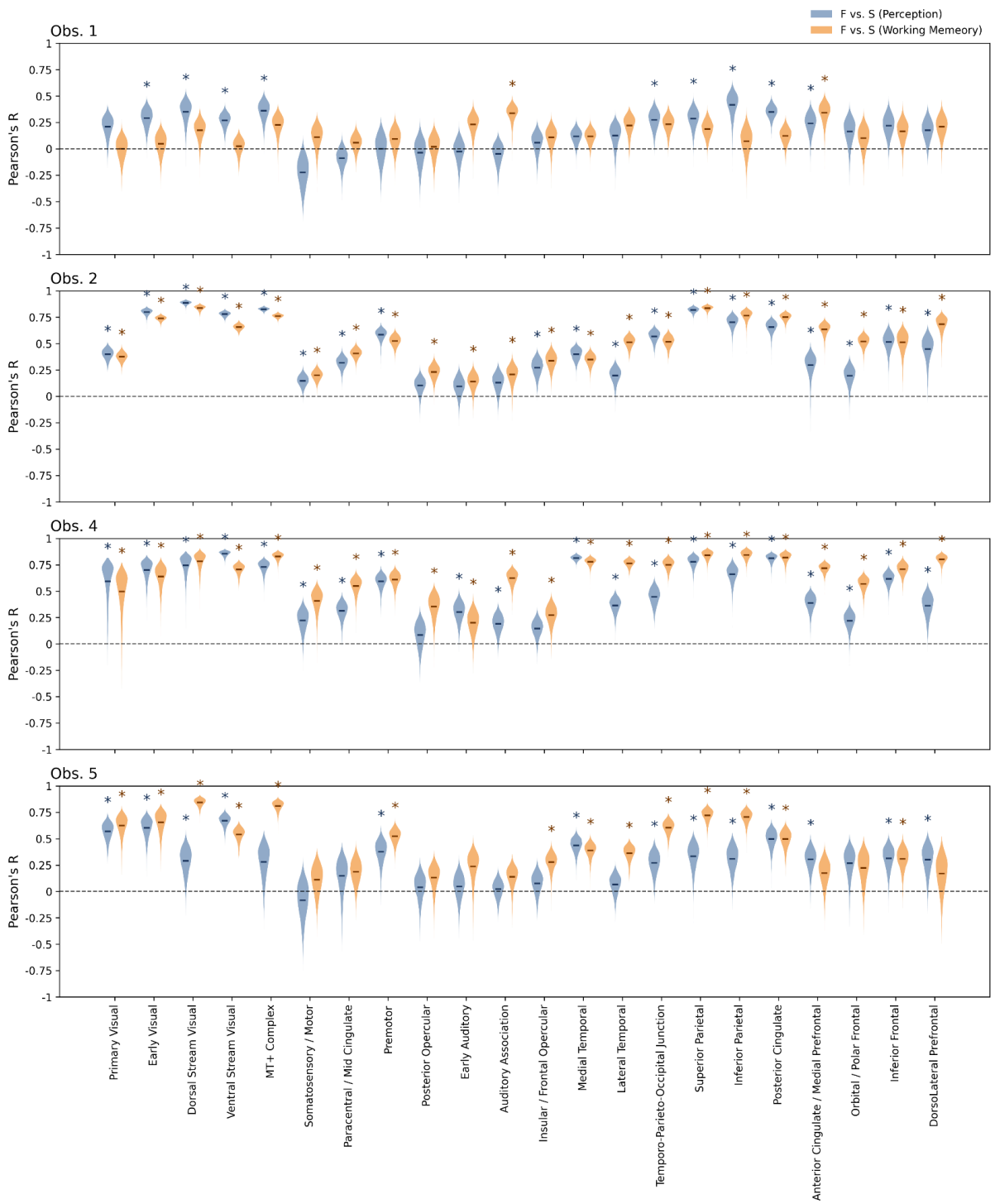

Figure A3. Test-retest reliability of face-vs-scene tuning. *Each panel indicates each observer. Asterisks indicate statistical significance after FDR-correction among all ROIs. Except for Obs. 1 (top), face-vs-scene tuning was reliable across different runs in the visual and some association areas (especially in parietal 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.*

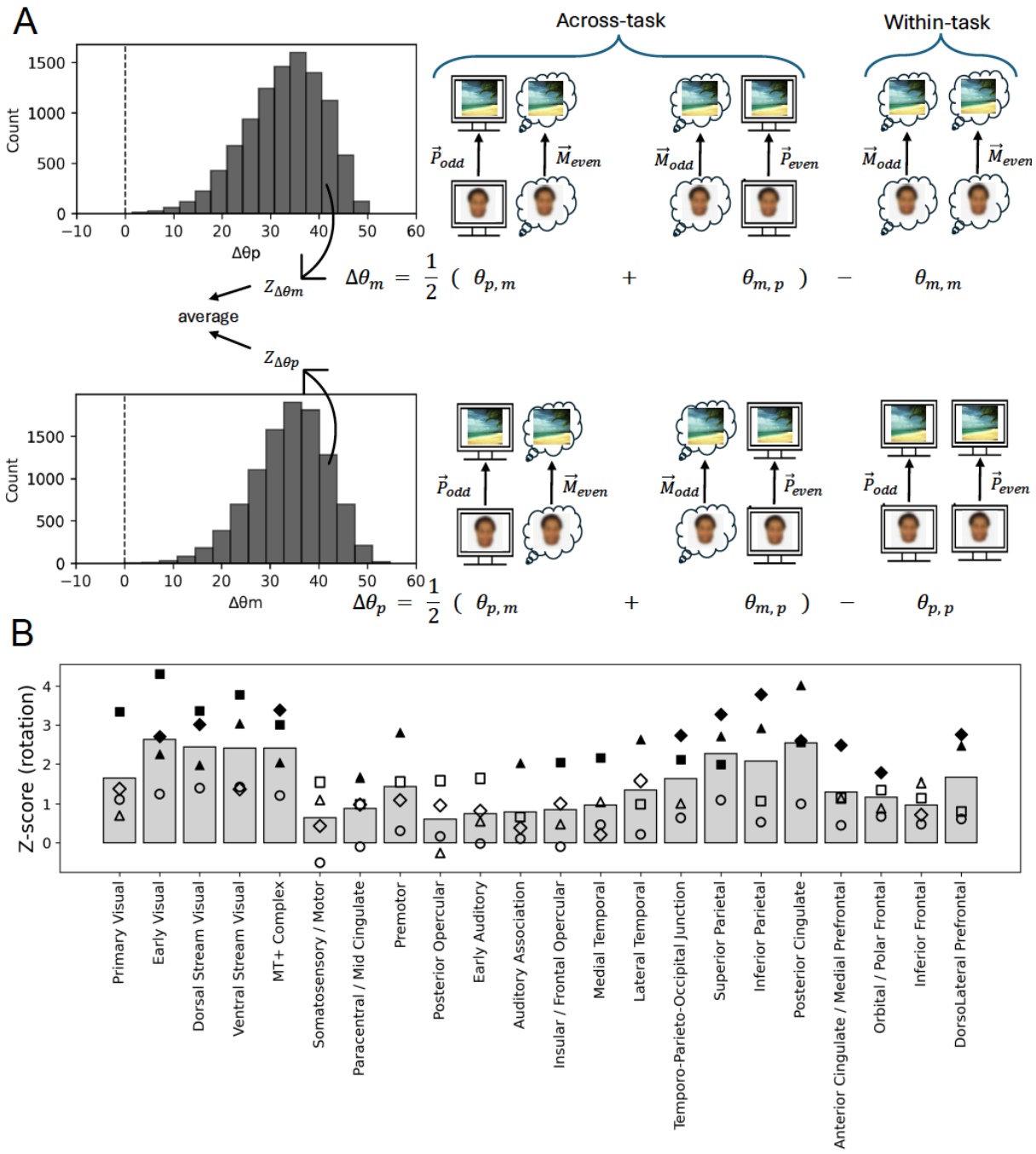

Figure A4. Angle rotation in a low-dimensional subspace. *A) To examine the rotation in a low-dimensional subspace, we applied a splitting-half approach (Johnston et al., 2024) and used principal component analysis (PCA) to get components explaining high amounts of variance from one half of the runs. First, 32 runs were randomly split into two halves. Within each ROI, 4 (conditions) × N (number of vertices) matrix was averaged across each half of the runs, in order to capture across-condition variance while minimizing across-run variability. PCA was applied to data from one half, retaining four components (the maximum attainable number of components). Then, data from the other half were projected onto this 4-dimensional PC space, which produced a 4 (conditions) × 4 (PC dimensions) loadings matrix for each half. For each split-half combination of runs, the angle*

differences were separately calculated using different noise floors (within-cognitive-process angle; $\theta_{p,p}$, $\theta_{m,m}$), which resulted in two distributions ($\Delta\theta_p$, $\Delta\theta_m$). Z-scores from the two distributions were averaged over 10,000 split combinations for each observer and ROI. Representative data from one ROI (posterior cingulate cortex) for a single observer (Obs. 4) is shown. B) Z-values (derived as in A) for all observers and ROIs. The convention for the markers is the same as Figure 4. In the majority of observers, the coding axis rotation across perceptual and mnemonic contents was evident in some visual and parietal association areas. Note that in Obs. 2, the early auditory cortex also showed a significant subspace rotation when the runs with poor image quality were excluded. The face and scene stimuli were reproduced with permission from Ma et al. (2015) and Konkle et al. (2010). Face stimuli were further blurred for posting on the repository.

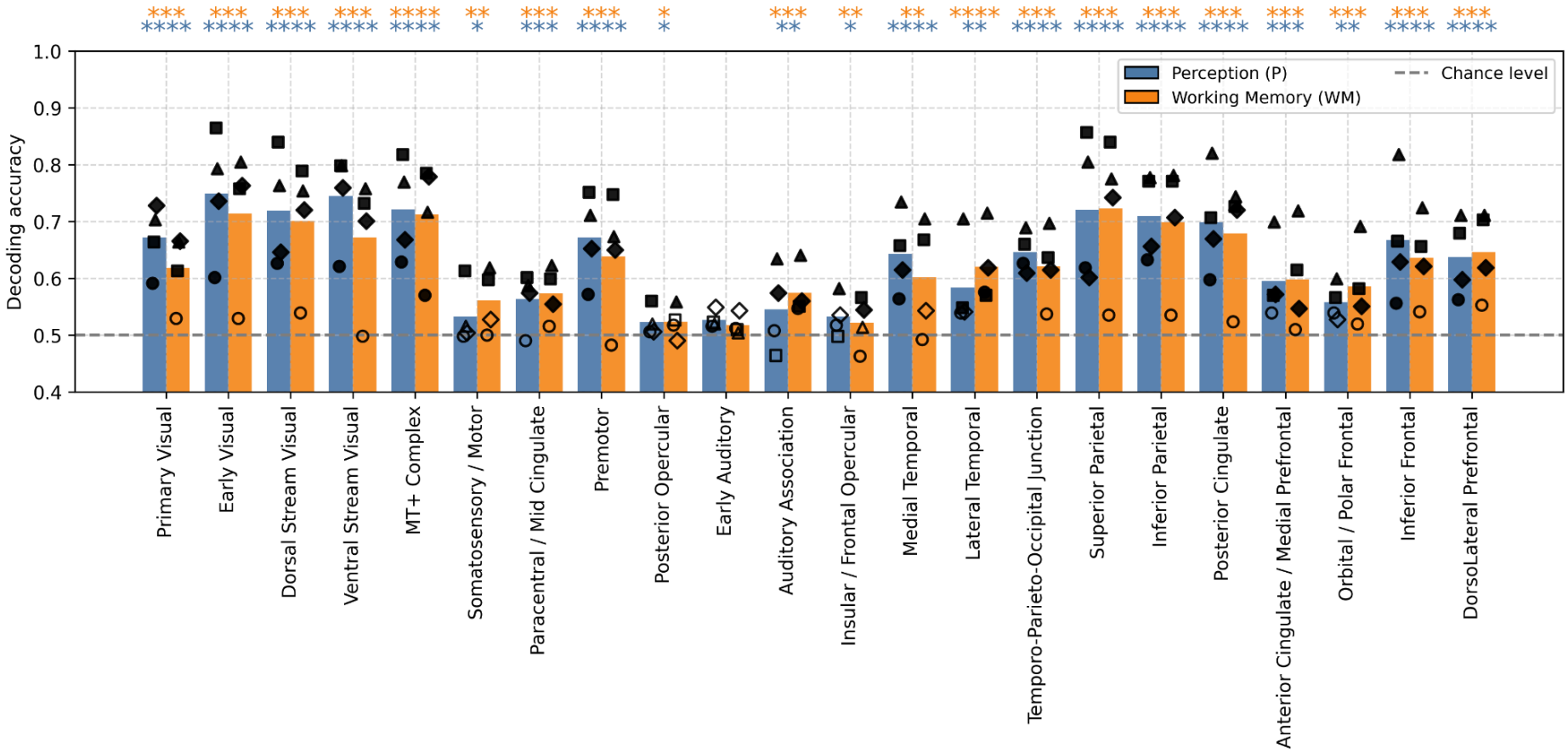

Figure A5. Decoding accuracy for face vs. scene assessed separately within the perception and working-memory phases. *The conventions for the markers and asterisks are the same as Figure 5. The dichotomy between face vs. scene was decodable both from perception and working memory phases separately in the entire visual cortex, the paracentral areas, and the temporal, parietal, and frontal association cortices (cf. Figure 5). These results are consistent with face-vs-scene tuning measured via test-retest reliability (Figure A3). Note that in Obs. 5, the auditory association cortex and the insular/frontal opercular cortex no longer showed significant face-vs-scene decoding during working memory when the runs with poor image quality were excluded.*