nature human behaviour

Article

https://doi.org/10.1038/s41562-025-02357-5

Combined evidence from artificial neural networks and human brain-lesion models reveals that language modulates vision in human perception

Received: 19 March 2025

Accepted: 14 October 2025

Published online: 15 December 2025

Check for updates

Haoyang Chen[1,2,11], Bo Liu[3,4,5,11], Shuyue Wang[2], Xiaosha Wang[2], Wenjuan Han[6], Xiaochun Wang[3,4,5,12] ✉, Yixin Zhu[1,7,8,9,12] ✉ & Yanchao Bi[1,2,7,9,10,12] ✉

Comparing information structures in between deep neural networks (DNNs) and the human brain has become a key method for exploring their similarities and differences. Recent research has shown better alignment of vision–language DNN models, such as contrastive language–image pretraining (CLIP), with the activity of the human ventral occipitotemporal cortex (VOTC) than earlier vision models, supporting the idea that language modulates human visual perception. However, interpreting the results from such comparisons is inherently limited owing to the 'black box' nature of DNNs. Here we combine model–brain fitness analyses with human brain lesion data to examine how disrupting the communication pathway between the visual and language systems causally affects the ability of vision–language DNNs to explain the activity of the VOTC to address this. Across four diverse datasets, CLIP consistently captured unique variance in VOTC neural representations, relative to both label-supervised (ResNet) and unsupervised (MoCo) models. This advantage tended to be left-lateralized at the group level, aligning with the human language network. Analyses of 33 patients who experienced a stroke revealed that reduced white matter integrity between the VOTC and the language region in the left angular gyrus was correlated with decreased CLIP–brain correspondence and increased MoCo–brain correspondence, indicating a dynamic influence of language processing on the activity of the VOTC. These findings support the integration of language modulation in neurocognitive models of human vision, reinforcing concepts from vision–language DNN models. The sensitivity of model–brain similarity to specific brain lesions demonstrates that leveraging the manipulation of the human brain is a promising framework for evaluating and developing brain-like computer models.

A full list of affiliations appears at the end of the paper. ✉e-mail: wangxiaochun@sydyy.com; yixin.zhu@pku.edu.cn; ybi@pku.edu.cn

Recent advancements in artificial neural networks have shown unprecedented capabilities in capturing the responses of the human brain, especially in the field of vision[1–4]. Image classification models based on deep neural networks (DNNs), such as AlexNet[5] and ResNet[6], trained on large image datasets have yielded object representations that are significantly associated with those in the human and macaque ventral occipitotemporal cortex (VOTC), a region crucial for visual object perception and recognition[7,8]. Such findings have inspired a line of model–brain fitness analyses that assessed the similarities between the performance of different models (in terms of architecture, parameters, goals and datasets) and the activity of the VOTC to unravel the representation of the VOTC and assess the biological fitness of DNNs[9–12].

Recently, multimodal vision models, particularly those trained by aligning image features with language captions, have demonstrated improved performance in brain fitting, brain decoding and brain-guided image synthesis[13–18]. These findings provide evidence of the influence of language on the visual cortex, a classic topic that remains highly controversial in the field of cognitive neuroscience[19,20]. Supporting cognitive evidence comes from prominent behavioural cross-linguistic studies and verbal learning experiments, which have demonstrated that verbal labels can influence visual colour categorization[21–24]. However, the robustness of these cognitive findings has been challenged[25,26], and even when such effects are actually observed, researchers are unable to determine whether they arise from the perceptual stages or other cognitive processes[27].

The results from the new vision–language model–brain fitness approach, although compelling, are inconclusive for several reasons. A prominent model used in these studies is contrastive language–image pretraining (CLIP), with the extensiveness of the dataset on which it was trained potentially contributing substantially to its performance[28]. Wang et al.[13] compared the performance of a CLIP-style model with an unsupervised model trained on the same dataset (that is, YFCC15M) and reported that, after controlling for the details of the training dataset, the CLIP-style model showed diminished performance in fitting the activity of the VOTC. The remaining effects were primarily localized to the bilateral extrastriate body area and fusiform face area, potentially reflecting specific influences regarding human interactions within the images (see discussion in Wang et al.[13]). In addition, the empirical robustness of these findings requires further validation, as current comparisons predominantly rely on a single functional magnetic resonance imaging (fMRI) dataset (Natural Scenes Dataset[29]).

More critically, the model–brain fitness approach suffers from an inherent challenge in interpretation because of the 'black box' nature of the learned representations in DNNs. Even when the CLIP models differ from control models solely by the addition of language supervision, their superior explanatory power for the activity of brain regions might stem from their ability to capture higher-order object-relational structures that are not attributable only to language, as language encodes relational structures that overlap with a range of non-linguistic high-level relations (for example, 'hammer' and 'hand' are similar in both linguistic and visual-action association spaces). This alternative hypothesis is supported by a study demonstrating that pure-language models perform comparably to vision models in predicting activity in both the macaque inferior temporal cortex and the human VOTC[30]. Therefore, more direct testing is needed to determine whether the unique effects of language–vision models genuinely reflect how the language system modulates visual perception in the human brain.

To address these issues, we investigated the influence of language experience on activity in the VOTC in response to object images through a combination of human brain lesion data, fMRI data and computational modelling approaches. Study 1 evaluates whether, compared with unsupervised vision models, vision models that align with different scales of language information (for example, sentence-level versus word-level alignment) consistently show unique model–brain correspondence in the human VOTC across four datasets featuring diverse picture stimuli, language experiences and tasks. Study 2 provides a crucial test of causality through human brain-lesion models. Specifically, in a group of patients with brain damage ($N = 33$), we examined whether reducing the structural integrity of the connections between the VOTC and language regions affects the ability of different computational models to explain the activity of the VOTC during visual processing.

Results

Computational models fitting four brain datasets from healthy populations with different language experiences (study 1)

To investigate whether and how language supervision influences the model–brain correspondence within the VOTC, we compared three computational vision models with different degrees of language supervision with representational similarity analysis[31] (RSA) across four distinct fMRI datasets. These datasets are characterized by different participant populations, tasks and language experiences, allowing us to examine the generalizability of the effects of language on visual perception. An overview of the datasets, computational models, and analysis workflow in study 1 is shown in Fig. 1.

Object neural representation in the four fMRI picture-viewing datasets. We included four fMRI datasets from healthy subjects viewing object pictures, encompassing different object types and tasks. The first three datasets were derived in-house (Methods), and the fourth was an open dataset. Object neural activity patterns in the VOTC are included in each fMRI dataset and used to construct the object representation dissimilarity matrices (neural representational dissimilarity matrices (RDMs)).

OPN95 (oral picture naming) and SPN95 (sign language picture naming). The participants viewed colour pictures of 95 objects across three categories (32 animals, 35 small, manipulable objects and 28 large, non-manipulable objects) and were instructed to name them. The OPN95 dataset includes the data from typically developed participants (that is, those with speech experience; $N = 26$ in the analyses) performing oral picture naming[32], whereas the SPN95 dataset includes the data from early deaf individuals (that is, those with sign language experience; $N = 32$ in the analyses) performing sign language naming with their non-dominant hand. The data from these two groups were analysed both separately and collectively. As OPN95 and SPN95 used an identical stimulus set (95 object pictures), hereafter we refer to this common set as the OPN95–SPN95 stimulus set.

FV14 (colour knowledge retrieval). The participants ($N = 33$ in the analyses) viewed greyscale pictures of 14 fruits/vegetables (one image per object) and were asked to judge whether each object's typical skin colour was red[33].

THINGS. An open dataset in which participants ($N = 3$) viewed 720 objects from various categories (for example, animals, small, manipulable objects, large, non-manipulable objects, plants and fruits), with each object represented by 12 different naturalistic images[34]. The images were presented in rapid succession while participants maintained fixation on a central point and performed an oddball detection task by pressing a button in response to occasionally presented artificially synthesized distractor images.

Object model representation in three computation vision models. We considered three computer vision models with different levels of language involvement during pretraining (Fig. 1). (1) CLIPvision is the visual encoder of CLIP, supervised by natural language (sentence descriptions). CLIP is trained with image–caption pairs to align features between the images and the text, and sentence-level descriptions containing verbal labels and object relationships are integrated into

https://doi.org/10.1038/s41562-025-02357-5

a

Dataset ID	Reference	Stimulus type	Object domain	Task	Subject (*N*)
OPN95	Fu et al., 2023	Isolated objects	Animal, large non-manipulable object, small manipulable object	Picture naming by oral language	Healthy, hearing (26)
SPN95	Unpublished	Isolated objects	Animal, large non-manipulable object, small manipulable object	Picture naming by sign language	Healthy, early deaf (32)
FV14	Liu et al., 2025	Isolated objects	Fruits, vegetables	Object colour knowledge judgment	Healthy, hearing (33)
THINGS	Hebart et al., 2023	Naturalistic objects	Animal, large non-manipulable object, small manipulable object, foods	Oddball detection	Healthy, hearing (3)

b

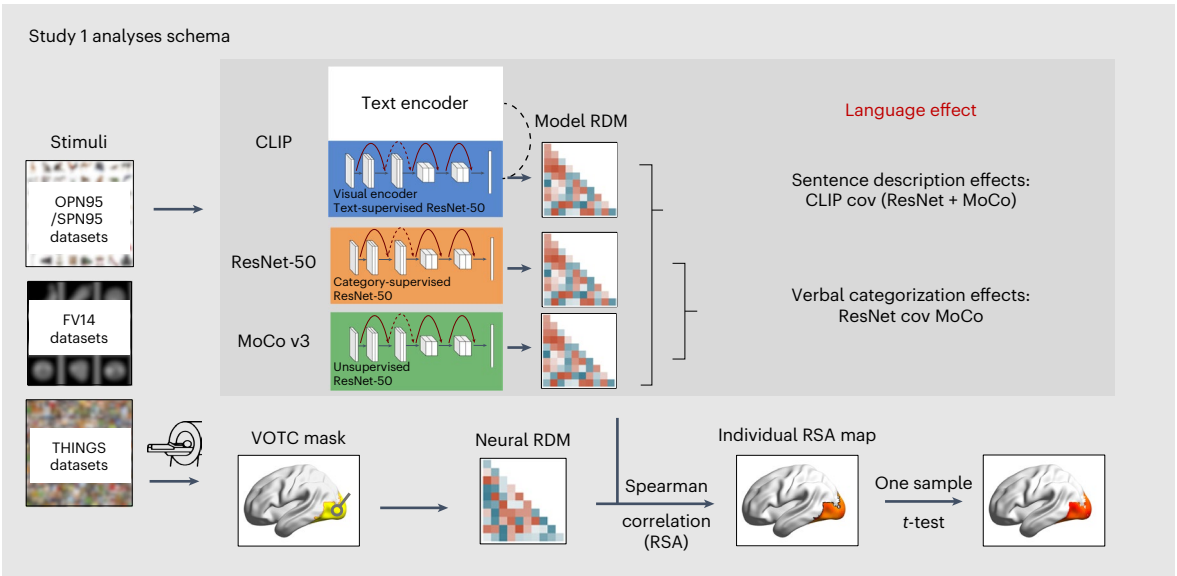


Fig. 1 | Overview of the fMRI datasets, vision models and study 1 analysis schema. a, The four fMRI datasets (OPN95, SPN95, FV14 and THINGS) used in this study. The table shows the references, stimulus types, object domains, tasks and participant numbers. Note that the participants in the FV14 dataset serve as healthy controls to the patient group in study 2. **b**, The analysis schema for study 1. With RSA, neural responses in the VOTC are compared with three vision models with a shared ResNet-50 architecture: the visual encoder of OpenAI's CLIP[44] (supervised by natural language text; hereinafter referred to as CLIPvision), ResNet-50[6] (supervised by human-generated category labels) and MoCo v3[64] (self-supervised). Searchlight RSA computes neural RDMs for spherical ROIs, and Spearman's partial correlations between neural and model RDMs are Fisher *z*-transformed and tested across participants using one-sample *t*-tests. Two language-related effects are defined as: (1) the sentence description effect—the partial correlation coefficient between the CLIP RDM and the neural RDMs, controlling for the ResNet and MoCo derived RDMs and (2) the verbal categorization effect—the partial correlation coefficient between the ResNet RDM and the neural RDMs, controlling for the MoCo RDM. See Methods for details of preprocessing procedures, ROI definitions and statistical analyses. cov, covarying.

the visual encoder of CLIP using a multimodal loss signal in the final layer. (2) ResNet is a supervised image classification model trained to predict human-generated verbal labels (word-level categorization) for input images. ResNet uses image-label pairs to minimize the classification error. (3) MoCo is a self-supervised visual representation learning model trained exclusively on images. The distance between an image and its transformed version is minimized, whereas the distance between different images is maximized.

These three models share the same backbone architecture (ResNet-50) and differ in pretraining objectives and training datasets. The pretraining objective differences represent our core experimental variable, allowing us to investigate how language feedback influences visual representations. We selected the original model implementations to ensure optimal performance, while conducting additional control analyses with matched training datasets ('Control analyses for model training dataset differences' section). The relevant model characteristics are summarized in Supplementary Table 1.

The identical architectures of CLIPvision, ResNet and MoCo enabled direct comparisons of the object set representations in the corresponding layers. The image features were extracted from the three models spanning 18 matched layers, ranging from early to late stages. The model RDMs were generated by computing the Pearson distances (1-Pearson's correlation) between feature vectors for each object pair. For the THINGS dataset, which includes multiple images per concept, the RDMs were calculated using the average features of the images corresponding to the same concept. The final pooling layer, which showed the lowest average intermodel correlation (Fig. 2a)—indicating more distinct and specialized representations across models—was selected for the subsequent brain-fitting analyses. The resulting RDMs of the final pooling layer illustrated moderate intercorrelations across the OPN95–SPN95, FV14 and THINGS datasets, with Spearman's correlation coefficients ranging from 0.43 to 0.63 (Bonferroni-corrected $P < 0.05$). A representative visualization of the RDMs from the THINGS dataset is provided in Fig. 2b, left.

To assess similarity between the model and human behavioural object-relatedness, we conducted correlation analyses between the model-generated and the human behavioural RDMs. The human behavioural RDMs were generated by averaging ratings across participants within distinct groups. For the OPN95–SPN95 and FV14 datasets, object-semantic similarity was evaluated through a 7-point Likert

https://doi.org/10.1038/s41562-025-02357-5

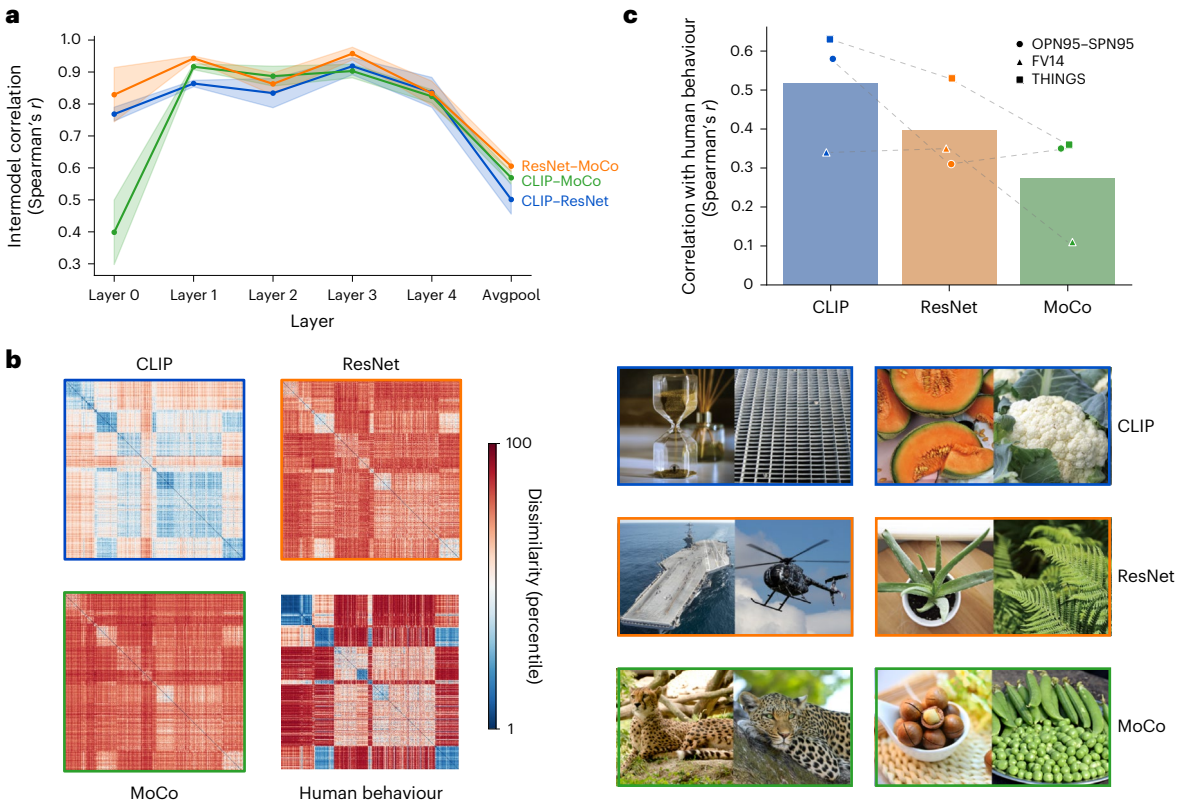


Fig. 2 | Intercorrelations among vision model RDMs and their alignment with human behaviour. a, The layer-wise intermodel Spearman's correlations between RDMs from CLIPvision, ResNet and MoCo, averaged across the datasets used in this study (OPN95 and SPN95, FV14 and THINGS). Each point represents the mean correlation coefficient, with shaded regions denoting the standard error across the datasets ($n = 3$ datasets). The average pooling layer ('Avgpool') yields the lowest intermodel correlations and is selected for the subsequent analyses. **b**, Left: the RDMs from the three vision models for the THINGS dataset, where red corresponds to the least similar object pairs and blue corresponds to the most similar object pairs. A human behavioural object-relatedness RDM derived from a three-alternative odd-one-out paradigm is also shown[34]. Right: the top two representative image pairs with the highest within-model similarity after controlling for the other two models, illustrating how these models group object concepts differently. The original images have been replaced with CC0-licensed examples from the THINGSplus database[71]. For the THINGS dataset, each object concept is represented by 12 images, across which the model-derived features are averaged to construct the concept-level RDMs (see Methods for details). **c**, Spearman's correlation coefficients between each model's RDM and the human behavioural object-relatedness RDM. The bar heights represent the mean correlation coefficient across datasets ($n = 3$ datasets: OPN95–SPN95, FV14, THINGS). The overlaid scatter points represent individual-dataset values and are connected by dotted lines for visual guidance. The human behavioural RDMs were derived from pairwise similarity ratings (OPN95–SPN95, FV14) or odd-one-out judgments (THINGS). Panel **b** reproduced from ref. 70 under a Creative Commons license CC BY 4.0.

scale. For the THINGS dataset, object similarity was assessed with an odd-one-out task, and the probability of participants selecting objects i and j as belonging together was quantified[34].

The results are summarized in Fig. 2c, where each bar represents one model (CLIP, ResNet and MoCo) and contains three data points (corresponding to the three datasets). As the degree of language involvement increased (from MoCo to ResNet and to CLIP), the alignment with human behaviour progressively improved, demonstrating the relative advantage of language information alignment in capturing semantic relationships. Figure 2b highlights the top two image pairs in each model's RDM within the THINGS dataset after partialling out the other two models. These examples illustrate the heavier emphasis of CLIP on semantic properties, the moderate semantic focus of ResNet and the predominant reliance of MoCo on lower-level visual attributes, thus revealing a continuum of semantic–visual encoding across the three models.

Object model representations fitting brain activity representations in the VOTC (RSA results). We next examined how the model representations relate to neural data with searchlight RSA in the VOTC mask, defined by contrasting pictures with the baseline in the OPN95 dataset (false discovery rate $q < 0.05$; see Supplementary Fig. 1 and Methods for details). For each participant, the Spearman's correlation coefficient was assessed between the model visual RDMs and the neural RDMs within a 10-mm radius searchlight sphere. We focused on testing whether CLIPvision and ResNet provided additional explanatory power over MoCo to assess the effects of sentence description and verbal categorization, respectively.

Specifically, to evaluate the relative specific contribution of the alignment with sentence descriptions (that is, the introduction of word relations), we computed the partial Spearman's correlation coefficient between the CLIPvision RDM and the neural RDMs, with the ResNet RDM and MoCo RDM serving as covariates. To evaluate the specific relative contribution of the alignment with verbal categorizations, we computed the partial Spearman's correlation coefficient between the ResNet RDM and the neural RDMs, with the MoCo RDM serving as a covariate. Below, the results are shown at a statistical voxel-level threshold $P < 0.001$, one-tailed, cluster-level family-wise error (FWE)-corrected $P < 0.05$, unless otherwise specified.

Sentence description effect (CLIPvision over ResNet and MoCo). In the OPN95 dataset, group-level one-sample *t*-tests on the Fisher *z*-transformed partial Spearman's correlation coefficients for all participants revealed two significant clusters: one in the left lateral

https://doi.org/10.1038/s41562-025-02357-5

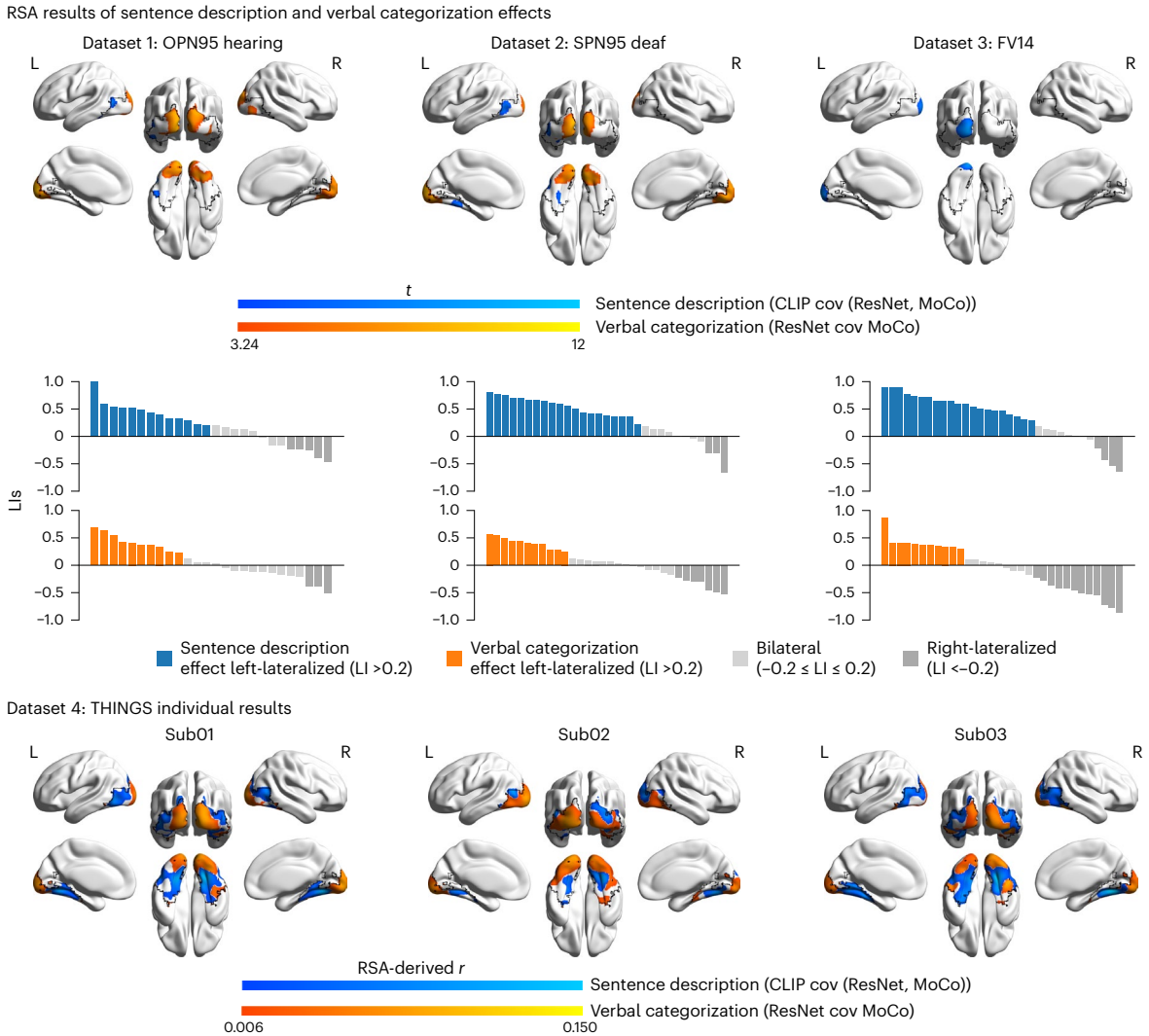


Fig. 3 | Language effect in VOTC across datasets. The group-level searchlight RSA results for sentence description effects (blue) and verbal categorization effects (orange) in three datasets: OPN95 (n = 26 hearing), SPN95 (n = 32 deaf) and FV14 (n = 33 hearing), with corresponding LIs for each subject displayed below each brain map. For THINGS dataset (n = 3 hearing), individual-participant results are shown. All the brain maps are thresholded at voxel-level $P < 0.001$, one-tailed and cluster-level FWE-corrected $P < 0.05$ in the VOTC mask (indicated by the black contours). L, left hemisphere; R, right hemisphere; cov, covarying.

occipital complex and another in the left fusiform gyrus extending into the left inferior temporal gyrus (Fig. 3, top left, blue). In the SPN95 dataset, similar significant clusters were found but with more medial peak locations (Fig. 3, top middle, blue). These effects were not affected by auditory experience (hearing versus deaf) or the language modality (oral versus sign language naming): the region of interest (ROI) analysis of these two clusters revealed statistically comparable effects in the deaf and hearing groups (hearing versus deaf: $t(55.43) = -0.666$, $P = 0.508$, Hedges' $g = -0.168$, 95% confidence interval (CI) −0.03 to 0.01; BF_{10} (Bayes factor for the alternative hypothesis relative to the null) of 0.322, moderate evidence in support of null hypothesis (H_0; deaf = hearing). In the *FV14* dataset, analyses revealed a significant cluster in the left occipital pole (Fig. 3, top right, blue). In the THINGS dataset, the individual-level analysis was conducted for each subject at a voxel-level $P < 0.001$, one-tailed and cluster-level FWE-corrected $P < 0.05$. Across all the subjects, the analysis revealed significant clusters in the bilateral lateral occipital complex, extending to the bilateral lingual gyrus, fusiform gyrus, parahippocampal gyrus and inferior temporal gyrus (Fig. 3, bottom, blue).

Verbal categorization effect (ResNet over MoCo). In the OPN95 dataset, group-level one-sample t-tests on the Fisher z-transformed partial Spearman's correlation coefficients for all participants revealed significant clusters in the bilateral occipital pole (OP), both extending to the lingual gyrus (LING), the lateral occipital complex and the posterior fusiform gyrus (Fig. 3, top left, orange). In the SPN95 dataset, the analyses revealed similar clusters. ROI analysis of these two clusters revealed no significant difference between the two groups (hearing versus deaf: $t(55.34) = 0.302$, $P = 0.764$, Hedges' $g = 0.078$, 95% CI −0.01 to 0.02; BF_{10} of 0.278, moderate evidence in support of H_0: deaf = hearing). In the FV14 dataset, no cluster survived at the conventional threshold (voxel-level $P < 0.001$, one-tailed and cluster-level FWE-corrected $P < 0.05$). In the THINGS dataset, across all subjects, the analysis revealed significant clusters in the bilateral lingual gyrus, extending to the occipital pole, occipital fusiform gyrus, lateral occipital complex and intracalcarine cortex (Fig. 3, bottom, orange).

RSA-based lateralization results. Consistent with the well-established left lateralization of language networks in the human brain (see review in ref. 35), the above searchlight results of the sentence description effect appear to reveal larger, positive clusters in the left VOTC than in the right VOTC. In this section, we quantified and examined these potential VOTC laterality effects. We computed the laterality index (LI) in the VOTC for each dataset with the LI toolbox[36] implemented

https://doi.org/10.1038/s41562-025-02357-5

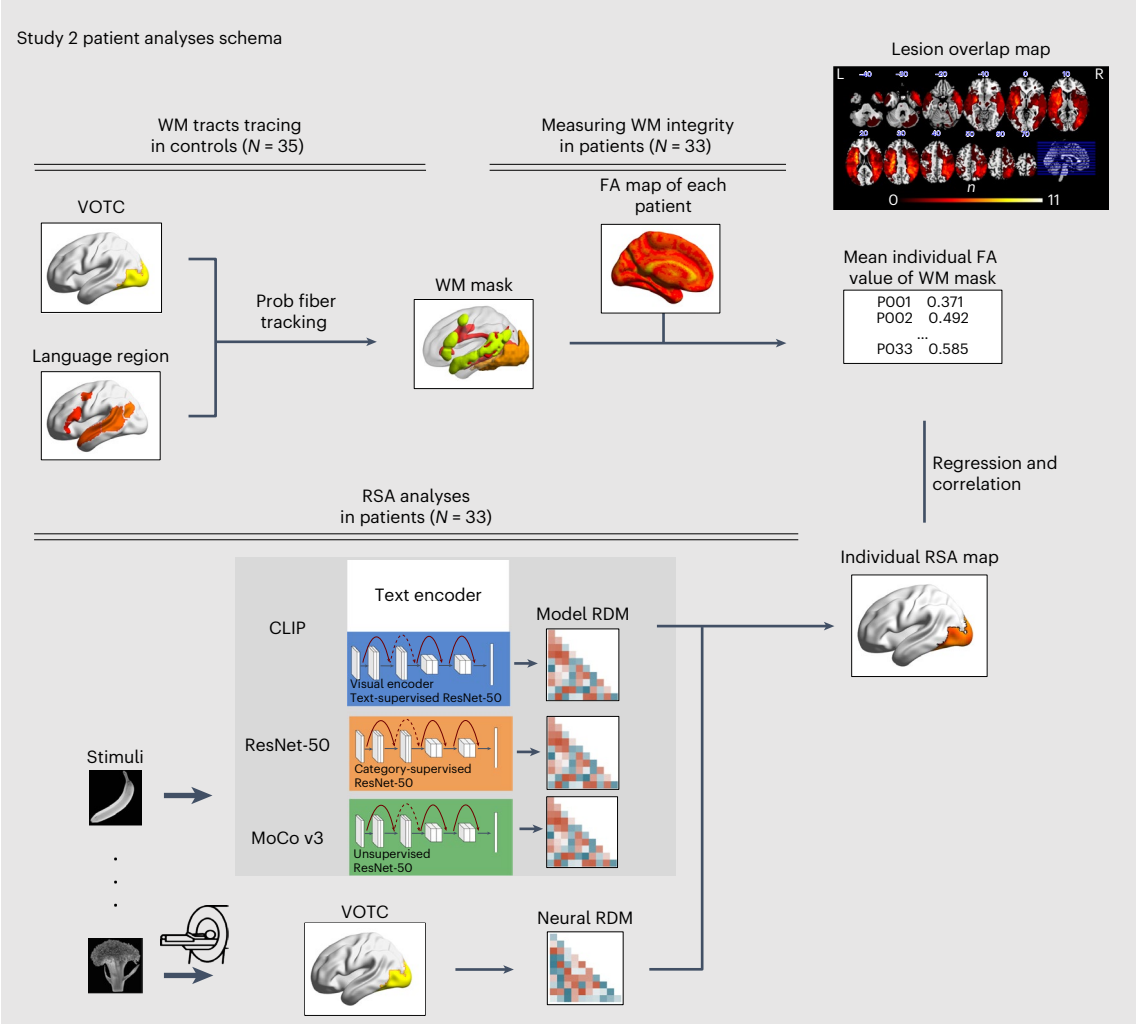


Fig. 4 | Study 2 analysis workflow linking WM integrity and model–brain correspondence in patients with chronic stroke. Top: the workflow for quantifying WM integrity in patients. A probabilistic tractography was first performed in the native spaces of the healthy controls ($N = 35$; the same participants as in the FV14 dataset of study 1), using six language-related ROIs, both individually and together as the overarching language mask[38], as seed points and the VOTC as the target. Each tractography-derived map was normalized, scaled to its maximum voxel intensity, binarized at 0.1 and then aggregated across controls to yield a group-level WM mask (within an explicit WM template; probability >0.4). In the patient group ($N = 33$), WM integrity between each language region and the VOTC was quantified by extracting mean FA values within the corresponding group-level WM mask. Bottom: the RSA workflow for task-fMRI data in these patients, following the same procedure as in study 1 but performed at the ROI level. Multiple linear regression and correlation analyses were then conducted across patients, relating WM integrity (that is, FA values) in specific tracts to each patient's RSA measures in the VOTC. Top right: the lesion distribution map of the 33 patients, in which the value of each voxel represents the number of patients with a lesion at that voxel. The analytical framework tests whether vision–language structural connectivity predicts the strength of language-related neural representations in VOTC following stroke. The original experimental stimulus images have been replaced owing to copyright restrictions. L, left hemisphere; R, right hemisphere; prob, probability.

in SPM12 (see Methods for details). At the group level, the LIs were computed by extracting t values generated from the one-sample t-tests against zero conducted above for RSA within the whole VOTC mask. At the individual level, the RSA-derived Fisher z-transformed rho values within the VOTC mask were used to compute the LIs, followed by one-sample t-tests against zero. In the THINGS dataset, individual subject LIs were calculated following the same procedure but without statistical testing. The group-level LI ranges from −1 to 1, where −1 indicates complete right lateralization, +1 indicates complete left lateralization and values between −0.2 and 0.2 are considered to indicate bilateralization[37].

Sentence description effect. We observed a significant left-lateralization trend for the sentence description effect in the VOTC at both the group and individual levels in our three in-house datasets with larger subject numbers. At the group level, the OPN95, SPN95 and FV14 datasets all showed strong left lateralization (OPN95, 0.64; SPN95, 0.68; FV14, 0.68). At the individual level, as shown in Fig. 3 (blue bar), the results revealed a significant left-lateralized sentence description effect for the three datasets (OPN95, $t(25) = 2.657$, $P = 0.014$, Cohen's $d = 0.521$, 95% CI 0.04 to 0.32; SPN95, $t(30) = 4.986$, $P < 0.001$, Cohen's $d = 0.896$, 95% CI 0.19 to 0.46; FV14, $t(32) = 4.766$, $P < 0.001$, Cohen's $d = 0.830$, 95% CI 0.20 to 0.49). For the THINGS dataset individual subjects, two subjects did not show laterization, and one showed tendency of right lateralization (Sub01, −0.18; Sub02, −0.38; Sub03, −0.17), indicating that the left lateralization may be statistically visible only when enough subjects are sampled.

Verbal categorization effect. The verbal categorization effect exhibited no clear lateralization pattern at either the group or individual level. At the group level, the results for the OPN95, SPN95 and FV14 datasets did not show any consistent lateralization pattern within the VOTC (OPN95, 0.04; SPN95, 0.03; FV14, −0.57). At the individual level, as shown in Fig. 3 (orange bar), the LIs for these datasets were relatively balanced, with

https://doi.org/10.1038/s41562-025-02357-5

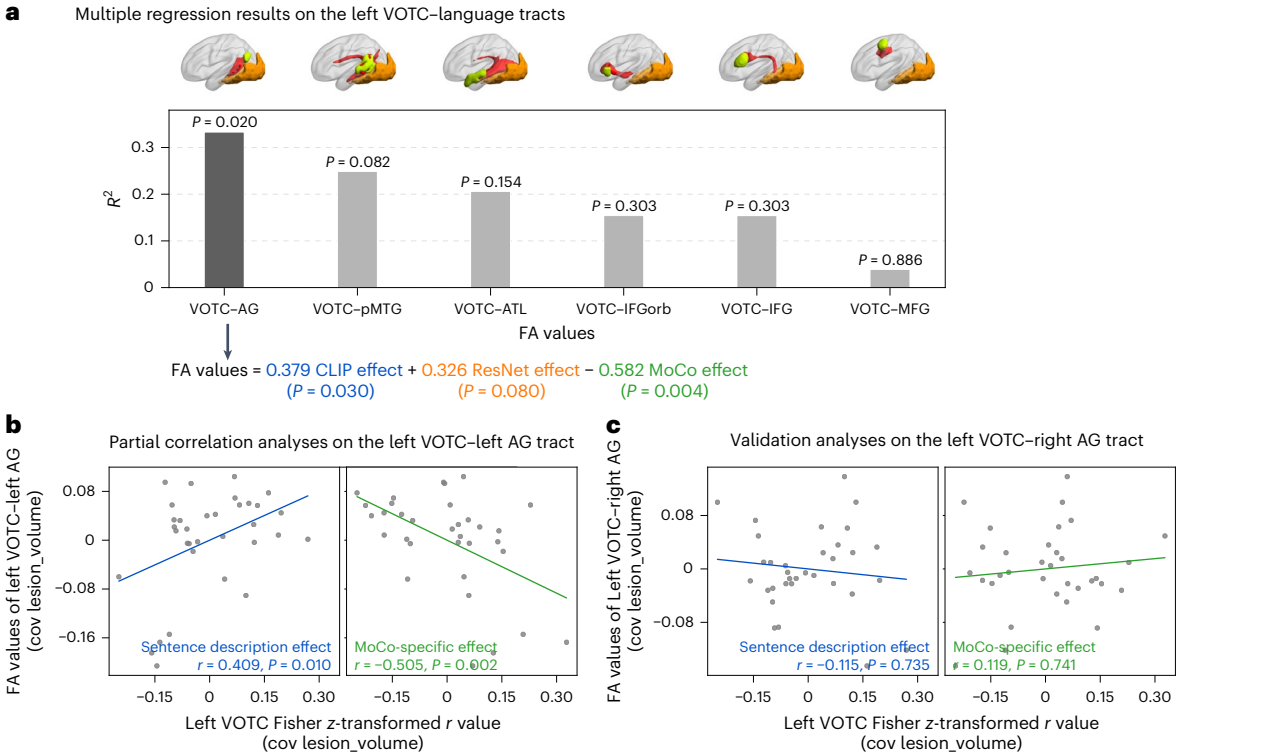


Fig. 5 | WM integrity of left VOTC–left AG tract predicts model–brain correspondence of CLIP and MoCo (n = 33 patients). a, The multiple linear regression predicting FA values for the six left VOTC–language WM tracts from three neural representation measures (CLIP effect, ResNet effect, MoCo effect) with lesion volume as a covariate. The bar shows the variance explained (R^2) across participants. The model was significant only for the left VOTC–left AG tract (labelled as 'VOTC–AG' on the x axis): $F(4, 28) = 3.495$, $P = 0.020$, $R^2 = 0.333$, 95% CI 0.19 to 0.66, with standardized coefficients (P values) indicating significant contributions of both CLIP effect ($t(28) = 2.288$, $P = 0.030$, $\beta = 0.379$, 95% CI 0.04 to 0.72) and MoCo effect ($t(28) = -3.118$, $P = 0.004$, $\beta = -0.582$, 95% CI −0.96 to −0.20). The multiple comparisons across six tracts were uncorrected. Brain visualizations show the VOTC mask (orange), the language ROI mask (yellow-green) and the traced white matter tract (red). **b**, The partial Pearson correlations (degrees of freedom (d.f.) of 30) between left VOTC–left AG tract integrity and model–brain correspondence. Correlation coefficients and one-tailed P values are displayed on the plot. Both sentence description effects (CLIP-specific) and MoCo-specific effects correlate significantly with WM tract integrity. **c**, A validation analysis using connections with right AG (d.f. of 30) reveals no significant relationships, confirming left-lateralized pathway specificity. The correlation coefficients and one-tailed P values are displayed on the plots. All the analyses control for total lesion volume. cov lesion_volume, covarying total lesion volume.

no statistically significant lateralization detected (OPN95, $t(25) = 1.270$, $P = 0.216$, Cohen's $d = 0.249$, 95% CI −0.05 to 0.21; SPN95, $t(31) = 1.109$, $P = 0.276$, Cohen's $d = 0.196$, 95% CI −0.05 to 0.17; FV14, $t(31) = -0.686$, $P = 0.498$, Cohen's $d = -0.121$, 95% CI −0.20 to 0.10). Similarly, in the THINGS dataset, the three subjects did not show stable lateralization pattern (Sub01, −0.38; Sub02, 0.34; Sub03, −0.58).

Causal testing of model-fitting–brain effects in patients with brain damage (study 2)

In study 1, we observed the specific relative advantages of CLIPvision over ResNet and MoCo in explaining VOTC responses, especially left VOTC, to object images across four datasets. In study 2, we causally tested whether such advantages were indeed related to vision–language alignment with brain lesion data. An overview of the study 2 analysis workflow is shown in Fig. 4. Specifically, we examined whether these effects were modulated by the integrity of the communication pathway between the VOTC and the higher-order language system. We reasoned that if the CLIPvision model's effect diminished following damage to vision–language communication centres, it would provide positive evidence for the language-alignment origin of the CLIPvision effect advantage regarding the representations within the VOTC. We tested this in a group of patients with varying degrees of brain damage in terms of lesions affecting the white matter (WM) structural connections between the language network and VOTC while sparing the ventral visual cortex (in-house data collected for another project investigating colour representations; see Methods for details). Analyses focusing on both left VOTC and bilateral VOTC were conducted, and results were highly convergent. In the main text below, we first report results of the left VOTC for simplicity, and the more detailed results with bilateral VOTC are further shown in Extended Data Fig. 1.

Patient picture-viewing fMRI data and WM structural data. A group of patients with brain damage ($N = 33$) underwent the same fMRI scans as did the healthy controls in the FV14 dataset described above. For the in-scanner experiment, we deliberately chose items that were familiar to patients and a task that was easy for patients to ensure a meaningful measurement of VOTC neural activity related to object processing. All the patients completed the task and achieved ceiling-level performance comparable to controls in the FV14 dataset (mean accuracy ± s.d., patients, 0.90 ± 0.09; controls, 0.92 ± 0.06; $t(64) = 1.131$, $P = 0.262$, Cohen's $d = 0.278$, 95% CI −0.02 to 0.06). Note that all three vision models could extract information relevant for the same behaviour task from FV14 images well (CLIP, 71.4%; ResNet, 86.6%; MoCo, 71.4%). In addition to fMRI data, we collected structural and high angular resolution diffusion imaging (HARDI) data from both the patient and healthy control groups to assess the integrity of the WM connections of interest. We investigated the associations between lower integrity in the language–vision system WM connections and different vision–model RSA effects. The distribution of the lesions in the 33 patients is shown in Fig. 4, revealing a typical middle

cerebral artery stroke pattern, with lesions widely distributed across grey matter regions and WM tracts.

To evaluate the integrity of the left vision–language WM connections in the patients, we first mapped the WM connections by establishing WM masks of interest with the HARDI data in the healthy control group ($N = 35$). We performed tractography between the VOTC mask and the cortical language regions defined by a commonly used language mask[38] (contrasting intact sentences to non-word lists in 220 subjects; Fig. 4, language region). The mask encompasses language-activated clusters located in the left angular gyrus (AG), left posterior middle temporal gyrus (pMTG), left anterior temporal lobe (ATL), left inferior frontal gyrus (IFG), left IFG orbital part (IFGorb) and left middle frontal gyrus (MFG). The WM connections between the VOTC mask and each language region, as well as the VOTC mask connecting all six regions together, were traced and binarized. We applied an individual-level probability threshold of 0.1, a group-level threshold of 0.5 and an explicit WM mask probability threshold of 0.4. For each patient who experienced a stroke, the mean fractional anisotropy (FA) values within each WM mask were computed to quantify the integrity of the left vision–language WM connections.

Association between vision–left AG WM integrity and sentence description effects on the VOTC. Following a similar RSA procedure to that used in study 1, we quantified how well each of the three vision models (CLIPvision, MoCo and ResNet) explained the neural activity patterns in the VOTC during picture viewing for each patient, as well as their specific effects.

We tested whether the structural integrity of the left vision–language WM connections (beyond total lesion volume) was associated with different vision model's model–brain correspondence. Multiple regression analyses were conducted with the patients' FA values of the left vision–language WM connections as the dependent variable. The independent variables included three vision models' brain-correspondence measures (Fisher z-transformed RSA Spearman's correlation coefficients within the left VOTC mask), plus total lesion volume as a covariate. As shown in Fig. 5a, the regression model for the connection linking the left VOTC and the left AG was significant ($F(4, 28) = 3.495$, $P = 0.020$, $R^2 = 0.333$, 95% CI 0.19 to 0.66; Table 1). Significant but opposite effects were observed for the CLIPvision model and MoCo model (CLIP, $t(28) = 2.288$, $P = 0.030$, $\beta = 0.379$, 95% CI 0.04 to 0.72; MoCo, $t(28) = -3.118$, $P = 0.004$, $\beta = -0.582$, 95% CI −0.96 to −0.20; see Table 1 for details of all tracts). That is, a greater integrity of the left VOTC–left AG brain connection was associated with better left VOTC activity fitness to the CLIPvision model and weaker fitness to the MoCo model, whereas disruption of this connection was associated with weaker CLIPvision model fitness and stronger MoCo model fitness. A parallel analysis, testing whether the connection between bilateral VOTC and left AG was associated with model–brain correspondence in bilateral VOTC, yielded similar results (full regression model $F(4, 28) = 3.761$, $P = 0.014$, $R^2 = 0.349$, 95% CI 0.21 to 0.69; CLIP, $t(28) = 2.369$, $P = 0.025$, $\beta = 0.412$, 95% CI 0.06 to 0.77; MoCo, $t(28) = -3.435$, $P = 0.002$, $\beta = -0.653$, 95% CI −1.04 to −0.26).

We also validated these results with partial correlation analyses, in which the correlations between each vision model's specific effect on the left VOTC to the left VOTC–left AG WM integrity was assessed while controlling for total lesion volume. As shown in Fig. 5b, these analyses aligned with the regression model findings: the left VOTC–left AG integrity was positively correlated with the specific model–brain correspondence measure of CLIPvision ($r = 0.409$, one-tailed $P = 0.010$, 95% CI 0.13 to 1.00) and negatively correlated with that of MoCo ($r = -0.505$, one-tailed $P = 0.002$, 95% CI −1.00 to −0.24). For the bilateral VOTC–left AG, the significant correlations persisted (CLIP specific effect, $r = 0.467$, one-tailed $P = 0.003$, 95% CI 0.20 to 1.00; MoCo-specific effect, $r = -0.525$, one-tailed $P = 0.001$, 95% CI −1.00 to −0.27; Extended Data Fig. 1a).

To further test whether the enhanced MoCo-specific correspondence following VOTC–left AG tract damage reflects increased dependence on low-level versus mid-to-high-level visual features, we considered the effect of a relatively lower-level visual property control model, the GIST model[39], which captures global scene properties and low-level visual statistics. We examined the VOTC correspondence of GIST model, as well as MoCo-specific effects, after controlling for the GIST RDM. The analyses revealed two key findings: first, the GIST effects in left VOTC showed a negative correlation trend with left VOTC–left AG tract integrity ($r = -0.258$, one-tailed $P = 0.077$, 95% CI −1.00 to 0.04; Extended Data Fig. 2a, left); second, after controlling for the GIST RDM, MoCo-specific effects in left VOTC maintained a significant negative correlation with tract integrity ($r = -0.505$, one-tailed $P = 0.002$, 95% CI −1.00 to −0.24; Extended Data Fig. 2a, right). For bilateral VOTC, similar correlation patterns were observed in relation to its connection with left AG (GIST effect, $r = -0.284$, one-tailed $P = 0.058$, 95% CI −1.00 to 0.01; MoCo-specific effect, $r = -0.525$, one-tailed $P = 0.001$, 95% CI −1.00 to −0.27; Extended Data Fig. 2b).

Taken together, these results show the unique brain-explanatory power of the CLIPvision model, with respect to ResNet and MoCo, is indeed related to its alignment with language processes. The damage to brain structural connections between the VOTC and language-activated left AG clusters weakens the advantage of the CLIPvision model while increasing the performance of the unsupervised visual models in explaining brain response patterns during object viewing.

Validation showing no effects of the right homologous VOTC–left AG WM connection. To confirm that the observed modulatory effects arose specifically from the language-related functions of the left AG rather than its non-linguistic (for example, non-verbal multimodal integration) capabilities, we examined the analogous WM connections in the right hemisphere. The right AG is also involved in multimodal integration but lacks the same degree of language specialization as the left AG. Therefore, if the observed effects were solely due to the non-linguistic, bilateral functions of the AG, we would expect to encounter similar correlations in the right hemisphere. No such effects were observed for the right AG: for the CLIPvision-specific effect, $r = -0.115$, one-tailed $P = 0.735$, 95% CI −0.40 to 1.00, in the left VOTC and for the MoCo-specific effect, $r = 0.119$, one-tailed $P = 0.741$, 95% CI −1.00 to 0.40, in the left VOTC (Fig. 5c). Similarly, no significant effects were found for the right VOTC–right AG tract: CLIPvision-specific effect, $r = -0.042$, one-tailed $P = 0.590$, 95% CI −0.33 to 1.00, in the right VOTC; MoCo-specific effect, $r = 0.023$, one-tailed $P = 0.549$, 95% CI −1.00 to 0.32, in the right VOTC. These null results for connection with the right AG support the interpretation that the modulation stems from the language-related properties of the left AG rather than from general non-verbal multimodal integration functions.

Whole-brain VFSM results. To validate our tract-specific findings at the whole-brain level, we conducted voxel-wise FA–symptom mapping (VFSM) analyses for both the CLIP-specific effect (sentence description effect) and MoCo-specific effect. For each voxel, we computed Pearson correlations between FA values across patients and their corresponding model–brain correspondence measures within VOTC (voxel-level $P < 0.005$, one-tailed, cluster-level FWE-corrected $P < 0.05$, with total lesion volume as a covariate). For the sentence description effect in VOTC, VFSM revealed three significant clusters where the FA value positively correlated with the sentence description effect: regions near the left AG and pMTG and a small cluster near the left frontal region. For the MoCo-specific effect, we identified a significant cluster in the vicinity of the left AG where reduced FA value was associated with stronger MoCo-specific effect (Extended Data Fig. 3). The overlap analyses with our predefined VOTC–language tract masks confirmed maximal overlap with the VOTC–left AG tract for both effects. In reference to the JHU WM tractography atlas[40], the sentence description

https://doi.org/10.1038/s41562-025-02357-5

Table 1 | Results of regression analysis showing effects of CLIP, ResNet and MoCo model–brain correspondence on FA values in left VOTC–language tracts

FA values of WM tract (dependent variable)	R^2	*F* stats (*P* value)	Standardized β (*P* value)		
			CLIP	ResNet	MoCo
Left VOTC–language	0.201	1.758 (0.165)	0.195 (0.291)	0.209 (0.296)	−0.437 (0.041)
Left VOTC–left AG	0.333	3.495 (0.020)	0.379 (0.030)	0.326 (0.080)	−0.582 (0.004)
Left VOTC–left ATL	0.206	1.815 (0.154)	0.308 (0.099)	0.261 (0.193)	−0.468 (0.029)
Left VOTC–left IFG	0.154	1.275 (0.303)	−0.034 (0.855)	0.094 (0.644)	−0.271 (0.207)
Left VOTC–left IFGorb	0.154	1.277 (0.303)	−0.099 (0.598)	0.009 (0.965)	−0.218 (0.308)
Left VOTC–left MFG	0.038	0.283 (0.886)	−0.046 (0.819)	−0.039 (0.857)	−0.127 (0.575)
Left VOTC–left pMTG	0.249	2.318 (0.082)	0.309 (0.090)	0.277 (0.156)	−0.489 (0.020)

The FA values in each tract are predicted from CLIP, ResNet and MoCo model–brain correspondence, with total lesion volume included as a covariate. The columns show variance explained (R^2), *F* statistics and standardized regression coefficients (β) for each predictor. All the models have the same degrees of freedom (d.f. 1 = 4, d.f. 2 = 28). All the tests are two-tailed.

effect overlapped with the left superior longitudinal fasciculus, left inferior longitudinal fasciculus, forceps minor and left anterior thalamic radiation, whereas the MoCo effect overlapped only with the left superior longitudinal fasciculus and left inferior longitudinal fasciculus (Supplementary Table 2).

Control analyses for model training dataset differences

The three models of interest were selected to contrast different training supervisions, with architecture (ResNet-50) and performance being comparable. To achieve such similarly optimal performance, their training datasets differed, with CLIP having the largest and most diverse training dataset. Language–vision align models such as CLIP inherently require larger datasets to achieve optimal performance compared with traditional visual models—a characteristic well-established in the language model literature[41,42]. This creates an inherent challenge for dataset-matched comparisons, as constraining CLIP to smaller datasets necessarily compromises its performance relative to its optimal functioning.

Nevertheless, to fully characterize results related to training dataset differences, we conducted supplementary analyses using CLIP and SimCLR (a self-supervised model) both trained on identical YFCC15M datasets[43]. This comparison tests whether language supervision contributes to enhanced VOTC correspondence when training datasets are held constant, while acknowledging that such constraints place CLIP at a performance disadvantage relative to its design specifications (zero-shot ImageNet accuracy of 76.2% versus 40.4% for OpenAI-CLIP[44] versus YFCC15M-CLIP[43] with identical architectures). We extracted final-layer features from ViT-B/32 versions of both models, generated corresponding RDMs and performed RSA analyses focusing on CLIPvision-specific effects relative to SimCLR in VOTC neural activity. In study 1, group-level analyses in the FV14 dataset revealed a significant cluster for CLIP's specific effect relative to SimCLR, located in the right occipital pole (Fig. 6a, left). Individual-level analyses using the THINGS dataset identified significant clusters in bilateral fusiform gyrus, extending to the bilateral lingual gyrus and the lateral occipital complex across all three participants (Fig. 6a, right). These results were significant at voxel-level $P < 0.001$ (one-tailed) with the cluster-level FWE correction at $P < 0.05$. No significant clusters emerged in the OPN95 and SPN95 datasets, and lateralization analyses showed no consistent left-lateralization across datasets (OPN95, $t(25) = 1.433$, $P = 0.164$, Cohen's $d = 0.281$, 95% CI −0.05 to 0.28; SPN95, $t(31) = 0.213$, $P = 0.833$, Cohen's $d = 0.038$, 95% CI −0.15 to 0.19; FV14, $t(31) = -0.053$, $P = 0.958$, Cohen's $d = -0.009$, 95% CI −0.18 to 0.17; THINGS, Sub01, 0.004; Sub02, −0.770; Sub03, 0.061).

In study 2's patient dataset, after controlling for total lesion volume, CLIPvision-specific effect in VOTC showed a positive correlation with VOTC–left AG tract integrity ($r = 0.342$, one-tailed $P = 0.028$, 95% CI 0.05 to 1.00; Fig. 6b, left). No such effect was observed for the right AG ($r = 0.177$, one-tailed $P = 0.166$, 95% CI −0.13 to 1.00; Fig. 6b, right). These findings suggest that even after controlling for training dataset differences that compromised performance of the CLIP model, CLIP continues to show tendency of capturing additional VOTC representations through language alignment, and these representations are modulated by the VOTC–left AG tract integrity. L, left hemisphere; R, right hemisphere; cov, covarying; lesion_volume, total lesion volume.

Discussion

Whether human visual cortex activity during visual perception is modulated by language experience remains debated. Our results show that vision DNN models with language alignment supervision—both sentence-level (CLIPvision) and category-label-level (ResNet)—consistently captured the unique variance of the human VOTC neural representation across four datasets with respect to an unsupervised model (MoCo). The sentence-level advantage exhibited by CLIPvision demonstrated left-hemisphere lateralization group-tendency, which aligns with the lateralization of the language network. The analyses of 33 patients who experienced a stroke indicated that the model–brain correspondence in VOTC depended on the structural integrity of its WM connection with the left AG. Reduced tract integrity weakened the CLIPvision-fitting effect while enhancing the MoCo-fitting effect, independent of the degree of overall brain damage.

Recent research comparing CLIP and ResNet in fitting the responses in the human VOTC with the NSD dataset suggested the presence of CLIP-specific effects, although the robustness and interpretation of their results are controversial[13,28]. Our findings validate the robustness of these effects across four datasets of differing stimulus types (isolated objects without backgrounds, FV14, OPN95 and SPN95; natural images, THINGS), populations (healthy typical populations with speech experience and early-deaf individuals with sign language experience) and task demands (image oddball detection, oral picture naming, sign language picture naming and colour knowledge retrieval). CLIP consistently demonstrated unique explanatory power for the activity of the VOTC across all four datasets with respect to the label-supervised (ResNet) and unsupervised (MoCo) vision models.

Critically, two pieces of evidence suggest that the advantage of CLIP in fitting the neural responses within the VOTC stems from the language system in the human brain rather than from the ability to capture some kind of non-verbal, higher-order relations. First, in patients who experienced a stroke, the compromised integrity of the WM tract connecting the VOTC and the left AG (language cluster) diminished the ability of CLIP and enhanced the ability of MoCo to fit the neural activity within the VOTC. This finding suggests that VOTC activity reflects the representations of CLIP or MoCo depending on the relevant communication efficiency with the language brain system. Notably, this effect was not attributable to the overall lesion severity or general AG functioning, as the connections to the right AG showed no modulation

https://doi.org/10.1038/s41562-025-02357-5

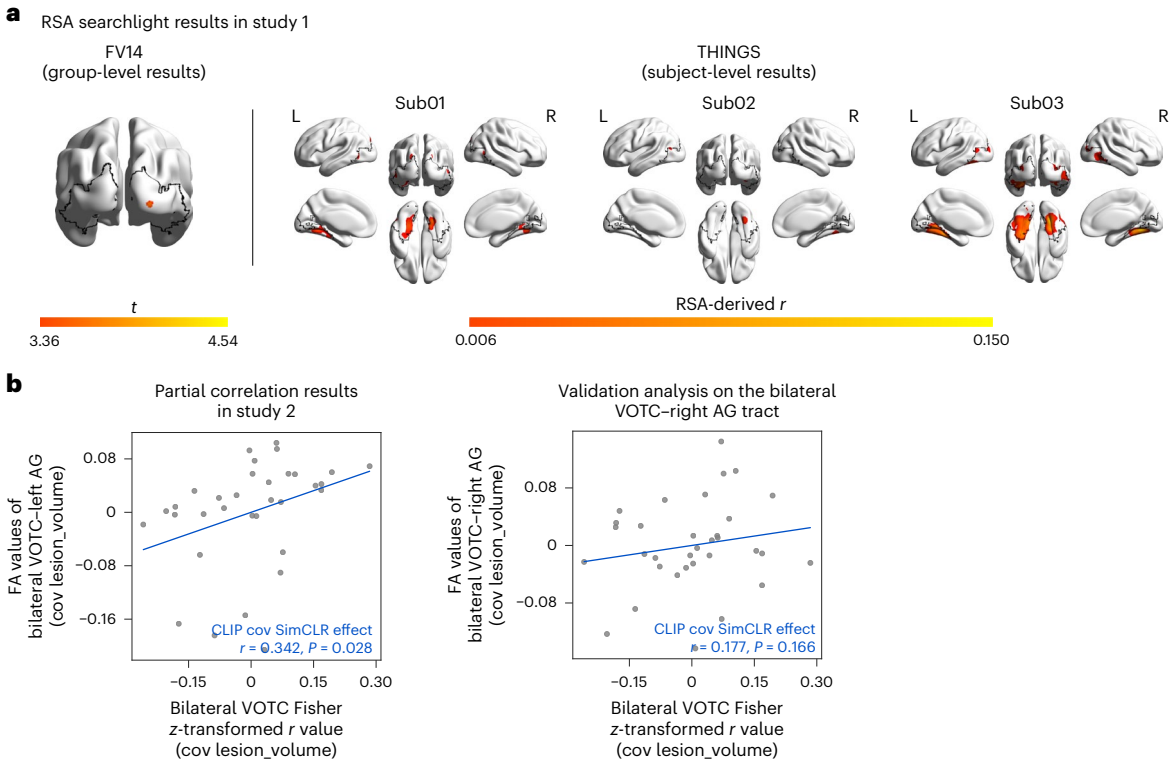


Fig. 6 | Validation analyses using vision models trained on the identical dataset. a, A representation similarity analysis showing CLIP-specific effects (controlling for SimCLR) in VOTC at the group-level for FV14 ($n = 33$ hearing) and at the individual-level for THINGS ($n = 3$ hearing) datasets. All the results are thresholded at voxel-level $P < 0.001$, one-tailed and cluster-level FWE-corrected $P < 0.05$ in the VOTC mask (indicated by the black contours). The coloured bars indicate t values (FV14) or Spearman's correlation coefficients (THINGS). **b**, The partial correlations (degrees of freedom (d.f.) of 30) between VOTC–left AG tract integrity and model–brain correspondence, controlling for lesion volume ($n = 33$ patients). The CLIP-specific effects significantly relate to VOTC–left AG tract integrity ($P < 0.05$), whereas the validation analysis using right AG connection (right) shows no significant relationship, confirming left-lateralized pathway specificity. The correlation coefficients and one-tailed P values are displayed on the plots. This analysis controls for the dataset-specific training effects by comparing CLIP and SimCLR (self-supervised) models both trained on the YFCC15M dataset.

effects. Second, the whole-brain VFSM analyses convergently revealed a highly specific anatomical pattern: only three significant clusters, all located in the left hemisphere with maximal overlap with the VOTC–left AG pathway. If CLIPvision's advantage originated from richer training data enabling superior non-language-specific semantic representations, we would expect modulation by damage to bilateral semantic networks across widespread perceptual systems. Instead, the selective involvement of left-hemisphere language–vision pathways suggests that the effect (at least partly) arises from language supervision during training.

How does language shape the activity of the VOTC during visual perception? Our findings provide several key clues to answering this classical question. First, we observed distinct effects across the three vision–language models: with respect to MoCo, CLIPvision showed consistent, positive and vision–language related effects, whereas ResNet's effects were more variable and unrelated to vision–language connection integrity. As explained earlier, these models reflect two potential pathways through which language may influence visual processing: verbal labels facilitate specific object categorization[45], and larger linguistic units introduce broader relational structures to visual concepts[46]. The robust advantages of CLIP over ResNet suggest a key role for relational structures among word combinations in larger linguistic units.

The second clue stems from the functionality of the left AG, whose connection with the VOTC modulates vision model-fitting patterns. The left AG is widely regarded as a cross-modal hub that integrates different semantic networks related to language and multisensory experiences[47,48]. It processes both taxonomic and thematic object relations[49,50] and supports semantic composition[51,52]. Its connection with VOTC thus is well situated to facilitate interactions between the language system's relational structures and the activity of the VOTC, resonating with CLIPvision (that is, sentence description supervision on visual processing). The absence of significant findings regarding connections between the VOTC and other language regions (for example, the anterior temporal lobe) should be interpreted with caution, as negative results are inherently difficult to parse. Note that our exploratory searchlight analysis within the established language regions[38] revealed inconsistent patterns across datasets when examining the sentence description effects of visual models (Supplementary Fig. 2). This is not surprising, as the target models were vision models, testing how visual representations might be shaped by language, and may not optimally pick up the signals from the language cortex. Future studies using different multimodal models, such as those with fusion-encoder (for example, FLAVA) or encoder–decoder (for example, MetaLM) architectures, may better illuminate these regions' roles.

A final key clue is the patient-level result showing reduced CLIPvision effects following the disruption of VOTC–language left AG connections. This observation challenges the view of language modulation as merely a 'pretraining' process that creates fixed object representations within the VOTC. Instead, our results suggest at least two alternative mechanisms. One is an online interactive process, in which the visual parsing of the VOTC communicates with language system representations, enabling real-time modulation of the activity of the VOTC (see Lupyan et al.[53] for a detailed discussion on online versus offline effects). The other is a poststroke plasticity process (approximately 3 months in our chronic patients), wherein the disconnection of the visual cortex

from the language system induces plastic changes in the VOTC, leading to a more visually driven state. Both interpretations emphasize the dynamic nature of language–vision interactions.

These findings indicate that CLIP's visual–language alignment training probably parallels processes in the human brain and addresses a classical question in human cognition regarding the role of language in visual processing. Although there is ample evidence of non-visual influences on the activity patterns of the VOTC, the effects of language have not been directly established. For example, when non-visual stimuli, such as object names or tactile inputs, are presented—even to congenitally blind individuals—the VOTC exhibits response profiles that are at least partially similar to those during visual perception, particularly in terms of object category preference distributions[54–56] (see Ricciardi et al.[57] and Bi et al.[58] for reviews). The origin of these non-visual effects might be multifaceted, including multimodal shape[59] and/or mappings between shape and non-visual object properties such as related actions[58,60]. Here, we show directly that language is at least one of the sources contributing to VOTC activity and naturally provides an additional mechanism for the similar blind-sighted results reported in literature. Our finding that language dysfunction associates with plasticity shifts toward enhanced visual processing in this region raises a further interesting possibility—blind individuals may even undergo shift in the complementary direction: the absence of visual input drives VOTC representations toward greater language dependence. This aligns with findings of enhanced resting-state functional connectivity between occipital and frontoparietal networks in congenitally blind individuals[61,62].

Two open questions remain. First, pinpointing the specific anatomical locations corresponding to CLIP effects across datasets remains challenging owing to differences in stimuli and tasks. The VOTC exhibits considerable heterogeneity in object category preferences and visual attribute preferences, along with sensitivity to task demands[56]. Indeed, our results showed that CLIP-specific effects varied in both location and magnitude across different datasets, potentially reflecting differences in stimulus content, including both the images' peak sensitivity distribution within VOTC (for example, relating to their categorical preferences) and their corresponding properties in language. Understanding how language experiences interact with these factors will require a systematic investigation of the complex dynamics underlying these interactions.

Our study has several limitations that warrant further investigation. First, CLIP was trained on larger and more diverse datasets than ResNet and MoCo, given language models' dependence on large-scale datasets[41,42]. Although the key finding in study 2—that CLIP superiority was diminished when vision–language tracts were compromised—suggests that CLIP's superiority was not readily explained by broader benefits of larger training sets, controlling for training dataset differences remains informative for understanding model comparisons. Our supplementary analyses using models trained a same smaller dataset (YFCC15M), but with compromised CLIP performance, reduced the robustness of results in study 1 but not study 2. Ideally, future studies should train ResNet-50 and MoCo on large-scale datasets comparable to OpenAI-CLIP to enable more definitive model comparisons. Second, patient studies are intrinsically constrained by the distribution of lesions—the effects of regions that do not have enough patients with lesions cannot be tested. We do not intend to exclude potential additional contributions of non-language-specific effects underlying the superiority of CLIP-like models in fitting VOTC functional profiles. Finally, our datasets tend to use controlled, isolated object images with simple tasks. Although this helps with interpretability, it may underestimate language–vision interactions that could be more pronounced in naturalistic scenarios, involving complex scenes, dynamic contexts or demanding cognitive tasks.

To conclude, our findings indicate that vision models using language supervision uniquely capture representations in the human visual cortex. The strong dependence of this effect on VOTC–left AG WM tract integrity suggests that the human VOTC actively integrates visual inputs with language experience during perception. The convergence of evidence from anatomical specificity, tract-specific association and control analyses of model training datasets supports that language supervision, rather than dataset scale alone, underlies the enhanced correspondence with human VOTC representations. These findings indicate that the human visual cortex engages in dynamic computations that align visual inputs with language experiences during perception. The sensitivity of model–brain similarity to brain lesions, which is specifically related to the properties of the model, further highlights that the leveraging of human brain manipulation is a promising framework for evaluating and developing brain-like machine models.

Methods

fMRI dataset collection (study 1)

Participants. All the participants in the fMRI tasks were right-handed except for two left-handed individuals (ID: D010, N001) and three ambidextrous individuals (ID: D017, D018, D022) in the SPN95. None of them had experienced psychiatric or neurological disorders or had sustained a head injury. The participants in the OPN95[32] and FV14[33] datasets were all native Mandarin speakers, and the deaf participants in the SPN95 were fluent Chinese Sign Language (CSL) users (11 native signers, the others' age of sign language acquisition <12 years), whereas those in the THINGS dataset were all native English speakers. The participants in SPN95 were all congenital or early deaf individuals, which are required to demonstrate hearing thresholds exceeding 85 dB hearing level, indicating profound hearing loss. The sample sizes of the OPN95, SPN95, THINGS and FV14 databases were 29 (18 female participants; median age, 20 years; range, 18–32 years), 36 (19 female participants; median age, 32 years; range, 22–45 years), 3 (2 female participants; mean age at the beginning of the study, 25 years) and 35 (14 female participants; median age, 54 years; range, 31–65 years), respectively.

For the behavioural object-relatedness rating tasks, independent healthy, hearing groups were recruited for OPN95–SPN95, FV14 and THINGS, with several participants in the FV14 dataset overlapping with those in the fMRI task. The sample sizes for the behavioural task were as follows: OPN95–SPN95 (100 participants, 71 female participants; median age, 21 years; range, 18–26 years), FV14 (35 participants, 14 female participants; median age, 54 years; range, 31–65 years) and THINGS (14,025 participants; see Hebart et al.[34] for details). All participants in the OPN95–SPN95 and FV14 datasets were native Mandarin speakers, and those in the THINGS dataset were native English speakers.

Ethics approval. All the protocols and procedures of the current study were approved by the Ethics Committee of the State Key Laboratory of Cognitive Neuroscience and Learning at Beijing Normal University (ICBIR_A_0040_008) and the Ethics Committee of the First Hospital of Shanxi Medical University (no. 2021-K035). Before participation, all the participants provided written informed consent. Compensation was provided in proportion to task completion. The study was conducted in accordance with the Declaration of Helsinki and adhered to all relevant ethical guidelines.

Stimuli and procedures for the task-fMRI experiment. *OPN95 and SPN95.* A total of 95 objects were chosen, encompassing three common categories (32 animals, 35 small manipulable objects and 28 large non-manipulable objects). Each object was presented by a single 400 pixel × 400 pixel colour image depicting a representative exemplar on a white background (10.55° × 10.55° of visual angle). All the participants were asked to name the pictures displayed with oral language (hearing group) or sign language (deaf group). In the sign language naming, the deaf participants were asked to respond with their left hand to reduce confounding activation due to hand movements with language-related activation in the left frontal cortex. The

whole experiment included six runs, with each item being repeated six times. Each run (8 min 45 s) consisted of 95 trials, with each item presented once per run. Each trial consisted of 0.5 s of fixation, 0.8 s of stimulus presentation and an intertrial interval ranging from 2.7 s to 14.7 s. Each run began and ended with a 10-s fixation.

THINGS. A total of 720 objects were chosen, including multiple categories such as animals, tools, fruits and bodies. The participants viewed images of 720 representative object concepts and were instructed to fixate on a central point and press a button upon viewing artificial images. Each concept includes 12 exemplars, that is, 8,640 unique images in total. The whole experiment included a total of 15–16 scanning sessions, among which the first one to two sessions were dedicated to testing the reliability of the head-fixation model and obtaining functional localizers and retinotopic maps. The subsequent 12 runs consisted of a unique sample of the aforementioned 720 concepts, in which each item was presented once. All the presented images subtended 10° of visual angle and were presented on a grey background for 0.5 s and overlaid with a fixation crosshair subtending 0.5 degrees, followed by a 4-s rest stage.

FV14. A total of 14 fruits and vegetables were chosen. Each object was presented by a single 400 pixel × 400 pixel greyscale image depicting a representative exemplar on a white background (6.22° × 6.22° of visual angle). The participants judged whether the typical skin colour of each presented fruit or vegetable was red, responding 'yes' with their right index finger and 'no' with their right middle finger. The experiment included four runs, each of which consisted of 32 1-s-long stimulus trials and 32 11-s-long null trials. Each image was presented twice within each run. The order of the 32 trials was pseudorandomized while ensuring that no two consecutive trials were identical. Each run began with a 12-s silence period and ended with a 4-s silence period.

Image acquisition. *OPN95 and SPN95*. Functional and anatomical MR images were collected at the MRI centre of Beijing Normal University with a 3-T Siemens Trio Tim scanner. The high-resolution three-dimensional (3D) structural images were collected with a 3D magnetization-prepared rapid gradient echo (MPRAGE) sequence in the sagittal plane (144 slices, repetition time (TR) of 2,530 ms, echo time (TE) of 3.39 ms, flip angle of 7°, matrix size of 256 × 256 and voxel size of 1.33 × 1 × 1.33 mm^3). Functional images were acquired with an echo-planar imaging (EPI) sequence (33 axial slices, TR of 2,000 ms, TE of 30 ms, flip angle of 90°, matrix size of 64 × 64 and voxel size of 3 × 3 × 3.5 mm^3 with a gap of 0.7 mm).

THINGS. Functional and anatomical MR images were collected at the National Institutes of Health in Bethesda, MD, USA, with a 3-T Siemens Magnetom Prisma scanner and a 32-channel head coil. The high-resolution 3D structural images were collected with an MPRAGE sequence (208 sagittal slices, voxel size of 0.8 × 0.8 × 0.8 mm^3, TR of 2.4 s, TE of 2.24 ms, matrix size of 320 × 300, field of view (FOV) of 256 × 40 mm^2, flip angle of 8°). The whole-brain functional MR data were collected in 2 mm isotropic resolution (60 axial slices, voxel size of 2 × 2 × 2 mm^3, TR of 1.5 s, TE of 33 ms, matrix size of 96 × 96, FOV of 192 × 192 mm^2, flip angle of 75°).

FV14. Functional and anatomical MR images were collected at the Department of Magnetic Resonance Imaging, First Hospital of Shanxi Medical University, with a 3-T Siemens Magnetom Skyra scanner. The high-resolution 3D structural images were acquired with an MPRAGE sequence (sagittal slices, TR of 2,530 ms, TE of 2.88 ms, flip angle of 7°, matrix size of 224 × 256, interpolated to 448 × 512, voxel size of 0.5 × 0.5 × 1 mm^3, FOV of 224 × 256 mm^2). The functional images were acquired with a multiband EPI sequence (axial slices, TR of 2,000 ms, TE of 30 ms, flip angle of 90°, matrix size of 72 × 72, voxel size of 2.5 × 2.5 × 2.5 mm^3, FOV of 180 × 180 mm^2, multiband factor of 2).

Preprocessing for the task-fMRI data. *OPN95 and SPN95*. The functional images were preprocessed and analysed with Statistical Parametric Mapping (SPM12; http://www.fil.ion.ucl.ac.uk/spm). For each participant, the first five volumes of each run were discarded to ensure signal equilibration. The remaining images were subsequently corrected for slice timing and head motion and then spatially normalized to the Montreal Neurological Institute (MNI) space via unified segmentation[63] (resampled to a 3 × 3 × 3 mm^3 voxel size). The data of three hearing participants and four deaf participants were excluded from the analyses because of excessive head motion (>3 mm or 3°). The object-relevant beta weights of the functional images of each participant were obtained with a general linear model (GLM) that contained an onset regressor for each of 95 items, six regressors of no interest corresponding to the six head motion parameters and a constant regressor for each run. Each item-relevant regressor was convolved with a canonical haemodynamic response function, and the high-pass filter cut-off was set as 128 s. The resulting *t* maps for each item with respect to the baseline were used to create neural RDMs.

THINGS. The preprocessing pipeline for the functional images included slice-timing correction, rigid-body head-motion correction, susceptibility-distortion correction on the basis of the field maps, spatial alignment to each participant's T1-weighted anatomical reference images and brain tissue segmentation and reconstruction of the pial and WM surfaces. The single-trial beta weights were estimated with a GLM. The mean beta maps of 12 samples for each concept were used to create neural RDMs.

FV14. The functional images were preprocessed and analysed with SPM12 following the similar procedure used for the OPN95 and SPN95 databases. For each participant, the first six volumes of each run were discarded for signal equilibration; the remaining images were subsequently corrected for time slicing and head motion and then spatially normalized to the MNI space via unified segmentation (resampling into a 2 × 2 × 2 mm^3 voxel size). Two subjects were excluded from the fMRI analysis: one self-reported an uncomfortable feeling in the head during scanning, and the other exhibited excessive head motion during the scans (>2.5 mm or 2.5°). One or two runs from four subjects showed excessive head motion (>2.5 mm or 2.5°) and were excluded from the analysis. The single-trial beta weights were estimated with a GLM, after which the *t* maps for each item versus baseline were estimated and used to create neural RDMs.

Behavioural object-relatedness rating (outside the scanner). For the OPN95–SPN95 and FV14 datasets, participants performed pairwise similarity ratings of the same objects presented in the fMRI task using a 7-point Likert scale (1 being least similar, 7 being most similar; note that for the FV14 dataset, the scale was reversed with 1 being most similar). For the THINGS dataset, the participants were presented with object triplets and asked to select the 'odd one out' in each triplet. Each triplet was evaluated by multiple but not all participants, resulting in a total of 5,517,400 triplet choices (see Hebart et al.[34] for further details).

Analysis procedures (study 1)

Model details and feature extraction. The models used in the analyses included (1) CLIP trained by OpenAI (with a ResNet-50 backbone)[44], (2) ResNet-50[6] pretrained on ImageNet-1k; (3) MoCo v3 pretrained on ImageNet-1k (with a ResNet-50 backbone)[64], and for control analyses, (4) CLIP and SimCLR pretrained on YFCC15M (with a ViT-Base/32 backbone)[43] and the (5) GIST model[39].

Each image was preprocessed according to the preprocessing parameters provided by the pretrained model and passed through the model to extract the outputs from each layer. As the differences in features among the models were most pronounced at the penultimate layer (Fig. 2), we selected the visual encoder output in CLIP and the

penultimate layer outputs of ResNet-50 and MoCo v3 for fitting the neural representations. For the GIST model, 512-dimensional feature vectors were extracted using a bank of Gabor filters with eight orientations and four spatial scales applied across a 4 × 4 spatial grid, capturing global spatial structure and spectral properties of each image. As none of the models performed well in recognizing or classifying greyscale images, we passed the coloured versions of the images in the FV14 dataset to the models. The dissimilarity between each item was computed as 1 − Pearson's *r* to generate the RDMs. For the THINGS dataset, the average features from the 12 images for each concept were used to generate the RDM.

ROI definition. The VOTC was defined by combining functionally defined regions (activation from hearing participants viewing pictures relative to baseline in the OPN95 dataset; false discovery rate-corrected $q < 0.05$) and anatomical parcels from the Harvard–Oxford Atlas (probability >0.2), specifically the posterior and temporooccipital divisions of the inferior temporal gyrus (15,16), the inferior division of the lateral occipital cortex (23), the posterior division of the parahippocampal gyrus (35), the lingual gyrus (36), the posterior division of the temporal fusiform cortex (38), the temporal occipital fusiform cortex (39), the occipital fusiform gyrus (40), the supracalcarine cortex (47) and the occipital pole (48). This definition resulted in the selection of 2,467 voxels in the left hemisphere and 2,420 voxels in the right hemisphere.

Representation similarity analysis. To identify the areas of the VOTC representing the sentence description effect and verbal categorization effect, we conducted RSA using a searchlight procedure[65] within the defined VOTC mask for each participant.

For the OPN95, SPN95 and FV14 datasets, the *t* value (corresponding to each object relative to the baseline) images of each item were used to calculate the neural RDMs, and for the THINGS dataset, we used beta-value images instead owing to the high signal-to-noise ratio in this dataset. For each voxel within the VOTC mask, multivariate activation patterns within a sphere (radius of 10 mm) centred at that voxel were extracted. Neural RDMs were computed with the Pearson distance within the searchlight sphere. Then, Spearman's correlation coefficients between the neural RDM and model-derived visual RDMs were computed, controlling for the effects of models with lower levels of language involvement to determine the sentence description effect (correlations with the visual RDMs derived from CLIPvision while controlling for the visual RDMs derived from ResNet-50 and MoCo) and the verbal categorization effect (correlations with the visual RDMs derived from ResNet-50 while controlling for the visual RDMs derived from MoCo). Correlation maps were obtained for each participant by moving the searchlight centre across the VOTC mask. These maps were Fisher *z*-transformed and spatially smoothed with a 6-mm (for OPN95 and SPN95) or a 4-mm (for FV14 and THINGS) full-width half-maximum Gaussian kernel. The correlation maps were compared with 0 with one-tailed one-sample *t*-tests.

For the ROI analysis, multivariate activity patterns for each stimulus within the ROI mask were extracted. Neural RDMs were computed on the basis of Pearson distances and then correlated with the RDM generated by CLIPvision while controlling for the RDMs generated by ResNet-50 and MoCo. The resulting correlation coefficients between the neural and model RDMs were Fisher *z*-transformed and compared with zero with one-tailed one-sample *t*-tests. For the comparison between OPN95 and SPN95, Bayesian independent samples *t*-tests were conducted with the Pingouin package[66] on the mean correlation values within the ROI, with a default Cauchy prior width of $r = 0.707$ for the effect size on the alternative hypothesis (H_1: hearing ≠ deaf).

For additional analyses of training datasets, the CLIP specific effect was defined as the correlations with the visual RDMs derived from CLIPvision (YFCC15M) while controlling for the visual RDM derived from SimCLR.

Computation of the LI. Functional lateralization for each dataset was assessed with the LI via the bootstrap method in the LI tool[36] for SPM12. For the group-level analysis, the *t* map of the group RSA-derived rho values versus zero in each dataset was entered as inputs to calculate the LI. For the individual-level analysis, the RSA-derived *r* value map of each subject was entered as the input to calculate the LI and compared with 0 using a two-tailed one-sample *t*-test. The Cohen's *d* effect sizes were also computed.

The LI was calculated via the bootstrap method with the following options: the bilateral VOTC mask defined above as an inclusive mask, no exclusive mask and the default bootstrapping parameters. This method involved the computation of 20 thresholds with equal step lengths ranging from 0 to the maximum *t* value. At each threshold, 100 bootstrapped samples were drawn from both the left and right ROIs, resulting in a total of 200 samples. From these samples, all 10,000 potential LI combinations (100 samples from the left ROIs multiplied by 100 samples from the right ROIs) were calculated for the surviving voxels with the formula $(L - R)/(L + R)$, where L and R represent the values from the left and right ROIs, respectively. To mitigate the influence of statistical outliers, only the central 50% of the data was retained and averaged. The group-level LI index ranged from −1 to 1, where a value of −1 indicates complete right lateralization, a value of 1 indicates complete left lateralization and values between −0.2 and 0.2 are classified as bilateralization[37].

Model-fitting–brain effects tested with lesion models (study 2)

MRI dataset of patients. *Participants*. A total of 33 patients who experienced a stroke (eight female patients; mean ± s.d. age, 51.55 ± 10.02 years; range, 30–65 years), demographically matched with the participants in the FV14 database, without major lesions affecting the VOTC, were recruited following the same task procedure used for the FV14 database (that is, the colour retrieval task) from the First Hospital of Shanxi Medical University. Note that patients with lesions affecting the left hemisphere responded with their left hand instead of right hand. The inclusion criteria for the patients who experienced a stroke were as follows: aged 20–65 years, right-handedness before stroke onset, normal or corrected-to-normal vision, at least 3 months after stroke, first symptomatic stroke of ischaemic or intraparenchymal haemorrhagic aetiology, lesion location involving the cortex and/or subcortical WM, no other neurological or psychiatric diseases, ability to perform simple cognitive tasks and understand instructions and intact object perception. Both the stroke group and the healthy group (participants in FV14) participated in all the following scans. All participants received payments and provided written informed consent. The study has been conducted according to the principles expressed in the Declaration of Helsinki and was approved by the Ethics Committee of First Hospital of Shanxi Medical University (approval no. 2021-K035).

Image acquisition. Functional and anatomical magnetic resonance images were collected at the Department of Magnetic Resonance Imaging, First Hospital of Shanxi Medical University, with a 3-T Siemens Magnetom Skyra scanner. The scans included task-fMRI, HARDI, high-resolution 3D T1-weighted imaging, 3D T2-weighted imaging and 3D fluid-attenuated inversion-recovery (FLAIR) T2-weighted imaging. Functional images were acquired with a multiband EPI sequence (axial slices, TR of 2,000 ms, TE of 30 ms, flip angle of 90°, matrix size of 72 × 72, voxel size of 2.5 × 2.5 × 2.5 mm^3, FOV of 180 ×180 mm^2, multiband factor of 2). The HARDI images were acquired with a multiband EPI sequence (axial slices, TR of 3,000 ms, TE of 100 ms, flip angle of 90°, matrix size of 112 × 112, voxel size of 2 × 2 × 2 mm^3, FOV of 224 × 224 mm^2, multiband factor of 2, diffusion gradient value *b* set to 0, 1,000 and 2,000 s mm^{-2}, b_0 repeated ten times and 64 directions each for $b_{1,000}$ and $b_{2,000}$). The 3D T1-weighted images were acquired with an MPRAGE sequence (sagittal slices, TR of 2,530 ms, TE of 2.88 ms, flip

angle of 7°, matrix size of 224 × 256, interpolated to 448 × 512, voxel size of 0.5 × 0.5 × 1 mm^3, FOV of 224 × 256 mm^2). The T2-weighted images were acquired with a sampling perfection with application optimized contrast using different flip angle evolution sequence (sagittal slices, TR of 3,200 ms, TE of 408 ms, flip angle of 120°, matrix size of 256 × 256, voxel size of 0.9 × 0.9 × 0.9 mm^3, FOV of 230 × 230 mm^2). The FLAIR T2-weighted images were acquired with a FLAIR sequence (sagittal slices, TR of 5,000 ms, TE of 394 ms, flip angle of 120°, matrix size of 256 × 256, interpolated to 512 × 512, voxel size of 0.5 × 0.5 × 1 mm^3, FOV of 250 × 250 mm^2).

Image analysis. The functional images were preprocessed and analysed as described for study 1 above. One or two runs from four patients who experienced a stroke demonstrated excessive head motion (>3 mm per 3°), and the runs were subsequently excluded from the analysis. For the HARDI data, preprocessing was performed using FSL (version 6.07) from Oxford University, including (1) Eddycorrect for correcting distortions and head movement, (2) BET for skull removal and (3) DTIFIT for building diffusion tensor models and calculating FA maps. FA images were registered to the T1 images with FLIRT and then to the MNI space with T1-to-MNI transformation, achieving a voxel size of 2 × 2 × 2 mm^3. The derived transformation parameters were used to warp the ROI from MNI space to the native diffusion space via nearest-neighbour interpolation. For the structural MRI data, lesions were drawn by a radiology resident and reviewed by a radiologist using ITK-SNAP. The protocol included (1) coregistration of the T2-weighted and FLAIR images to the 3D T1-weighted images, (2) lesion delineation on the axial FLAIR images, including glioses, (3) lesion image registration to MNI space via normalization parameters from fMRI preprocessing, with a voxel size of 1 × 1 × 1 mm^3 and (4) application of a 3-mm smoothing kernel to obtain the brain injury map.

Defining the WM mask of interest in healthy controls. WM connections of the healthy individuals were mapped with probabilistic tractography. This technique examines the probability distributions of the fibre orientations in the brain, allowing the representation of uncertainty and the delineation of crossing fibres. For datasets with multiple *b* values, this approach enhances fibre orientation accuracy in the WM and regions near the cortex.

ROI selection. We selected the left language-specific regions as defined by Fedorenko et al.[38] and the left VOTC mask (defined in the ROI definition section). For complementary analyses, a bilateral VOTC mask was also included. For the controls, the right-hemisphere homologues of the left-hemisphere language regions were also included. All the ROIs were transformed into the diffusion native space for each participant for probabilistic tractography.

Probabilistic tractography. Using the HARDI data from the healthy controls with FMRIB's Diffusion Toolbox in FSL, we performed tractography between each pair of ROIs. This procedure focused on (1) connections between the VOTC mask (left or bilateral) and each left language region and (2) connections between the VOTC mask (left or bilateral) and the right hemisphere homologue of each left-hemisphere language region. Fibre tracking was performed using FSL's BEDPOSTX with default parameters, modelling WM fibre orientations and crossing fibres. Fibre tracking was initiated in both directions, with 5,000 streamlines drawn from each voxel in the ROI with PROBTRACKX 2.0. A cerebrospinal fluid mask was used as the exclusion mask. The connectivity distributions were normalized to MNI space and standardized. At the individual participant level, the path images were thresholded at a value of 0.1 to remove low-connectivity probability voxels. At the group level, fibre projections present in more than 50% of the individuals within the explicit WM mask (probability >0.4 in the WM probability map[67]) were retained for analysis.

Correlation analysis of the WM–vision-model effects. To assess the associations between VOTC–language WM connection integrity and the effect patterns of the different vision models, we performed multiple regression analyses. In these analyses, RSA-derived rho values (Fisher *z*-transformed) of CLIPvision, ResNet-50 and MoCo were used as predictors of the WM connection integrity values between the VOTC mask (left or bilateral) and each language region, with the total lesion volume as a covariate. WM integrity was quantified as the mean FA value within the WM tract mask for each patient. The model significance was measured with *F* tests and *t*-tests.

To confirm whether the effect of interest observed in the VOTC was related to language, partial Pearson's correlation analyses were subsequently conducted, with total lesion volume as a covariate, to analyse the relationship between the average integrity of the WM tracts (as measured by the mean FA values) of interest and the effect of interest within the VOTC mask (left or bilateral). One-sample *t*-tests were performed on the Fisher *z*-transformed correlation coefficients.

To rule out the possibility that the sentence description effect in the VOTC was related to other high-order cognitive control functions rather than language, the analyses conducted for the left language regions were also performed on their right-hemisphere homologues.

To assess whether the three models extract information relevant for the in-scanner task from FV14 images, we trained linear classification layers for all three models to judge whether input images had red surfaces, using leave-one-out cross-validation.

Whole-brain VFSM. For each voxel within the whole-brain WM template, FA values were extracted and correlated with model–brain correspondence across patients using Pearson correlation, with total lesion volume included as a covariate. This approach generated whole-brain VFSM correlation maps (*r* maps) for each specific model effect.

The statistical significance was determined using a voxel-level threshold of $P < 0.005$ combined with cluster-level FWE correction at $P < 0.05$. The significant clusters were then assessed for spatial overlap with both the JHU WM atlas and our custom-traced six VOTC–language WM masks to identify the specific WM pathways contributing to each neural effect.

Brain visualization. The brain results were projected onto the MNI brain surface for visualization with BrainNet Viewer[68] (version 1.7; https://www.nitrc.org/projects/bnv/; Research Resource Identifier, SCR_009446) with the default 'interpolated' mapping algorithm, unless stated explicitly otherwise. The slice view results were visualized using DPABI[69] and xjView toolbox (https://www.alivelearn.net/xjview).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available via figshare at https://doi.org/10.6084/m9.figshare.29531288.v3 (ref. 70). The original neuroimaging data are not publicly available owing to ethical constraints. De-identified data may be accessed by researchers who meet the criteria upon reasonable request: study 1 via the corresponding author (ybi@pku.edu.cn); study 2 via the Ethics Committee of the First Hospital of Shanxi Medical University (phone: +86 351 4639242) or the corresponding author (ybi@pku.edu.cn). Eligible requests will receive a response within 2 weeks.

Code availability

The custom codes that support the findings of this study are available via figshare at https://doi.org/10.6084/m9.figshare.29531288.v3 (ref. 70).

References

1. Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
2. Schrimpf, M. et al. The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl Acad. Sci. USA* **118**, e2105646118 (2021).
3. Schrimpf, M. et al. Brain-score: which artificial neural network for object recognition is most brain-like? Preprint at *bioRxiv* https://doi.org/10.1101/407007 (2018).
4. Yamins, D. L. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
5. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
6. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
7. Kriegeskorte, N. et al. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–1141 (2008).
8. Ungerleider, L. G. & Haxby, J. V. 'What' and 'where'in the human brain. *Curr. Opin. Neurobiol.* **4**, 157–165 (1994).
9. Dobs, K., Martinez, J., Kell, A. J. E. & Kanwisher, N. Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci. Adv.* **8**, eabl8913 (2022).
10. Konkle, T. & Alvarez, G. A. A self-supervised domain-general learning framework for human ventral stream representation. *Nat. Commun.* **13**, 491 (2022).
11. Vinken, K., Prince, J. S., Konkle, T. & Livingstone, M. S. The neural code for 'face cells' is not face-specific. *Sci. Adv.* **9**, eadg1736 (2023).
12. Prince, J. S., Alvarez, G. A. & Konkle, T. Contrastive learning explains the emergence and function of visual category-selective regions. *Sci. Adv.* **10**, eadl1776 (2024).
13. Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J. & Wehbe, L. 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).
14. Zhou, Q., Du, C., Wang, S. & He, H. CLIP-MUSED: CLIP-guided multi-subject visual neural information semantic decoding. In *Proc. 12th International Conference on Learning Representations* (eds Kim, B. et al.) https://openreview.net/pdf?id=lKxL5zkssv (ICLR, 2024).
15. Doerig, A. et al. High-level visual representations in the human brain are aligned with large language models. *Nat. Mach.Intell.* **7**, 1220–1234 (2025).
16. Conwell, C., Prince, J. S., Hamblin, C. J. & Alvarez, G. A. Controlled assessment of CLIP-style language-aligned vision models in prediction of brain and behavioral data. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models* (ME-FoMo, 2023).
17. Luo, A. F., Henderson, M. M., Wehbe, L. & Tarr, M. J. Brain diffusion for visual exploration: cortical discovery using large-scale generative models. In *Proc. 37th International Conference on Neural Information Processing Systems* (eds Oh, A. et al.) 75740–75781 (Curran Associates, 2023).
18. Luo, A. F., Henderson, M. M., Tarr, M. J. & Wehbe, L. BrainSCUBA: fine-grained natural language captions of visual cortex selectivity. In *Proc. 12th International Conference on Learning Representations* (eds Kim, B. et al.) https://openreview.net/pdf?id=mQYHXUUTkU (ICLR, 2024).
19. Lupyan, G. The centrality of language in human cognition. *Lang. Learn.* **66**, 516–553 (2016).
20. Thierry, G. Neurolinguistic relativity: how language flexes human perception and cognition. *Lang. Learn.* **66**, 690–713 (2016).
21. Gilbert, A. L., Regier, T., Kay, P. & Ivry, R. B. Whorf hypothesis is supported in the right visual field but not the left. *Proc. Natl Acad. Sci. USA* **103**, 489–494 (2006).
22. Drivonikou, G. V. et al. Further evidence that Whorfian effects are stronger in the right visual field than the left. *Proc. Natl Acad. Sci. USA* **104**, 1097–1102 (2007).
23. Winawer, J. et al. Russian blues reveal effects of language on color discrimination. *Proc. Natl Acad. Sci. USA* **104**, 7780–7785 (2007).
24. Ting Siok, W. et al. Language regions of brain are operative in color perception. *Proc. Natl Acad. Sci. USA* **106**, 8140–8145 (2009).
25. Martinovic, J., Paramei, G. V. & MacInnes, W. J. Russian blues reveal the limits of language influencing colour discrimination. *Cognition* **201**, 104281 (2020).
26. Fedorenko, E., Piantadosi, S. T. & Gibson, E. A. Language is primarily a tool for communication rather than thought. *Nature* **630**, 575–586 (2024).
27. Maier, M. & Abdel Rahman, R. No matter how: top-down effects of verbal and semantic category knowledge on early visual perception. *Cogn. Affect. Behav. Neurosci.* **19**, 859–876 (2019).
28. Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A. & Konkle, T. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nat. Commun.* **15**, 9383 (2024).
29. Allen, E. J. et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).
30. Conwell, C. et al. Monkey See, model knew: large language models accurately predict visual brain responses in humans and non-human primates. Preprint at *bioRxiv* https://doi.org/10.1101/2025.03.05.641284 (2025).
31. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 249 (2008).
32. Fu, Z. et al. Different computational relations in language are captured by distinct brain systems. *Cereb. Cortex.* **33**, 997–1013 (2023).
33. Liu, B. et al. Object knowledge representation in the human visual cortex requires a connection with the language system. *PLoS Biol.* **23**, e3003161 (2025).
34. Hebart, M. N. et al. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife* **12**, e82580 (2023).
35. Güntürkün, O., Ströckens, F. & Ocklenburg, S. Brain lateralization: a comparative perspective. *Physiol. Rev.* **100**, 1019–1063 (2020).
36. Wilke, M. & Lidzba, K. LI-tool: a new toolbox to assess lateralization in functional MR-data. *J. Neurosci. Methods* **163**, 128–136 (2007).
37. Seghier, M. L. Laterality index in functional MRI: methodological issues. *Magn. Reson. Imaging* **26**, 594–601 (2008).
38. Fedorenko, E., Hsieh, P. J., Nieto-Castañón, A., Whitfield-Gabrieli, S. & Kanwisher, N. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J. Neurophysiol.* **104**, 1177–1194 (2010).
39. Oliva, A. & Torralba, A. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **42**, 145–175 (2001).
40. Hua, K. et al. Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. *Neuroimage* **39**, 336–347 (2008).
41. Brown, T. B. et al. Language models are few-shot learners. In *Proc. 34th International Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems* Vol. 33 (eds Larochelle, H. et al.) 1877–1901 (Curran Associates, Inc., 2020).

42. Kaplan, J. et al. Scaling laws for neural language models. Preprint at https://arxiv.org/abs/2001.08361 (2020).
43. Mu, N., Kirillov, A., Wagner, D. & Xie, S. SLIP: self-supervision meets language-image pre-training. In *European Conference on Computer Vision* (eds Avidan, S. et al.) 529–544 (Springer, 2022).
44. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning: Proc. Machine Learning Research* Vol. 139 (eds Meila, M. & Zhang, T.) 8748–8763 (PMLR, 2021).
45. Gelman, S. A. & Roberts, S. O. How language shapes the cultural inheritance of categories. *Proc. Natl Acad. Sci. USA* **114**, 7900–7907 (2017).
46. Unger, L. & Fisher, A. V. The emergence of richly organized semantic knowledge from simple statistics: a synthetic review. *Dev. Rev.* **60**, 100949 (2021).
47. Xu, Y., He, Y. & Bi, Y. A tri-network model of human semantic processing. *Front. Psychol.* **8**, 1538 (2017).
48. Seghier, M. L. The angular gyrus: multiple functions and multiple subdivisions. *Neuroscientist* **19**, 43–61 (2013).
49. Xu, Y. et al. Doctor, teacher, and stethoscope: neural representation of different types of semantic relations. *J. Neurosci.* **38**, 3303–3317 (2018).
50. Schwartz, M. F. et al. Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. *Proc. Natl Acad. Sci. USA* **108**, 8520–8524 (2011).
51. Zhang, W., Xiang, M. & Wang, S. The role of left angular gyrus in the representation of linguistic composition relations. *Hum. Brain Mapp.* **43**, 2204–2217 (2022).
52. Price, A. R., Bonner, M. F., Peelle, J. E. & Grossman, M. Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *J. Neurosci.* **35**, 3276–3284 (2015).
53. Lupyan, G., Rahman, R. A., Boroditsky, L. & Clark, A. Effects of language on visual perception. *Trends Cogn. Sci.* **24**, 930–944 (2020).
54. Mattioni, S. et al. Categorical representation from sound and sight in the ventral occipito-temporal cortex of sighted and blind. *Elife* **9**, e50732 (2020).
55. van den Hurk, J., Van Baelen, M. & Op de Beeck, H. P. Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proc. Natl Acad. Sci. USA* **114**, E4501–E4510 (2017).
56. Wang, X. et al. How visual is the visual cortex? Comparing connectional and functional fingerprints between congenitally blind and sighted individuals. *J. Neurosci.* **35**, 12545–12559 (2015).
57. Ricciardi, E., Bonino, D., Pellegrini, S. & Pietrini, P. Mind the blind brain to understand the sighted one! Is there a supramodal cortical functional architecture?. *Neurosci. Biobehav. Rev.* **41**, 64–77 (2014).
58. Bi, Y., Wang, X. & Caramazza, A. Object domain and modality in the ventral visual pathway. *Trends Cogn. Sci.* **20**, 282–290 (2016).
59. Peelen, M. V. & Downing, P. E. Category selectivity in human visual cortex: beyond visual object recognition. *Neuropsychologia* **105**, 177–183 (2017).
60. Mahon, B. Z. et al. Action-related properties shape object representations in the ventral stream. *Neuron* **55**, 507–520 (2007).
61. Striem-Amit, E. et al. Functional connectivity of visual cortex in the blind follows retinotopic organization principles. *Brain* **138**, 1679–1695 (2015).
62. Burton, H., Snyder, A. Z. & Raichle, M. E. Resting state functional connectivity in early blind humans. *Front. Syst. Neurosci.* **8**, 51 (2014).
63. Ashburner, J. & Friston, K. J. Unified segmentation. *Neuroimage* **26**, 839–851 (2005).
64. Chen, X., Xie, S. & He, K. An empirical study of training self-supervised vision transformers. In *Proc. IEEE/CVF International Conference on Computer Vision* 9640–9649 (IEEE, 2021).
65. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl Acad. Sci. USA* **103**, 3863–3868 (2006).
66. Vallat, R. Pingouin: statistics in Python. *J. Open Source Softw.* **3**, 1026 (2018).
67. Fonov, V. et al. Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage* **54**, 313–327 (2011).
68. Xia, M., Wang, J. & He, Y. BrainNet Viewer: a network visualization tool for human brain connectomics. *PLoS ONE* **8**, e68910 (2013).
69. Yan, C. G., Wang, X. D., Zuo, X. N. & Zang, Y. F. DPABI: data processing and analysis for (resting-state) brain imaging. *Neuroinformatics* **14**, 339–351 (2016).
70. Chen, H. Language modulates vision: evidence from neural networks and human brain-lesion models. *figshare* https://doi.org/10.6084/m9.figshare.29531288.v3 (2025).
71. Stoinski, L. M., Perkuhn, J. & Hebart, M. N. THINGSplus: new norms and metadata for the THINGS database of 1854 object concepts and 26,107 natural object images. *Behav. Res.* **56**, 1583–1603 (2024).

Acknowledgements

This research was supported by grants from the STI2030-Major Project 2021ZD0204100 (grant no. 2021ZD0204104 to Y.B.); the National Natural Science Foundation of China (grant nos. 31925020 and 82021004 to Y.B.; grant no. 62376009 to Y.Z.; grant no. 32171052 to Xiaosha Wang; 62406020 to W.H.); the Fundamental Research Funds for the Central Universities (Y.B.); and the PKU-BingJi Joint Laboratory for Artificial Intelligence (Y.Z.). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank H. Yang, Z. Xiong, Z. Fu and H. Wen for their valuable comments on earlier drafts of the manuscript.

Author contributions

Y.B., Y.Z. and W.H. conceived the study. H.C., B.L., Xiaosha Wang and Xiaochun Wang designed the experiment. H.C., B.L. and S.W. implemented and conducted the experiments. H.C. and B.L. analysed the data. H.C. and Y.B. wrote the initial draft. All authors reviewed and edited the Article.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-025-02357-5.

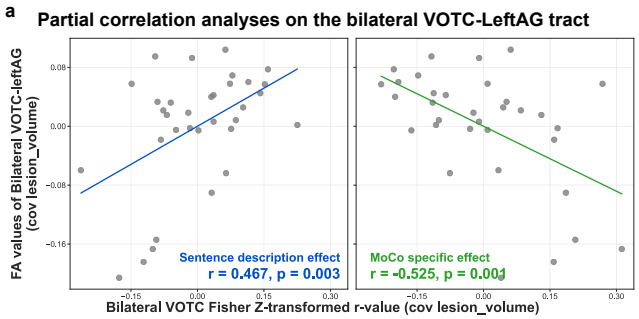
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-025-02357-5.

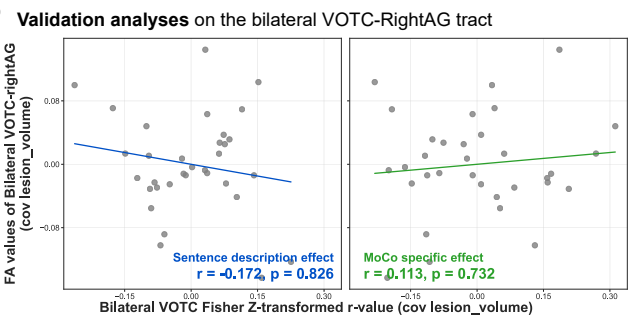
Correspondence and requests for materials should be addressed to Xiaochun Wang, Yixin Zhu or Yanchao Bi.

Peer review information *Nature Human Behaviour* thanks Guadalupe Dávila, Francesca Setti and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

[1]School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China. [2]State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China. [3]Department of Radiology, First Hospital of Shanxi Medical University, Taiyuan, China. [4]College of Medical Imaging, Shanxi Medical University, Taiyuan, China. [5]Shanxi Key Laboratory of Intelligent Imaging, First Hospital of Shanxi Medical University, Taiyuan, China. [6]School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China. [7]Institute for Artificial Intelligence, Peking University, Beijing, China. [8]State Key Laboratory of General Artificial Intelligence, Peking University, Beijing, China. [9]Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China. [10]IDG/McGovern Institute for Brain Research, Peking University, Beijing, China. [11]These authors contributed equally: Haoyang Chen, Bo Liu. [12]These authors jointly supervised this work: Xiaochun Wang, Yixin Zhu, Yanchao Bi. ✉e-mail: wangxiaochun@sydyy.com; yixin.zhu@pku.edu.cn; ybi@pku.edu.cn

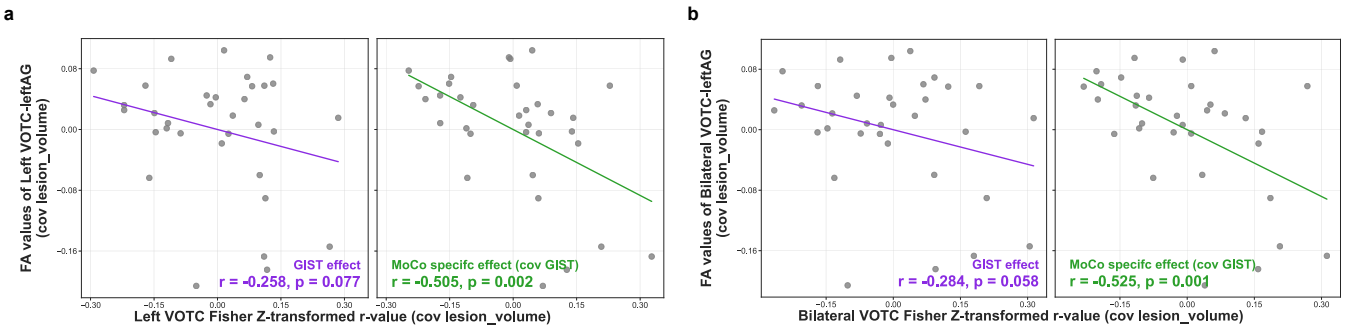
https://doi.org/10.1038/s41562-025-02357-5





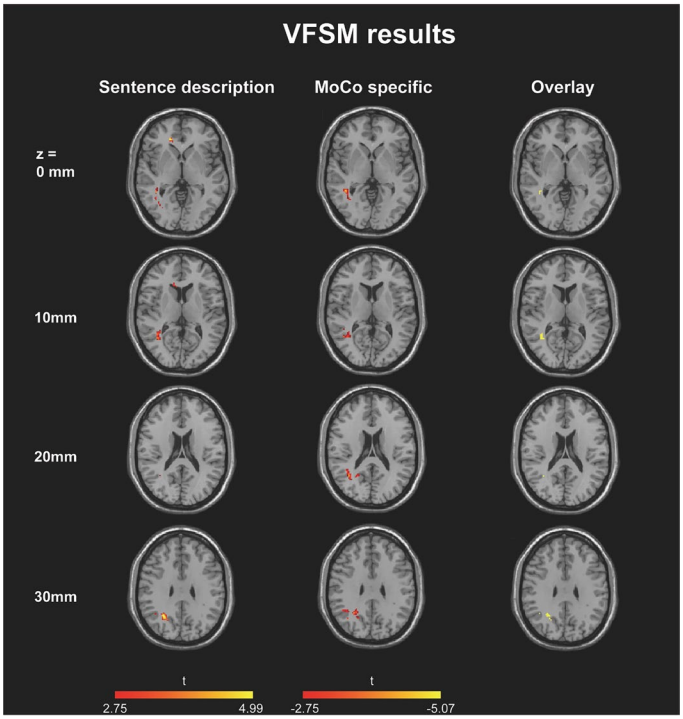
Extended Data Fig. 1 | White matter integrity of the bilateral VOTC–LeftAG tract predicts model–brain correspondence of CLIP and MoCo (n = 33 patients). Correlation coefficients and one-tailed P values are displayed on the plots (d.f.=30). **a**. Partial correlations between bilateral VOTC–LeftAG tract integrity and model–brain correspondence, controlling for lesion volume. Both sentence description effects (CLIP-specific) and MoCo-specific effects correlate significantly with VOTC–LeftAG tract integrity. **b**. Validation analysis using connections with right AG shows no significant relationships, confirming left-lateralized pathway specificity.

https://doi.org/10.1038/s41562-025-02357-5



Extended Data Fig. 2 | Low-level versus higher-level visual feature dependencies in WM–neural representation relationships (n = 33 patients). Pearson correlations (d.f.=30) between left (**a**) or bilateral (**b**) VOTC–AG tract FA values and model–brain correspondence, controlling for lesion volume. The left panel shows that GIST effects (low-level visual features) exhibit a negative trend with tract integrity, whereas the right panel demonstrates that MoCo-specific effects (controlling for CLIP, ResNet, and GIST) correlate significantly with WM integrity. Correlation coefficients and one-tailed P values are displayed on the plots.

https://doi.org/10.1038/s41562-025-02357-5



Extended Data Fig. 3 | Voxel-based FA–symptom mapping (VFSM) results for model–brain correspondence (n = 33 patients). Whole-brain correlation analyses (Pearson's correlations) examine the relationships between FA values of each voxel and Fisher *z*-transformed RSA correlations in VOTC across patients. Left column shows correlations with sentence description effects (CLIP-specific effect); middle column displays MoCo-specific effects; right column indicates voxels with significant correlations for both conditions. Colour bars represent t-statistics from correlation analyses controlling for lesion volume. Results are thresholded at voxel-level $P < 0.005$, one-tailed, and cluster-level FWE-corrected $P < 0.05$. Axial slices displayed in MNI coordinate space.

nature portfolio

Corresponding author(s): Yixin Zhu, Xiaochun Wang and Yanchao Bi

Last updated by author(s): Sep 29, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collection was conducted using E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA)(Schneider et al. 2002).
Data analysis	The fMRI data were preprocessed using the Statistical Parametric Mapping software (SPM12; http://www.fil.ion.ucl.ac.uk/spm/) and DPABI V3.0 (Yan et al., 2016). The HARDI data were preprocessed using FSL (version 6.07) from Oxford University. The probabilistic tractography was performed with FMRIB's Diffusion Toolbox in FSL. After preprocessing, data were analysed using Python (version 3.10) and Matlab (R2023b). Bayesian analysis was conducted using the Pingouin package (Vallat, 2018). All codes for analysis are available from figshare (https://doi.org/10.6084/m9.figshare.29531288.v3).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data that support the findings of this study are available via figshare (https://doi.org/10.6084/m9.figshare.29531288.v3). The original neuroimaging data are not publicly available owing to ethical constraints. De-identified data may be accessed by researchers who meet the criteria upon reasonable request: Study 1 via the corresponding author (ybi@pku.edu.cn); Study 2 via the Ethics Committee of the First Hospital of Shanxi Medical University (phone: +86 351 4639242) or the corresponding author (ybi@pku.edu.cn). Eligible requests will receive a response within two weeks.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Our findings apply to both sexes and genders. Genders and sexes were not considered in our studies. We performed no sex- or gender-based analyses, because there was no sufficient evidence about differences in neural correlates of the language-related influence on visual perception between sexes or genders.
Reporting on race, ethnicity, or other socially relevant groupings	Our findings do not involve any racial or ethnic classification.
Population characteristics	For the fmri tasks in study 1, all the participants were right-handed except for two left-handed individuals (ID: D010, N001) and three ambidextrous individuals (ID: D017, D018, D022) in the SPN95. The participants in the OPN95 and FV14 datasets were all native Mandarin speakers, and the deaf participants in the SPN95 were fluent Chinese Sign Language (CSL) users (11 native signers, others' age of sign language acquisition < 12 years), whereas those in the THINGS dataset were all native English speakers. The participants in SPN95 were all congenital or early-deaf individuals. The sample sizes of the OPN95, SPN95, THINGS, and FV14 databases were 26 (18 females; median age: 20 years; range: 18–32 years), 36 (19 females; median age: 32 years; range: 22–45 years), 3 (2 females; mean age at the beginning of the study: 25 years), and 33 (12 females; mean ± SD age: 50.23 ± 9.94 years; range: 31–65 years), respectively. For the behavioural tasks, independent groups were recruited for OPN95/SPN95, FV14, and THINGS. Some participants in the FV14 dataset overlapped with those in the fMRI task. Specifically, the sample sizes for the behavioural task were as follows: OPN95/SPN95 (100 participants, 71 females; median age: 21 years; range: 18–26), FV14 (36 participants, 14 females; median age: 36 years; range: 20–65), and THINGS (14,025 participants; for details, see Hebart et al., 2023). All participants in the OPN95/SPN95 and FV14 datasets were native Mandarin speakers, whereas those in the THINGS dataset were native English speakers. For study 2, 33 stroke patients (8 females; mean ± SD age: 51.55 ± 10.02 years; range: 30–65 years), demographically matched with the participants in the FV14 database, were recruited following the same task procedure used for the FV14 database (i.e., colour retrieval task).
Recruitment	For Study1, except for the open dataset THINGS, all participants were recruited online from adults in Beijing and Taiyuan. Participant should be native Mandarin (or Chinese Sign Language) users. None of them had experienced psychiatric or neurological disorders or had sustained a head injury. The participants in SPN95 were all congenital or early-deaf individuals, required to demonstrate hearing thresholds exceeding 85 dB HL (hearing level), indicating profound hearing loss. For Study 2, the patients were recruited from the First Hospital of Shanxi Medical University. The additional inclusion criteria for the stroke patients were as follows: aged 20–65 years; right-handedness before stroke onset; normal or corrected-to-normal vision; at least 3 months after stroke; first symptomatic stroke of ischaemic or intraparenchymal haemorrhagic aetiology; lesion location involving the cortex and/or subcortical WM; no other neurological or psychiatric diseases; ability to perform simple cognitive tasks and understand instructions; and intact object perception; no major lesions affecting the VOTC. Due to the adult participants, the research results may not generalize to other populations (e.g., children).
Ethics oversight	All protocols and procedures of the current study were approved by the Ethics Committee of the State Key Laboratory of Cognitive Neuroscience and Learning at Beijing Normal University (ICBIR_A_0040_008), and the Ethics Committee of the First Hospital of Shanxi Medical University (No. 2021-K035).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This is a quantitative experimental research involving human subjects.
Research sample	Except for the open dataset THINGS (Hebart et al., 2023), the rest of participants are all native Chinese Mandarin (or Chinese Sign Language for SPN95) adult users in Beijing and Taiyuan. There were 124 participants in total. TThe sample sizes of the OPN95, SPN95, THINGS, and FV14 databases were 29 (18 females; median age: 20 years; range: 18–32 years), 36 (19 females; median age: 32 years; range: 22–45 years), 3 (2 females; mean age at the beginning of the study: 25 years), and 35 (14 females; median age: 54 years; range: 31–65 years), respectively. The sample size in Study 2 was 33 stroke patients (8 females; mean ± SD age: 51.55 ± 10.02 years; range: 30–65 years), who were recruited from the First Hospital of Shanxi Medical University. The participants in this research are all adults, so they may not fully represent other groups (e.g., children).
Sampling strategy	For Study 1, none of them should have experienced psychiatric or neurological disorders or had sustained a head injury. And the participants in SPN95 should be all congenital or early-deaf individuals, required to demonstrate hearing thresholds exceeding 85 dB HL (hearing level), indicating profound hearing loss. Except for the open dataset THINGS (Hebart et al., 2023), the rest of participants should be all native Chinese Mandarin (or Chinese Sign Language for SPN95) adult users in Beijing and Taiyuan. The participants in THINGS are all native English speakers. For Study 2, the participants should be: aged 20–65 years; right-handedness before stroke onset; normal or corrected-to-normal vision; at least 3 months after stroke; first symptomatic stroke of ischaemic or intraparenchymal haemorrhagic aetiology; lesion location involving the cortex and/or subcortical WM; no other neurological or psychiatric diseases; ability to perform simple cognitive tasks and understand instructions; and intact object perception; no major lesions affecting the VOTC. The sampling procedure was random designed. Sample sizes were determined by the previous model-fMRI fitting and HARDI-based correlation studies on the language-vision effects, and sample availability.
Data collection	In all fMRI experiments, the ongoing brain activity during the task was recorded using a MRI scanner; the participants' button responses were recorded with a computer and picture naming responses were not recorded. In HARDI, participants were required to keep quiet, while the DW signal was recorded. No one was present in the room together with the participants during the experiments. The researcher was aware of the experimental conditions and the study hypothesis during data collection.
Timing	OPN95 dataset was collected in 2019-2020. SPN95 was collected in 2023-2024. FV14 and patients' data was collected in 2022-2023.
Data exclusions	Study 1: the data of three hearing participants (OPN95) and four deaf participants (SPN95) were excluded from the analyses because of excessive head motion (> 3 mm/3°). And two subjects were excluded from the fMRI analysis in FV14 dataset: one self-reported an uncomfortable feeling in the head during scanning, and the other exhibited excessive head motion during the scans (> 2.5 mm/2.5°). One or two runs from four subjects (FV14) showed excessive head motion (> 2.5 mm/2.5°) and were excluded from the analysis. Study2: four stroke patients showed excessive head motion in 1 or 2 runs (> 3mm/3°), and those unusual runs were excluded.
Non-participation	No participants declined participation or dropped out.
Randomization	Participants were not allocated into experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.

Magnetic resonance imaging

Experimental design

Design type

All the experiment were task state with event-related design.

Design specifications

OPN95/SPN95: there were 6 runs for each participant; each run contained 95 trials; Each trial consisted of 0.5s of fixation, 0.8s of stimulus presentation and an intertrial interval (ITI) ranging from 2.7s-14.7s. Each run began and ended with a 10-s fixation.

FV14 and Study 2: there were 4 runs for each participant, each of which consisted of 32 1s-long stimulus trials and 32 11s-long null trials. Each image was presented twice within each run. The order of the 32 trials was pseudorandomized while ensuring that no two consecutive trials were identical. Each run began with a 12-s silence period and ended with a 4-s silence period.

THINGS: The whole experiment included a total of 15–16 scanning sessions, among which the first 1–2 sessions were dedicated to testing the reliability of the head fixation model and obtaining functional localizers and retinotopic maps. The subsequent 12 runs consisted of a unique sample of the aforementioned 720 concepts, in which each item was presented once. All the presented images subtended 10 degrees of visual angle and were presented on a grey background for 0.5 s and overlaid with a fixation crosshair subtending 0.5 degrees, followed by a 4-s rest stage.

Behavioral performance measures

For the OPN95/SPN95 and FV14 datasets, the participants were asked to rate the similarity of the same objects presented during the fMRI task pairwise on a 7-point Likert scale (1 = least similar, 7 = most similar). Because there was no theoretically correct response, inter-subject correlation were used to establish that the participants were performed the task as expected.

For the THINGS dataset, participants were presented with object triplets and asked to select the "odd-one-out" in each triplet. Each triplet was evaluated by multiple but not all participants, resulting in a total of 5,517,400 triplet choices (for further details, see Hebart et al., 2023).

Acquisition

Imaging type(s)

Functional; Diffusion; Structural

Field strength

3 T

Sequence & imaging parameters

OPN95/SPN95. High-resolution 3D structural images were collected with a 3D magnetisation prepared-rapid gradient echo (MPRAGE) sequence in the sagittal plane (144 slices, TR=2530 ms, TE=3.39 ms, flip angle=7°, matrix size=256 × 256, and voxel size=1.33 × 1 × 1.33 mm). Functional images were acquired with an echo–planar imaging (EPI) sequence (33 axial slices, TR=2000 ms, TE=30 ms, flip angle=90°, matrix size=64 × 64, and voxel size=3 × 3 × 3.5 mm with a gap of 0.7 mm).

THINGS. High-resolution 3D structural images were collected with an MPRAGE sequence (208 sagittal slices, voxel size =

0.8 × 0.8 × 0.8 mm, TR = 2.4 s, TE = 2.24 ms, matrix size = 320 × 300, FOV = 256 × 40 mm, flip angle = 8°). The whole-brain functional MR data were collected in 2 mm isotropic resolution (60 axial slices, voxel size = 2 × 2 × 2 mm, TR = 1.5 s, TE = 33 ms, matrix size = 96×96, FOV = 192 × 192 mm, flip angle = 75°).

FV14 and Study 2. Functional images were acquired with a multiband EPI sequence (axial slices, TR=2000 ms, TE=30 ms, flip angle=90°, matrix size=72 × 72, voxel size=2.5 × 2.5 × 2.5 mm3, FOV = 180 mm × 180 mm, multiband factor = 2). HARDI images were acquired with a multiband EPI sequence (axial slices, TR = 3000 ms, TE = 100 ms, flip angle = 90°, matrix size = 112 × 112, voxel size = 2 × 2 × 2 mm^3, FOV = 224 × 224 mm^2, multiband factor = 2, diffusion gradient value b set to 0, 1000, and 2000 s/mm^2, b0 repeated 10 times, and 64 directions each for b1000 and b2000). 3D T1-weighted images were acquired with an MPRAGE sequence (sagittal slices, TR = 2530 ms, TE = 2.88 ms, flip angle = 7°, matrix size = 224 × 256, interpolated to 448 × 512, voxel size = 0.5 × 0.5 × 1 mm^3, FOV = 224 × 256 mm^2). T2-weighted images were acquired with a Sampling Perfection with Application optimized Contrast using different flip angle Evolution (SPACE) sequence (sagittal slices, TR = 3200 ms, TE = 408 ms, flip angle = 120°, matrix size = 256 × 256, voxel size = 0.9 × 0.9 × 0.9 mm^3, FOV = 230 × 230 mm^2). FLAIR T2-weighted images were acquired with a FLAIR sequence (sagittal slices, TR = 5000 ms, TE = 394 ms, flip angle = 120°, matrix size = 256 × 256, interpolated to 512 × 512, voxel size = 0.5 × 0.5 × 1 mm^3, FOV = 250 × 250 mm^2).

Area of acquisition: Whole brain scans.

Diffusion MRI ☒ Used ☐ Not used

Parameters: Diffusion gradient value b set to 0, 1000, and 2000 s/mm^2, b0 repeated 10 times, and 64 directions each for b1000 and b2000. Multi shell; No cardiac gating.

Preprocessing

Preprocessing software: The fMRI data were preprocessed using the Statistical Parametric Mapping software (SPM12; http://www.fil.ion.ucl.ac.uk/spm/) and DPABI V3.0 (Yan et al., 2016). The HARDI data were preprocessed using FSL (version 6.07) from Oxford University. The probabilistic tractography was performed with FMRIB's Diffusion Toolbox in FSL.

Normalization: For each participant, structural image was segmented using a unified segmentation module (Ashburner & Friston 2005). Next, a custom, study-specific template was generated by applying diffeomorphic anatomical registration through exponentiated lie algebra (DARTEL; Ashburner, 2007). The parameters obtained during segmentation were used to normalize the functional images of each participant

Normalization template: MNI305

Noise and artifact removal: For the preprocessing of the task fMRI data, the first five volumes of each functional run were discarded to reach signal equilibrium. Slice timing and 3-D head motion correction were performed. After normalization, the functional images were spatially smoothed using a 6-mm full-width-half-maximum Gaussian kernel for univariate analysis but not for multivariate pattern analysis.

Volume censoring: None

Statistical modeling & inference

Model type and settings: Multivariate pattern analysis (MVPA); GLM analysis was first performed to obtain results of each regressors; during GLM analysis, six head motion parameters were included as nuisance regressors, and a high-pass filter (128 s) was used to remove low-frequency signal drift for each run; then MVPA were performed; the results of MVPA were entered into second-level (between-subject) random-effect analysis.

Effect(s) tested: Representational similarity analyses; partial Spearman's correlation: 1) the partial correlation between CLIP and brain > 0, controlling the ResNet and the MoCo; 2) the partial correlation between ResNet and brain > 0, controlling the MoCo.

Specify type of analysis: ☐ Whole brain ☒ ROI-based ☐ Both

Anatomical location(s): Based on the previous fMRI studies about visual and language regions:

1. The VOTC was defined by combining functionally defined regions (activation from hearing participants viewing pictures relative to baseline in the OPN95 dataset) and anatomical parcels from the Harvard–Oxford Atlas (probability > 0.2), specifically the posterior and temporooccipital divisions of the inferior temporal gyrus (15#, 16#), the inferior division of the lateral occipital cortex (23#), the posterior division of the parahippocampal gyrus (35#), the lingual gyrus (36#), the posterior division of the temporal fusiform cortex (38#), the temporal occipital fusiform cortex (39#), the occipital fusiform gyrus (40#), the supracalcarine cortex (47#), and the occipital pole (48#). This definition resulted in the selection of 2467 voxels in the left hemisphere and 2420 voxels in the right hemisphere.

2. The left language-specific regions as defined by Fedorenko et al. (2010) and the left VOTC mask (defined in the ROI definition section). For the controls, the right-hemisphere homologues of the left-hemisphere language regions were also included.

Statistic type for inference

(See Eklund et al. 2016)

All effects in VOTC or langauge mask level were tested by one sample t-tests and cluster-wise FWE correction as implemented in SPM12.
Effects at ROI level were tested by null-hypothesis one sample ttests and Bayesian one sample t-tests, simultaneously.

Correction

For searchlight analysis, multiple comparison corrections were conducted using cluster-level FWE correction (p <.05) as implemented in SPM12 (voxel-wise p <.001).

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis

Features in MVPA were voxel-based t (beta for THINGS dataset) values of regressors.

We conducted searchlight MVPA within a ventral occipitotemporal cortex (VOTC) mask. This mask was defined on the basis of regions showing stronger activation to all pictures relative to baseline in hearing participants from the OPN95 dataset ($q < 0.05$, FDR-corrected). The functional mask was further constrained by the following anatomical parcels (Harvard–Oxford Atlas, probability > 0.2): the posterior and temporooccipital divisions of the inferior temporal gyrus (15#, 16#), the inferior division of the lateral occipital cortex (23#), the posterior division of the parahippocampal gyrus (35#), the lingual gyrus (36#), the posterior division of the temporal fusiform cortex (38#), the temporal occipital fusiform cortex (39#), the occipital fusiform gyrus (40#), the supracalcarine cortex (47#), and the occipital pole (48#). This definition resulted in the selection of 2467 voxels in the left hemisphere and 2420 voxels in the right hemisphere.

For each voxel within the VOTC mask, multivariate activation patterns within a sphere (radius = 10 mm) centred at that voxel were extracted. Neural RDMs were computed with the Pearson distance within the searchlight sphere. Then, Spearman's rank correlation coefficients between the neural RDM and model-derived visual RDMs was computed, controlling for the effects of models with lower levels of language involvement to determine the sentence description effect (correlations with the visual RDMs derived from CLIPvision while controlling for the visual RDMs derived from ResNet and MoCo) and the verbal categorization effect (correlations with the visual RDMs derived from ResNet while controlling for the visual RDMs derived from MoCo). Correlation maps were obtained for each participant by moving the searchlight centre across the VOTC mask. These maps were Fisher z-transformed and spatially smoothed with a 6 mm (OPN95 and SPN95) or 4 mm (FV14, THINGS and Study 2) full-width half-maximum (FWHM) Gaussian kernel. The correlation maps were compared to 0 with one-tailed one-sample t tests.

ROI-based MVPA was conducted with the same VOTC mask as ROI. Specifically, multivariate activity patterns for each stimulus within the ROI mask were extracted. Neural RDMs were computed on the basis of Pearson distances and then correlated with the RDM generated by CLIPvision while controlling for the RDMs generated by ResNet and MoCo. The resulting correlation coefficients between the neural and model RDMs were Fisher z-transformed and compared to zero with one-tailed one-sample t tests.