A tale of three types of internal object representations in the human brain

Shuang Tian[1,2], Xiaomin Mao[2], Rui Feng[2], Xiaoying Wang[2 #], Dahui Wang[1,2,3 #], Yanchao Bi[4,5,6,7,8,2 #]

Affiliations:

1 School of Systems Science, Beijing Normal University, Beijing, China, 100875

2 State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China, 100875

3 Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, Beijing, China, 100875

4 School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing 100871, China

5 Institute for Artificial Intelligence, Peking University, Beijing 100871, China

6 IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China

7 Chinese Institute for Brain Research, Beijing 102206, China

8 Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China

Correspondence to: Yanchao Bi (email: ybi@pku.edu.cn); Dahui Wang (email: wangdh@bnu.edu.cn); Xiaoying Wang (email: wangxiaoying@bnu.edu.cn)

Abstract

Both object knowledge retrieval and sensory imagery generation engage perceptual regions in the brain and exhibit representational overlap with perception. However, whether they rely on common or distinct representational mechanisms remains unclear. In this study, we conducted an fMRI visual working memory experiment in individuals with (visualizers) and without (aphantasics) visual imagery, combined with support vector machine decoding to identify regions that contained shared neural representations across the presence and absence of visual input. Critically, we applied representational similarity analysis using a multimodal deep neural network (DNN) model comprising text and image encoders to dissociate language-structured and visual-structured knowledge representations. Comparisons between visualizers and aphantasics allowed us to test the relevance of voluntary imagery experience to these representations. We found that while ventral occipitotemporal regions contained shared representations across visual presence and absence, anatomically distinct regions supported different representational structures with differential associations to imagery experience: 1) the left ventral lateral occipitotemporal cortex (LOTC) encoded visual-structured knowledge representations and was linked to imagery experience; 2) the bilateral fusiform gyrus, left dorsal LOTC, and right ventral LOTC encoded language-structured knowledge representations, irrelevant to imagery experience; 3) the left superior parietal lobule encoded priorless knowledge-absent visual maintenance representation, also irrelevant to imagery experience. Together, these findings suggest that shared neural representations between the presence and absence of sensory input arise from multiple, functionally and computationally distinct mechanisms, which vary in their reliance on prior knowledge and imagery experience.

Introduction

Visual perception, conceptual knowledge, and mental imagery are closely related but functionally distinct cognitive processes. Perception involves processing external sensory input (perceiving the image of a *cat*); conceptual knowledge supports abstract, conceptual understanding of that object even without sensory input (understanding what *cat* is); and imagery generates internal visual experiences (visualizing in mind a *cat*). Substantial evidence suggests that both conceptual knowledge and visual imagery are (partly) rooted in perceptual systems.

In studies on object knowledge retrieval, visual regions are activated when participants answer questions given object names or cues. For instance, understanding object names activate the ventral occipitotemporal cortex respecting object domain organization (Bi et al., 2016; Chao et al., 1999; Devlin et al., 2005; He et al., 2013; B. Z. Mahon et al., 2009; M. V. Peelen et al., 2013; van den Hurk et al., 2017), and retrieving object color or shape properties activate regions in the fusiform gyrus and lateral occipital cortex, overlapping with those perceiving shape and color (Bannert & Bartels, 2013; Conway & Tsao, 2009; Hsu et al., 2011; M. V. Peelen et al., 2014; Xu et al., 2023). Even early visual cortex has been implicated in semantic representations (Vetter et al., 2014, 2020).

Parallel findings have emerged from studies on visual imagery. Participants activate early visual cortex when maintaining or reconstructing visual stimuli in the absence of external input (Breedlove et al., 2020; Kosslyn et al., 1999, 2001; Pearson & Kosslyn, 2015). Even when there is no increase in mean activity strength, specific visual content—such as orientation gratings or object identity—can still be decoded from multivoxel patterns in early visual cortex (e.g., Dijkstra et al., 2017; Harrison & Tong, 2009; Lee et al., 2012). Beyond early visual cortex, imagery-related representations have also been observed in higher-level regions including the fusiform gyrus (Spagna et al., 2021), lateral occipital cortex (Bannert & Bartels, 2018; Lee et al., 2012), and parietal cortex (Li et al., 2023; Xu & Chun, 2006).

These findings together suggest a potentially shared representations between object perception, knowledge, and imagery. This joint convergence raises a critical question: do imagery and knowledge retrieval rely on the same neural code, differing only in conscious access requirement, or do they engage distinct representations at overlapping regions?

Some evidence from congenital aphantasia (Zeman, 2024; Zeman et al., 2015)—a condition marked by a lifelong absence of visual imagery experience — has hinted that visual imagery and knowledge representation may have (at least partly) different neural substrates. Individuals with aphantasia exhibit preserved semantic knowledge despite lacking conscious imagery (J. Liu & Bartolomeo, 2023). Recent fMRI studies reveal that aphantasics retain latent early visual cortex representations to support decoding for natural sounds during passive tasks (e.g., listening to dog barking vs. traffic noise) but not during active (attempted) imagery generation for these sounds (Cabbai et al., 2024; Montabes de la Cruz et al., 2024). Furthermore, in a region of the left fusiform gyrus thought to be involved in imagery, the neural representational space of (attempted) object imagery in aphantasics, while resembling that of visualizers, is less similar to perception than that of visualizers (J. Liu et al., 2025). That is, the neural representations during perception, imagery generation, and knowledge access may not be identical.

To understand the relationship between perception, imagery, and knowledge retrieval, we need to be more specific about the types of potential representations that may underlie the observed activation overlap or cross-task decoding success, and test them accordingly. We articulate the following possible types of internal object representations when visual stimuli are absent: 1) **a visual-experience-structured object knowledge representation** (**visual-knowledge** for short), i.e., object neural representations derived from past visual experiences; 2) a **language-experience-structured (nonvisual semantic) object knowledge representation** (**language-knowledge** for short), i.e., object representations derived from language experience; 3) a **visual maintenance representation not necessarily dependent on prior knowledge** (**priorless visual** for short) – when a visual stimulus is presented (with or without associated internal knowledge), there may be a period of maintenance in memory after it disappears. Note that the first two types of representations have been proposed based on behavioral and neural evidence (Bi, 2021; Paivio, 1991). While visual-knowledge representations are commonly assumed to reside in the visual cortex—given the observed overlap with perception (see above)—there is no evidence against the co-existence of nonvisual semantic object representations, especially considering that language and visual systems intrinsically interact (see discussions in H. Chen et al., 2025; B. Liu et al., 2025; X. Wang et al., 2020). Notably, congenitally blind people lack effects in visual cortex when retrieving pure visual properties (e.g., color; X. Wang et al., 2020) due to the complete absence of visual input, whereas congenitally aphantasic individuals have normal vision but are only 'blind' in mind. Orthogonal to the three representational types is the dimension of **conscious access**—any of these representations may or may not associate with imagery experience. Thus, six representational candidates are postulated in total (see Fig. 1 for this two-dimensional hypothesis space varying along representational structure and voluntary imagery relevance). These representational structures tend to correlate – for example, *cats* are more similar to *dogs* than to *tables* across all of them – and thus, cross-task decoding may be supported by any one or even multiple of these representations.

Here we carefully leverage the state-of-the-art large computation models, combined with populations with different imagery experience (visualizers vs. aphantasics), to examine these candidate internal neural representations. Specifically, we adopted CLIP (Contrastive Language–Image Pretraining) -- a multimodal neural network trained on paired images and captions (Radford et al., 2021). CLIP comprises two parallel encoders: an image-encoder that captures visual similarity, and a text-encoder that captures semantic similarity. By correlating brain activity patterns with the representational structures derived from each encoder, we can infer whether neural representations are more aligned with vision-derived-structure or language-derived-structure. We adopted CLIP for the main analyses because the paired training architecture and training datasets make the comparisons between vision and language easier, and have also used other language and vision computation models for validation.

Building on prior findings, we conducted an fMRI experiment using a retro-cue paradigm (Dijkstra et al., 2017; Harrison & Tong, 2009) in both visualizers and aphantasics, with the following predictions along the two representation dimensions. Regarding information content, ***1) visual-knowledge representation)***: If visual-experience-structured object knowledge supports shared representations, neural patterns should align with CLIP image-encoder in both perception and imagery; ***2) language-knowledge representation)***: If language-experience-structured object knowledge underlies shared representations, neural patterns should align with CLIP text-encoder, regardless of visual input; ***3) priorless visual representation)***: If low-level visual maintenance is sufficient to support shared

representations, we should observe above-chance decoding of meaningless objects—which lack semantic associations—across perception and imagery. Regarding ***imagery experience*** relevance, we asked whether voluntary imagery is associated with these particular neural representations. If so, we should observe group differences between visualizers and aphantasics. If not, shared representations and specific representational structure should emerge in both groups. Together, this framework allowed us to probe both the structural format (visual vs. language) and cognitive basis (imagery-relevant vs. irrelevant) of shared neural representations that bridge the presence and absence of sensory input.

Methods

Participants

Seventeen healthy individuals with congenital aphantasia (mean age ± SD: 23 ± 3 years; 14 females) and twenty-two control participants with typical visual imagery (22 ± 2 years; 17 females) took part in the study. Aphantasia was identified using the Vividness of Visual Imagery Questionnaire (VVIQ, Marks, 1973), with scores below 32. Follow-up interviews ensured participants fully understood the instructions and confirmed the absence or near absence of visual imagery experience. All participants had normal or corrected-to-normal vision and hearing, and no history of psychiatric or neurological disorders. Informed consent was obtained from all participants, who received monetary compensation. The study was approved by the Institutional Review Board of the State Key Laboratory of Cognitive Neuroscience and Learning at Beijing Normal University, and adhered to the Declaration of Helsinki.

Stimuli

Three categories of images were used (Supplementary Fig. 1): two meaningful domains—animals (e.g., cat, tortoise, elephant, crab, panda, bee) and large objects (e.g., tea table, dressing table, bathtub, fence, sailboat, children's slide)—and one meaningless object class. Meaningless objects were drawn arbitrarily in Photoshop, filled with 10 × 10 pixel patches randomly sampled from the meaningful images, followed by 2D Gaussian blurring in MATLAB R2020a (v.2020a, MathWorks).

All images were converted to grayscale and normalized for luminance and contrast using the SHINE toolbox (Willenbockel et al., 2010). Familiarity and pixel-based inter-image dissimilarity between animal and large object images were matched (two-sample t-test ps > .08), as were those between meaningless and real objects ($p = .36$). The test images (used in the test phase; see Procedure) were visually similar but distinct exemplars and were excluded from further analyses.

Procedure

We employed a retro-cue paradigm (Dijkstra et al., 2017; Harrison & Tong, 2009; Fig. 2a). In each trial, two images were shown sequentially for 1 s each (perception phase), separated by a jittered interstimulus interval of 2–4 s. A numeric cue ('1' or '2') then indicated which image participants should visualize during a subsequent 4 s blank interval (imagery phase). Finally, a test image was presented (test phase), and participants judged whether it was exactly the same image, not just the same object,

as the cued one. Responses were made with the index fingers (left = "yes", right = "no"). Each trial lasted 17 s.

To optimize the experimental design, trial sequences were counterbalanced: the two sequentially presented stimuli always came from different categories, each had an equal probability of appearing first or second, and each was equally likely to be cued. Image pairings were unique across every four runs. The numbers of “yes” and “no” trials were also balanced. Each run contained 18 trials, such that within each run every item was imagined once and perceived twice.

Thirty-seven participants completed all 8 runs; two (one from each group) completed only 4 runs due to personal reasons.

Image acquisition

MRI data were collected on a Siemens Prisma 3T scanner using a 64-channel phased-array head coil at Beijing Normal University Neuroimaging Center. Functional images were acquired with a multi-band echo-planar imaging sequence: 64 axial slices, 2 mm thickness, phase encoding direction from posterior to anterior; 0.2 mm gap; multi-band factor = 2; TR = 2000 ms; TE = 30 ms; flip angle = 90°; matrix size = 104 × 104; FoV = 208*208; voxel size = 2 × 2 × 2 mm^3. Each task run lasted 5 minutes and 34 seconds (167 volumes). High-resolution T1-weighted anatomical images were collected with a 3D MPRAGE sequence: 192 sagittal slices; 1 mm thickness; TR = 2530 ms; TE = 2.98 ms; TI = 1100 ms; flip angle = 7°; FoV = 224 × 256 mm^2; voxel size = 0.5 × 0.5 × 1 mm^3, interpolated; matrix size = 224 × 256.

Data preprocessing

Functional data preprocessing was conducted in SPM12 (SPM12 Software - Statistical Parametric Mapping) using MATLAB R2020a (v.2020a, MathWorks). The first 6 volumes, acquired during the initial fixation period, were discarded. The remaining data underwent slice-timing correction, motion correction, and high-pass filtering (cutoff: 0.008 Hz). Data were normalized to Montreal Neurological Institute (MNI) space using unified segmentation and resampled to 2 mm isotropic voxels

Data analysis

Behavioral Performance. For each participant and stimulus class, we computed accuracy and performed a repeated-measures ANOVA with group and stimulus class as factors. For reaction times (RT), we included only correct trials and excluded outliers beyond ±3 standard deviations. RTs were then averaged across trials for each participant and class, followed by another repeated-measures ANOVA with the same factors. All statistical analyses were conducted using the rstatix package in R (Kassambara, 2021; R Core Team, 2019).

GLM Construction. We modeled each subject’s data using a general linear model (GLM) for each run in SPM12. Predictors of interest included all items during perception, imagery, and test, with six motion parameters included as nuisance regressors. All predictors were convolved with the canonical hemodynamic response function. Since the test phase was only used to ensure task engagement, corresponding data were excluded from further analyses. For multivariate decoding analyses, beta

images were spatially smoothed using a 2-mm FWHM Gaussian kernel (Hendriks et al., 2017; Op de Beeck, 2010).

Searchlight SVM Decoding Analysis. Searchlight analysis was restricted to a whole-brain gray matter mask, defined as voxels with probability > 1/3 in the SPM12 gray matter template and located within cerebral regions (1–90) of the Automated Anatomical Labeling (AAL) atlas, yielding 122,694 voxels (981,552 mm^3). For each voxel, we defined an 8-mm radius sphere (257 voxels) and extracted beta estimates for each item and run, normalized within the sphere. We employed repeated split-half cross-validation(Valente et al., 2021), exhaustively iterating through all possible data splits, yielding 70 combinations for participants with 8 runs and 6 combinations for those with 4 runs. In each split, perception data from one half were used to train an SVM classifier, which was then tested on imagery data from the other half—and vice versa (Fig. 2a). The mean accuracy across both directions and all splits was assigned to the sphere's center voxel. Group-level statistical inference was performed using permutation-based Statistical nonParametric Mapping (SnPM) (NISOx: SnPM), with 5000 permutations and no variance smoothing. Cluster-level FWE correction was applied for multiple comparisons, as it is the only method implemented in the software for cluster-level inference. All searchlight decoding results were thresholded at voxelwise $p < .001$, cluster-level FWE-corrected $p < .05$, unless stated otherwise. We Domain-level decoding classified animals vs. large objects; item-level decoding classified among individual object identities.

Searchlight CLIP Model RSA. CLIP is a neural network trained on a variety of image-text pairs, being adapted by the alignment between images and their corresponding text descriptions (Radford et al., 2021). We utilized a Chinese version of CLIP (Yang et al., 2023), pretrained on around 200 million publicly available image-text pairs data in Chinese, using the ViT-L/14 architecture for the image-encoder and RoBERTa-wwm-Base for the text-encoder. We extracted embeddings for each image and its label (e.g., "cat," "elephant," "dressing table" in Chinese) from the final output layer of the model, and computed representational dissimilarity matrices (RDMs) based on cosine distance, yielding image-RDM and text-RDM, respectively.

Searchlight RSA was conducted within the cross-decoding mask (Fig. 2c-ii), defined by a one-sample t-test across all subjects (voxelwise $p < .001$, cluster-level FWE-corrected $p < .05$). Neural RDMs were computed for each subject by extracting T-value vectors from searchlight spheres (as in the decoding analysis), and measuring dissimilarities as 1 − Pearson correlation between stimulus pairs. We then computed Spearman partial correlations between neural RDMs and one encoder RDM, controlling for the other, and used commonality analysis (Nalborczyk, 2025) to quantify shared variance among all three RDMs. Coefficients were assigned to the center voxel of each sphere. To improve the sensitivity of statistical inferences, we applied threshold-free cluster enhancement (TFCE) with 10,000 permutations, along with FDR correction (as implemented in DPABI; Chen et al., 2018; Smith & Nichols, 2009; Yan et al., 2021) for multiple comparisons. All searchlight RSA results were thresholded at FDR-corrected TFCE-p < .05, unless stated otherwise.

ROI-based RSA with Unimodal DNN Models. To evaluate the role of encoder alignment, we repeated the RSA using two unimodal models: a Vision Transformer (ViT) model pretrained on LVD-142M with self-supervised DINOv2 method (ViT-Dino; Oquab et al., 2024); a Chinese RoBERTa model pretrained on Chinese Wikipedia and extended data (Cui et al., 2021), both architecturally matched to the CLIP

encoders. Their RDM correlations are shown in Supplementary Fig. 8. Neural RDMs were constructed using T-value vectors extracted from each ROI defined based on the searchlight results (see Results). For each ROI, we computed Spearman correlations between neural- and model- RDMs. To enable direct comparison with the main results, which were based on partial correlations, we also computed partial correlations for visual models (CLIP image-RDM and DINO-RDM) while regressing out the same CLIP text-RDM, and for language models (CLIP text-RDM and RoBERTa-RDM) while regressing out the same CLIP image-RDM. All correlation coefficients were Fisher-Z transformed prior to statistical analysis in R.

ROI-based Layerwise CLIP Model RSA. To extract layerwise features from the Chinese-CLIP model, we registered forward hooks on each residual block of the visual transformer and each encoder layer of the BERT-based text encoder. During a forward pass, we collected the [CLS] token embeddings from each layer for both image and text inputs. For image processing, preprocessed images were passed through the visual encoder, and for text, tokenized captions were passed through the text encoder. At each layer, [CLS] embeddings were detached, stored, and stacked across stimuli. Cosine similarity matrices were computed from L2-normalized features, and converted into RDMs by subtracting similarity from one. Correlations and partial correlations between neural RDMs and encoder RDMs were computed within the defined ROIs. For partial correlations, the final output layer RDM of the non-target encoder was regressed out.

Individual Differences in Visual Imagery. To assess whether imagery vividness modulated neural representations among visualizers at the individual level, we extracted the averaged partial correlations to the CLIP encoder in the ROI that showed sensitivity to imagery experience (i.e., Visualizers > Aphantasics; see Results) from the participants, and correlated with their VVIQ scores.

Results

To explore the representational mechanisms underlying seemingly shared neural patterns across the presence and absence of visual input, and to examine their relevance to voluntary imagery, we conducted an fMRI experiment using a retro-cue visual working memory paradigm (Fig. 2a) in individuals with (visualizers) and without (aphantasics) visual imagery. We first analyzed behavioral performance to ensure that both groups performed the task comparably. Next, we conducted searchlight SVM cross-decoding to identify brain regions supporting (at least some) shared representations across perception and imagery phases for both meaningless and real objects. Within the regions identified for real objects (i.e., with internal knowledge representations), we applied searchlight RSA using a Chinese CLIP model to dissociate visual- vs. language-derived representational structures, and to examine their relevance for conscious imagery experience by comparing visualizers and aphantasics. Finally, ROI-based RSA was conducted to verify the searchlight results by assessing the necessity of encoder alignment, the layerwise representational profiles, and the relationship between neural representations and individual imagery vividness in visualizers.

Behavioral results

To test whether aphantasics could perform the retro-cue working memory task as well as visualizers, we analyzed participants' behavioral performance in the scanner. As shown in Fig. 2b, repeated-measures ANOVA on accuracy revealed no significant main effect of group ($F_{(1, 37)} = 1.17$, $p = .29$, $\eta^2 = .03$), stimulus class ($F_{(2, 74)} = 0.99$, $p = .38$, $\eta^2 = .005$), or their interaction ($F_{(2, 74)} = 0.78$, $p = .46$, $\eta^2 = .004$).

For reaction time (RT), repeated-measures ANOVA showed no significant main effect of group ($F_{(1, 37)} = 0.05$, $p = .82$, $\eta^2 = .001$) or interaction ($F_{(2, 74)} = 2.96$, $p = .06$, $\eta^2 = .001$), but a significant main effect of stimulus class ($F_{(2, 74)} = 6.01$, $p = .004$, $\eta^2 = .002$). Pairwise comparisons indicated that responses to meaningless objects were significantly slower than responses to animals or large objects (ps < .02, FDR-corrected). That is, performance differences were driven by stimulus properties rather than group, indicating that aphantasic participants performed the task as accurately and efficiently as visualizers.

Searchlight SVM Decoding Results: Shared neural representations across perception-imagery phases identified

To identify brain regions supporting (some) shared neural representations between perception and imagery phases, we conducted whole-brain searchlight SVM cross-decoding analyses for meaningless objects and real objects, respectively. A repeated split-half cross-validation approach was used, where an SVM classifier was trained on perception data from half of the runs and tested on imagery data from the remaining half, and vice versa. This procedure was repeated across a grey matter mask. For each searchlight sphere, the decoding accuracy was averaged across all split-half combinations and both train–test directions, and the resulting value was assigned to the center voxel. Group-level statistical inference was conducted using permutation test. Results were thresholded at voxelwise $p < .001$, cluster-level FWE-corrected $p < .05$, unless stated otherwise.

The left SPL contained shared representations across perception and imagery phases for meaningless objects (priorless visual representation) in both aphantasics and visualizers.

To assess whether visual maintenance alone is sufficient to support shared neural representations between perception and imagery phases, we conducted searchlight SVM cross-decoding for meaningless objects (6-way classification). We first pooled all subjects together to perform group-level inference to maximize sensitivity to identify the areas containing shared neural representations across perception and imagery phases in all subjects. The result showed a significant cluster in the left superior parietal lobule (SPL; Fig. 2c-i; voxelwise $p < .005$, cluster-level FWE-corrected $p < .05$; not surviving multiple correction at voxelwise $p < .001$). Mean decoding accuracy extracted from this SPL cluster did not differ between groups ($t_{(36.9)} = 0.03$, $p = .97$), indicating that the perception-like representations for maintaining meaningless objects were contained in the dorsal stream irrespectively of voluntary imagery.

When tested separately for each group, no significant above-chance decoding was found in either group after multiple comparisons correction. An uncorrected trend showed that the effects in visualizers distributed across both ventral and dorsal regions, whereas in aphantasics mainly in dorsal regions (see Supplementary Fig. 2 for the uncorrected maps).

The VOTC areas contained shared representations across perception and imagery phases for real objects (Knowledge-Scaffolded representation) in both aphantasics and visualizers.

For real objects with prior knowledge, we examined brain regions that contained common representations of domains (animal vs. large object) across perception and imagery phases. As shown in Fig. 2c-ii, the result of group-level analysis combining all subjects revealed significant cross-decoding in the bilateral ventral occipitotemporal cortex (VOTC), with extensions into the left lateral occipital cortex (LOC), as well as in the right inferior frontal gyrus (IFG). Mean decoding accuracy extracted from these regions did not differ between groups ($t_{(27.9)} = -0.29$, $p = .77$).

The group-level result of each group (Supplementary Fig. 3) showed that visualizers showed successful cross-decoding for domains in bilateral VOTC, extending to the LOC. By contrast, the aphantasic group showed significant decoding only in the right Fusiform Gyrus (FG). All these areas were included in the regions identified by the combined group-level result. We then tested group difference within these regions. The result showed a trend toward higher accuracy in the left ventral lateral occipitotemporal cortex (LOTC) in the visualizers; however, it could not survive multiple comparisons correction (see Supplementary Fig. 3 for an uncorrected result).

We further conducted a more stringent analyses, i.e., item-level cross-decoding (12-way classification) for real objects. This analysis revealed similar but spatially more restricted results (Supplementary Fig. 4) than domain-level decoding. Pooled group-level result showed a distribution across VOTC that largely overlapped with the domain-level decoding result. The visualizer group showed significant decoding in the left FG, and the aphantasic group showed significant decoding in the right lateral FG. However, no significant group differences were observed, even under a liberal uncorrected threshold.

These results suggest that the VOTC regions contained shared representations across perception and imagery phases in both groups, with only a trend of effect of imagery experience in the left ventral LOTC. Given that successful cross-decoding alone does not reveal what types of underlying representational structures support such decoding, we performed searchlight RSA to dissociate representational structures within the combined group-level domain cross-decoding mask (i.e., the regions shown in Fig. 2c-ii)

Searchlight CLIP-model-based RSA results: Multiple representation types identified

To dissociate these representational structures, we leveraged a Chinese version of the multimodal deep neural network CLIP (Radford et al., 2021; Yang et al., 2023), which consists of an image-encoder and a text-encoder aligned in a shared embedding space. We employed searchlight RSA in the cross-decoding mask (Fig. 2c-ii) to examine the similarity between neural representations (during perception phase and during imagery phase) and the representational structures of each encoder. We input the stimuli images into the image-encoder and the names of objects into the text-encoder, extracted the features from the final output layers, and transformed them into RDMs based on cosine distance (Fig. 3a; see Supplementary Fig. 9 for layer-specific results). There was a medium-level, significant correlation between the RDMs of the two encoders ($r = 0.43$, $p = .0004$). Therefore, we partitioned the variance into components specific to each model, as well as the shared variance contributing to the explanation of the neural RDMs, using partial correlations and commonality analyses (Nalborczyk, 2025). We used the threshold-free cluster enhancement (TFCE; Chen et al., 2018; Smith & Nichols, 2009; Yan et al., 2021) method to improve the sensitivity of statistical inferences, and searchlight RSA results were all thresholded at FDR-corrected TFCE-$p < .05$, unless stated otherwise.

Both language- and visual-knowledge representations existed in visualizers.

In visualizers, during perception phase (Fig. 3b, top), significant partial correlations between neural-RDM and CLIP text-encoder-RDM (blue regions) spanned the entire mask—covering bilateral VOTC, extending into the left LOC and right IFG. During imagery phase, CLIP text-encoder-RDM effects remained widespread but notably bypassed the left ventral LOTC and were spatially restricted in the right FG and IFG.

Partial correlations with the CLIP image-encoder-RDM (orange-red regions) were also broadly distributed during perception phase, with reductions observed in the bilateral medial FG and left LOC. During imagery phase, significant CLIP image-encoder-RDM effects were primarily localized to the left LOC, left FG, and bilateral ventral LOTC.

Overlap maps across perception and imagery are shown in Supplementary Fig. 5. Commonality analysis (green regions) revealed shared variance among neural-, CLIP image-encoder-, and CLIP text-encoder-RDMs throughout the mask during perception, but no significant clusters emerged during imagery (see Supplementary Fig. 6 for uncorrected results).

Only language-structured representations existed in Aphantasics.

In the aphantasic group, during both perception and imagery phases (Fig. 3b, bottom), partial correlations with the CLIP text-encoder-RDM were also widely distributed across the mask (see Supplementary Fig. 5 for overlap maps). No significant correlations with the CLIP image-encoder-RDM were observed either during perception or during imagery phase. In uncorrected maps (Supplementary Fig. 6) during perception, significant effects of CLIP image-encoder-RDM appeared in bilateral FG; during imagery, significant effects were still absent except for a small area in right dorsal LOTC (voxel number = 22). Commonality effects were mainly restricted to bilateral lateral FG during perception.

Group difference in visual-structured representations existed for both perception and imagery.

Two-sample t-tests across the mask (Fig. 3b, right; Supplementary Fig. 7) revealed significant group differences specifically in the effects of CLIP image-encoder-RDM, with visualizers showing greater model-neural similarity than aphantasics during both perception and imagery phases. During perception, group differences (visualizers > aphantasics) were distributed across the left hemisphere of the mask, right medial FG, right ventral LOTC, and right IFG during perception; During imagery, such group difference was robust specifically in the left ventral LOTC.

No significant group differences emerged in correlations with the CLIP text-encoder-RDM, either during perception phase or imagery phase. Even under an uncorrected threshold, effects were absent during perception, and minimal during imagery: very small clusters appeared in left FG (visualizers > aphantasics; voxel number = 94) and left LOC (aphantasics > visualizers; voxel number = 52).

Summary

To summarize the CLIP-model-based RSA findings above, two types of regions with different representational properties emerged (see brain maps in Fig. 4): one type of regions where both groups

showed reliable similarity to the CLIP text-encoder (blue areas) during both perception and imagery phases, covering the left dorsal LOTC, bilateral FG, and right ventral LOTC. That is, neural activity patterns in these regions during both the object perception and object imagery stages, in both visualizers and aphantasics, showed significant specific correlation with CLIP text-encoder-derived representation. This region likely internally represents object by how they are related in the language space supporting both perception and imagery processes, and is not relevant to voluntary imagery experiences (having representations here does not relate to if an individual can mentally experience the object visually). We coin these regions "Language-knowledge-imagery-irrelevant". The other is the left ventral LOTC (orange-red areas), which shows significant effects of the CLIP image-encoder during both perception and imagery in only visualizers and not aphantasics (with significant group differences). This region likely internally represents objects by how they are experienced during visual experiences in typical populations, and the absence of such representations associates with the lack of voluntary visual imagery experience. We coin this region "Visual-knowledge-imagery-relevant".

ROI-based RSA results: In-depth-analyses with different model validations

In this section we zoom into the two types of regions identified above -- "Language-knowledge-imagery-irrelevant" (blue in Fig. 4, i.e., left dorsal LOTC, bilateral FG, and right ventral LOTC) and "Visual-knowledge-imagery-relevant" (red-orange in Fig. 4, i.e., left ventral LOTC) -- to further characterize their representational profiles.

The results of representational structures were replicated using unimodal DNN models.

To test whether the observed neural similarity to one CLIP encoder was necessarily driven by its alignment with the other, we repeated RSA using ViT-DINO (unimodal visual model) and RoBERTa (unimodal language model, pretrained on Chinese texts) in ROIs. Pairwise model RDM correlations are shown in Supplementary Fig. 8.

In the language- knowledge-imagery-irrelevant ROI (Fig. 4a), RoBERTa showed significant correlations and partial correlations with neural-RDMs during both perception and imagery phases in both groups ($ps < .03$, FDR-corrected), without group differences ($ps > .707$, FDR-corrected). No differences were found between RoBERTa and the CLIP text-encoder ($ps > .51$, FDR-corrected), supporting that the observed effects of text were not specific to CLIP.

In the visual- knowledge-imagery-relevant ROI (Fig. 4b), ViT-DINO showed marginally significant correlation and partial correlation with neural-RDMs during perception in visualizers ($ps < .09$, FDR-corrected), but with no group differences ($ps > .13$, FDR-corrected). During imagery, significant correlation and partial correlation were found in visualizers ($ps < .02$, FDR-corrected), with group differences favoring visualizers ($ps < .02$, FDR-corrected), replicating the pattern found with the CLIP image-encoder. Although a trend toward weaker effects for ViT-DINO was observed, the differences between models were not statistically significant ($ps > .88$, FDR-corrected).

That is, multimodal alignment between encoders was not essential for the observed representational structures of language vs. vision.

Visualizers and aphantasics converged across multiple layers in similarity to the text-encoder, but diverged to the image-encoder.

To examine the layerwise similarity between neural representations and those of the CLIP model, we extracted features from different layers of both CLIP encoders and performed RSA within the language-knowledge-imagery-irrelevant ROI and visual-knowledge-imagery-relevant ROI (Supplementary Fig. 9). In the language-knowledge-imagery-irrelevant ROI, the similarity to the text-encoder-RDM increased across layers in both groups, during both perception and imagery phases, consistent with the hierarchical abstraction in the language model. In the visual-knowledge-imagery-relevant ROI, group differences in image-encoder similarity began to emerge at higher layers during perception, with aphantasics and visualizers showing similar patterns at lower and mid-level layers but diverging at higher ones. In contrast, during imagery phase, aphantasics showed no significant similarity to the image-encoder at any layer, whereas visualizers maintained significant alignment across multiple levels. These results further elucidate that the visual-structured representation observed in the visual-knowledge-imagery-relevant ROI is imagery-relevant, particularly in higher-level layers. Both the correlation and partial correlation results showed consistent patterns, indicating that the effects attributed to each encoder were not driven by shared variance with the other.

The similarity to the image-encoder was modulated by individual vividness of imagery in visualizers.

The only group difference was observed in neural representational similarity to the image-encoder-RDM during both perception and imagery phases, overlapping in the left ventral LOTC. To further validate the imagery-related group effect in this area at the level of individuals, we calculated correlations between VVIQ scores and neural representational similarity to the CLIP image-encoder within the visualizer group. As shown in Fig. 4c, consistent with the group-level results, the similarity to the image-encoder was significantly correlated with VVIQ scores during imagery ($r = 0.61$, $p = 0.005$, FDR-corrected), but not during perception ($r = 0.13$, $p = 0.57$, FDR-corrected). That is, how well the object neural activity pattern fits with those derived from visual experience (aligning with CLIP visual-encoder) explains how well a person experience vivid visual imagery of objects, both in a continuous fashion for so-called visualizers and in extreme cases, i.e., aphantasics.

Discussion

We investigated how conceptual knowledge representation and mental imagery relate to the shared neural representations across the presence and absence stage of visual input (as defined by areas supporting cross-process decoding). Specifically, what are the relationships among neural representations for seeing a *cat*, imagining a *cat*, and the internal memory storage of *cat*? Using a retro-cue fMRI paradigm separating the seeing phase and maintaining phase, we leveraged language- and vision-DNNs to test the object neural knowledge representation content (language-structured vs. visual-structured), and the contrast between individuals with and without visual imagery to test their relevance for conscious imagery experience (imagery-relevant vs. imagery-irrelevant).

We identified three types of neural representations that are common between visual-present and -absent phases (Fig. 5): 1) One type of object neural representation that is specifically fitted by large

vision models, is present in visualizers (as a function of subjective imagery vividness) but absent in aphantasics, i.e., a **visual-knowledge-imagery-relevant representation**, and is localized to left ventral LOTC; 2) Another type of object neural representation that is specifically fitted by large language models, is present in both visualizers and aphantasics, i.e., a **language-knowledge-imagery-irrelevant representation**, and is localized to the left dorsal LOTC, bilateral FG, and right ventral LOTC; 3) A final type supporting novel meaningless objects decoding in both visualizers and aphantasics, i.e., a **priorless-imagery-irrelevant visual representation**, and is localized to the left SPL**.**

Multifaceted object knowledge representations in the ventral visual cortex

The nature of object representation in the ventral visual cortex is one of the classical topics in cognitive neuroscience, and has been difficult to reconcile (see reviews in Martin, 2016). Overlap between knowledge retrieval and perception, along with parsimonious principle, has led to the default assumption that it is visual and/or multimodal computations that converge with shape or other geometrical properties that vary by object domains/categories (see discussions in Martin, 2016; and about results with congenitally blind individuals in Bi et al., 2016). Here we showed two further dimensions of heterogeneity – object content organization more fitted with vision or language, and the involvement in subjective imagery experience. Specifically, the FG showed strong similarity with CLIP's text-encoder, consistent with its role in multimodal and abstract object representations (Bi et al., 2016; Caramazza & Mahon, 2003; Mahon & Caramazza, 2011), irrelevant to imagery experience. The LOTC showed subregional dissociation: Left dorsal LOTC aligned with the text-encoder, suggesting a language-structured content. The left ventral LOTC aligned with the image-encoder and tracked imagery vividness, suggesting a visual-structured, imagery-related format. These results enrich our understanding of LOTC as a heterogeneous convergence zone, where representations of varied content reside (e.g., tool objects, action, shape knowledge; Chao et al., 1999; Martin et al., 1996; Wurm & Caramazza, 2019, 2022; Xu et al., 2023), suggesting that future studies on LOTC should consider the role of imagery processes and representational structures.

It is important to note that better fitting with CLIP text-encoder (and other large language models) by itself does not necessarily imply that the neural representation is processing language by nature. It is possibly still representing nonlinguistic aspects of objects that respects relations that are better captured by language-derived higher-level semantics (see similar discussions in H. Chen et al., 2025; Conwell et al., 2025; A. Y. Wang et al., 2023). Here the results simply suggest that the neural structures for objects are better explained by specific relational structures parallel those derived from language than from vision, and this type of representation is similarly present for visualizers and aphantasics, irrelevant to subjective imagery experience.

Functional localization of object imagery experience

Theoretical models of mental imagery propose a top-down reconstruction process, such that higher-level stored content is transformed into visual detail in lower-level sensory cortices (see reviews in Pearson, 2019, 2020). However, the timing and location of this reconstruction remain unclear. Our results show that while object knowledge is broadly represented in regions supporting shared

perception–imagery patterns, imagery-specific content was confined to left ventral LOTC. How well the neural representation here during imagery fits with CLIP image decoder (and other vision models) is predictive of how much subjective vivid imagery an individual experience. This region is close to the so-called fusiform imagery node (FIN) proposed to support domain-general internal visualization (Spagna et al., 2021, 2024). Prior work has shown reduced imagery-perception similarity, measured by the correlation between objects imagery-RDM and perception-RDM, in aphantasics, and positive associations with imagery vividness in visualizers in this node(J. Liu et al., 2025). Our findings support the critical role of this area in imagery function, and further revealed that in aphantasics not only the neural representations for imagery, but also those underlying perception, are altered to less "visual"-like (see discussion in the next section). However, this imagery effect may be limited to object-level stimuli. Earlier studies using low-level visual features (e.g., gratings, lines) have localized imagery effects to early visual cortices, which were not observed in the current study using object-level materials, suggesting that the neural basis of imagery may vary by stimulus complexity—a question for future research.

For meaningless objects that lack prior knowledge, left SPL emerged as the key site of shared representation between perception and imagery. This is consistent with its role in visual working memory (Koenigs et al., 2009; Li et al., 2023; Xu, 2018) and sub-conscious processing (Almeida et al., 2008; Fang & He, 2005; Maclean et al., 2023; Suzuki et al., 2014). In particular, aphantasics showed comparable SPL representations despite lacking subjective imagery, reinforcing the idea that consciousness-independent visual coding can be implemented via dorsal-stream mechanisms.

Diverse representations during perception

Intriguingly, in the aphantasic group, visual-structured representations at low to mid-levels were preserved during perception, but group differences emerged at higher layers, suggesting that aphantasics may adopt alternative coding strategies for incoming visual input. This implies that aphantasia may not only reflect a deficit in internal imagery, but a broader reorganization of perceptual encoding strategies. While visualizers and aphantasics perceived the same stimuli, their underlying neural organizations diverged.

However, we found no correlation between the vision-aligned representation during perception and VVIQ in visualizers, implying that this perceptual group difference is not directly modulated by conscious imagery vividness. Instead, it may reflect a compensatory shift toward language-based processing in aphantasics—a finding with implications for how we understand cognitive diversity in perception. Additionally, both visual-structured and language-structured representations—along with their shared variance—were significant during perception, emphasizing the integration of both bottom-up and top-down processes during perception.

The relationship between object perception, knowledge, and imagery experience

Although both object knowledge access and imagery generation engage the perceptual systems, our findings indicate that they rely on both shared and distinct representational mechanisms. Object knowledge access and perception aligned in both the language-structured and visual-structured-representational space, while subjective imagery experience and perception aligned in the visual-structured space. These distinctions highlight that each process entails multiple types of representations, and the seemingly overlapping neural representations with a same process can arise from different types of representations. Note that for meaningless shapes without knowledge priors, while the left SPL

was the strongest site of cross-decoding, we observed a trend the in ventral visual regions among visualizers. Whether the visual-structured representations in the left ventral LOTC only applies to those with internal stores (memory), or can also arise from temporary bottom-up visual inputs, remains an open question for future studies to directly test.

Conclusion

Our findings reveal that shared neural representations across the presence and absence of sensory input arise from multiple, heterogeneous mechanisms: Visual-structured object knowledge representations emerged specifically in the left ventral LOTC and were related to imagery experience. Language-structured knowledge representations were distributed across the VOTC, invariant to imagery experience. Knowledge-absent and imagery-irrelevant visual representations were localized to the left SPL. Together, these results offer a more nuanced account of how object knowledge and mental imagery interact to support internal representations in human mind. They also underscore the value of integrating computational modeling, individual differences, and multivariate neuroimaging to uncover the representational architecture of the human brain.

Acknowledgement

We thank Jiahui Lu, Haojie Wen, and Ziyi Xiong for their assistance with data collection; Ze Fu for his advice on the analysis; Haoyang Chen for his assistance with the utility of the DNN models; Ze Fu, Haojie Wen, and Huichao Yang, for their advice on the figures; Yiyang Cai and Ziyi Xiong for their advice on the manuscript.

Funding

STI2030-Major Project 2021ZD0204100 (2021ZD0204104 to Y.B.); National Natural Science Foundation of China (31925020 and 82021004 to Y.B.); the Fundamental Research Funds for the Central Universities (Y.B.); China Postdoctoral Science Foundation funded project (2025M773448 to S.T.)

DNN models were acquired from:

Chinese-CLIP: https://github.com/OFA-Sys/Chinese-CLIP

ViT: https://huggingface.co/timm/vit_large_patch14_reg4_dinov2.lvd142m

RoBERTa: https://github.com/ymcui/Chinese-BERT-wwm, RoBERTa-wwm-ext

References

Almeida, J., Mahon, B. Z., Nakayama, K., & Caramazza, A. (2008). Unconscious processing dissociates along categorical lines. *Proceedings of the National Academy of Sciences*, *105*(39), 15214–15218. https://doi.org/10.1073/pnas.0805867105

Bannert, M. M., & Bartels, A. (2013). Decoding the yellow of a gray banana. *Current Biology*, *23*(22), 2268–2272. https://doi.org/10.1016/j.cub.2013.09.016

Bannert, M. M., & Bartels, A. (2018). Human V4 activity patterns predict behavioral performance in imagery of object color. *Journal of Neuroscience*, *38*(15), 3657–3668. https://doi.org/10.1523/JNEUROSCI.2307-17.2018

Bi, Y. (2021). Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, *25*(10), 883–895. https://doi.org/10.1016/j.tics.2021.07.006

Bi, Y., Wang, X., & Caramazza, A. (2016). Object Domain and Modality in the Ventral Visual Pathway. *Trends in Cognitive Sciences*, *20*(4), 282–290. https://doi.org/10.1016/j.tics.2016.02.002

Breedlove, J. L., St-Yves, G., Olman, C. A., & Naselaris, T. (2020). Generative Feedback Explains Distinct Brain Activity Codes for Seen and Mental Images. *Current Biology*, *30*(12), 2211-2224.e6. https://doi.org/10.1016/j.cub.2020.04.014

Cabbai, G., Racey, C., Simner, J., Dance, C., Ward, J., & Forster, S. (2024). Sensory representations in primary visual cortex are not sufficient for subjective imagery. *Current Biology*, *34*(21), 5073-5082.e5. https://doi.org/10.1016/j.cub.2024.09.062

Caramazza, A., & Mahon, B. Z. (2003). The organization of conceptual knowledge: The evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, *7*(8), 354–361. https://doi.org/10.1016/S1364-6613(03)00159-1

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*, *2*(10), 913–919. https://doi.org/10.1038/13217

Chen, H., Liu, B., Wang, S., Wang, X., Han, W., Zhu, Y., Wang, X., & Bi, Y. (2025). *Language modulates vision: Evidence from neural networks and human brain-lesion models*. http://arxiv.org/abs/2501.13628

Chen, X., Lu, B., & Yan, C. G. (2018). Reproducibility of R-fMRI metrics on the impact of different strategies for multiple comparison correction and sample sizes. *Human Brain Mapping*, *39*(1), 300–318. https://doi.org/10.1002/HBM.23843

Conway, B. R., & Tsao, D. Y. (2009). Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(42), 18034–18039. https://doi.org/10.1073/pnas.0810943106

Conwell, C., MacMahon, E., Jagadeesh, A., Vinken, K., Sharma, S., Prince, J. S., Alvarez, G. A., Konkle, T., Livingstone, M., & Isik, L. (2025). *Monkey See, Model Knew: Large Language Models Accurately*

Predict Visual Brain Responses in Humans and *Non-Human Primates*. https://doi.org/10.1101/2025.03.05.641284

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-Training with Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *29*, 3504–3514. https://doi.org/10.1109/TASLP.2021.3124365

Devlin, J. T., Rushworth, M. F. S., & Matthews, P. M. (2005). Category-related activation for written words in the posterior fusiform is task specific. *Neuropsychologia*, *43*(1), 69–74. https://doi.org/10.1016/j.neuropsychologia.2004.06.013

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. J. (2017). Vividness of visual imagery depends on the neural overlap with perception in visual areas. *Journal of Neuroscience*, *37*(5), 1367–1373. https://doi.org/10.1523/JNEUROSCI.3022-16.2016

Fang, F., & He, S. (2005). Cortical responses to invisible objects in the human dorsal and ventral pathways. *Nature Neuroscience*, *8*(10), 1380–1385. https://doi.org/10.1038/nn1537

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

He, C., Peelen, M. V., Han, Z., Lin, N., Caramazza, A., & Bi, Y. (2013). Selectivity for large nonmanipulable objects in scene-selective visual cortex does not require visual experience. *NeuroImage*, *79*, 1–9. https://doi.org/10.1016/j.neuroimage.2013.04.051

Hendriks, M. H. A., Daniels, N., Pegado, F., & de Beeck, H. P. O. (2017). The effect of spatial smoothing on representational similarity in a simple motor paradigm. *Frontiers in Neurology*, *8*(MAY), 1–11. https://doi.org/10.3389/fneur.2017.00222

Hsu, N. S., Kraemer, D. J. M., Oliver, R. T., Schlichting, M. L., & Thompson-Schill, S. L. (2011). Color, Context, and Cognitive Style: Variations in Color Knowledge Retrieval as a Function of Task and Subject Variables. *Journal of Cognitive Neuroscience*, *23*(9), 2544–2557. https://doi.org/10.1162/jocn.2011.21619

Kassambara, A. (2021). *rstatix: Pipe-friendly framework for basic statistical tests. R package version 0.7. 0*.

Koenigs, M., Barbey, A. K., Postle, B. R., & Grafman, J. (2009). Superior Parietal Cortex Is Critical for the Manipulation of Information in Working Memory. *The Journal of Neuroscience*, *29*(47), 14980–14986. https://doi.org/10.1523/JNEUROSCI.3706-09.2009

Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience*, *2*(9), 635–642. https://doi.org/10.1038/35090055

Kosslyn, S. M., Pascual-Leone, A., Felician, O., Camposano, S., Keenan, J. P., Thompson, W. L., Ganis, G., Sukel, K. E., & Alpert, N. M. (1999). The Role of Area 17 in Visual Imagery: Convergent Evidence from PET and rTMS Published by: American Association for the Advancement of Science Stable URL: http://www.jstor.org/stable/2899164 Linked references are available on JSTOR for this article: T. *Science*, *284*(5411), 167–170.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1038/nn.2303

Lee, S. H., Kravitz, D. J., & Baker, C. I. (2012). Disentangling visual imagery and perception of real-world objects. *NeuroImage*, *59*(4), 4064–4073. https://doi.org/10.1016/j.neuroimage.2011.10.055

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

Liu, B., Wang, X., Wang, X., Li, Y., Han, Y., Lu, J., Zhang, H., Wang, X., & Bi, Y. (2025). Object knowledge representation in the human visual cortex requires a connection with the language system. *PLOS Biology*, *23*(5). https://doi.org/10.1371/journal.pbio.3003161

Liu, J., & Bartolomeo, P. (2023). Probing the unimaginable: The impact of aphantasia on distinct domains of visual mental imagery and visual perception. *Cortex*, *166*, 338–347. https://doi.org/10.1016/j.cortex.2023.06.003

Liu, J., Zhan, M., Hajhajate, D., Spagna, A., Dehaene, S., Cohen, L., & Bartolomeo, P. (2025). Visual mental imagery in typical imagers and in aphantasia: A millimeter-scale 7-T fMRI study. *Cortex*, *185*, 113–132. https://doi.org/10.1016/j.cortex.2025.01.013

Maclean, M. W., Hadid, V., Spreng, R. N., & Lepore, F. (2023). Revealing robust neural correlates of conscious and unconscious visual processing: Activation likelihood estimation meta-analyses. *NeuroImage*, *273*, 120088. https://doi.org/10.1016/j.neuroimage.2023.120088

Mahon, B. Z., Anzellotti, S., Schwarzbach, J., Zampini, M., & Caramazza, A. (2009). Category-Specific Organization in the Human Brain Does Not Require Visual Experience. *Neuron*, *63*(3), 397–405. https://doi.org/10.1016/j.neuron.2009.07.012

Mahon, B. Z. B. Z., & Caramazza, A. (2011). What drives the organization of object knowledge in the brain? *Trends in Cognitive Sciences*, *15*(3), 97–103. https://doi.org/10.1016/j.tics.2011.01.004

Marks, D. F. (1973). VISUAL IMAGERY DIFFERENCES IN THE RECALL OF PICTURES. *British Journal of Psychology*, *64*(1), 17–24. https://doi.org/10.1111/j.2044-8295.1973.tb01322.x

Martin, A. (2016). GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic Bulletin and Review*, *23*(4), 979–990. https://doi.org/10.3758/s13423-015-0842-3

Martin, A., Wiggs, C. L., Ungerleider, L. G., & Haxby, J. V. (1996). Neural correlates of category-specific knowledge. *Nature*, *379*(6566), 649–652. https://doi.org/10.1038/379649a0

Montabes de la Cruz, B. M., Abbatecola, C., Luciani, R. S., Paton, A. T., Bergmann, J., Vetter, P., Petro, L. S., & Muckli, L. F. (2024). Decoding sound content in the early visual cortex of aphantasic participants. *Current Biology*. https://doi.org/10.1016/j.cub.2024.09.008

Nalborczyk, L. (2025, January 7). *Making Sense of Commonality Analysis*. https://lnalborczyk.github.io/blog/2025-01-07-commonality

Op de Beeck, H. P. (2010). Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? *NeuroImage*, *49*(3), 1943–1948. https://doi.org/10.1016/j.neuroimage.2009.02.047

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., … Bojanowski, P. (2024). *DINOv2: Learning Robust Visual Features without Supervision*. http://arxiv.org/abs/2304.07193

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *45*(3), 255.

Pearson, J. (2019). The human imagination: the cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, *20*(10), 624–634. https://doi.org/10.1038/s41583-019-0202-9

Pearson, J. (2020). Reply to: Assessing the causal role of early visual areas in visual mental imagery. *Nature Reviews Neuroscience*, *21*(9), 517–518. https://doi.org/10.1038/s41583-020-0349-4

Pearson, J., & Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the National Academy of Sciences*, *112*(33), 10089–10092. https://doi.org/10.1073/pnas.1504933112

Peelen, M. V. M. V., Bracci, S., Lu, X., He, C., Caramazza, A., & Bi, Y. (2013). Tool selectivity in left occipitotemporal cortex develops without vision. *Journal of Cognitive Neuroscience*, *25*(8), 1225–1234. https://doi.org/10.1162/jocn_a_00411

Peelen, M. V., He, C., Han, Z., Caramazza, A., & Bi, Y. (2014). Nonvisual and Visual Object Shape Representations in Occipitotemporal Cortex: Evidence from Congenitally Blind and Sighted Adults. *The Journal of Neuroscience*, *34*(1), 163–170. https://doi.org/10.1523/JNEUROSCI.1114-13.2014

R Core Team. (2019). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. *3*. https://www.r-project.org/

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. http://arxiv.org/abs/2103.00020

SMITH, S., & NICHOLS, T. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Spagna, A., Hajhajate, D., Liu, J., & Bartolomeo, P. (2021). Visual mental imagery engages the left fusiform gyrus, but not the early visual cortex: A meta-analysis of neuroimaging evidence. *Neuroscience and Biobehavioral Reviews*, *122*(December 2020), 201–217. https://doi.org/10.1016/j.neubiorev.2020.12.029

Spagna, A., Heidenry, Z., Miselevich, M., Lambert, C., Eisenstadt, B. E., Tremblay, L., Liu, Z., Liu, J., & Bartolomeo, P. (2024). Visual mental imagery: Evidence for a heterarchical neural architecture. *Physics of Life Reviews*, *48*(December 2023), 113–131. https://doi.org/10.1016/j.plrev.2023.12.012

Suzuki, M., Noguchi, Y., & Kakigi, R. (2014). Temporal dynamics of neural activity underlying unconscious processing of manipulable objects. *Cortex*, *50*, 100–114. https://doi.org/10.1016/j.cortex.2013.09.005

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*. https://doi.org/10.1016/j.neuroimage.2021.118145

van den Hurk, J., Van Baelen, M., & Op de Beeck, H. P. (2017). Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences*, *114*(22), E4501–E4510. https://doi.org/10.1073/pnas.1612862114

Vetter, P., Bola, Ł., Reich, L., Bennett, M., Muckli, L., & Amedi, A. (2020). Decoding Natural Sounds in Early "Visual" Cortex of Congenitally Blind Individuals. *Current Biology*, *30*(15), 3039-3044.e2. https://doi.org/10.1016/j.cub.2020.05.071

Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology*, *24*(11), 1256–1262. https://doi.org/10.1016/j.cub.2014.04.020

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, *5*(12), 1415–1426. https://doi.org/10.1038/s42256-023-00753-y

Wang, X., Men, W., Gao, J., Caramazza, A., & Bi, Y. (2020). Two Forms of Knowledge Representations in the Human Brain. *Neuron*, *107*(2), 1–11. https://doi.org/10.1016/j.neuron.2020.04.010

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods*, *42*(3), 671–684. https://doi.org/10.3758/BRM.42.3.671

Wurm, M. F., & Caramazza, A. (2019). Lateral occipitotemporal cortex encodes perceptual components of social actions rather than abstract representations of sociality. *NeuroImage*, *202*. https://doi.org/10.1016/j.neuroimage.2019.116153

Wurm, M. F., & Caramazza, A. (2022). Two 'what' pathways for action and object recognition. *Trends in Cognitive Sciences*, *26*(2), 103–116. https://doi.org/10.1016/j.tics.2021.10.003

Xu, Y. (2018). The Posterior Parietal Cortex in Adaptive Visual Processing. *Trends in Neurosciences*, *41*(11), 806–822. https://doi.org/10.1016/j.tins.2018.07.012

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, *440*(7080), 91–95. https://doi.org/10.1038/nature04262

Xu, Y., Vignali, L., Sigismondi, F., Crepaldi, D., Bottini, R., & Collignon, O. (2023). Similar object shape representation encoded in the inferolateral occipitotemporal cortex of sighted and early blind people. *PLOS Biology*, *21*(7), e3001930. https://doi.org/10.1371/journal.pbio.3001930

Yan, C.-G., Wang, X.-D., Zuo, X.-N., & Zang, Y.-F. (2021). DPABI: Data Processing & Analysis for (Resting-State) Brain Imaging. *Neuroinformatics*. https://doi.org/10.1007/s12021-016-9299-4

Yang, A., Pan, J., Lin, J., Men, R., Zhang, Y., Zhou, J., & Zhou, C. (2023). *Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese*. http://arxiv.org/abs/2211.01335

Zeman, A. (2024). Aphantasia and hyperphantasia: exploring imagery vividness extremes. *Trends in Cognitive Sciences*, *28*(5), 467–480. https://doi.org/10.1016/j.tics.2024.02.007

Zeman, A., Dewar, M., & Della Sala, S. (2015). Lives without imagery – Congenital aphantasia. *Cortex*, *73*, 378–380. https://doi.org/10.1016/j.cortex.2015.05.019

Figures

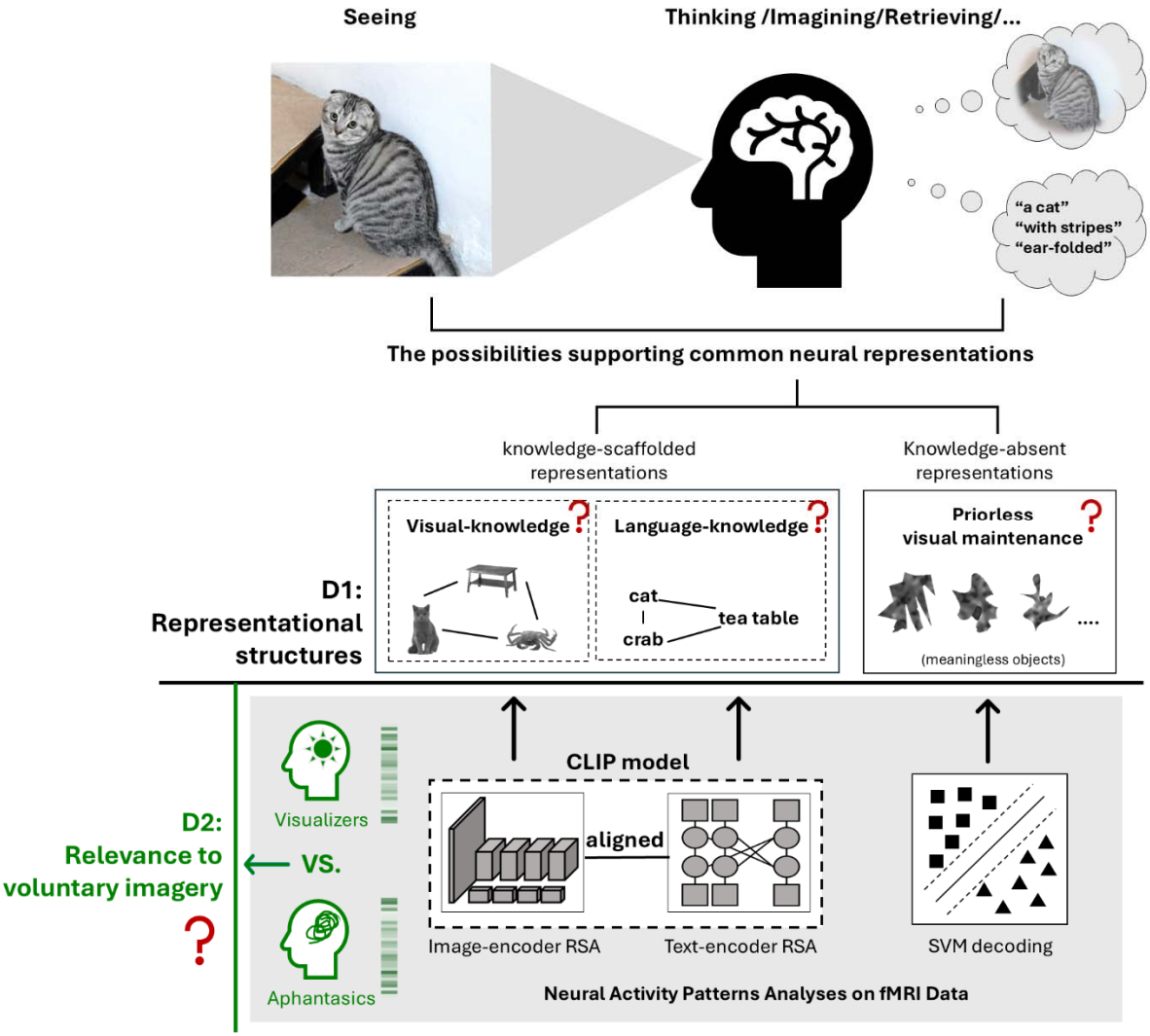


Fig. 1 | Hypotheses framework. We hypothesized that multiple representational mechanisms contribute to shared neural representations across the presence and absence of visual input, and tested the mechanisms along two key dimensions: **D1-representational structures**, including visual-knowledge representation, language-knowledge representation, and priorless visual maintenance representation. **D2-voluntary imagery relevance**. D1 was assessed through the combination of cross-decoding and multimodal deep neural network (DNN) model RSA; and D2 was assessed via group comparisons between participants with and without conscious visual imagery experience (i.e., visualizers vs. aphantasics).

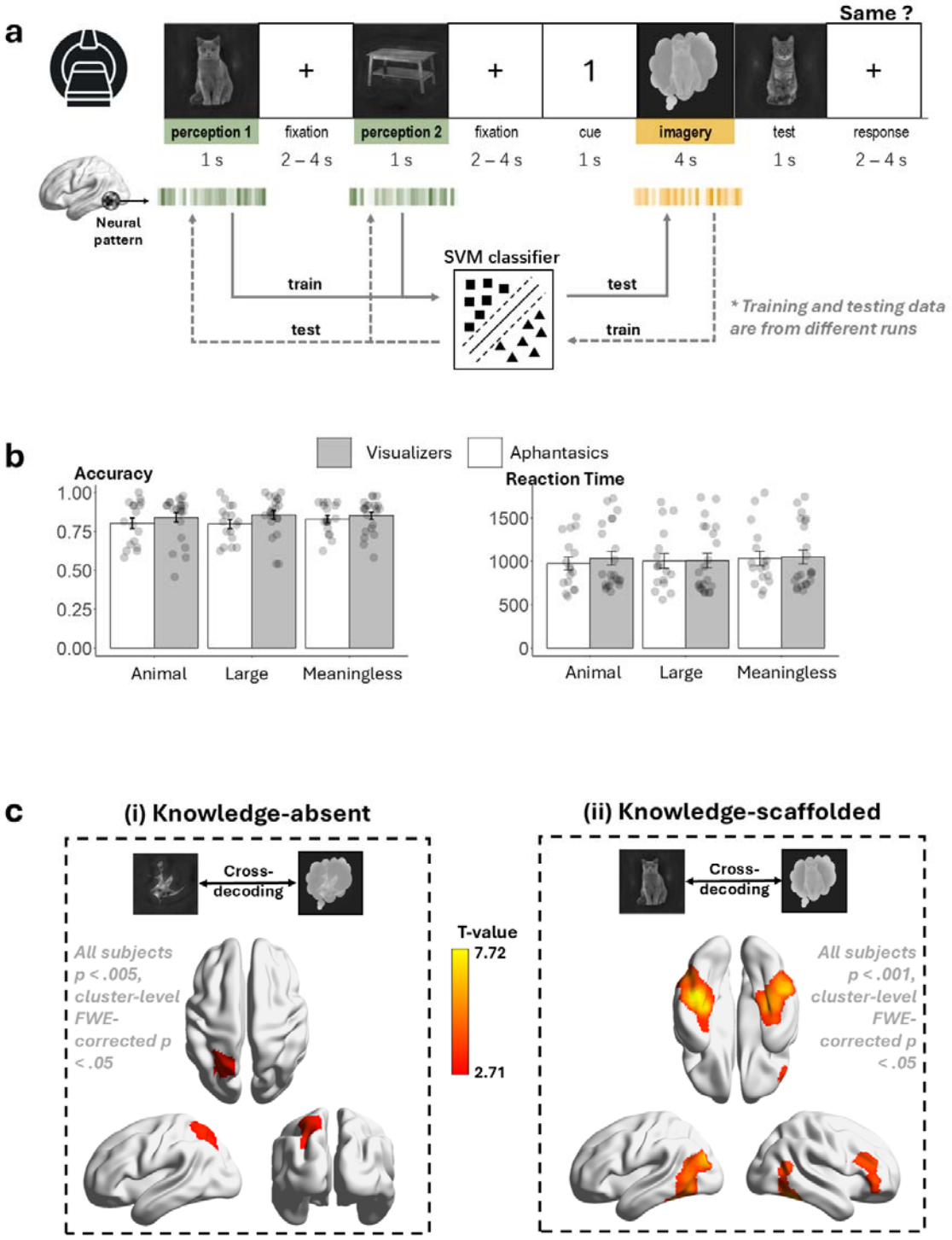


Fig. 2 | Experimental design, behavioral performance, and cross-decoding results. a, Schematic of the fMRI experiment and cross-modal SVM decoding. A retro-cue paradigm was used during scanning. Each trial began with sequential presentation of two images (perception phase), followed by a numeric cue

indicating which image to visualize (imagery phase). A test image was then presented, and participants judged whether it was exactly the same image (not just the same object) as the cued one. SVM decoding was performed across perception and imagery phases from different runs using a repeated split-half cross-validation approach (see 'Searchlight SVM Decoding Analysis' in Methods for details). **b**, Behavioral performance. Bar plots show mean accuracy and reaction times (RTs) for each group (gray = visualizers, white = aphantasics), with error bars representing ±1 SEM. Dots indicate individual participant data. **c**, Searchlight SVM cross-decoding results. (i) Six-way item-level decoding map of meaningless objects (voxelwise $p < .005$, cluster-level FWE-corrected $p < .05$). (ii) Two-way object domain-level decoding map (voxelwise $p < .001$, cluster-level FWE-corrected $p < .05$). Full item-level results are provided in Supplementary Fig. 4.

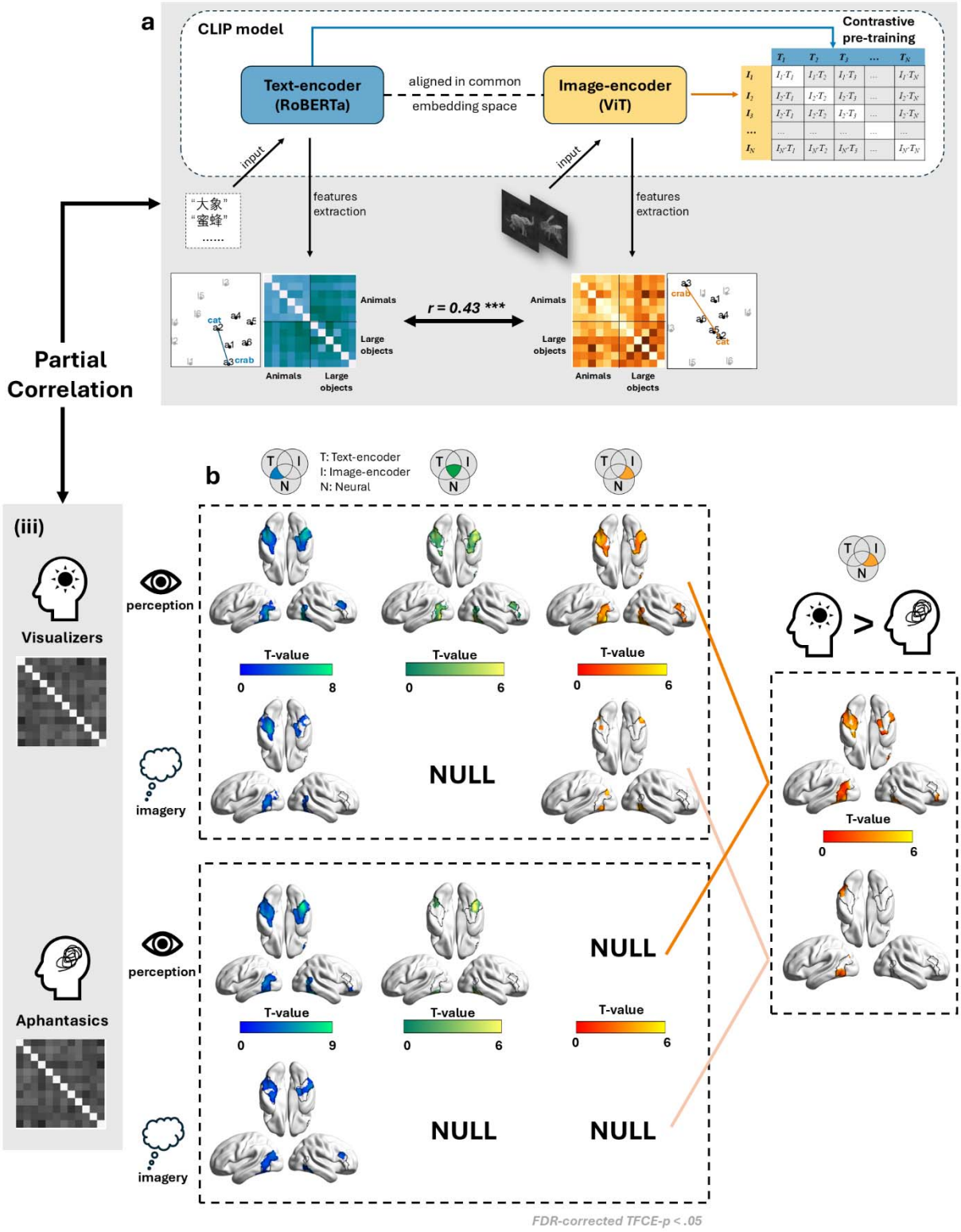


Fig. 3 | Searchlight RSA with CLIP encoders. a, We used a multimodal deep neural network—Contrastive Language-Image Pre-Training (CLIP, Chinese version)—to examine neural representational

structures. It jointly trains an image-encoder and a text-encoder to align images and their corresponding textual descriptions in a shared embedding space using contrastive learning. Stimulus labels were input into the text-encoder and stimulus images into the image-encoder; embeddings from the final output layer were used to compute representational dissimilarity matrices (RDMs) based on cosine distance. Although the RDMs from the image and text encoders were significantly correlated ($r = 0.43$, $p < .001$), they captured distinct representational properties. For example, “cat” and “crab” were closer in the text embedding space but more distant in the visual embedding space. Searchlight RSA was conducted within the mask showing significant cross-decoding for real object domains (from Fig. 2c-ii), which was demarcated by black boundaries on the brain maps. For each searchlight sphere, we computed partial correlations between the neural RDM and one encoder RDM while controlling for the other, as well as the shared variance (commonality) between the neural RDM and the two encoder RDMs. **b**, Searchlight RSA results. Blue regions show significant partial correlations with the text-RDM; orange-red regions show significant partial correlations with the image-RDM; green regions show significant commonality across neural, image, and text-RDMs. Top row: visualizer group; bottom row: aphantasic group; right panel: regions showing significant group differences, observed only in the partial correlations between neural- and image-RDM. Maps are thresholded at FDR-corrected TFCE-$p < .05$. Uncorrected results are shown in Supplementary Fig. 6.

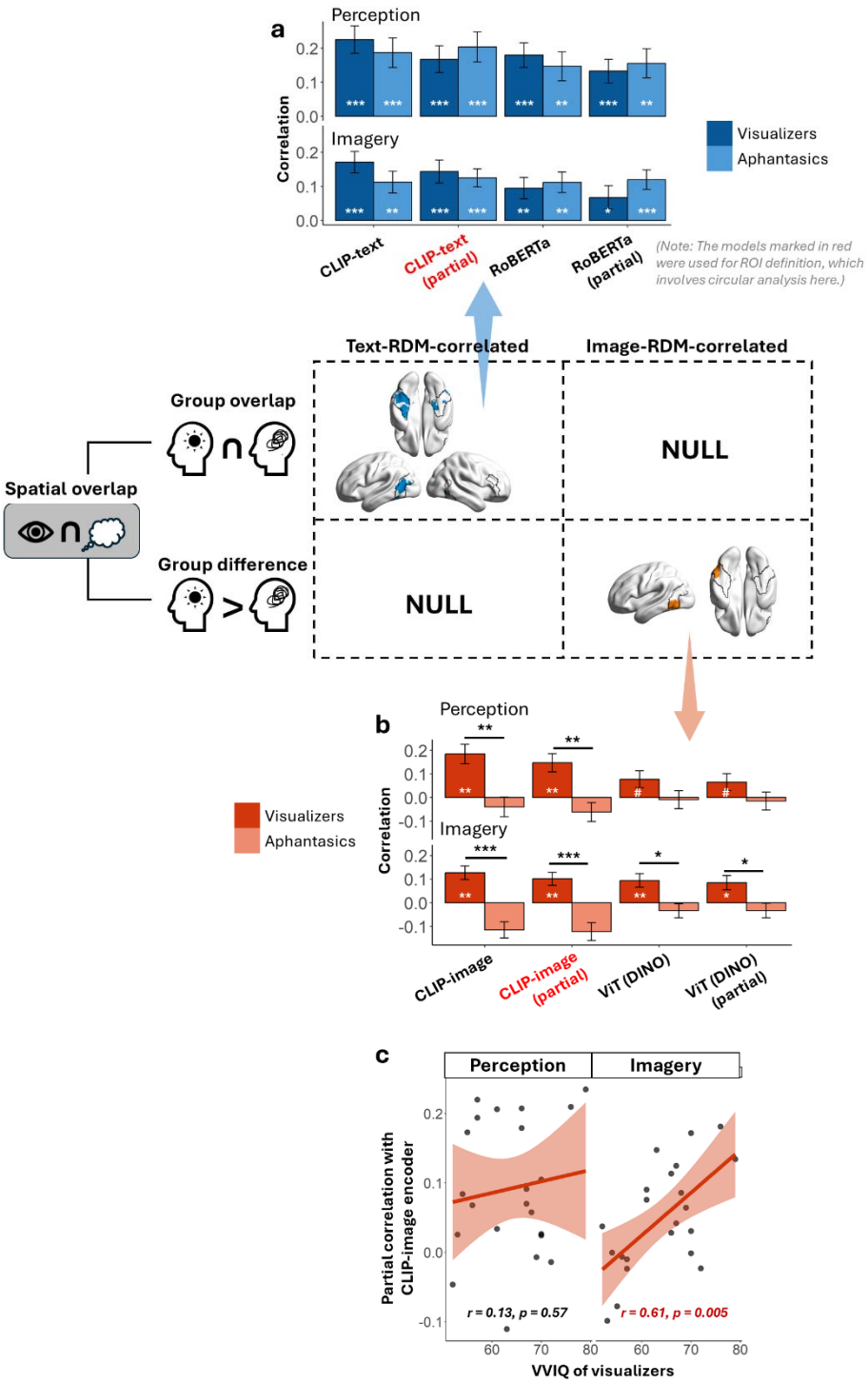


Fig. 4 | ROI-based RSA and individual differences. Two regions of interest (ROIs) were defined based on searchlight RSA results (Fig. 3b): the language-knowledge-imagery-irrelevant ROI was the overlap of regions showing significant similarity to the text-RDM across both perception and imagery phases, and across both groups; the visual-knowledge-imagery-relevant ROI was defined by overlapping regions that showed significant group differences in similarity to the image-RDM across both phases. Two unimodal

models that architecturally matched to the respective CLIP encoders were used for RSA: **a**, RoBERTa (language model) for the language-knowledge-imagery-irrelevant ROI; and **b**, ViT-DINO (vision model) for the visual-knowledge-imagery-relevant ROI. Partial correlations for these models were computed by regressing out the same CLIP encoder RDM as was done in the CLIP analysis. Note: Red-labeled models on the x-axis indicate the measurements used to define the ROIs, and thus reflect double-dipping (Kriegeskorte et al., 2009); they are shown here for visualization purposes only. Significance markers: #$p < .1$, *$p < .05$, **$p < .01$, ***$p < .001$. **c**, Correlation between VVIQ scores and neural representational similarity to the image-RDM within visualizers. Partial correlations (y-axis) were values extracted and averaged from the visual-knowledge-imagery-relevant ROI. All statistical tests shown were FDR-corrected.

Hypotheses		D2: Voluntary-imagery-relevance	
		imagery-relevant	imagery-irrelevant
D1: Representational structures	Visual-structured knowledge	Left ventral LOTC	—
	Language-structured knowledge	—	Left dorsal LOTC Bilateral FG Right ventral LOTC
	Priorless visual maintenance	—	Left SPL

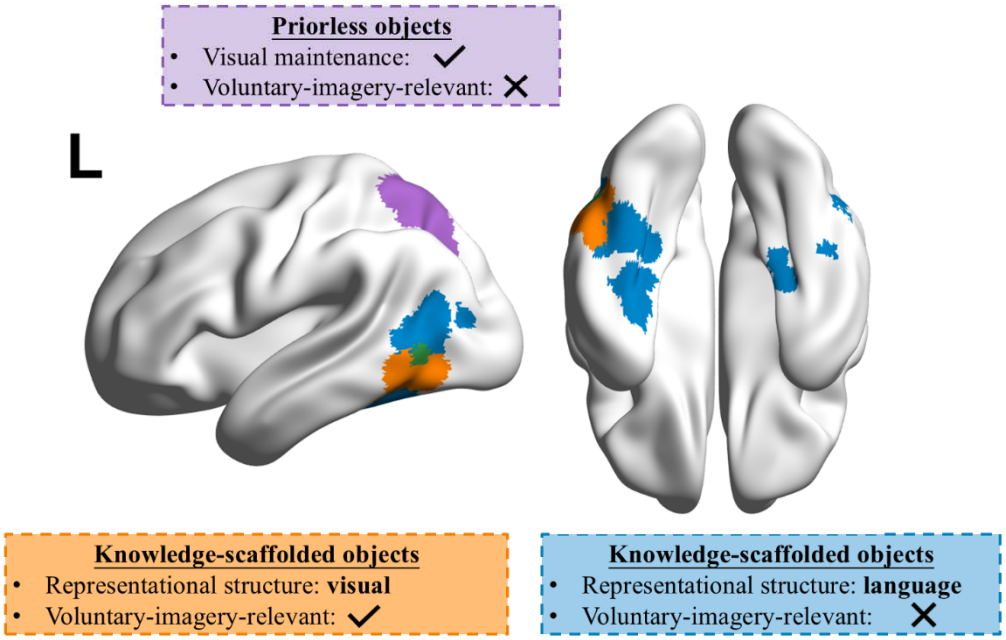
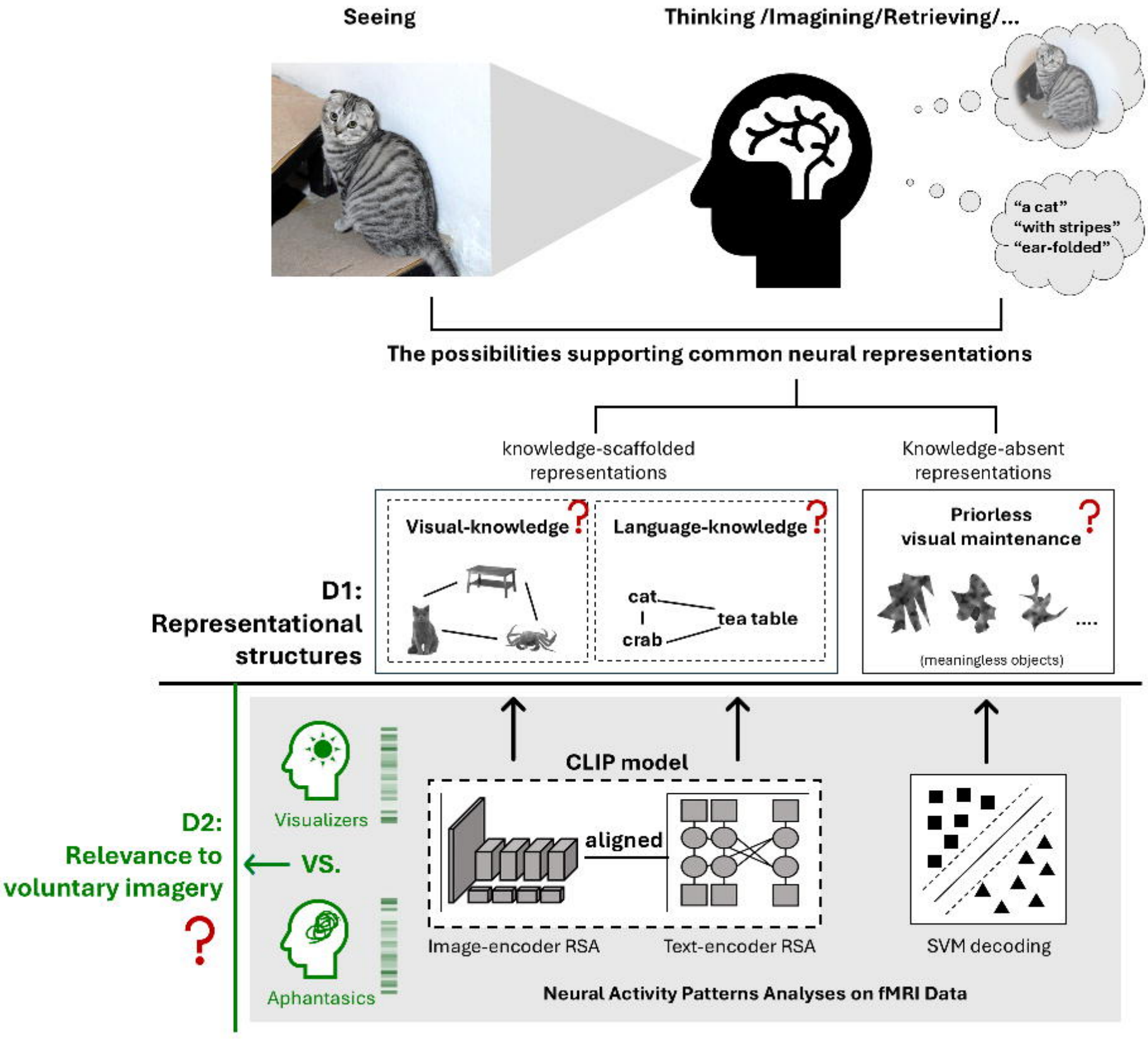
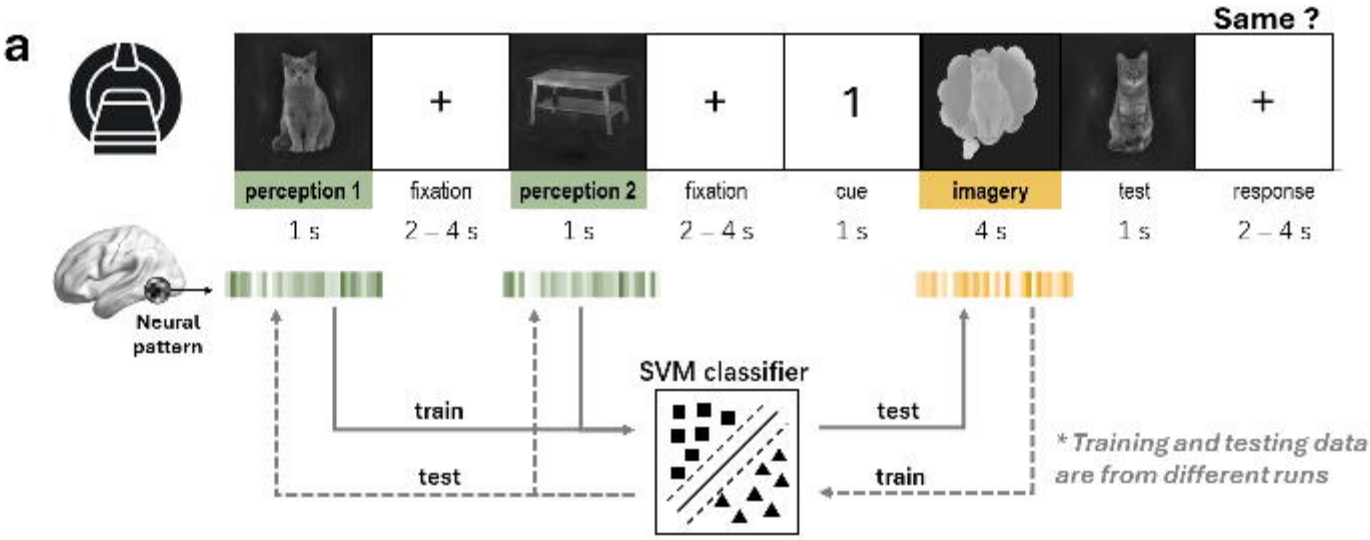
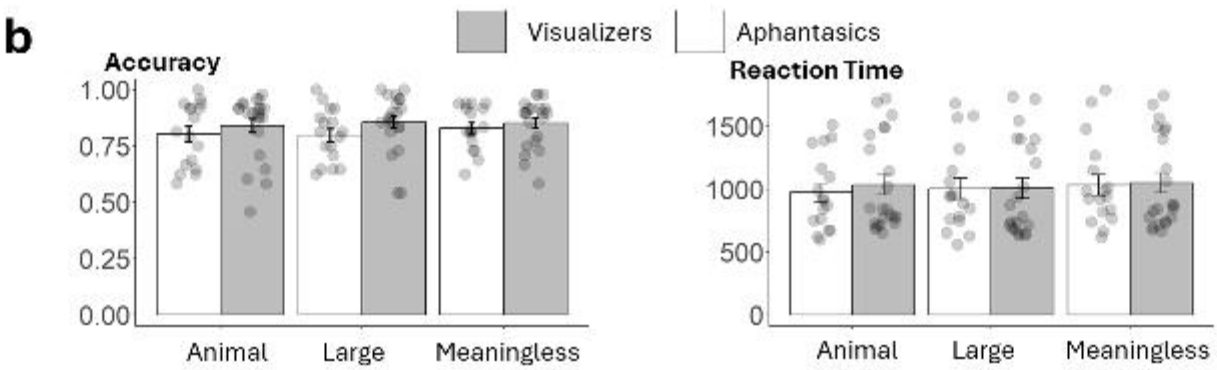
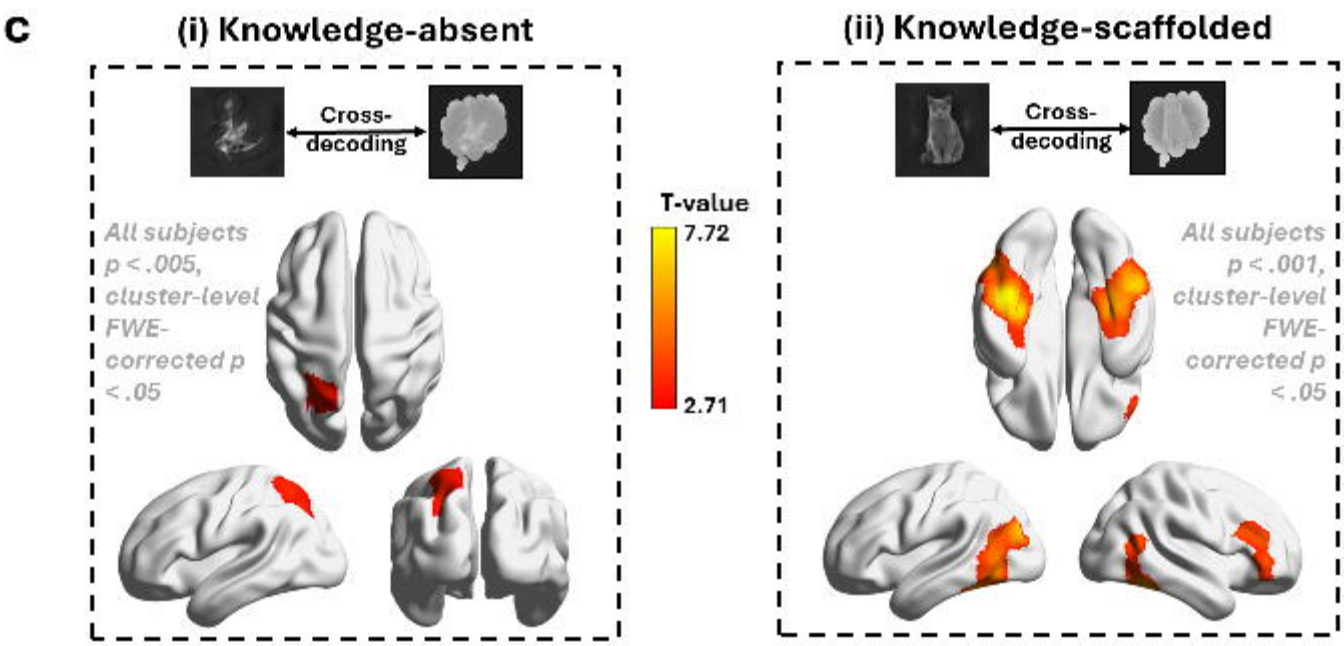


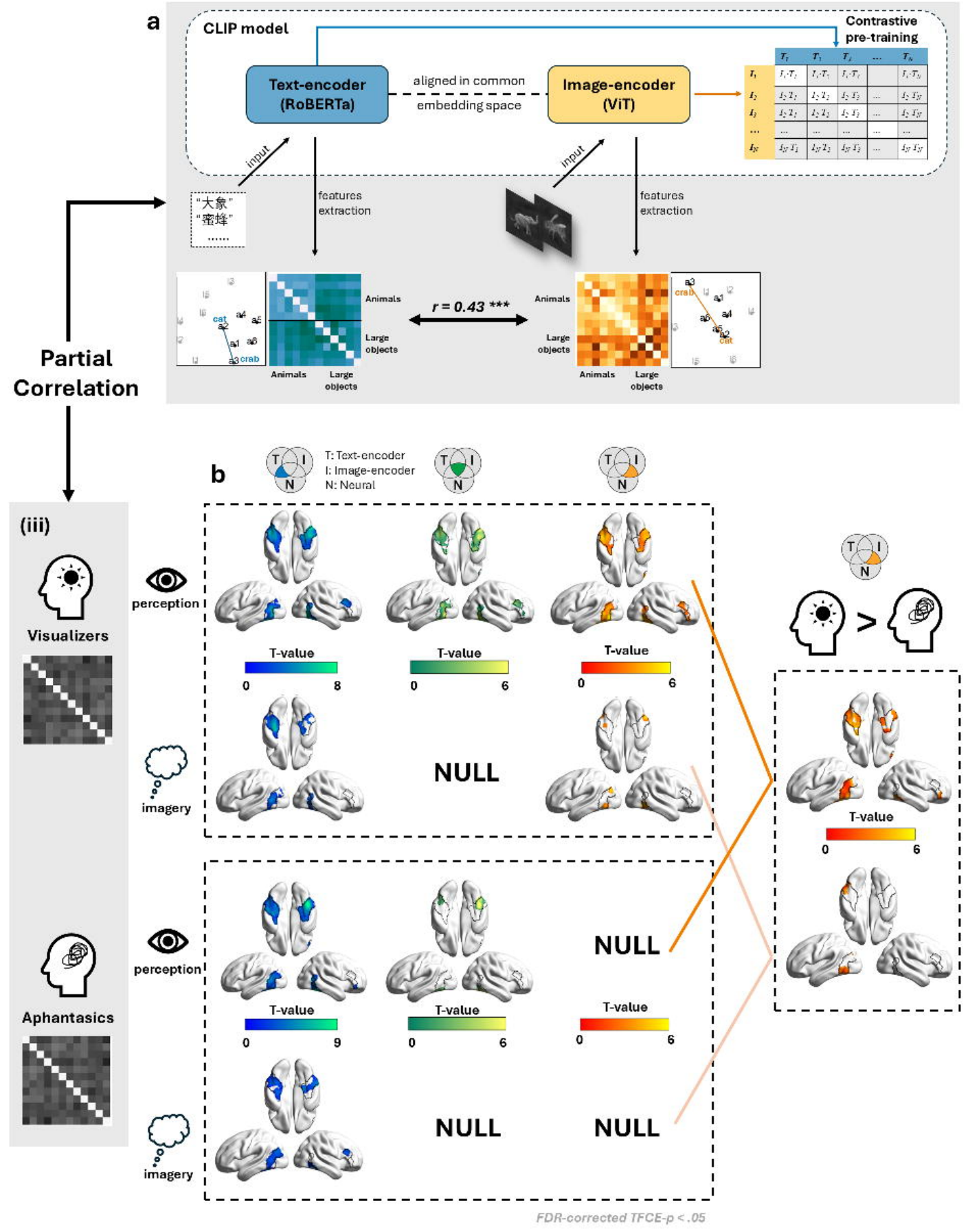
Fig. 5 | Summary of results mapped onto hypothesis framework. The results reveal distinct brain regions supporting shared neural representations across perception and imagery, each associated with specific representational structures: The left ventral LOTC (orange) encodes visual-structured knowledge representation that is sensitive to voluntary imagery experience. The bilateral FG, left dorsal LOTC, and right ventral LOTC (blue) encode language-structured knowledge representation that is irrelevant to imagery experience. The left SPL (purple) supports priorless visual representation that is unaffected by imagery experience. Green indicates overlapping areas.

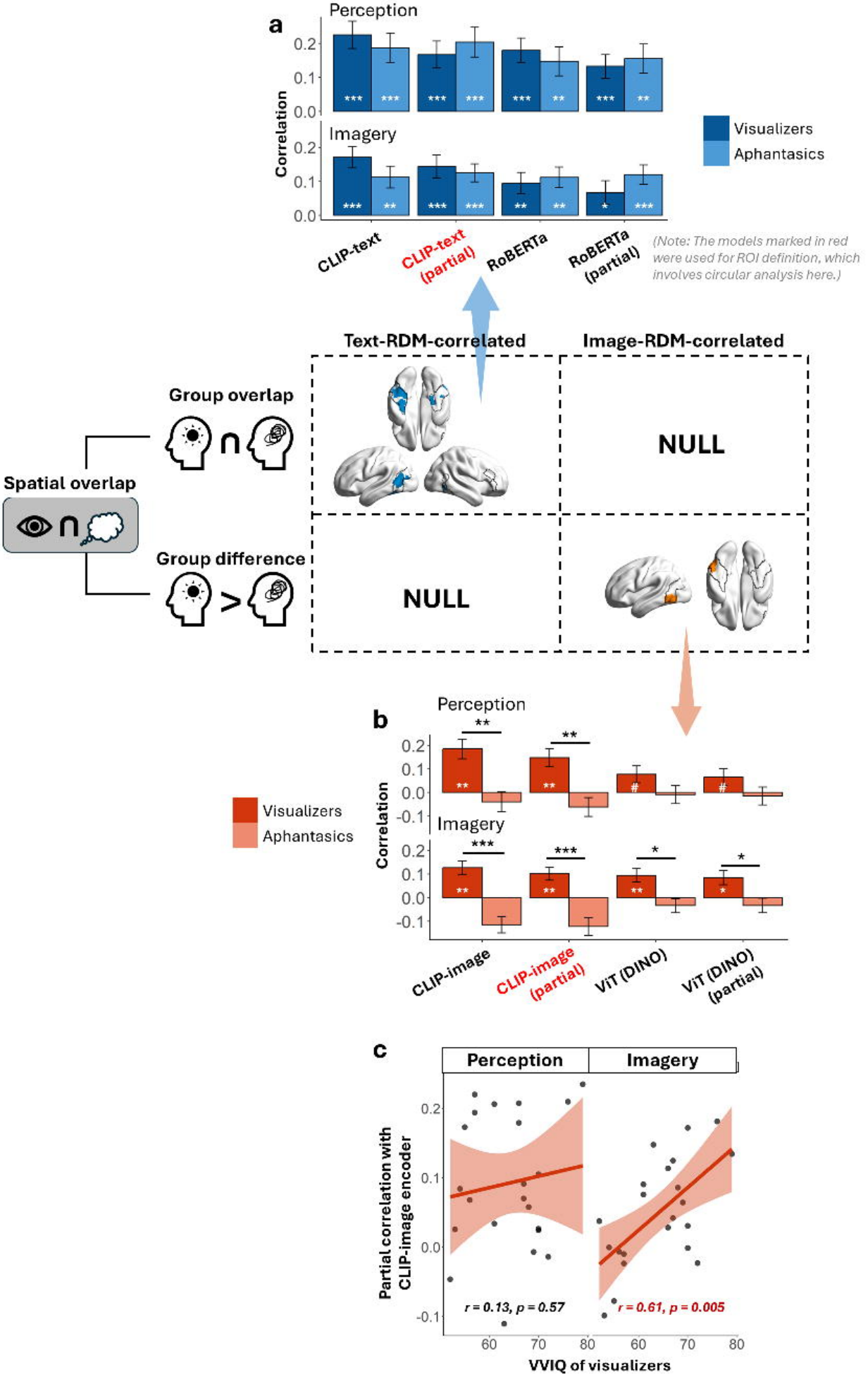
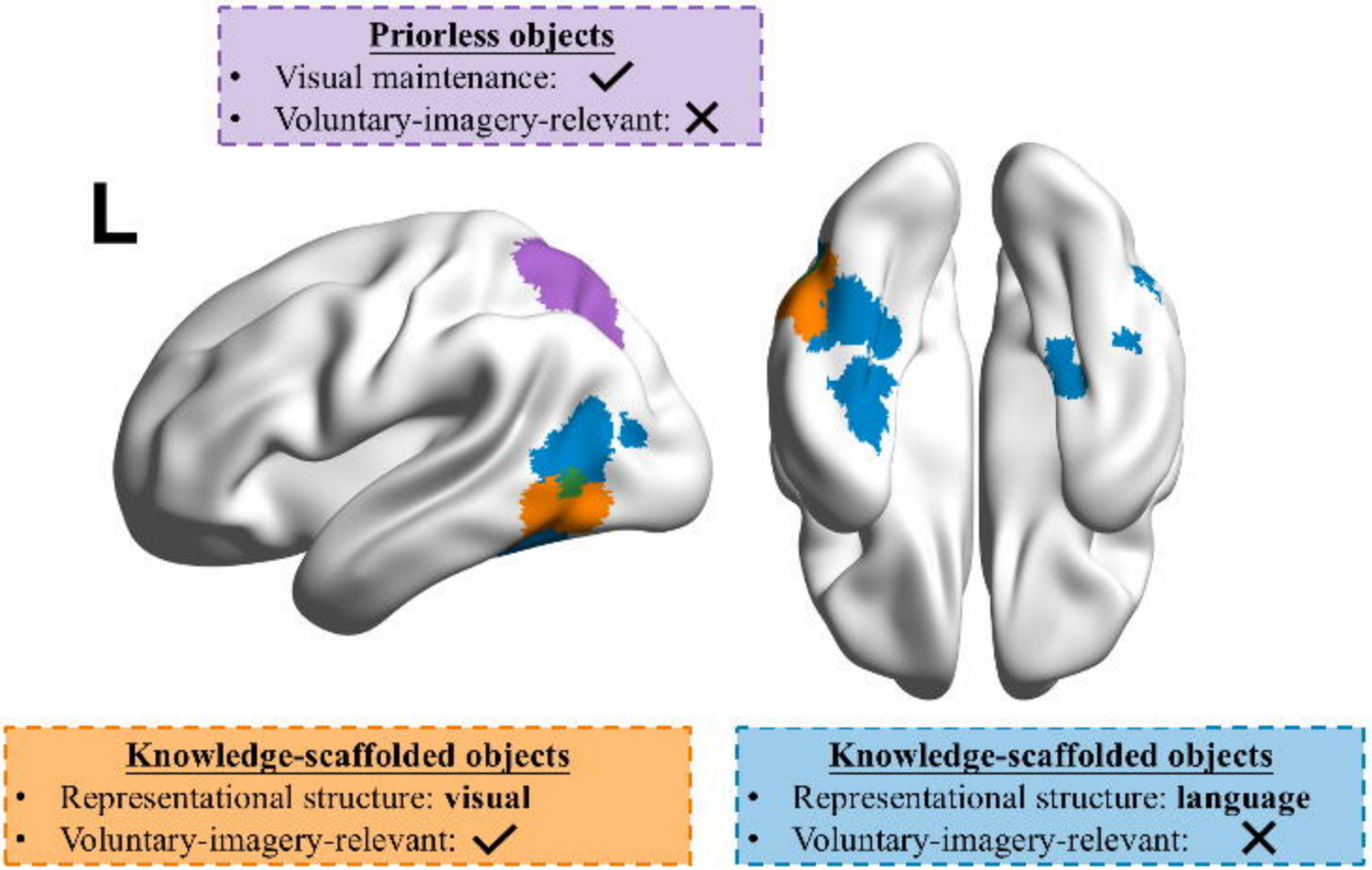

Seeing
Thinking /Imagining/Retrieving/...
"a cat"
"with stripes"
"ear-folded"
The possibilities supporting common neural representations
knowledge-scaffolded representations
Knowledge-absent representations
Visual-knowledge
Language-knowledge
Priorless visual maintenance
cat
crab
tea table
....
(meaningless objects)
D1: Representational structures
D2: Relevance to voluntary imagery
Visualizers
VS.
Aphantasics
CLIP model
aligned
Image-encoder RSA
Text-encoder RSA
SVM decoding
Neural Activity Patterns Analyses on fMRI Data


a
Same ?
+
+
1
+
perception 1
fixation
perception 2
fixation
cue
imagery
test
response
1 s
2 – 4 s
1 s
2 – 4 s
1 s
4 s
1 s
2 – 4 s
Neural pattern
SVM classifier
train
test
test
train
* Training and testing data are from different runs


b
Visualizers
Aphantasics
Accuracy
1.00
0.75
0.50
0.25
0.00
Animal
Large
Meaningless
Reaction Time
1500
1000
500
0
Animal
Large
Meaningless


c
(i) Knowledge-absent
Cross-decoding
All subjects p < .005, cluster-level FWE-corrected p < .05
T-value
7.72
2.71
(ii) Knowledge-scaffolded
Cross-decoding
All subjects p < .001, cluster-level FWE-corrected p < .05


a
CLIP model
Contrastive pre-training
Text-encoder (RoBERTa)
aligned in common embedding space
Image-encoder (ViT)
input
features extraction
"大象"
"蜜蜂"
.....
r = 0.43 ***
Animals
Large objects
cat
crab
Partial Correlation
(iii)
Visualizers
Aphantasics
b
T: Text-encoder
I: Image-encoder
N: Neural
perception
imagery
T-value
NULL
FDR-corrected TFCE-p < .05


a
Perception
Imagery
Correlation
Visualizers
Aphantasics
CLIP-text
CLIP-text (partial)
RoBERTa
RoBERTa (partial)
(Note: The models marked in red were used for ROI definition, which involves circular analysis here.)
Text-RDM-correlated
Image-RDM-correlated
Group overlap
Spatial overlap
Group difference
NULL
NULL
b
Perception
Imagery
Correlation
Visualizers
Aphantasics
CLIP-image
CLIP-image (partial)
ViT (DINO)
ViT (DINO) (partial)
c
Perception
Imagery
Partial correlation with CLIP-image encoder
r = 0.13, p = 0.57
r = 0.61, p = 0.005
VVIQ of visualizers

Hypotheses		D2: Voluntary-imagery-relevance	
		imagery-relevant	imagery-irrelevant
D1: Representational structures	Visual-structured knowledge	Left ventral LOTC	—
	Language-structured knowledge	—	Left dorsal LOTC Bilateral FG Right ventral LOTC
	Priorless visual maintenance	—	Left SPL