Dissociating internal representations underlying knowing and experiencing

Shuang Tian,[1,2] Xiaomin Mao,[2] Dahui Wang,[1,2,3] Xiaoying Wang,[2]* 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); Xiaoying Wang (email: wangxiaoying@bnu.edu.cn)**

Abstract

How do internal brain representations bridge seeing an object and thinking about it after it disappears? Both object knowledge and mental imagery are involved in this process, engaging overlapping perceptual regions, yet whether their neural codes are shared or distinct remains unknown. We compared people with ("visualizers") and without voluntary visual imagery ("aphantasics") using fMRI, to examine experience of imagery sensation, and a multimodal deep neural network model, to examine representational contents (encoding text vs. image). We found distinct types of internal representations: (1) the left lateral occipitotemporal cortex (LOTC) encoded visual-structured knowledge linked to imagery sensation; (2) the bilateral fusiform gyrus, left dorsal LOTC, and right inferior frontal gyrus encoded language-structured knowledge independent of imagery sensation; and (3) the left superior parietal lobule maintained visual representation without prior knowledge, also independent of imagery. These findings reveal functionally and computationally distinct neural mechanisms that bridge seeing and thinking of objects, differing in their reliance on knowing and internal experiencing.

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 internal representations derived from the 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 (*1*–*7*), and retrieving object color or shape properties activate regions in the fusiform gyrus and lateral occipital cortex, overlapping with those perceiving color and shape (*8*–*12*). Even early visual cortex has been implicated in semantic representations (*13*, *14*).

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 (*15*–*18*). 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., 19–21). Beyond early visual cortex, imagery-related representations have also been observed in higher-level regions including the fusiform gyrus (*22*), lateral occipital cortex (*20*, *23*), and parietal cortex (*24*, *25*).

These findings together suggest a potentially shared representations among object perception, knowledge, and imagery. This convergence raises the question: do imagery and knowledge retrieval rely on the same neural code, differing only in conscious access requirement (see discussions in 26–28), or do they engage distinct representations at overlapping regions? Are these representations “visual”-related in nature by the virtual of overlapping with visual perception, i.e., aligning with the commonly held “embodied” (or “grounded”) notion (see reviews in 28, 29)

Some evidence from congenital aphantasia (*30*, *31*)—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 (*32*). 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 (*33*, *34*). 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 (*35*). 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 propose 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 (*36*, *37*). 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 38–40). Notably, congenitally blind people lack effects in visual cortex when retrieving pure visual properties (e.g., color; 39) 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 (*41*). 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 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.

Building on prior findings, we conducted fMRI experiments using a retro-cue paradigm (*19*, *21*) 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 in both perception and imagery; ***3) priorless visual representation***: If low-level visual maintenance is sufficient to support shared representations, then even meaningless shapes—lacking semantic associations—should be decoded above chance 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.

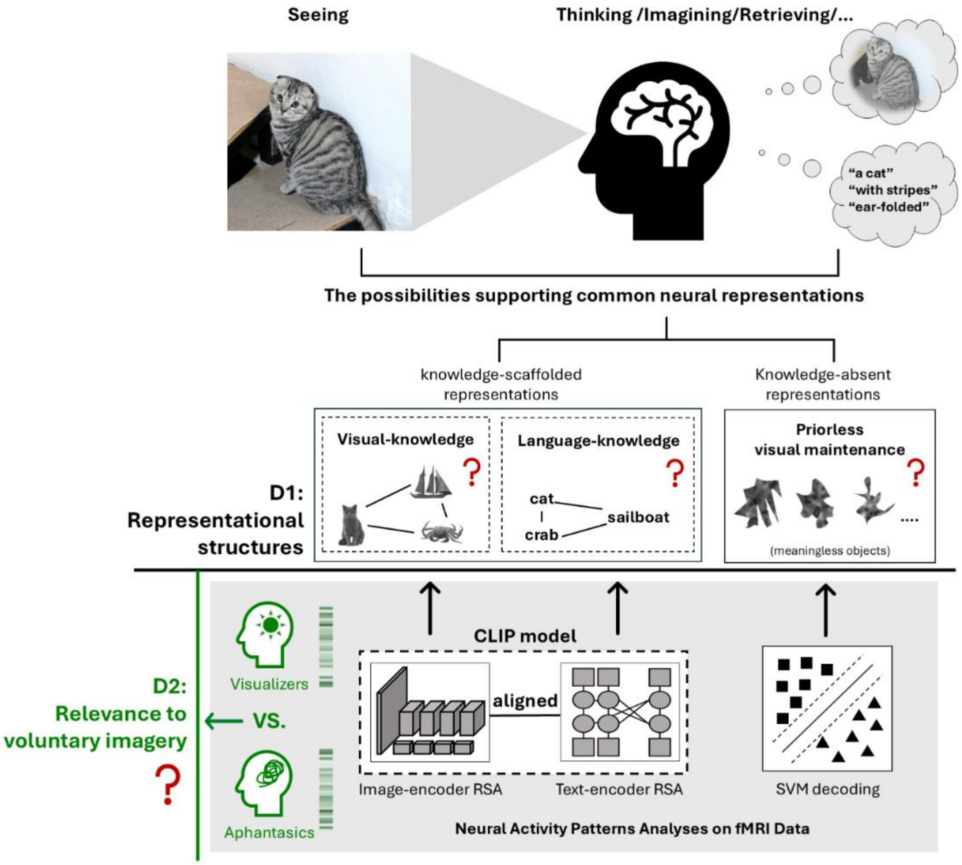

Seeing
Thinking /Imagining/Retrieving/...
"a cat"
"with stripes"
"ear-folded"
The possibilities supporting common neural representations
knowledge-scaffolded representations
Knowledge-absent representations
Visual-knowledge
Language-knowledge
cat
sailboat
crab
Priorless visual maintenance
(meaningless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

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).

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 well. 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, layer-wise ROI-based RSA was conducted to explore the neural representational similarity profiles across different processing hierarchies of the encoders.

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 shapes 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.

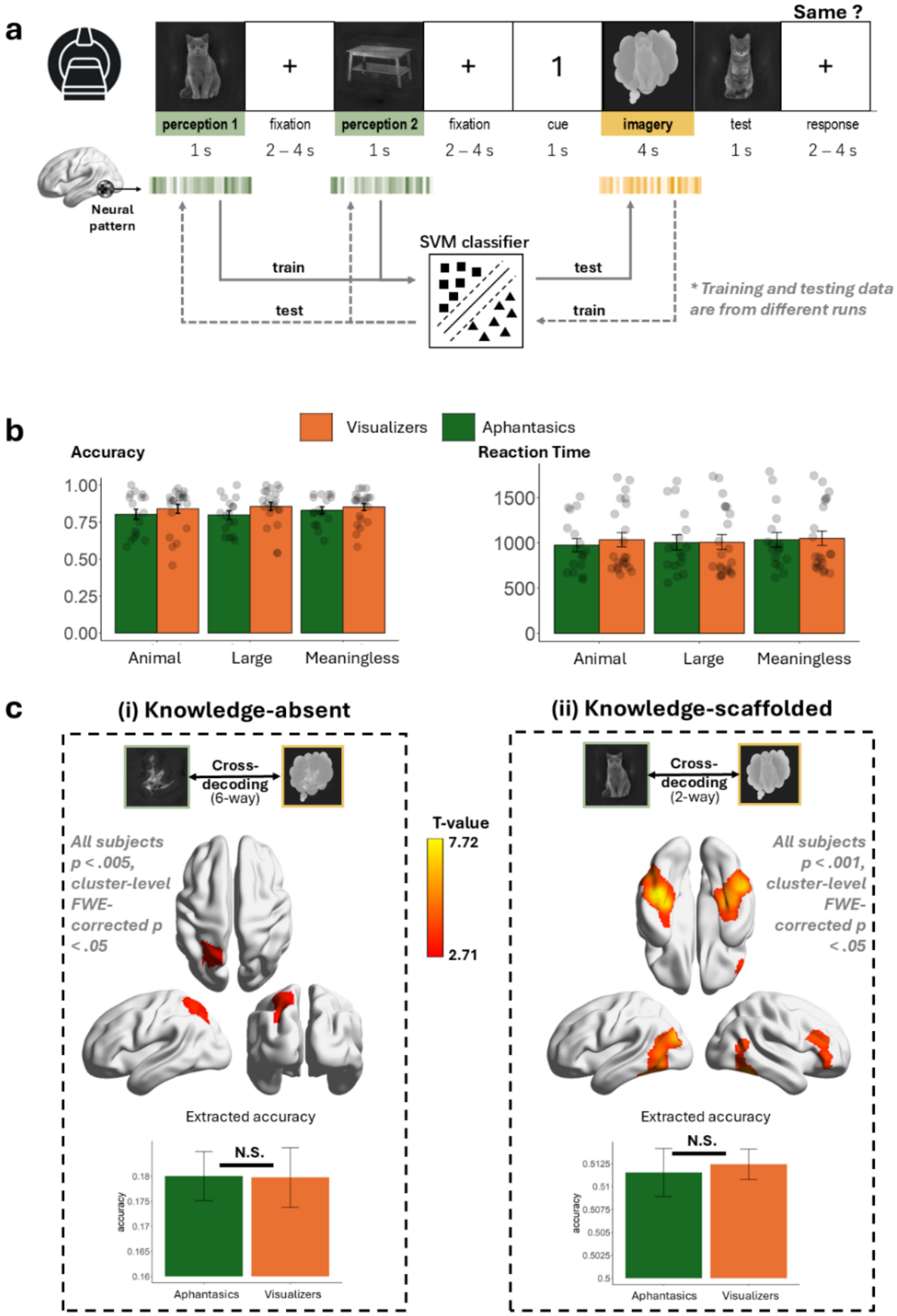


Fig. 2. Experimental design, behavioral performance, and cross-decoding results. a, Schematic of the fMRI experiment and cross-modal SVM decoding. A retro-cue paradigm was used during scanning. Each trial began with sequential presentation of two images (perception phase), followed by a numeric cue indicating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

(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 Fig. S4. Bar plots display the mean accuracy extracted from the significant clusters shown above for each group, with error bars representing ±1 SEM.

Whole-brain 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 shapes 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 the perception and imagery phases for meaningless shapes (priorless visual representation) in both aphantasics and visualizers***.**

To assess whether visual maintenance alone is sufficient to support shared neural representations between the perception and imagery phases, we conducted searchlight SVM cross-decoding for meaningless shapes (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 the 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 shapes 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. Group comparison did not reveal statistically survived clusters (see Fig. S2 for the uncorrected maps).

*The VOTC areas contained shared representations across the 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 the 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 (Fig. S3) showed that visualizers had successful cross-decoding for domains in bilateral VOTC, extending to the left 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 differences by voxels 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 Fig. S3 for an uncorrected result).

We further conducted an item-level cross-decoding (12-way classification) for real objects. This analysis revealed similar but spatially more restricted results (Fig. S4) 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 of $p < .01$.

These results suggest that the VOTC regions contained shared representations across the 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 (*41*, *42*), 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 the objects into the text-encoder, extracted the features from the *penultimate layers*, and transformed them into RDMs based on cosine distance (Fig. 3a). There was a medium-level, significant correlation between the RDMs of the two encoders ($r = 0.40$, $p < .001$). 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 analysis (*43*). Because the shared variance among image-, text-, and neural-RDMs cannot be unambiguously assigned to either representational structure, we focused on the unique contributions of the image- and text-encoders to the neural representations.

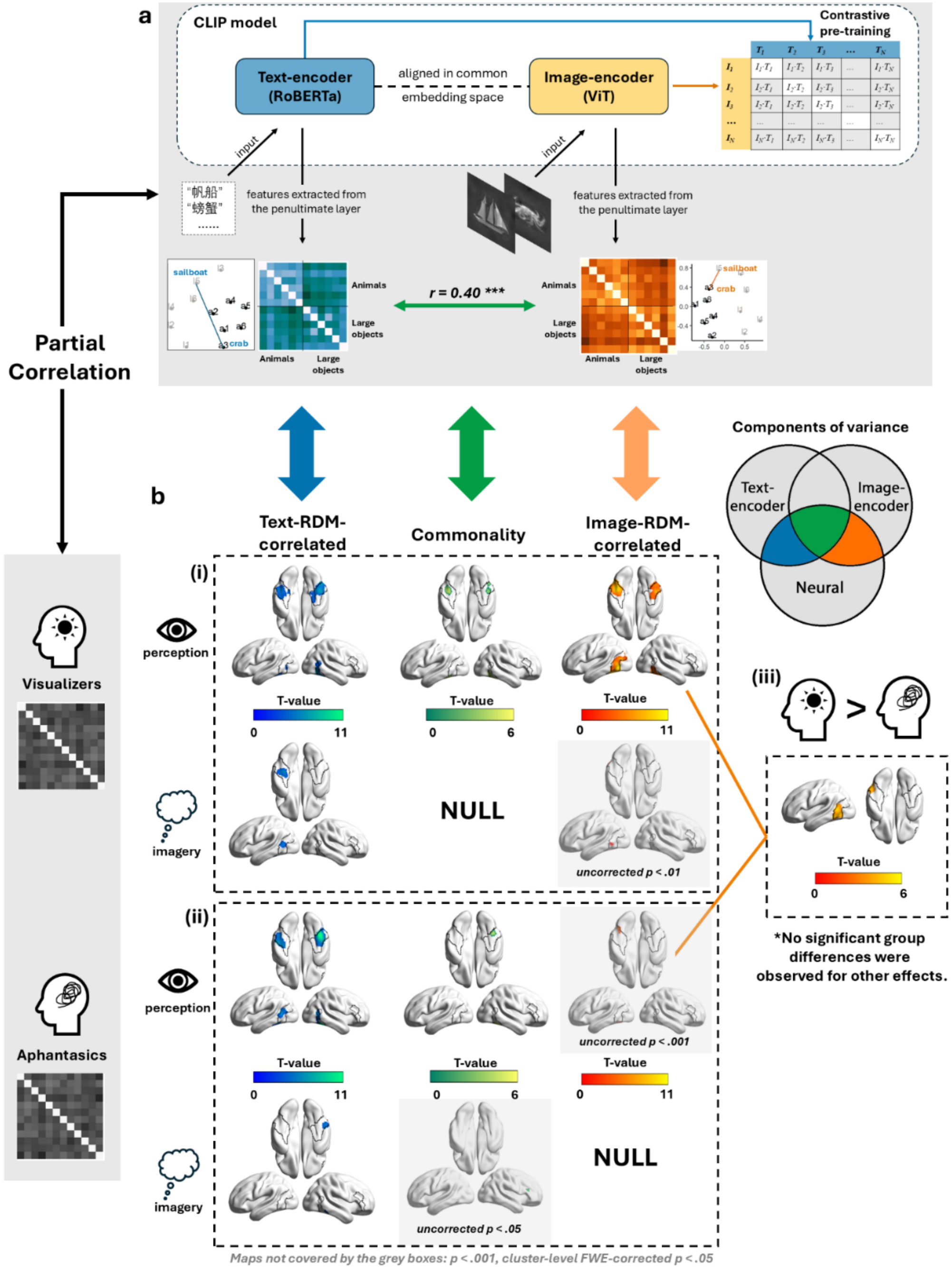


Fig. 3. Searchlight RSA with CLIP encoders. a, We used a multimodal deep neural network—Contrastive Language-Image Pre-Training (CLIP, Chinese version)—to examine neural representational structures. It jointly trains an image-encoder and a text-encoder to align images and their corresponding textual 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 penultimate 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 = 0.40, p < .001), they captured distinct representational properties. For example, "sailboat" and "crab" were farther apart in the text embedding space but closer together 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. The Venn diagram indicates the components of variance. (i) results in the visualizer group; (ii) results in the aphantasic group; (iii) regions showing significant group differences, observed only in the partial correlations between neural- and image-RDM during the perception phase. Maps not covered by the grey boxes are thresholded at voxelwise p < .001, cluster-level FWE-corrected p < .05. Maps within the grey boxes did not survive multiple-comparisons correction and are displayed at the most stringent threshold they reached. All uncorrected results are shown in Fig. S6-7.

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

In visualizers (Fig. 3b, top), during the perception phase, partial correlations between neural and CLIP text-encoder RDMs (blue) appeared in the bilateral FG, left dorsal LOTC, and right LOTC. During the imagery phase, effects were restricted to the left FG and left dorsal LOTC.

Partial correlations with the CLIP image-encoder RDM (orange-red) during the perception phase were distributed across the bilateral FG and extended into the bilateral LOTC. No significant effects emerged during the imagery phase (see the map covered by the grey box in Fig. 3b and Fig. S5 for uncorrected results).

Commonality analysis (green) revealed shared variance among neural, image-encoder, and text-encoder RDMs in the bilateral FG during the perception phase, with no significant clusters during the imagery phase even at the highly lenient threshold of p < .05.

Only unique language-structured representations existed in aphantasics.

In the aphantasic group (Fig. 3b, bottom), partial correlations with the CLIP text-encoder RDM were observed during the perception phase in the bilateral FG, left dorsal LOTC, and right LOTC, consistent with the visualizer group. During the imagery phase, only a cluster in the right lateral FG was observed. No significant correlations with the CLIP image-encoder RDM were observed during either the perception or imagery phase. Commonality effects were observed in the right FG during the perception phase but not observed during the imagery phase (see maps covered by the grey boxes in Fig. 3b and Fig. S5 for uncorrected results).

Group difference existed in visual-structured representations during the perception phase.

Two-sample t-tests within the mask (Fig. 3b, right) revealed significant group differences specifically for the effect of CLIP image-encoder RDM, with visualizers showing greater model-neural similarity than aphantasics in the left LOTC. No significant group differences emerged in similarity to the CLIP text-encoder RDM, either during the perception phase or the imagery phase (see Fig. S6 for uncorrected results).

Summary

To summarize the searchlight CLIP RSA findings, two types of representations with different relevance to voluntary imagery experience emerged:

One type was observed in both groups, showing reliable similarity to the CLIP text-encoder (penultimate layer representations) during both the perception and imagery phases, without significant group differences. We therefore combined all subjects (visualizers and aphantasics) to perform group-level inference, enhancing the common effect of the CLIP text-encoder, and overlaid the statistical maps from the perception and imagery phases (Fig. 4, blue areas; Fig. S7). Overlapping regions were observed in the bilateral FG, left dorsal LOTC, and right IFG. These regions likely represent objects in terms of their relations within the language space, supporting both perception and imagery, and their involvement does not depend on voluntary imagery experience. That is, having representations here does not determine whether one can mentally visualize objects. We refer to these regions as **“Language-knowledge-imagery-irrelevant.”**

Another type was observed only in the visualizer group, showing significant similarity to the CLIP image-encoder (penultimate layer representations) during the perception phase, with a significant group difference in the left LOTC (Fig. 4, orange area). This region likely represents objects in terms of their visual experiential properties in typical populations, and the absence of such representations is associated with the lack of voluntary visual imagery. We refer to this region as **“Visual-knowledge-imagery-relevant”.**

ROI-based RSA results: In-depth-analyses with different hierarchies of CLIP model encoders

In this section we zoom into the two types of regions of interest (ROIs) identified above -- “Language-knowledge-imagery-irrelevant” ROIs, including the left dorsal LOTC, bilateral FG, and right IFG (blue areas in Fig. 4); “Visual- knowledge-imagery-relevant” ROI, i.e., the left LOTC (orange area in Fig. 4) -- to characterize their representational profiles, referencing to representations at different layers of the target DNN models.

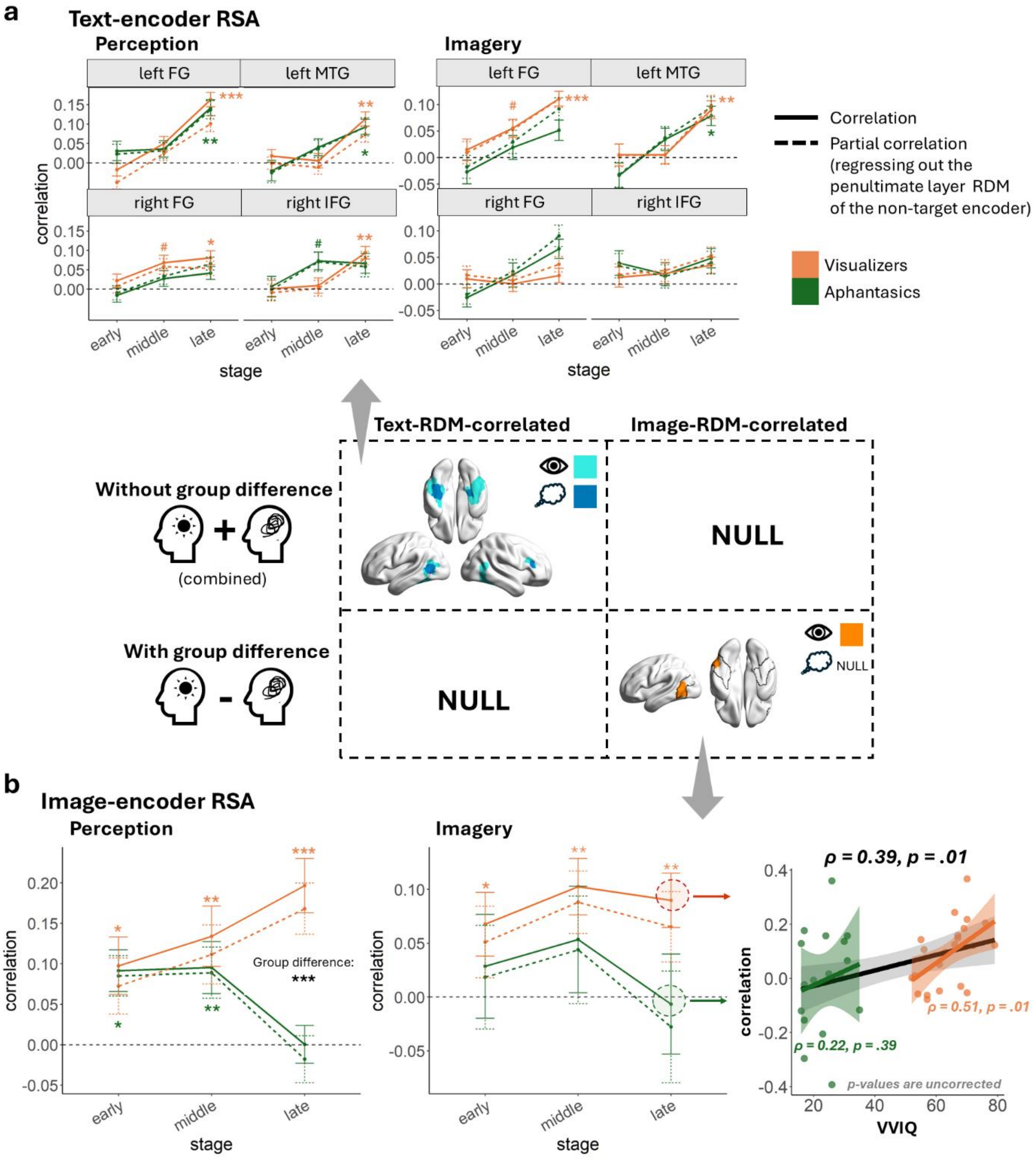


Fig. 4. ROI-based RSA with different processing hierarchies of the CLIP model. Two types of regions of interest (ROIs) were defined based on searchlight RSA results: the **language-knowledge-imagery-irrelevant ROIs** were defined as the overlap of regions showing significant similarity to the text-RDM in the group-level analysis combining all subjects across both the perception and imagery phases (see statistical maps in Fig. S7); the **visual-knowledge-imagery-relevant ROI** was defined as the region showing significant group differences in similarity to the image-RDM, which was only observed during the perception phase in the searchlight analysis. **a**, The layers of the text-encoder were grouped into three parts: early (layers 1-4), middle (5-8), and late (9-12). Correlations (solid lines) and partial correlations (dashed lines) were computed using RDMs from all layers and were averaged within each layer-group. Partial correlations were computed while regressing out the penultimate-layer RDM of the image-encoder. **b**, The layers of the image-encoder were grouped as early (layers 1–8), middle (9–16), and late (17–24). Partial correlations were computed while regressing out the penultimate-layer RDM of the text-encoder. The right panel shows the correlations between VVIQ

scores and the mean correlations of the late layers across all subjects (black line), aphantasics (green line), and visualizers (orange line). Asterisks and pound sign indicate different significance levels of correlations: #p < .1, *p < .05, **p < .01, ***p < .001. All statistical tests shown were FDR-corrected, except for the right panel in **b**.

Visualizers and aphantasics converged across multiple hierarchies in similarity to the CLIP text-encoder, during both the perception and imagery phases.

To test whether representational similarity to the text-encoder differed between groups at earlier representational hierarchies, we grouped the text-encoder layers into three parts: early (layers 1 - 4), middle (5 - 8), and late (9 - 12) (see layerwise results in Fig. S8). Features from all layers were extracted and converted into RDMs, which were then used to compute correlations and partial correlations with neural RDMs of the language-knowledge-imagery-irrelevant ROIs (i.e., left dorsal LOTC, bilateral FG, and right IFG). As regressing out a specific RDM (i.e., the penultimate layer RDM of the non-target encoder) was not fair for all layers of the target encoder, statistical inferences were performed directly on the mean Fisher-Z-transformed correlations, instead of partial correlations, of the layer-groups. As shown in Fig. 4a, both visualizers and aphantasics exhibited an increasing trend in similarity from early to late layers in all ROIs during both the perception and imagery phases.

During the perception phase, a mixed ANOVA with factors subject-group (visualizers, aphantasics), layer-group (early, middle, late), and ROI revealed a significant main effect of layer-group ($F_{(2,74)} = 23.41$, $p = 1.32 \times 10^{-8}$, $\eta^2 = 0.08$) and a significant interaction between layer-group and ROI ($F_{(6,222)} = 2.3$, $p =0.04$, $\eta^2 = 0.02$). Simple effects test indicated significant layer-group effect in the left FG and left MTG ($Fs > 6.33$, $ps < .004$, FDR-corrected), with the late layers being significantly higher than the former two layer-groups ($ps < 0.01$, FDR-corrected) but no difference between the early and middle layers ($ps > 0.22$, FDR-corrected). The right FG and right IFG showed marginally significant simple effects of layer-group ($Fs < 2.57$, $ps > 0.08$, FDR-corrected), with only a trend of significant difference between the late and early layers ($ps > .07$, FDR-corrected). No effects of subject-group, ROI, or their interactions were observed ($Fs < 1.11$, $ps > .36$).

During the imagery phase, the mixed ANOVA revealed a significant main effect of layer-group ($F_{(2,74)} = 11.58$, $p = 4.21 \times 10^{-5}$, $\eta^2 = 0.05$). Pairwise comparisons showed late > middle and late > early ($ps < .004$, FDR-corrected), with no early-middle difference ($p = .13$, FDR-corrected). No effects of group, ROI, or interactions were observed ($Fs < 1.45$, $ps > .23$).

We also grouped layers by clustering for the RDMs derived from all layers (Fig. S8). Results were consistent with the main analyses. During the perception phase, there was a main effect of layer-group ($F_{(2,74)} = 17.83$, $p = 4.77 \times 10^{-7}$, $\eta^2 = 0.08$), with late > middle and early ($ps < 3.44 \times 10^{-5}$, FDR-corrected), and early not differing with middle ($p = .09$, FDR-corrected). During the imagery phase, a main effect of layer-group was also observed ($F_{(2,74)} = 11.49$, $p = 4.5 \times 10^{-5}$, $\eta^2 = 0.06$), with late and middle > early ($ps < 1.69 \times 10^{-4}$, FDR-corrected), but no significant difference between late and middle ($p = .35$, FDR-corrected).

Together, these findings validate that language-knowledge-imagery-irrelevant representations (i.e., left dorsal LOTC, bilateral FG, and right IFG) in visualizers and aphantasics converged across multiple layers, with representational similarity to the text-encoder increasing at later, more abstract layers.

Visualizers and aphantasics differed at later layers in similarity to the CLIP image-encoder, during both the perception and imagery phases.

For RSA in the visual-knowledge-imagery-relevant ROI (i.e., left LOTC), we also grouped the layers of the image-encoder into three parts: early (layers 1 - 8), middle (9 - 16), and late (17 - 24) (see the layerwise result in Fig. S9). As shown in Fig. 4b, during the perception phase, similarity to the image-encoder was significantly above zero in both visualizers and aphantasics at the early and middle layers (ts > 2.73, ps < .02, FDR-corrected) but diverged at the late layers. A mixed ANOVA with factors subject-group (visualizers, aphantasics) and layer-group (early, middle, late) revealed a significant interaction effect ($F_{(2,74)} = 15.39$, $p = 2.58 \times 10^{-6}$, $\eta^2 = 0.08$). Simple effects showed a significant subject-group difference at the late layers ($t_{(34.5)} = 4.75$, $p = 3.46 \times 10^{-5}$, FDR-corrected), but not at the early or middle layers (ts < 0.83, ps > 0.62, FDR-corrected).

During the imagery phase, similarities to the image-encoder were significant in visualizers across all layer-groups (ts > 2.27, ps < .04, FDR-corrected) but not in aphantasics (ts < 1.11, ps > .21, FDR-corrected). However, no main effects of subject-group or subject-group × layer-group interactions were observed (Fs < 1.56, ps > .22). We then took another approach to test the effect of imagery experience on the neural representational similarity to the image-encoder by calculating correlations between the mean similarity values (i.e., the correlation between the neural RDM and image-encoder RDM) and VVIQ scores across individuals. Mean similarity at the late layers significantly correlated with VVIQ scores across all subjects (Spearman's $\rho = 0.39$, $p = .04$, FDR-corrected; Fig. 4b, right), but not at the early or middle layers (Spearman's ρs < 0.28, ps > 0.08, FDR-corrected). Within-subject-group analysis revealed a significant correlation between the similarity to the image-encoder and VVIQ scores in visualizers (Spearman's $\rho = 0.51$, $p = .03$, FDR-corrected) but not in aphantasics (Spearman's $\rho = 0.22$, $p = .39$, FDR-corrected).

We also tested whether similarity to the image-encoder correlated with VVIQ scores in visualizers during the perception phase, but no association was found (Spearman's $\rho = 0.11$, $p = .61$, uncorrected), indicating that subjective imagery vividness does not tract the visual-structured representation during object perception, and instead is a property for the imagery process.

An alternative layer-group partitioning approach based on clustering (Fig. S9) yielded consistent results. During the perception phase, a mixed ANOVA revealed a significant subject-group × layer-group interaction ($F_{(2,74)} = 14.72$, $p = 3.89 \times 10^{-8}$, $\eta^2 = 0.09$), with significant subject-group differences at the mid-late and late layers (ts > 3.70, ps < .002, FDR-corrected). During the imagery phase, no significant effects were observed in the mixed ANOVA. Correlation analysis showed that similarity to the image-encoder correlated with VVIQ scores across all subjects at the mid-late and late layers (Spearman's ρs > 0.36, ps < .05, FDR-corrected), with a significant correlation in visualizers at the mid-late layers (Spearman's $\rho = 0.49$, $p =0.04$, FDR-corrected).

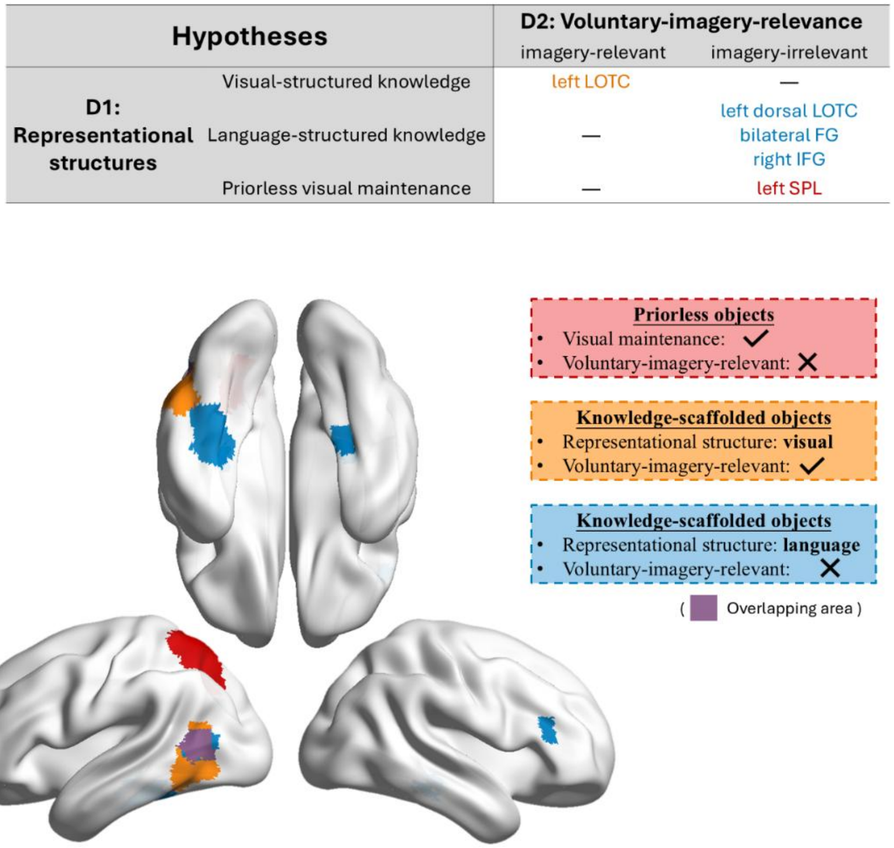
Although no significant clusters were observed in the searchlight analysis with the penultimate layer in visualizers during the imagery phase (Fig. 3b, top), the late layers overall showed significant similarity to the image-encoder in the ROI analysis. Inspection of the layerwise correlation profile (Fig. S9) revealed a drop in correlation at the last two layers. To examine whether other areas contained image-encoder-like representations during the imagery phase, we selected layer 18—the later layer with the highest average correlation across all subjects in the ROI analysis—and used its RDM for whole-brain searchlight RSA. This analysis (Fig. S10) revealed significant clusters in the bilateral ventral LOTC and left LOC, again with a significant group difference in the left ventral LOTC (Visualizers > Aphantasics). Importantly, the cluster location in the left LOTC could not be attributed to the choice of the optimal layer within the ROI of left LOTC, as a trend of such group difference was already present in this area in the searchlight RSA with the penultimate layer (Fig. S6).

These results further elucidate that the visual-structured representation observed in the visual-knowledge-imagery-relevant ROI is related to conscious imagery experience particularly at later vision model layers, during both the perception and imagery phases.

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



Hypotheses		D2: Voluntary-imagery-relevance	
		imagery-relevant	imagery-irrelevant
D1: Representational structures	Visual-structured knowledge	left LOTC	—
	Language-structured knowledge	—	left dorsal LOTC bilateral FG right IFG
	Priorless visual maintenance	—	left SPL

Priorless objects
- Visual maintenance: ✓
- Voluntary-imagery-relevant: ✗

Knowledge-scaffolded objects
- Representational structure: **visual**
- Voluntary-imagery-relevant: ✓

Knowledge-scaffolded objects
- Representational structure: **language**
- Voluntary-imagery-relevant: ✗

(🟪 Overlapping area)

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 LOTC (orange) encodes visual-structured knowledge representation that is sensitive to voluntary imagery experience. The bilateral FG, left dorsal LOTC, and right IFG (blue) encode language-structured knowledge representation that is irrelevant to imagery experience. The left SPL (red) supports priorless visual representation that is unaffected by imagery experience. Purple indicates overlapping area.

Multifaceted object knowledge representations

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 29). 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 29; and about results with congenitally blind individuals in 1). 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 text-encoder, consistent with its role in multimodal and abstract object representations (*1*, *44*, *45*), irrelevant to imagery experience. The LOTC showed subregional characteristics: while it aligned with the image-encoder and tracked imagery vividness, suggesting a visual-structured, imagery-relevant format, the left dorsal subregion additionally aligned with the text-encoder, indicating a language-structured content. 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; 3, 11, 46–48), suggesting that future studies on LOTC should consider the role of imagery processes and representational structures.

The right IFG is not typically discussed relating to conceptual knowledge but is frequently implicated in tasks requiring auditory working memory (e.g., maintaining tones or semantic information 49–53), and is also considered a critical region for cognitive inhibitory control (*54*, *55*). In the retro-cue task of the current experiment, although participants may have needed to suppress non-target stimuli during the imagery phase after the cue, inhibitory control alone cannot account for the significant cross-decoding observed for the target stimulus. Instead, our finding that the right IFG encoded language-structured object representations aligns more closely with prior auditory working memory studies, suggesting that internal representations do not necessarily mirror the modality of external input when maintaining information.

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 reflect relations that are better captured by language-derived higher-level semantics (see similar discussions in 38, 56, 57). 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 58, 59). 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 the left LOTC. How well the neural representation here during imagery fits with the CLIP image-decoder is predictive of how vivid an individual experiences subjective imagery. This region is close to the so-called fusiform imagery node (FIN) proposed to support domain-general internal visualization (*22*, *60*). 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(*35*). 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 (*61*, *62*), 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 shapes 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 (*24*, *63*, *64*) and sub-conscious processing (*65*–*68*). 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 of left LOTC at low to mid-levels were preserved during perception, but group differences emerged at higher layers. 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 in the ventral visual cortices in visualizers during perception, emphasizing the integration of both bottom-up and top-down (language-modulated knowledge structures) 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 association between voluntary imagery experience and visual-structured representations in the left 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.

In 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 LOTC and were related to imagery experience. Language-structured knowledge representations were distributed in multiple areas across 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.

Materials and Methods

Experimental Design

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; 69), with scores below 32 (mean VVIQ ± SD: 23.53 ± 6.26). Follow-up interviews ensured participants fully understood the instructions and confirmed the absence or near absence of visual imagery experience. One participant with a VVIQ score of 35 was also included in the aphantasic group, as their subjective report closely aligned with the characteristics of aphantasia. 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 (Fig. S1): 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 shape class. Meaningless shapes 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 (*70*). 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 and preprocessing

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.

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.

Statistical 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.

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 (*71*, *72*).

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 mm³). 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(*73*), 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. 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 (*41*). We utilized a Chinese version of CLIP (*42*), 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 penultimate layers of the encoders (i.e., 23rd out of 24 layers for the image-encoder, and 11th layer for the text-encoder), and computed representational dissimilarity matrices (RDMs) based on cosine distance, yielding image-RDM and text-RDM, respectively. We used embeddings from the penultimate layer rather than the output layer (as in Chen et al., 2025; Shoham et al., 2024), because they preserve rich modality-specific representational structure while avoiding potential distortions from the final alignment step optimized for contrastive learning.

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 (*43*) to quantify shared variance among all three RDMs. Coefficients were assigned to the center voxel of each sphere. Group-level analyses were also performed using permutation-based Statistical nonParametric Mapping (SnPM) (NISOx: SnPM), with 5000 permutations and no variance smoothing. Results were thresholded at voxelwise p < .001, cluster-level FWE-corrected p < .05, unless stated otherwise.

ROI-based RSA with CLIP Model Layers

To explore whether the encoders' effects differ in different processing stages in the two groups, we conducted ROI-based RSA across multiple stages of the encoders. We extract layerwise features from the Chinese CLIP model by registering 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 (see Results). For partial correlations, the penultimate layer RDM of the non-target encoder was regressed out. We grouped the layers into three equal parts as early, middle, and late layers, averaged layer values within each layer-group for each subject, and performed statistical analyses at the layer-group level. We also conducted agglomerative clustering for the RDMs derived from the layers of each encoder to group them into different processing stages to repeat this analysis (see Fig. S8-9). All correlation coefficients were Fisher-Z transformed prior to statistical analysis in R.

All statistical analyses of behavioral and ROI data were conducted in R (*75*) using the rstatix package (*76*). Two-sample comparisons were performed with Welch's t-test, which adjusts for unequal sample sizes(*77*).

References

1. Y. Bi, X. Wang, A. Caramazza, Object Domain and Modality in the Ventral Visual Pathway. *Trends Cogn Sci* **20**, 282–290 (2016).
2. J. van den Hurk, M. Van Baelen, H. P. Op de Beeck, Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences* **114**, E4501–E4510 (2017).
3. L. L. Chao, J. V. Haxby, A. Martin, Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nat Neurosci* **2**, 913–919 (1999).
4. J. T. Devlin, M. F. S. Rushworth, P. M. Matthews, Category-related activation for written words in the posterior fusiform is task specific. *Neuropsychologia* **43**, 69–74 (2005).
5. B. Z. Mahon, S. Anzellotti, J. Schwarzbach, M. Zampini, A. Caramazza, Category-Specific Organization in the Human Brain Does Not Require Visual Experience. *Neuron* **63**, 397–405 (2009).
6. C. He, M. V. Peelen, Z. Han, N. Lin, A. Caramazza, Y. Bi, Selectivity for large nonmanipulable objects in scene-selective visual cortex does not require visual experience. *Neuroimage* **79**, 1–9 (2013).
7. M. V. M. V. Peelen, S. Bracci, X. Lu, C. He, A. Caramazza, Y. Bi, Tool selectivity in left occipitotemporal cortex develops without vision. *J Cogn Neurosci* **25**, 1225–1234 (2013).
8. B. R. Conway, D. Y. Tsao, Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex. *Proc Natl Acad Sci U S A* **106**, 18034–18039 (2009).
9. N. S. Hsu, D. J. M. Kraemer, R. T. Oliver, M. L. Schlichting, S. L. Thompson-Schill, Color, Context, and Cognitive Style: Variations in Color Knowledge Retrieval as a Function of Task and Subject Variables. *J Cogn Neurosci* **23**, 2544–2557 (2011).
10. M. M. Bannert, A. Bartels, Decoding the yellow of a gray banana. *Current Biology* **23**, 2268–2272 (2013).
11. Y. Xu, L. Vignali, F. Sigismondi, D. Crepaldi, R. Bottini, O. Collignon, Similar object shape representation encoded in the inferolateral occipitotemporal cortex of sighted and early blind people. *PLoS Biol* **21**, e3001930 (2023).
12. M. V. Peelen, C. He, Z. Han, A. Caramazza, Y. Bi, Nonvisual and Visual Object Shape Representations in Occipitotemporal Cortex: Evidence from Congenitally Blind and Sighted Adults. *The Journal of Neuroscience* **34**, 163–170 (2014).
13. P. Vetter, Ł. Bola, L. Reich, M. Bennett, L. Muckli, A. Amedi, Decoding Natural Sounds in Early "Visual" Cortex of Congenitally Blind Individuals. *Current Biology* **30**, 3039-3044.e2 (2020).
14. P. Vetter, F. W. Smith, L. Muckli, Decoding sound and imagery content in early visual cortex. *Current Biology* **24**, 1256–1262 (2014).
15. S. M. Kosslyn, G. Ganis, W. L. Thompson, Neural foundations of imagery. *Nat Rev Neurosci* **2**, 635–642 (2001).
16. S. M. Kosslyn, A. Pascual-Leone, O. Felician, S. Camposano, J. P. Keenan, W. L. Thompson, G. Ganis, K. E. Sukel, N. M. Alpert, 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 (1979)* **284**, 167–170 (1999).

17. J. L. Breedlove, G. St-Yves, C. A. Olman, T. Naselaris, Generative Feedback Explains Distinct Brain Activity Codes for Seen and Mental Images. *Current Biology* **30**, 2211-2224.e6 (2020).
18. J. Pearson, S. M. Kosslyn, The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the National Academy of Sciences* **112**, 10089–10092 (2015).
19. S. A. Harrison, F. Tong, Decoding reveals the contents of visual working memory in early visual areas. *Nature* **458**, 632–635 (2009).
20. S. H. Lee, D. J. Kravitz, C. I. Baker, Disentangling visual imagery and perception of real-world objects. *Neuroimage* **59**, 4064–4073 (2012).
21. N. Dijkstra, S. E. Bosch, M. A. J. van Gerven, Vividness of visual imagery depends on the neural overlap with perception in visual areas. *Journal of Neuroscience* **37**, 1367–1373 (2017).
22. A. Spagna, D. Hajhajate, J. Liu, P. Bartolomeo, Visual mental imagery engages the left fusiform gyrus, but not the early visual cortex: A meta-analysis of neuroimaging evidence. *Neurosci Biobehav Rev* **122**, 201–217 (2021).
23. M. M. Bannert, A. Bartels, Human V4 activity patterns predict behavioral performance in imagery of object color. *Journal of Neuroscience* **38**, 3657–3668 (2018).
24. S. Li, X. Zeng, Z. Shao, Q. Yu, Neural Representations in Visual and Parietal Cortex Differentiate between Imagined, Perceived, and Illusory Experiences. *The Journal of Neuroscience* **43**, 6508–6524 (2023).
25. Y. Xu, M. M. Chun, Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature* **440**, 91–95 (2006).
26. L. W. Barsalou, W. Kyle Simmons, A. K. Barbey, C. D. Wilson, Grounding conceptual knowledge in modality-specific systems. *Trends Cogn Sci* **7**, 84–91 (2003).
27. A. Martin, The Representation of Object Concepts in the Brain. *Annu Rev Psychol* **58**, 25–45 (2007).
28. L. W. Barsalou, Perceptual symbol systems. *Behavioral and Brain Sciences* **22**, 577–660 (1999).
29. A. Martin, GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychon Bull Rev* **23**, 979–990 (2016).
30. A. Zeman, M. Dewar, S. Della Sala, Lives without imagery – Congenital aphantasia. *Cortex* **73**, 378–380 (2015).
31. A. Zeman, Aphantasia and hyperphantasia: exploring imagery vividness extremes. *Trends Cogn Sci* **28**, 467–480 (2024).
32. J. Liu, P. Bartolomeo, Probing the unimaginable: The impact of aphantasia on distinct domains of visual mental imagery and visual perception. *Cortex* **166**, 338–347 (2023).
33. B. M. Montabes de la Cruz, C. Abbatecola, R. S. Luciani, A. T. Paton, J. Bergmann, P. Vetter, L. S. Petro, L. F. Muckli, Decoding sound content in the early visual cortex of aphantasic participants. *Current Biology*, doi: 10.1016/j.cub.2024.09.008 (2024).
34. G. Cabbai, C. Racey, J. Simner, C. Dance, J. Ward, S. Forster, Sensory representations in primary visual cortex are not sufficient for subjective imagery. *Current Biology* **34**, 5073-5082.e5 (2024).
35. J. Liu, M. Zhan, D. Hajhajate, A. Spagna, S. Dehaene, L. Cohen, P. Bartolomeo, Visual mental imagery in typical imagers and in aphantasia: A millimeter-scale 7-T fMRI study. *Cortex* **185**, 113–132 (2025).
36. Y. Bi, Dual coding of knowledge in the human brain. *Trends Cogn Sci* **25**, 883–895 (2021).
37. A. Paivio, Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie* **45**, 255 (1991).
38. H. Chen, B. Liu, S. Wang, X. Wang, W. Han, Y. Zhu, X. Wang, Y. Bi, Language modulates vision: Evidence from neural networks and human brain-lesion models. (2025).
39. X. Wang, W. Men, J. Gao, A. Caramazza, Y. Bi, Two Forms of Knowledge Representations in the Human Brain. *Neuron* **107**, 1–11 (2020).
40. B. Liu, X. Wang, X. Wang, Y. Li, Y. Han, J. Lu, H. Zhang, X. Wang, Y. Bi, Object knowledge representation in the human visual cortex requires a connection with the language system. *PLoS Biol* **23** (2025).
41. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision. (2021).
42. A. Yang, J. Pan, J. Lin, R. Men, Y. Zhang, J. Zhou, C. Zhou, Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. (2023).
43. L. Nalborczyk, Making Sense of Commonality Analysis (2025). https://lnalborczyk.github.io/blog/2025-01-07-commonality.
44. A. Caramazza, B. Z. Mahon, The organization of conceptual knowledge: The evidence from category-specific semantic deficits. *Trends Cogn Sci* **7**, 354–361 (2003).
45. B. Z. B. Z. Mahon, A. Caramazza, What drives the organization of object knowledge in the brain? *Trends Cogn Sci* **15**, 97–103 (2011).
46. A. Martin, C. L. Wiggs, L. G. Ungerleider, J V Haxby, Neural correlates of category-specific knowledge. *Nature* **379**, 649–652 (1996).

47. M. F. Wurm, A. Caramazza, Lateral occipitotemporal cortex encodes perceptual components of social actions rather than abstract representations of sociality. *Neuroimage* **202** (2019).

48. M. F. Wurm, A. Caramazza, Two 'what' pathways for action and object recognition. *Trends Cogn Sci* **26**, 103–116 (2022).

49. G. Shivde, S. L. Thompson-Schill, Dissociating semantic and phonological maintenance using fMRI. *Cogn Affect Behav Neurosci* **4**, 10–19 (2004).

50. I. Uluç, T. T. Schmidt, Y. hao Wu, F. Blankenburg, Content-specific codes of parametric auditory working memory in humans. *Neuroimage* **183**, 254–262 (2018).

51. S. Kumar, S. Joseph, P. E. Gander, N. Barascud, A. R. Halpern, T. D. Griffiths, A brain system for auditory working memory. *Journal of Neuroscience* **36**, 4492–4505 (2016).

52. S. Koelsch, K. Schulze, D. Sammler, T. Fritz, K. Müller, O. Gruber, Functional architecture of verbal and tonal working memory: An FMRI study. *Hum Brain Mapp* **30**, 859–873 (2009).

53. J. A. Schneiders, B. Opitz, H. Tang, Y. Deng, C. Xie, H. Li, A. Mecklinger, The impact of auditory working memory training on the fronto-parietal working memory network. *Front Hum Neurosci*, doi: 10.3389/fnhum.2012.00173 (2012).

54. A. R. Aron, T. W. Robbins, R. A. Poldrack, Inhibition and the right inferior frontal cortex: one decade on. *Trends Cogn Sci* **18**, 177–185 (2014).

55. A. R. Aron, T. W. Robbins, R. A. Poldrack, Inhibition and the right inferior frontal cortex. [Preprint] (2004). https://doi.org/10.1016/j.tics.2004.02.010.

56. A. Y. Wang, K. Kay, T. Naselaris, M. J. Tarr, L. Wehbe, Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nat Mach Intell* **5**, 1415–1426 (2023).

57. C. Conwell, E. MacMahon, A. Jagadeesh, K. Vinken, S. Sharma, J. S. Prince, G. A. Alvarez, T. Konkle, M. Livingstone, L. Isik, Monkey See, Model Knew: Large Language Models Accurately Predict Visual Brain Responses in Humans *and* Non-Human Primates. [Preprint] (2025). https://doi.org/10.1101/2025.03.05.641284.

58. J. Pearson, The human imagination: the cognitive neuroscience of visual mental imagery. *Nat Rev Neurosci* **20**, 624–634 (2019).

59. J. Pearson, Reply to: Assessing the causal role of early visual areas in visual mental imagery. *Nat Rev Neurosci* **21**, 517–518 (2020).

60. A. Spagna, Z. Heidenry, M. Miselevich, C. Lambert, B. E. Eisenstadt, L. Tremblay, Z. Liu, J. Liu, P. Bartolomeo, Visual mental imagery: Evidence for a heterarchical neural architecture. *Phys Life Rev* **48**, 113–131 (2024).

61. S. Chang, X. Zhang, Y. Cao, J. Pearson, M. Meng, Imageless imagery in aphantasia revealed by early visual cortex decoding. *Current Biology*, doi: 10.1016/j.cub.2024.12.012 (2025).

62. R. Keogh, J. Bergmann, J. Pearson, Cortical excitability controls the strength of mental imagery. *Elife* **9**, 1–33 (2020).

63. Y. Xu, The Posterior Parietal Cortex in Adaptive Visual Processing. *Trends Neurosci* **41**, 806–822 (2018).

64. M. Koenigs, A. K. Barbey, B. R. Postle, J. Grafman, Superior Parietal Cortex Is Critical for the Manipulation of Information in Working Memory. *The Journal of Neuroscience* **29**, 14980–14986 (2009).

65. F. Fang, S. He, Cortical responses to invisible objects in the human dorsal and ventral pathways. *Nat Neurosci* **8**, 1380–1385 (2005).

66. J. Almeida, B. Z. Mahon, K. Nakayama, A. Caramazza, Unconscious processing dissociates along categorical lines. *Proceedings of the National Academy of Sciences* **105**, 15214–15218 (2008).

67. M. W. Maclean, V. Hadid, R. N. Spreng, F. Lepore, Revealing robust neural correlates of conscious and unconscious visual processing: Activation likelihood estimation meta-analyses. *Neuroimage* **273**, 120088 (2023).

68. M. Suzuki, Y. Noguchi, R. Kakigi, Temporal dynamics of neural activity underlying unconscious processing of manipulable objects. *Cortex* **50**, 100–114 (2014).

69. D. F. Marks, VISUAL IMAGERY DIFFERENCES IN THE RECALL OF PICTURES. *British Journal of Psychology* **64**, 17–24 (1973).

70. V. Willenbockel, J. Sadr, D. Fiset, G. O. Horne, F. Gosselin, J. W. Tanaka, Controlling low-level image properties: The SHINE toolbox. *Behav Res Methods* **42**, 671–684 (2010).

71. H. P. Op de Beeck, Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? *Neuroimage* **49**, 1943–1948 (2010).

72. M. H. A. Hendriks, N. Daniels, F. Pegado, H. P. O. de Beeck, The effect of spatial smoothing on representational similarity in a simple motor paradigm. *Front Neurol* **8**, 1–11 (2017).

73. G. Valente, A. L. Castellanos, L. Hausfeld, F. De Martino, E. Formisano, Cross-validation and permutations in MVPA: Validity of permutation strategies and power of cross-validation schemes. *Neuroimage* **238** (2021).

74. A. Shoham, I. D. Grosbard, O. Patashnik, D. Cohen-Or, G. Yovel, Using deep neural networks to disentangle visual and semantic information in human perception and memory. *Nat Hum Behav* **8**, 702–717 (2024).

75. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. **3** (2019).

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

77. M. Delacre, D. Lakens, C. Leys, Why psychologists should by default use welch's t-Test instead of student's t-Test. *International Review of Social Psychology* **30**, 92–101 (2017).

Acknowledgments

We thank Rui Feng, 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 (YB)
National Natural Science Foundation of China 31925020 (YB)
National Natural Science Foundation of China 82021004 (YB)
Fundamental Research Funds for the Central Universities (YB)
China Postdoctoral Science Foundation funded project 2025M773448 (ST)

Competing interests:

All other authors declare they have no competing interests.

Data and materials availability:

All data in the main text or the supplementary materials will be deposited in a public repository and made available upon publication. All data have been fully anonymized in accordance with institutional and ethical guidelines.

DNN models were acquired from: https://github.com/OFA-Sys/Chinese-CLIP.