scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A large-scale fMRI dataset for vision-language semantic association

Shurui Li[1,2,3,11], Zheyu Jin[1,2,11], Shi Gu[4,5] ✉, Ru-Yuan Zhang[6,7,8,9] ✉ & Yuanning Li[1,2,3,10] ✉

Understanding the neural coding and association of visual and language information benefits from the development of deep learning models and the collection of massive datasets with extensive sampling of brain activity. Large-scale functional magnetic resonance imaging (fMRI) datasets with naturalistic stimuli provide more ecologically relevant experimental conditions and promote more reproducible research into the neural basis of sensory perception. Here, unlike most previous datasets restricted to isolated modalities, we present the Caption Scene Dataset (CSD), a large-scale fMRI dataset for vision-language semantic association, in which neural responses to 4,400 pairs of Chinese captions and naturalistic scenes were acquired from eight healthy participants. The participants were instructed to determine whether the semantics in the caption and the image are consistent. To illustrate the utility of the CSD dataset, we demonstrated that deep neural encoding models effectively predicted neural responses to both caption and image stimuli across different cortical regions. This dataset provides a platform for the investigation of the neural basis of semantic association across vision and language, facilitating cross-disciplinary advances between vision neuroscience and artificial intelligence.

Background & Summary

Understanding how the brain transforms diverse sensory inputs, including visual and linguistic signals, into unified and abstract semantic representations is essential for explaining how humans derive meaning from the world[1–3]. Vision and language are the two dominant channels through which we perceive, interpret, and communicate, yet the neural computations mediating the semantic associations between these modalities remain poorly understood[4–8]. In artificial intelligence, these same modalities underlie modern multimodal systems, and elucidating their association in the human brain offers critical insight into building models that more closely approximate human-like understanding[9–12].

With the rapid development in computer vision and AI, it has become feasible to utilize novel AI models for the study of neural encoding and decoding models[12–19]. AI models provide powerful computational analogs for studying cognitive processes, such as vision and language. For example, deep convolutional networks trained on natural images develop representational hierarchies that closely mirror the organization of the human visual cortex[20–25]. By using encoding models to map features from these networks onto neural responses, we can reveal how cortical areas progressively transform sensory input into increasingly abstract representations, offering insights into the computational principles shared by biological and artificial systems[12].

In order to promote the integration of cognitive neuroscience and artificial intelligence, there exists a growing need for large-scale brain data acquired with naturalistic stimuli. Most recently, large-scale neuroscience

[1]School of Biomedical Engineering, ShanghaiTech University, Shanghai, China. [2]State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China. [3]Shanghai Artificial Intelligence Laboratory, Shanghai, China, [4]College of Computer Science and Technology, Zhejiang University, Hangzhou, China. [5]State Key Laboratory of Brain Machine Intelligence, Zhejiang University, Hangzhou, China. [6]School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China. [7]IDG/McGovern Institute for Brain Research, Peking University, Beijing, China. [8]Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China. [9]Brain Health Institute, National Center for Mental Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine and School of Psychology, Shanghai, China. [10]Shanghai Clinical Research and Trial Center, Shanghai, China. . [11]These authors contributed equally: Shurui Li, Zheyu Jin. ✉e-mail: gus@zju.edu.cn; ruyuanzhang@gmail.com; liyn2@shanghaitech.edu.cn

datasets[26–30] using naturalistic stimuli have been collected with attempts to investigate complex neural activity and integrate neuroscience and deep learning models. Naturalistic stimuli are essential for understanding how the brain operates in real-world contexts[10,31]. Simplified or artificial stimuli reveal isolated mechanisms, but they miss the rich, dynamic structure of natural sensory input that drives cognition in everyday life. Using naturalistic visual and linguistic stimuli allows us to probe how the brain integrates information across space, time, and modality, revealing principles of prediction, abstraction, and semantic integration that cannot be captured by reductionist paradigms.

Functional magnetic resonance imaging (fMRI) is a safe and non-invasive neuroimaging technique that can be used to reveal functional responses of the human brain under various behavioral tasks[4,32–35]. In neuroscience, fMRI is often used to help understand various brain functions such as the processing of visual[4,27,36,37] and auditory information[10,17,31,33,38], working memory[35,39–43] and long-term memory[43–45]. Several massive neuroscience datasets primarily based on fMRI have been made public over recent years, including the Natural Scenes Dataset (NSD)[26], the THINGS-data[28], the Natural Object Dataset (NOD)[29] and the Brain, Object, Landscape Dataset (BOLD5000)[27]. These datasets were mainly designed for the study of the neural basis of human visual processing under naturalistic scenes. In each dataset, a large amount of fMRI data was collected while the subjects were viewing naturalistic images and performing certain cognitive tasks.

Although large-scale fMRI datasets have greatly advanced our understanding of cortical organization, they have been almost entirely unimodal, centered on either visual stimuli (images, objects, movies)[26,29,46] or linguistic materials (text, speech)[38,47]. While such datasets are invaluable for characterizing modality-specific pathways, they offer limited leverage for addressing a central question in cognitive neuroscience: how the human brain establishes semantic associations between visual and linguistic representations. To uncover the neural computations that support this process, it is essential to collect datasets featuring both modalities, allowing researchers to measure cross-modal semantic associations and investigate how linguistic information provides a context for visual interpretation.

To address these challenges, here we introduce the Caption Scene Dataset (CSD), a large-scale, rapid event-related human fMRI dataset, which contained fMRI responses from eight healthy subjects, each viewed more than 4,400 pairs of Chinese captions and naturalistic scenes, resulting in a total of 210 hours of functional scanning. The text-image pairs used in the CSD dataset were selected from COCO-CN[48], a cross-lingual learning dataset that extends the MS-COCO[49] dataset with manually written sentences and tags in Chinese. By selecting natural images from one of the most commonly used computer vision datasets and the corresponding text captions, we hope to facilitate the investigation of the neural mechanisms of semantic representations across different modalities.

It is noteworthy that a concurrent study[50] similarly collected fMRI responses to both images and text within a single dataset. However, our dataset offers several distinct experimental advantages. While their paradigm utilized a 1-back task with randomized, single-modality presentations, our design features a paired-stimulus structure where a caption is followed by its corresponding image. This sequential pairing specifically emphasizes the semantic associations between modalities, making the CSD uniquely suited for investigating language-guided visual perception. Furthermore, whereas their stimuli were presented once, our stimuli were presented twice to ensure higher signal stability and reliability.

In this paper, we provide a detailed description of the design of our experiment paradigms, the preparation of stimulus materials, the recruitment of participants, the acquisition and the processing of MRI data, and basic quality control results. We demonstrate neural encoding models based on deep neural networks to validate the feasibility of using the CSD dataset for further study in vision neuroscience and multi-modal semantic processing.

Beyond basic validation, we believe that the Caption Scene Dataset (CSD) offers several advantages for both cognitive neuroscience and artificial intelligence. First, it enables investigation of cross-modal modulation, particularly how linguistic priors shape visual representations and top-down semantic influences in the visual cortex. Second, it provides a valuable bridge between multimodal AI and human neurobiology, allowing researchers to evaluate the biological plausibility of vision–language models and to develop brain-inspired multimodal architectures. Third, its high-quality repeated-trial design makes it a useful benchmark for multimodal neural decoding and reconstruction, serving as an important complement to existing large-scale datasets such as NSD for studying complex semantic content in the brain.

Methods

Participants. All eight participants (four females), aged between 19 and 25 years at the start of the study (mean $\pm$ standard deviation: 22.62 ± 1.58 years), were recruited from the student population at ShanghaiTech University. All participants were healthy, right-handed, and had normal or corrected-to-normal vision. None reported a history of neurological or psychiatric disorders. The study was approved by the Institutional Review Board of ShanghaiTech University (IRB protocol number: Q2023-007). Data collection spanned approximately one year, from May 2023 to July 2024. Participants were informed about the long-term nature of the study prior to enrollment and were naive to the specific experimental hypotheses. Written informed consent was obtained from all participants in accordance with institutional guidelines. Participants were financially compensated for each completed session, and were informed of their right to withdraw from the study at any time without penalty. They were also made aware that their data would be used for scientific research and, following anonymization, would be shared publicly.

The standard operation procedure (SOP) for the CSD experiment was formalized before fMRI data collection. The participants were scheduled to take part in the experiments once or twice every week and the scan time for each participant was kept as consistent as possible across sessions. One day before the scanning procedure, the participants were reminded to avoid staying up late and ensure adequate sleep, in order to optimize

cognitive performance in the experiments. The participants were also instructed not to intake alcohol, caffeine or any other stimulant substances, within 6 hours before the start of fMRI scanning. If the participant felt unwell or experienced any symptoms such as coughing or fever on the day of the experiment, the session would be rescheduled.

Stimuli preparation. The image-text pairs used in the CSD dataset were selected from COCO-CN[48], a cross-lingual learning dataset that extends the MS-COCO[49] (Microsoft Common Objects in Context) dataset with manually written Chinese sentences and tags. The COCO-CN dataset has 20,342 images annotated with 27,218 Chinese sentences and 70,993 tags. For the convenience of stimulus presentation and further analysis using neural network models, based on the availability of images provided in COCO-CN, the images were prepared in RGB format with dimensionality of 480 pixels × 480 pixels. This image size was chosen because the short sides of a large portion of the images from the original dataset were set to 480 pixels.

The screening criteria for our stimuli preparation are summarized as follows. Firstly, only square images or rectangular images whose shorter sides were at least 424 pixels in length were retained for further processing. In order to convert the stimuli to a square aspect ratio, we utilized binary masks from the COCO API to guide the cropping process. These masks indicated the locations of salient objects within each image and served as a proxy for semantic content. In the cropping process, we applied a sliding window along the long side of each image. The window size was dynamically determined based on the length of the short side, ensuring that the resulting crop preserved the original image's semantic content as much as possible. For each image, we selected the window position that minimized the exclusion of masked (i.e., semantically meaningful) regions. Secondly, all squared images were then rescaled to 480 pixels x 480 pixels, either by image downsampling or upsampling, if their side lengths were not already 480 pixels. The average mask exclusion rate, quantified as the proportion of mask area removed during cropping and rescaling, was approximately 1.2% across the entire set of 14,084 images. Thirdly, only those image-text pairs that have Chinese captions with no more than 20 characters were selected, to ensure that the participants would be able to fully comprehend the sentence in 3 seconds, without too much eye movement. Finally, we also conducted manual inspection to make sure that there was no violent or inappropriate content in the stimulus materials and no significant semantic loss due to image cropping. We also manually checked if there is any image blurring or ambiguity in image-caption consistency. After the screening process, a total of 10,448 image-caption pairs were kept for further usage (Fig. 1c). In COCO-CN sessions, which is the core sessions in CSD, nine thousand pairs of matched Chinese captions and natural images were randomly sampled as the stimulus material for our CSD dataset.

In COCO-CN sessions, a total of 9,117 unique natural images were utilized as stimulus material. To facilitate both group-level reliability and broad semantic coverage, the dataset follows a hybrid distribution design: 1,000 core images were shared across all eight subjects, providing a basis for inter-subject consistency analysis. For the remaining stimuli, each subject was assigned 1,000 unique images. It should be noted that due to the inclusion of mismatched trials (where images were randomly sampled for the matching task), some images originally unique to one subject may have appeared as lures in the trials of others. This overlap results in the final distribution of 9,117 unique images across the entire dataset.

In addition to COCO-CN sessions where stimuli were selected from the COCO-CN dataset, retinotopic and functional localizer experiments were conducted to help define low-level visual areas and category-selective cortical regions in individual subjects. The stimuli for the retinotopic experiment came from Human Connectome Project (HCP) 7T Retinotopy Dataset[51] and consisted of slowly moving apertures filled with a dynamic colorful texture. There were two types of run design, referred to as 'multibar' and 'wedgering'. The former included bars sweeping in multiple directions, whereas the latter involved a combination of rotating wedges, and expanding and contracting rings. The stimuli for the functional localizer experiment came from the Grill-Spector laboratory[52] and consisted of different categories of grayscale images, including faces, body parts, scenes, words and objects.

Experimental design. The CSD experiment consists of three types of fMRI scanning sessions, including COCO-CN, population receptive field (pRF) mapping, and functional localizer (fLoc). All eight participants took part in one pRF session, one fLoc session, one resting-state fMRI session, and at least twenty-five COCO-CN sessions. Some of the eight participants completed the task with more scanning sessions due to uncomfortableness or reduced performance in the middle of scanning. However, the total number of runs each participant went through was consistent, which was 200 for COCO-CN sessions. A typical fMRI session consisted of 8 runs, each run lasting 5 minutes 22 seconds. One minute of rest was allowed between runs. In total, 210 hours of fMRI data were collected from all eight participants combined.

In terms of experiment scheme, a delayed-matching paradigm was employed, in which participants viewed a short text description followed by an image and were instructed to judge whether the two were semantically matched or not (Fig. 1a). Specifically, for each pair, the caption was presented first, for 3 seconds, and there was a 3-second blank before the presentation of the corresponding image. Each trail was also separated by a 3-second blank, so a complete trial took 12 seconds. The 3-second presentation durations and intervals were selected to allow for accurate estimation of event-specific beta weights using a general linear model. While previous studies such as NSD employed a rapid 3s ON / 1s OFF design under 7T acquisition, we used a 3T MRI scanner and thus extended the inter-stimulus intervals to improve the separation of hemodynamic responses elicited by individual events. The participants were also instructed to press the button in response after the image was presented to them, if they found that the caption and the image were not matched. The results of eye-tracking and the behavior task were constantly monitored during fMRI scanning to help improve the overall data quality.

Captions were consistently presented before images to maintain homogeneity in neural response patterns across the dataset. Varying the order within the same experiment would likely introduce distinct cognitive processes, as caption-first trials are expected to engage predictive semantic imagery, whereas image-first trials may

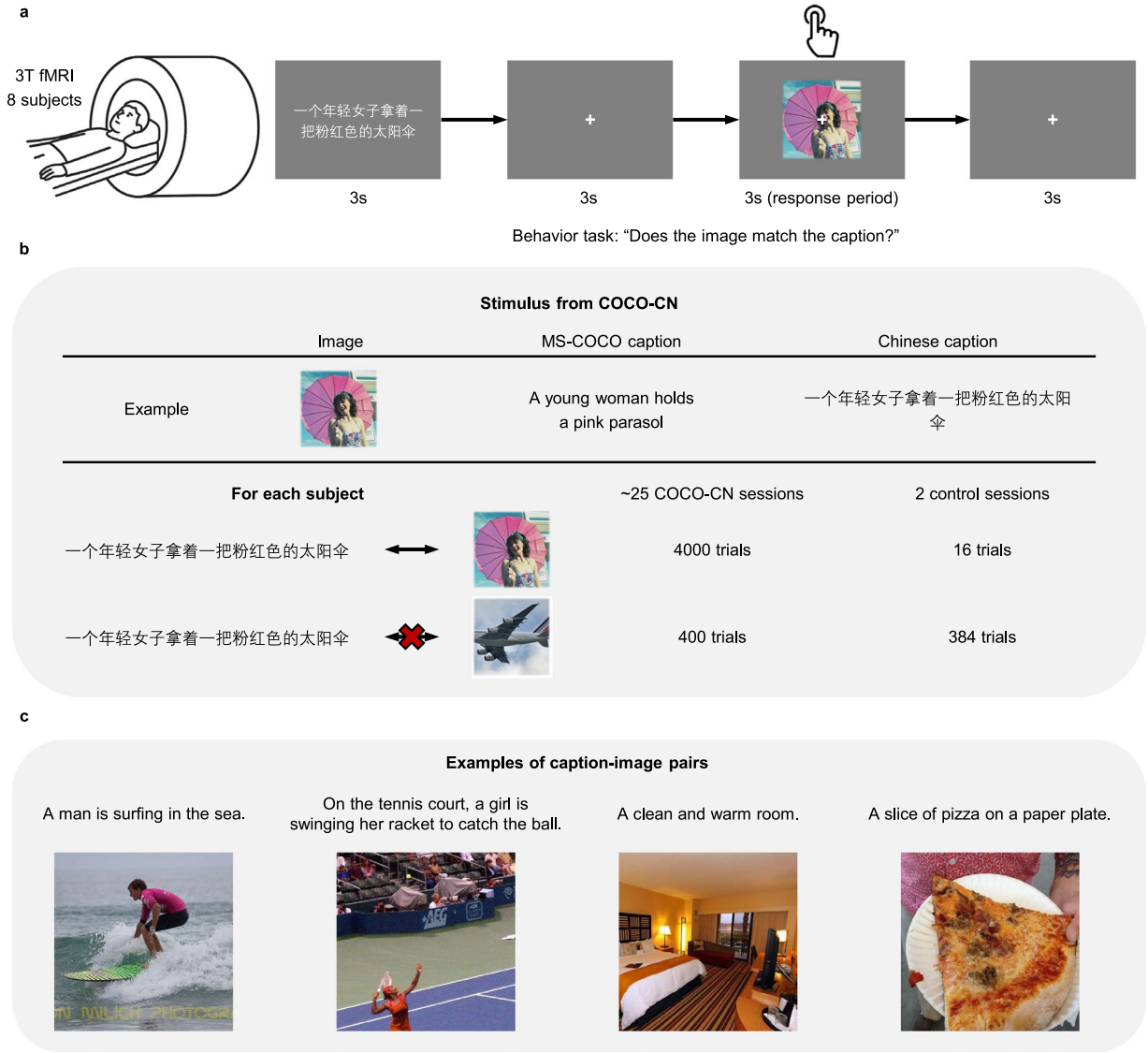


Fig. 1 Experimental paradigm of the COCO-CN delayed matching task. (**a**) The delayed-matching paradigm used in the experiment. fMRI data were collected from 8 participants using a 3 T scanner. Each trial consisted of a 3-second caption display, a 3-second fixation, a 3-second image presentation, and a final 3-second fixation. After the image appeared, participants were asked to judge whether the caption and image matched. (**b**) Caption-scene pairs used in the experiment. The image is taken from MS-COCO, shown with its original English caption and the corresponding Chinese translation. Below: each subject completed approximately 25 COCO-CN sessions (containing ~4000 matched trials and 400 mismatched trials) and 2 control sessions (384 mismatched trials and 16 matched trials). (**c**) Several examples of caption-image pairs.

instead evoke memory-based retrieval or verification. By adopting a fixed caption-then-image structure, we ensured that participants could fully read and integrate the linguistic information before viewing the image, allowing the visual stimulus to be processed within a well-defined semantic context. This design is particularly advantageous for studying high-level semantic representations and their influence on subsequent visual processing.

The following details for trial and run design apply for every session and every participant. At the beginning and the end of each run, a central red dot was shown on a blank screen for 8 seconds and 4 seconds, respectively. In each run, there were 20 trials (12 seconds each) where the Chinese captions and the natural images are matched (Fig. 1a). Two unmatched trials (12 seconds each) and four blank trials (6 seconds each) were randomly positioned in each run. This made the overall ratio between matched and unmatched trials 10:1. Unlike a balanced 1:1 design, which may encourage independent processing of modalities, our 10:1 ratio establishes a strong linguistic prior. This ensures that captions actively guide visual perception, shifting the participant's state from simple category matching to deep cross-modal semantic association. The stimuli for the unmatched trials were selected from a larger sample that contained all the caption-image pairs used in our COCO-CN sessions. The distribution of unmatched trials and blank trials were generated according to the following rules: 1) The first trial and the last trial of each run were matched trials. 2) There were no consecutive unmatched trials and

Session name	Number of runs	Matched trials	Unmatched trials
COCO-CN	200	200 × 20	200 × 2
Control	16	16 × 2	16 × 23

Table 1. Summary of experiment session design.

no consecutive blank trials. However, a blank trial could appear before or after an unmatched trial. 3) The blank trials were randomly inserted after the positions of the unmatched trials were determined. As a result, each run lasted 300 seconds and each session contained eight runs in our routine experiments.

In addition, the central dot was displayed throughout scanning, except during the 3-second presentation of Chinese captions. The participants were instructed to maintain central fixation when they were watching the images and when the screen was blank. All participants were naïve to the precise details of the CSD experiment, including the number of repetitions for each caption-image pair, as well as the total number and distribution of unmatched trials and blank trials.

The following details for session design apply for every participant. Before the COCO-CN sessions, three additional sessions were arranged to collect structural images and functional localizers. The first session was a fLoc session, consisting of four runs of structural T1-weighted imaging and four runs of category-selective localizer, conducted in an alternating manner. Each run of T1-weighted imaging was followed by a 5-minute run of the visual fLoc experiment, where a novel stimulus sequence that counterbalanced the ordering of five different stimulus domains (characters, bodies, faces, places, and objects) was generated and the participants were instructed to perform a 1-back behavioral task. The second session was a pRF session, consisting of two runs of structural T2-weighted imaging, two runs of diffusion MRI and four runs of the retinotopic mapping. In the pRF experiment, slowly moving apertures filled with a dynamic colorful texture were displayed with a central fixation dot and the participants were instructed to press the button when the color of the dot changed. The third session was a resting-state fMRI session, four runs of COCO-CN experiment were conducted before the acquisition of resting-state data. The remaining twenty-five sessions were COCO-CN sessions, each consisting of eight task-fMRI runs with an optional one-minute rest period between runs. If uncomfortableness was reported by the participant or reduced behavioral performance was observed, fewer runs (normally 6 out of 8 scheduled runs) were acquired in those sessions. However, the number of sessions would increase, if necessary, to ensure that all participants completed a total of 200 runs in the COCO-CN experiment (Table 1). In all, for each participant, 2,000 caption-image pairs were repeated twice, resulting in 4,000 matched trails. Moreover, the second repetition of the 2,000 caption-image pairs for each participant was presented after the entire first repetition was completed. Two distinct sequences of the 2,000 pairs were randomly generated and there was no overlap. A total of 400 unmatched trials were randomly inserted. The participants were blinded to the design of stimulus presentation and the number of repetitions.

In addition to the normal COCO-CN sessions, each participant also completed two additional control sessions, each consisting of 8 runs. The control sessions serve as a functional baseline, by comparing the noise ceilings of the control sessions with those of the COCO-CN sessions, we can quantitatively assess how the matching paradigm and its associated cognitive dependencies influence the reliability of trial-wise neural responses. The control sessions employed the same delayed-matching paradigm, but with a key modification in trial composition: the ratio of matched to unmatched trials was reversed to be 1:10. Specifically, each run contained 23 unmatched trials and only 2 matched trials, and no blank trials were included. The stimuli used in the control sessions were drawn from image-caption pairs that had never appeared in any previous sessions. From the full pool of 14,084 candidate pairs, we excluded the 9,000 used as matched stimuli and the 954 captions that had been used in unmatched trials, leaving a total of 494 unique, unused pairs. From these, 216 matched pairs were selected to serve as stimuli for the control sessions. Each participant viewed the same 32 matched pairs, randomly distributed across the 16 runs (Fig. 1b). The unmatched trials were constructed by randomly shuffling the captions and images from the remaining stimuli, independently for each participant and run.

The decision to use a 10:1 ratio in COCO-CN sessions and a 2:23 ratio in control sessions, was driven by two primary considerations: First, a 1:1 ratio would likely lead participants to treat captions and images as independent entities, as the predictive value of the caption would be minimal. By contrast, a 10:1 ratio establishes a strong prior expectation, encouraging participants to actively integrate the caption's semantic information into their visual processing. This design effectively shifts the task from simple "category matching" to a more profound "cross-modal semantic association" state. Second, the 2:23 ratio in the control sessions was strategically chosen to yield a comparable number of unmatched trials to the unmatched trials in COCO-CN sessions.

MRI data acquisition. Magnetic Resonance Imaging (MRI) data were collected on a 3 Tesla United Imaging uMR890 scanner at the main campus of ShanghaiTech University, Shanghai, China, using a 64-channel head coil. The MRI experiments consisted of both anatomical (T1, T2 and diffusion) and functional (resting-state and task-based) imaging. Structural images and resting-state fMRI data were acquired prior to the COCO-CN experiments within three sessions. Retinotopic experiment was conducted to localize the early visual cortex (V1-V3) of each participant. Functional localizer was used to define category-selective cortical regions, such as fusiform face area (FFA) and visual word form area (VWFA). In addition, eye-tracking data were collected for most part of our task-fMRI experiment to help monitor whether the participants were able to maintain central fixation during the scan. Physiological data were acquired for some part of the experiment.

During all MRI scanning, ear plugs were provided to help reduce the acoustic noise. In fMRI experiments, participants viewed the stimuli presented on the monitor through a mirror mounted on the RF coil and the

participants who required vision correction were provided with MRI-compatible glasses. The monitor was positioned at the head of the scanner bed and placed flush against the scanner bore. Visual stimuli were presented on a screen measuring 89.0 cm (width) × 50.5 cm (height) with a resolution of 1920 × 1080 pixels and a refresh rate of 60 Hz. The approximate distance from the mirror to the screen was 162 cm and the viewing distance from the mirror to the participants' eyes was 12 cm. The stimuli were adjusted to fill 8.2° of visual angle, in order to ensure that the captions and images were presented within a comfortable field of view for the participants. In addition, the text stimuli were displayed at a fixed size of 24 points to improve readability given the viewing distance and visual angle. All stimuli were presented using MATLAB with Psychtoolbox (version 3.0.18)[53]. The stimulus computer was configured with GStreamer version 1.18.6 to support Psychtoolbox operations.

Whole-brain functional MRI data were acquired using a T2*-weighted gradient-echo, multi-band accelerated echo-planar imaging (EPI) sequence provided by Shanghai United Imaging Healthcare. The imaging parameters for the task-based functional MRI are listed as follows: number of axial slices = 60, in-plane resolution = 2.5 × 2.5 mm, matrix size = 80 × 80, field-of-view (FOV) = 200 × 200 mm, slice thickness = 2.5 mm, no slice gap, repetition time (TR) = 2,000 ms, echo time (TE) = 30 ms, flip angle = 81 degrees, echo spacing = 0.52 ms, bandwidth = 2,320 Hz per pixel, multiband factor = 2, partial phase encoding = 90%. The phase encoding direction involves alternating between anterior-to-posterior (AP) and posterior-to-anterior (PA) for different runs (one AP and one PA, repeated four times for each session).

Resting-state functional MRI data were collected using a gradient-echo EPI sequence with the following parameters: number of axial slices = 72, in-plane resolution = 1.8 × 1.8 mm, matrix size = 116 × 116, FOV = 209 × 209 mm, slice thickness = 1.8 mm, no slice gap, TR = 800 ms, TE = 35 ms, flip angle = 52 degrees, echo spacing = 0.61 ms, bandwidth = 1,960 Hz per pixel, multiband factor = 8,partial phase encoding = 90%. Two separate runs of resting-state fMRI were conducted with different phase encoding directions, one AP and one PA.

Structural T1-weighted MRI images were acquired using a GRE-FSP sequence on the United Imaging scanner with 1 mm isotropic resolution (slabs number = 1, number of sagittal slices = 240, FOV = 256 × 256 mm, TR = 8.3 ms, TE = 2.2 ms, inversion time (TI) = 1,060 ms, flip angle = 8 degrees, bandwidth = 260 Hz per pixel, phase encoding anterior-to-posterior, partial phase encoding = 90%, ACS acceleration level = 2, TA = 3.73 min per scan). T2-weighted MRI images were acquired using an FSE-MX sequence with 1 mm isotropic resolution (slabs number = 1, number of sagittal slices = 240, FOV = 256 × 256 mm, TR = 3,000 ms, TE = 452.4 ms, echo spacing = 3.9 ms, bandwidth = 550 Hz per pixel, partial phase encoding = 90%, ACS acceleration level = 2, TA = 3.95 min per scan). Four T1-weighted runs and two T2-weighted runs were conducted. Two scans of diffusion MRI (dMRI) were obtained for each participant using protocols from Chinese Baby Connectome Project (CBCP). Imaging parameters included number of transversal slices = 92, in-plane resolution = 1.5 × 1.5 mm, matrix size = 140 × 140, FOV = 210 × 210 mm, slice thickness = 1.5 mm, no slice gap, TR = 3,016 ms, TE = 77.1 ms, echo spacing = 0.68 ms, bandwidth = 1,750 Hz per pixel, multiband factor = 4, TA = 4.22 min per scan. Diffusion-weighted images were acquired with 72 gradient directions, distributed across three b-values: 500, 1,000, and 3,000 s/mm^2. To facilitate distortion correction, two separate runs were collected with opposite phase encoding directions (AP and PA). Additionally, B0 map was acquired under the same imaging parameters with AP phase encoding, providing reference scans for preprocessing and registration.

No individual trials or sessions were excluded post-acquisition. Instead, data quality was controlled in real time during scanning using the native quality control (QC) interface of the uMR 890 system, which continuously monitored head motion and task engagement. Sessions showing excessive motion or sustained behavioral noncompliance were terminated and discarded immediately. This acquisition-stage quality control ensured that all retained data met high standards of physiological stability and behavioral performance, eliminating the need for post-hoc trial exclusion.

Eye-tracking data acquisition. Monocular eye movements were recorded at a sampling rate of 1000 Hz using an MRI-compatible EyeLink 1000 system (SR Research). The eye-tracker was positioned at the side of the scanner bore to capture the participant's eye position via a mirror mounted on the head coil. Due to technical constraints and scheduling during the early stages of data collection, eye-tracking data were not acquired for subject 1 to subject 3. For the remaining five subjects, the amount of valid recorded data varied due to intermittent calibration stability in the MRI environment, with the most extensive data available for subject 7.

Data preprocessing pipeline. Some basic preprocessing pipelines were adopted to the MRI data we collected, including slice-timing correction, distortion correction, motion correction for the functional data and cortical surface reconstruction based on T1-weighted images.

Each participant underwent four T1-weighted and two T2-weighted structural MRI scans during the study. To improve anatomical image quality, all T1 and T2 volumes were first aligned using FSL's FLIRT[54] tool and then averaged to enhance the signal-to-noise ratio. The averaged T1-weighted image was then used as the input for FreeSurfer's recon-all pipeline[55] to reconstruct individual cortical surfaces.

For the functional data, we applied a series of minimal preprocessing steps. Slice timing correction was first performed using FSL. To correct for susceptibility-induced distortions, we used the topup tool in FSL with pairs of runs acquired in opposite phase-encoding directions (AP and PA). Motion correction was then performed by rigidly aligning each run to the distortion-corrected mean image of the first fMRI run for each participant. The resulting images were further aligned to the averaged and corrected T1-weighted structural image using boundary-based registration (BBR). All transformations, including distortion correction, motion correction, and registration, were combined and applied in a single resampling step to minimize interpolation artifacts. After preprocessing in volume space, functional data were projected to the cortical surface using FreeSurfer's mri_vol2surf,

mapping the data from volume to surface space based on each participant's reconstructed cortical anatomy. At each preprocessing stage, we conducted manual quality control to ensure the accuracy and integrity of data processing.

Then, referring to the algorithms proposed in the NSD paper[26], we adopted GLMsingle[56] to obtain trial-wise beta estimation for both caption and image responses. Because our experimental design did not include repeated stimuli within a single session, we followed the strategy used in the original GLMsingle implementation for the BOLD5000 dataset[27]. Specifically, we concatenated data across multiple sessions to ensure that a subset of stimuli appeared more than once, enabling the model to estimate voxel-level reliability through repeated items. To account for session-wise intensity variability (e.g., due to scanner drift or baseline shifts), we first computed the mean fMRI volume within each session. Each session's data were then linearly rescaled using a two-parameter fit (a, b) to minimize differences in global signal intensity. After normalization, the time series from all sessions were concatenated and used as input to GLMsingle. This procedure yielded voxel-wise beta estimates for each individual caption and image event.

Eye-tracking data were preprocessed through a series of steps to ensure signal quality and alignment with visual stimuli. First, blink periods were removed, along with 100 ms segments before and after each blink, to eliminate unreliable data caused by blink-induced gaze drift. The raw gaze coordinates were then converted to degrees of visual angle, and each participant's empirical fixation center was estimated and redefined as the screen center, correcting for residual calibration errors. At each time point, we calculated the distance between the gaze position and the estimated center; data points with distances exceeding 6° of visual angle were excluded as noise. This threshold was chosen based on the stimulus layout, where the image subtended 8.2° in width ($\pm$4.1° from center), and captions were constrained to be within this range. Gaze data were then downsampled to 100 Hz, linearly detrended to remove slow drift, and re-centered such that the fixation center was aligned to the origin. Finally, we applied temporal smoothing using a sliding window of 50 ms to reduce high-frequency noise in the signal. To ensure data quality, only data with a valid tracking duration exceeding one-third of the total run time are considered valid and included in the dataset.

Region of interest definitions. To define regions of interest (ROIs), we used multiple localizer experiments in addition to task-based responses. Retinotopic mapping data from a population receptive field (pRF) experiment were used to delineate early visual areas V1, V2, and V3. Category-selective regions in higher-level visual cortex—including face-, body-, word-, and place-selective areas—were localized using the fLoc functional localizer.

In addition to these standard ROIs, we also defined task-driven ROIs based on data from the COCO-CN experiment. Specifically, we identified voxels whose responses could be reliably predicted by the estimated beta weights from the GLM. Voxels with cross-validated R^2 values greater than 1% were included in this ROI, indicating that a nontrivial portion of their response variance was explained by the model. This set of voxels thus represents regions with stable and interpretable responses to the caption and image stimuli.

Data Records

The whole dataset can be accessed from ScienceDB[57]. The dataset includes both raw and processed data from eight participants. Raw data for each participant are stored in folders named "*sub-AA*", where AA denotes the participant ID. Each "*sub-AA*" folder contains three subdirectories: "*anat/*": includes four T1-weighted scans and two T2-weighted scans. "*func/*": contains functional scans from all task runs, including 4 runs of the fLoc localizer, 2 runs of retinotopic multibar mapping, 2 runs of retinotopic wedgering, and 200 runs from the main CSD experiment. "*dwi/*": contains diffusion-weighted images acquired from two runs with opposite phase encoding directions. Functional scans in the raw "*func/*" folder follow the naming convention: *sub-AA_ses-BB_task-CC_run-DDD_AP/PA_bold.nii.gz*, where BB is the session index, CC specifies the task (e.g., CSD, fLoc), DDD is the run number, and AP/PA indicates the phase encoding direction.

In parallel with the raw data folders, the top-level "*derivatives/*" directory contains processed data organized into the following subfolders: "*freesurfer_results/*": manually-corrected FreeSurfer reconstructions based on averaged T1-weighted images. "*pp_betas_vol/*": volume-based single-trial beta estimates derived from GLMsingle. "*pp_betas/*": surface-based single-trial beta estimates derived from GLMsingle. "*pp_data/*": minimally preprocessed fMRI volume data including slice timing correction, distortion correction, motion correction, and alignment to individual anatomical scans. "*pp_eyetrack/*": preprocessed eye-tracking data aligned with fMRI runs. "*pycortex_db/*": subject-specific flatten map for pycortex database. Within each of these "*derivatives/*" subfolders, data are organized by participant (sub-AA) to match the raw data structure. Files in "*pp_data/*" follow the naming convention: *sub-AA_ses-BB_task-CC_run-DDD.nii.gz*, with distortion correction applied and no phase encoding direction label, reflecting harmonized preprocessing. Files in "*pp_eyetrack/*" are named as: *sub-AA_ses-BB_task-CC_run-DDD.edf*, corresponding to individual runs for which valid eye-tracking data were recorded. Some runs may be missing due to no eye-tracking acquisition. This standardized and hierarchical organization ensures consistency between raw and processed data, and facilitates flexible subject-level and run-level analysis across multiple modalities.

Technical Validation

To evaluate the reliability and overall quality of the dataset, we performed a series of validation analyses using multiple complementary measures. These included assessments of raw data quality (temporal signal-to-noise ratio and head motion), stimulus-evoked activation patterns (on–off R^2), reliability of GLM-based signal estimation (noise ceiling) and task engagement (behavioral performance). In addition, we replicated findings from prior encoding model studies to verify the utility of the dataset for model-based analyses. Together, these validations confirm the robustness and usability of the dataset for cognitive and computational neuroscience research.

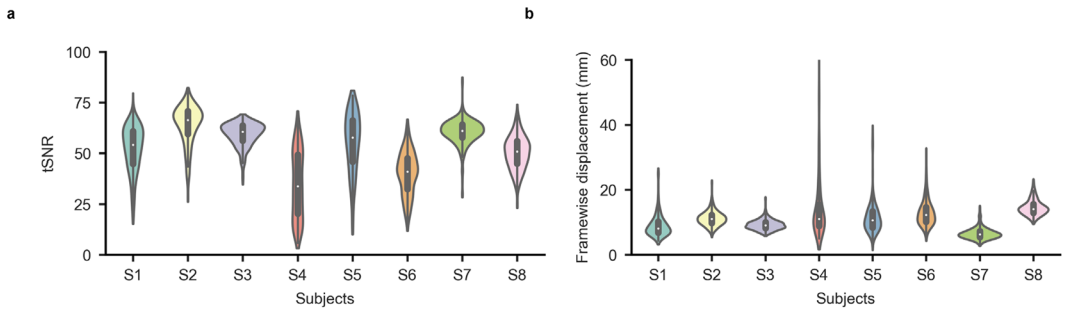


Fig. 2 Quality control of raw fMRI data. (**a**) Distribution of temporal signal-to-noise ratio (tSNR) across 200 runs for each of the 8 participants. Error bars represent the 25th to 75th percentile range. (**b**) Distribution of framewise displacement (FD) across the same 200 runs for each participant.

Temporal signal-noise ratio (tSNR). tSNR is a commonly used metric for assessing fMRI signal quality[58]. In this study, we computed tSNR using the raw, unprocessed fMRI volume data. Prior to computation, linear detrending was applied to remove potential signal drifts. For each voxel, tSNR was defined as the mean signal intensity divided by the standard deviation across time within a run. This computation was performed separately for each run in the CSD experiment, yielding a total of 200 tSNR values per participant. As shown in Fig. 2a, participants S4 and S5 exhibited relatively high variability in tSNR across runs, with standard deviations of 16.43 and 14.78, respectively. Additionally, S4 and S6 showed lower average tSNR values compared to other participants (S4: 34.67; S6: 40.10), indicating potential differences in raw signal stability.

Head motion. Head motion was quantified using framewise displacement (FD)[59], a standard metric that accounts for both translational (x, y, z) and rotational (pitch, yaw, roll) movements between consecutive fMRI volumes. For each run, we computed the mean FD value, resulting in 200 measurements per participant for the CSD experiment. As shown in Fig. 2b, participants S4, S5, and S6 exhibited greater variability in head motion across runs, with S4 showing particularly unstable motion patterns. Although participant S8 demonstrated relatively higher overall head motion, the distribution was tightly clustered, indicating consistent movement characteristics. In contrast, the remaining participants showed low and tightly distributed FD values, suggesting excellent motion stability throughout the scanning sessions. These results are consistent with the tSNR findings, further confirming the overall high quality of the raw fMRI data for most participants.

Stimulus-evoked activation. To evaluate whether the experimental stimuli reliably evoked neural responses, we computed ON–OFF R^2, a metric that quantifies the proportion of variance in voxel-wise activity explained by a block-level task regressor. This measure reflects the stability and strength of stimulus-evoked activation across runs. The ON–OFF R^2 was computed on minimally preprocessed fMRI data, including slice timing correction, distortion correction, head motion correction, and registration to each participant's anatomical image. The data were then projected onto the cortical surface for analysis.

In the computation, all caption trials within a run were treated as one condition, and all image trials as another. For each run, we estimated the voxel-wise responses to the two conditions using a simple general linear model and computed the coefficient of determination (R^2) between the predicted and observed response time series. A higher R^2 value indicates that a larger proportion of the voxel's signal variance can be explained by the experimental design, reflecting stronger and more reliable stimulus-driven activity.

In this analysis, all caption trials within a run were treated as one condition, and all image trials as another. For each run, we estimated voxel-wise responses to these two conditions using a simple general linear model and computed the coefficient of determination (R^2) between the predicted and observed time series. A higher R^2 indicates that a larger proportion of the voxel's response can be explained by the experimental design. For each participant, we averaged the R^2 values across all 200 runs and projected the result onto the fsaverage cortical surface. Group-level maps were obtained by averaging these surface maps across the eight participants. As shown in Fig. 3a, only regions with ON–OFF R^2 greater than 1% are displayed. Strong and reliable stimulus-driven responses were observed across large portions of the visual cortex, as well as in the inferior frontal sulcus (IFS), suggesting robust activation patterns induced by the task. Here, we explicitly modeled all trials as 2 separate conditions because we hypothesized that the neural responses to linguistic (Caption) and visual (Image) stimuli are fundamentally different. Treating them as a single, unified "ON-OFF" condition would force the model to fit a single set of parameters to two distinct neural processes, leading to a loss of variance explanation and a significant drop in model accuracy.

In addition to ON–OFF R^2, we also examined the activation patterns estimated by the GLM for the caption and image conditions separately. As shown in Fig. 3b,c, these maps represent the average activation across 200 runs per participant, further averaged across all eight participants and projected to the fsaverage cortical surface. The results are consistent with prior studies on visual and language processing. Caption stimuli elicited stronger responses than image stimuli in the ventral occipitotemporal cortex, particularly in regions near the visual word form area (VWFA). Moreover, caption-related activations were largely confined to the ventral visual stream, whereas image stimuli evoked robust responses in both ventral and dorsal pathways. These patterns

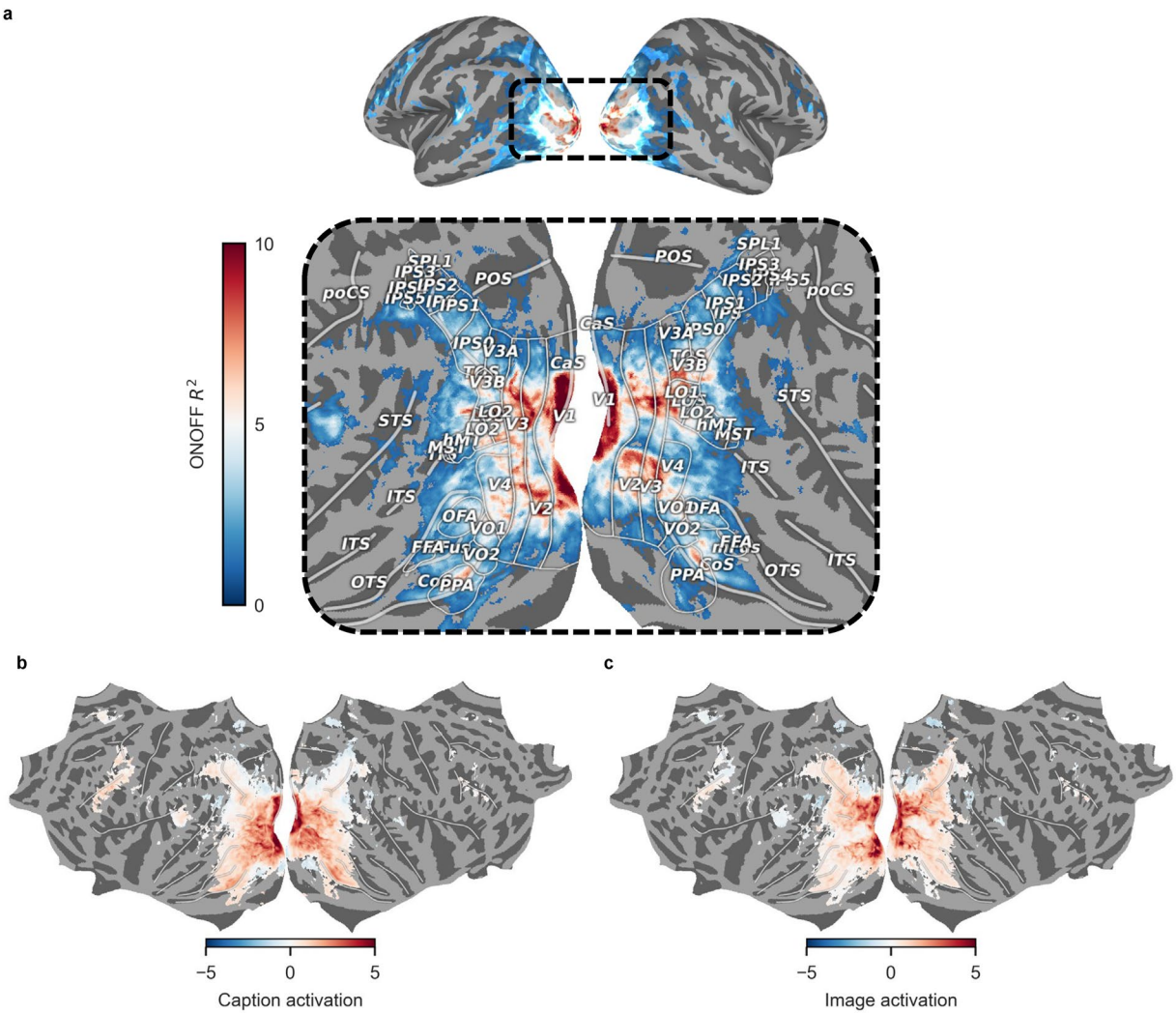


Fig. 3 Validation of stimulus-evoked BOLD responses. (**a**) Voxel-wise ONOFF R^2 map. Voxel-wise variance explained by the presence versus absence of stimulus, derived from a GLM. Only voxels with $R^2 > 1\%$ are shown. The early visual cortex (EVC) is highlighted with a black dashed box and zoomed in. (**b**) Caption condition activation. GLM beta weights for the caption condition show stronger responses along the ventral visual pathway compared to the dorsal pathway. (**c**) Image condition activation. Beta weights for the image condition reveal widespread activation across the visual cortex, with relatively weaker responses in ventral regions than in the caption condition.

further support the reliability of the data and the effectiveness of the experimental paradigm in driving distinct modality-specific activation.

Moreover, we contrasted the activation patterns evoked by caption and image stimuli to further validate the neural responses. As shown in Fig. 4a,b, these group-average surface maps (n = 8) revealed distinct topographies for the two modalities. Specifically, as shown in Fig. 4a, caption stimuli elicited significantly stronger activation in the ventral visual pathway, particularly in the vicinity of the fusiform area, as well as in the inferior frontal sulcus (IFS). In contrast, image stimuli evoked more prominent responses in the dorsal visual pathway (e.g., intraparietal sulcus) compared to captions (Fig. 4b).

Behavior performance. During the experiment, participants were instructed to press a button whenever they found a caption–image pair to be mismatched. To assess task engagement and compliance, we analyzed the behavioral error rate, defined as trials in which participants either failed to respond when a mismatch occurred or responded when no mismatch was present.

As shown in Fig. 5a, the distribution of error rates across runs in COCO-CN sessions was centered near zero for most participants, indicating high overall task accuracy. However, a subset of runs from participants S2, S4, and S5 exhibited relatively higher wrong response rates, suggesting occasional lapses in attention or task compliance. To further examine temporal trends, we grouped every 8 runs into a session (25 sessions per participant), and computed the mean wrong response rate per session. Consistent with these observations, the results from the control sessions exhibited a similar pattern (Fig. 5b). As shown in Fig. 5c, participant S4 showed

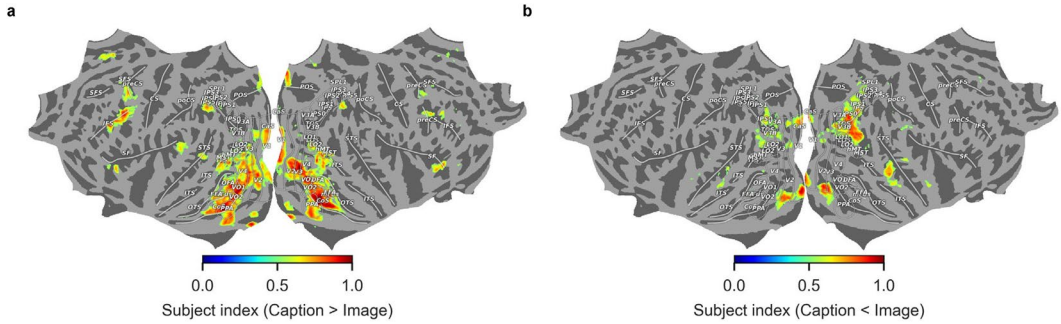


Fig. 4 Validation of modality-specific activation patterns. (**a**) Subject index map of caption-specific activations (Caption > Image). Only vertices exhibiting significantly greater activation for captions compared to images ($p < 0.00001$, cluster size > 50) at the subject level were projected to the fsaverage cortical surface. Vertices with SI > 0.5 and cluster size > 50 are displayed. (**b**) Subject index map of image-specific activations (Caption < Image). Only vertices exhibiting significantly greater activation for images compared to captions ($p < 0.00001$, cluster size > 50) at the subject level were projected to the fsaverage cortical surface. Vertices with SI > 0.5 and cluster size > 50 are displayed.

elevated error rates during the initial COCO-CN sessions, which then stabilized at approximately 5% in later sessions. In contrast, participant S1 and S2 displayed low error rates in early sessions, followed by a gradual increase to around 5% in the latter half of the experiment. The remaining participants maintained low and stable wrong response rates throughout the scanning period, indicating consistent task performance across sessions. Moreover, we calculate the confusion matrix for each participant to fully validate participant engagement (Fig. 6). The distribution of behavioral responses closely matched the prior probability of each trial type. Critically, we observed no 'near-zero' cells in the minority response categories across all participants, indicating that subjects remained attentive to all trial types and did not adopt a strategy of neglecting infrequent stimuli.

Overall, behavioral results indicate that the captions provide clear and representative descriptions of the visual stimuli, as participants were able to reliably distinguish matched from unmatched pairs. We also note that unmatched trials may involve greater subjective interpretation, since partial semantic overlap between a distractor caption and an image can introduce ambiguity. Such variability likely reflects sensitivity to fine-grained semantic relationships rather than simple inattention.

ROI-based analysis. To evaluate the reliability of our functionally defined ROIs and further validate the data quality, we performed a three-stage ROI-based analysis. We first performed a qualitative inspection of the top-activating stimuli to ensure functional alignment. Taking Subject 3 as a representative example, we identified the five stimuli that evoked the strongest neural responses for both caption and image conditions within the Face-selective and Place-selective ROIs. As illustrated in Table 2 and Table 3, the top stimuli in the Face-selective ROI predominantly featured portraits or social contexts, whereas the Place-selective ROI was primarily driven by scenic and architectural descriptions. This consistency across both modalities confirms that our ROI localization effectively captures high-level semantic preferences.

Next, we investigated whether the representational geometry of our ROIs aligns with computational semantic models. We selected 1000 shared trials (averaged across two repetitions per subject) to construct brain RDMs[60–62] (Representation Dissimilarity Matrix) for six target ROIs (V1, V2, V3, Face-selective, Place-selective, and a combined Face + Place ROI). For the model RDM, we utilized embeddings from the final layer of the CLIP-image encoder.

Our results revealed a clear hierarchical gradient (Fig. 7a): correlations were minimal in early visual areas (V1–V3), which do not primarily encode high-level semantic structures. In contrast, RSA correlations increased progressively in the Face-selective and Place-selective ROIs. This alignment between neural RDMs in the ventral visual stream and the CLIP semantic space further validates the biological plausibility of our dataset.

Finally, we quantified the spatial and semantic consistency across participants by calculating the lower and upper noise ceilings[61,63] across a broad range of Regions of Interest (ROIs) (Fig. 7b). These included early visual areas (V1, V2, and V3), category-selective regions such as the Place-selective area (Place), Face-selective area (Face), and Body-selective area (Body), as well as higher-order association areas including the Inferior Frontal Sulcus (IFS), Middle Frontal Gyrus (MFG), Inferior Occipital Gyrus (IOG), and Intraparietal Sulcus (IPS). High noise ceilings were observed across nearly all visual and associative regions, indicating robust inter-subject reliability. While the noise ceilings were comparatively lower in the IFS and MFG, likely due to the higher inter-individual functional variability inherent in prefrontal executive networks, the overall high consistency across the majority of ROIs reinforces the technical soundness.

Noise ceiling. To evaluate the reliability of trial-wise beta weights estimated by GLMsingle, we computed voxel-wise noise ceiling values using the mismatched trials from the control sessions. This analysis was based specifically on the caption and image responses within these trials. The control task design is more suitable for noise ceiling estimation, as it aligns with the underlying assumption that each stimulus evokes an independent response with minimal temporal correlation between adjacent events.

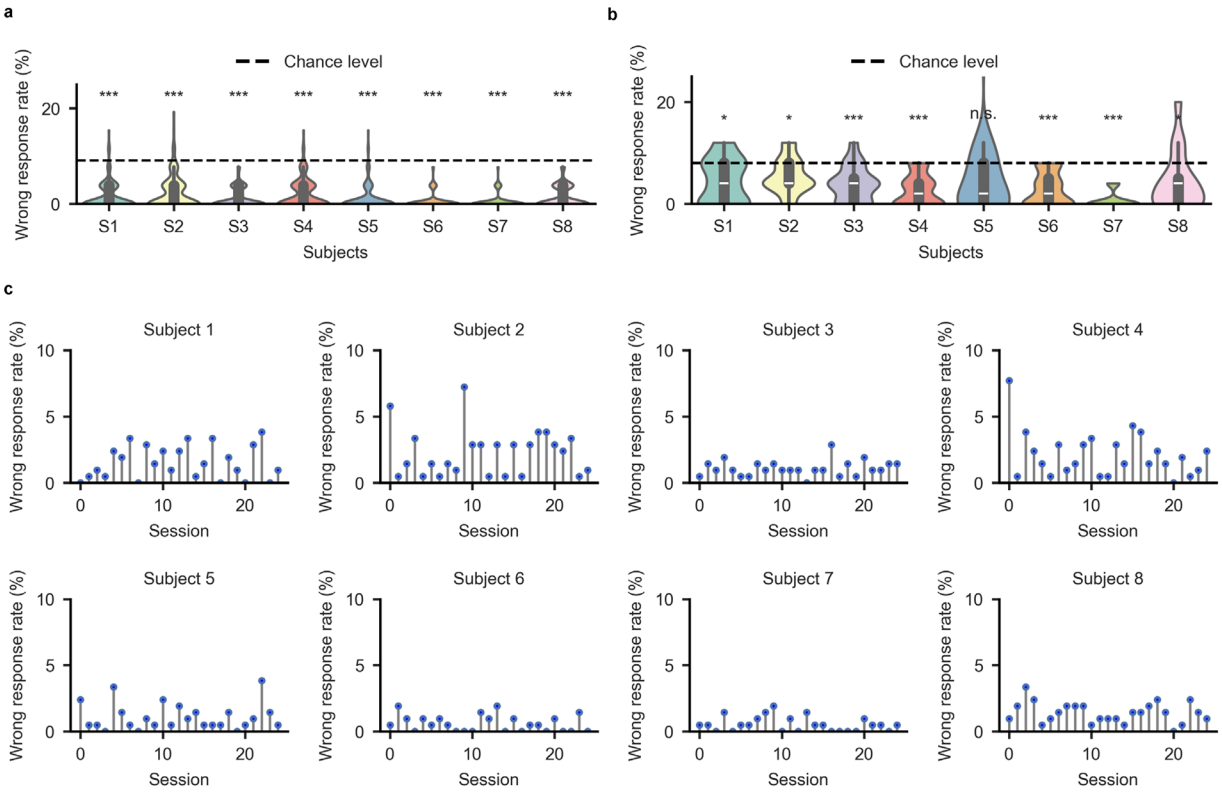


Fig. 5 Behavioral error rates across participants. (**a**) Distribution of error rates across runs for 8 participants in COCO-CN sessions. For each of the 8 participants, error rates are shown across 200 runs. Most runs exhibited error rates below 5%, with participants S1, S2, S4, and S5 showing relatively larger variability. (**b**) Distribution of error rates across runs for 8 participants in control sessions. For each of the 8 participants, error rates are shown across 200 runs. Most runs exhibited error rates below 5%, with participants S1, S2, S5, and S8 showing relatively larger variability. (**c**) Within-subject error rate trends. Each participant's error rates were averaged in non-overlapping 8-run blocks across the 200-run session. The resulting 25-point stem plots reflect performance consistency over time.

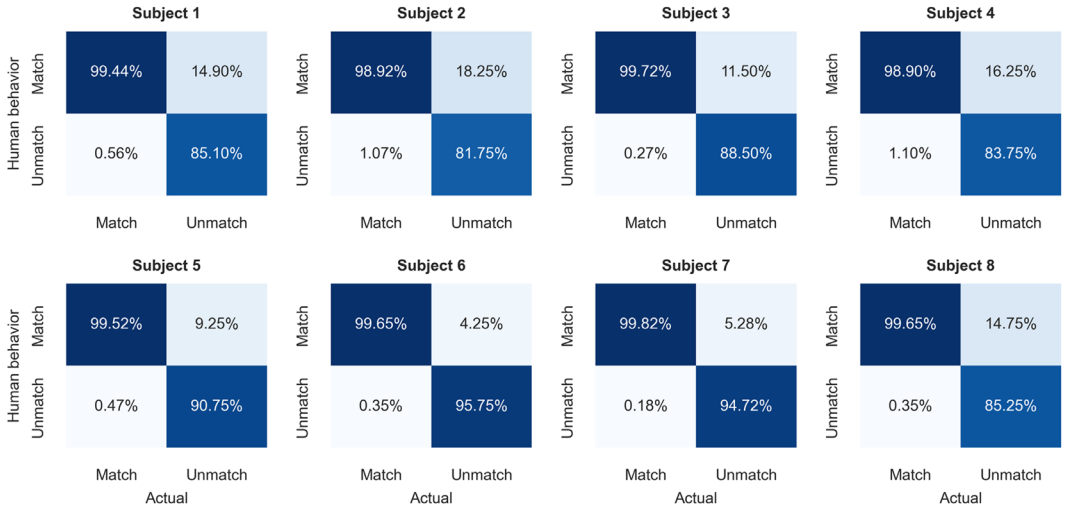


Fig. 6 Behavior confusion matrix across participants in COCO-CN sessions. The matrix displays the distribution of behavioral responses across different trial types, aggregated across all participants in COCO-CN sessions. The diagonal elements represent the hit rates for each category, while off-diagonal elements indicate the patterns of misclassification.

In contrast, the matched trials in the COCO-CN sessions may introduce cognitive dependencies between the caption and image phases. Since the two stimuli are semantically related and presented in close temporal

ROI	1st	2nd	3rd	4th	5th
Face	A green teddy bear is sitting on the bed	Several dogs are lying next to a clock	There are fruits such as apples and oranges on the fruit stand	A cat lying half open its eyes	A squirrel ate a piece of green food on the ground
Place	There is a bridge and other buildings near Big Ben	A kitchen has furniture such as counters, sinks, and tables	A garden with park benches and red umbrellas	There is an empty bench in the densely vegetated park	Two people skiing on the street next to the house

Table 2. Category-selective responses for caption stimuli.

ROIs	1st	2nd	3rd	4th	5th
Face					
Place					

Table 3. Category-selective responses for image stimuli.

succession, participants may integrate them into a single semantic episode. As a result, the BOLD response may reflect the joint cognitive effect of both stimuli rather than separable, event-specific responses, thereby violating the independence assumption required for accurate noise ceiling estimation.

As shown in Fig. 8, the control data yielded robust noise ceiling values across early visual and higher-level category-selective regions, indicating stable, stimulus-driven responses and confirming the integrity of the fMRI signal and preprocessing workflow. In contrast, the COCO-CN sessions exhibited lower noise ceiling values, particularly in early visual areas, which is consistent with the expected impact of interdependent cognitive processing.

A numerical gap between our noise ceilings and model prediction performance and those reported for the NSD is likely attributable to differences in acquisition hardware rather than intrinsic data quality. NSD was collected using a 7T scanner, which provides substantially higher signal-to-noise ratio than the 3T system used here and therefore enables higher achievable noise ceilings. Consistent with prior benchmarking studies, our results closely match those reported for other large-scale 3T datasets, such as BOLD5000, and fall within the expected performance range for 3T fMRI. These hardware-related differences should be considered when making cross-dataset comparisons.

These results validate that, under simplified and well-controlled task conditions, our dataset supports reliable trial-wise modeling and subsequent analyses. Moreover, the contrast between control and COCO-CN sessions highlights that while our preprocessing pipeline and GLM estimation are robust, the task design in COCO-CN may induce cognitive dependencies across temporally adjacent stimuli, which should be considered when interpreting or modeling the resulting neural responses.

Encoding model results. To validate the utility and quality of the dataset for model-based analyses, we replicated findings reported in Wang *et al.*[9]. Specifically, we used the text encoder from the CLIP[64] model (CLIP-BERT) to predict brain activity based on the final-layer embedding. The [CLS] token (1024-dimensional) $c \in \mathbb{R}^{1024}$ from the last transformer layer was first projected to a 50-dimensional representation $z \in \mathbb{R}^{50}$ via a linear transformation.

$$z = W_{proj}c + b_{proj},$$

This 50-dimensional vector was then used to train voxel-wise encoding models. For each voxel v, we modeled its BOLD response $y_v \in \mathbb{R}^N$ using a linear regression of the form

$$y_v = Zw_v + b_v,$$

where N is the total number of trials, $Z \in \mathbb{R}^{N\times 50}$ is the design matrix whose rows are the projected features z, w_v is the voxel-specific weight vector.

Both the projection parameters (W_{proj}, b_{proj}) and voxel-wise regression parameters (w_v, b_v) were jointly optimized under an ℓ_2 weight-decay (ridge) penalty. For a given weight-decay parameter λ, the optimization problem for each voxel can be written as:

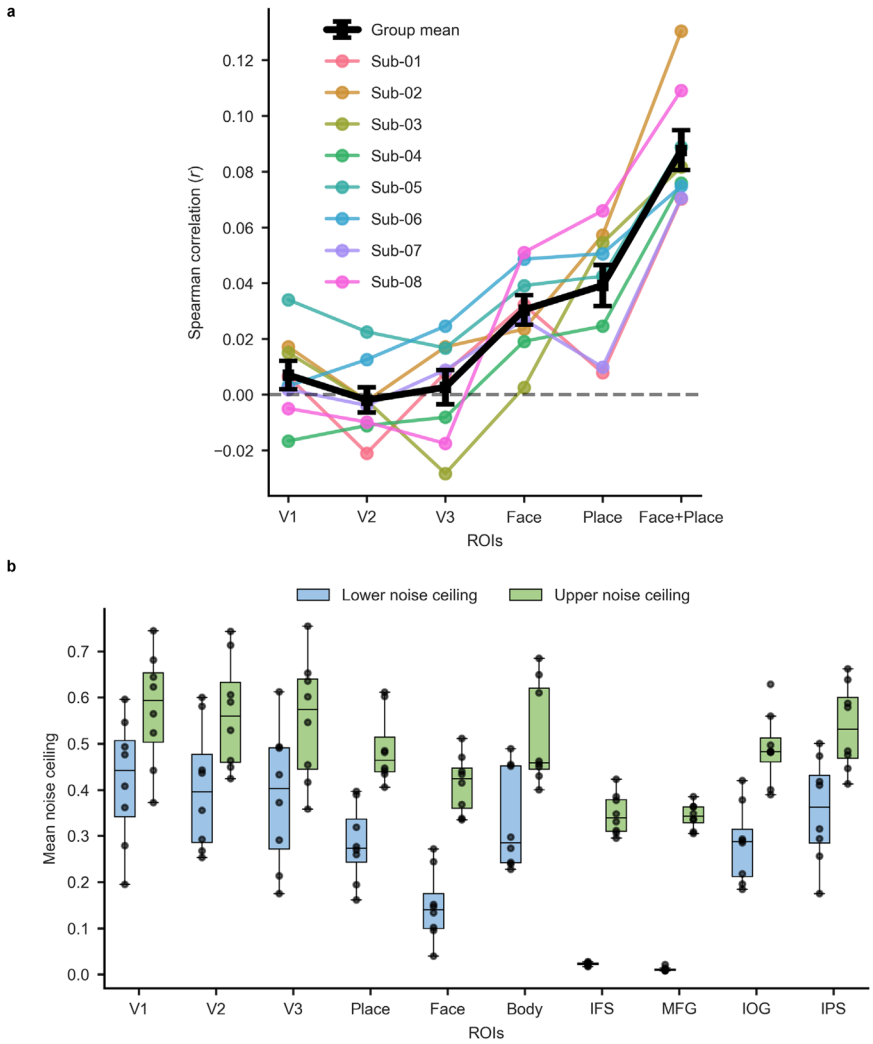


Fig. 7 ROI-based representation similarity analysis for COCO-CN sessions. (**a**) Spearman correlation between neural RDMs and model RDMs across 6 ROIs. Early visual areas (V1, V2, and V3) showed very weak correspondence with the model (group-mean correlation < 0.01), whereas higher-level category-selective regions (combined Face- and Place-selective ROIs) exhibited substantially stronger alignment (group mean $= 0.088$, SD $= 0.007$). (**b**) ROI-level group noise ceilings. For each of the 10 ROIs, upper and lower bounds of the noise ceiling were estimated across participants to quantify the maximum explainable variance given inter-subject variability.

$$\hat{w}_v(\lambda) = \arg\min_{w_v} \|y_v - Zw_v\|_2^2 + \lambda\|w_v\|_2^2$$

To optimize performance, we evaluated different weight decay values λ and selected the best parameter for each voxel. Model performance was assessed using five-fold cross-validation. For each voxel v, model performance was quantified as the squared Pearson correlation between the predicted response $\hat{y}_{v,k}$ and the observed response $y_{v,k}$:

$$R_{v,k}^2 = (corr(y_{v,k}, \hat{y}_{v,k}))^2$$

The final performance for each voxel was obtained by averaging $R_{v,k}^2$ across all five test folds.

As shown in Fig. 9a, the encoding model achieved robust performance across high-level visual and language-related areas, including the inferior temporal cortex and ventral temporal lobe. Notably, strong encoding performance was also observed in regions near Broca's area, closely matching the spatial patterns reported in previous work. These results confirm that our dataset reliably supports neural encoding modeling and replicates prior findings, thereby providing further evidence of its quality and validity.

To evaluate the robustness of our encoding framework across different computational architectures, we compared a Transformer-based vision model (CLIP-ViT) with a CNN-based model (AlexNet), using their final-layer embeddings to predict brain activity. All models achieved reliable encoding performance across the visual

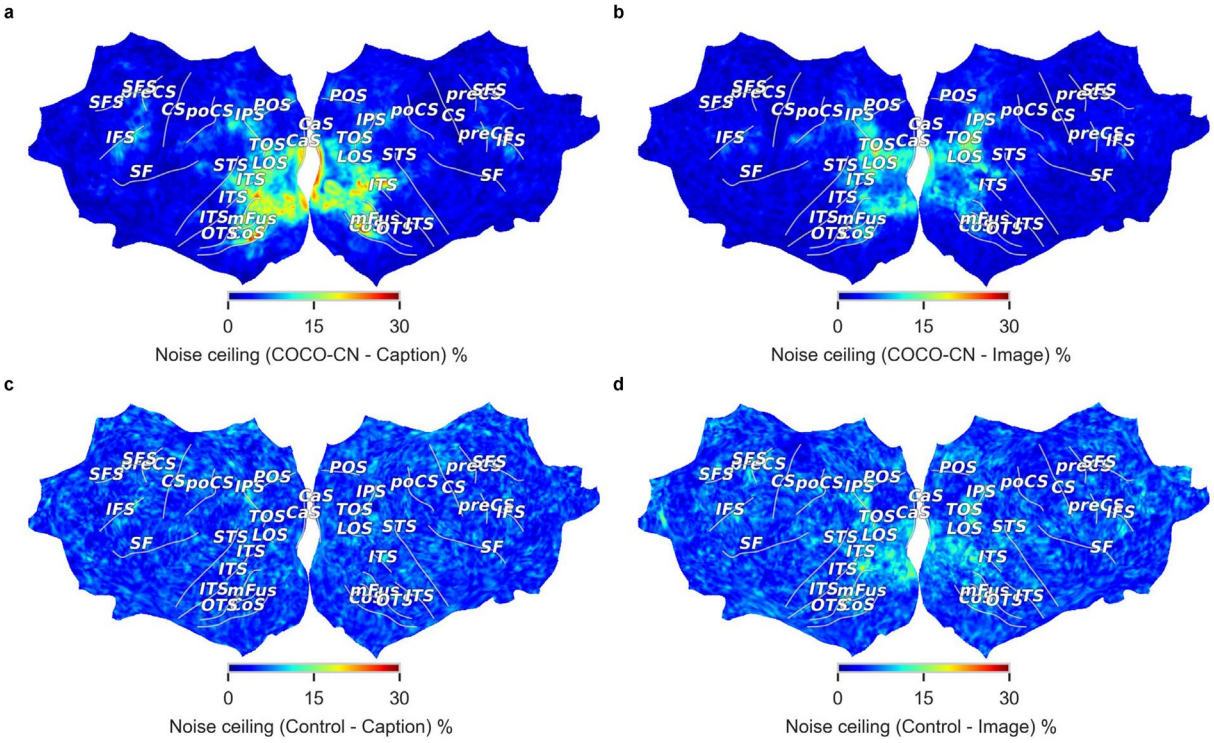


Fig. 8 Noise ceiling estimates for caption and image conditions in COCO-CN and control sessions. Only voxels with ON–OFF $R^2 > 1\%$ are shown in all panels. (**a**) COCO-CN session – caption condition. Noise ceiling values computed for the caption condition during the COCO-CN session. (**b**) COCO-CN session – image condition. Noise ceiling for the image condition in the same session. (**c**) Control session – caption condition. Caption-related noise ceiling values obtained in the control session. (**d**) Control session – image condition. Noise ceiling for the image condition in the control session.

cortex (Fig. 9b,d). Notably, AlexNet exhibited substantially higher performance than CLIP-ViT, while CLIP-ViT showed a modest advantage over CLIP-BERT. These differences were most pronounced in frontal regions and along the dorsal visual pathway, whereas encoding accuracy in the ventral visual stream was largely comparable across models (Fig. 9c,e).

Importantly, the overall pattern revealed a clear trend of increasing encoding performance with greater availability of image-derived information. The lower performance of CLIP-ViT relative to AlexNet is likely attributable to our use of only the global class token (1024 dimensions), which contains less spatially detailed visual information and substantially fewer features than the high-dimensional convolutional representations in AlexNet. Together, these findings indicate that encoding accuracy is strongly influenced by the richness and granularity of visual features captured by computational models, while also demonstrating the robustness of the COCO-CN dataset across diverse architectures.

Usage Notes

In the raw data, the phase encoding direction generally follows the rule that runs with odd-numbered indices correspond to anterior-to-posterior (AP) encoding, while even-numbered runs correspond to posterior-to-anterior (PA). However, there are exceptions in two participants: for sub-01 (runs 43–47) and sub-05 (runs 13–18), the pattern is reversed—odd-numbered runs are PA, and even-numbered runs are AP. This inconsistency has been corrected in the preprocessed data, where all images have been distortion-corrected and phase-encoding direction is no longer relevant.

Each run in the stimulus protocol includes 44 trials, consisting of alternating caption and image presentations. Captions are always presented before images. Accordingly, in the GLM-derived beta estimates, the first beta corresponds to the caption, the second to the image, and this alternation continues throughout the run (i.e., odd-numbered betas correspond to captions, even-numbered betas to images).

In both *"pp_betas_vol"* and *"pp_betas" folders*, we provide two versions of beta estimates: the raw beta values (without z-scoring) and the z-scored beta values. In *"pp_betas_vol"*, each NIfTI file is stored as a five-dimensional array with dimensions (X, Y, Z, T, flag), where the first four dimensions represent the brain volume and trial index, and the fifth dimension indicates the beta type: 0 for the raw beta and 1 for the z-scored beta. This organization allows users to choose either unnormalized or standardized beta estimates depending on their analytical requirements.

Note that participants S1-S3 did not participate in eye-tracking data collection.

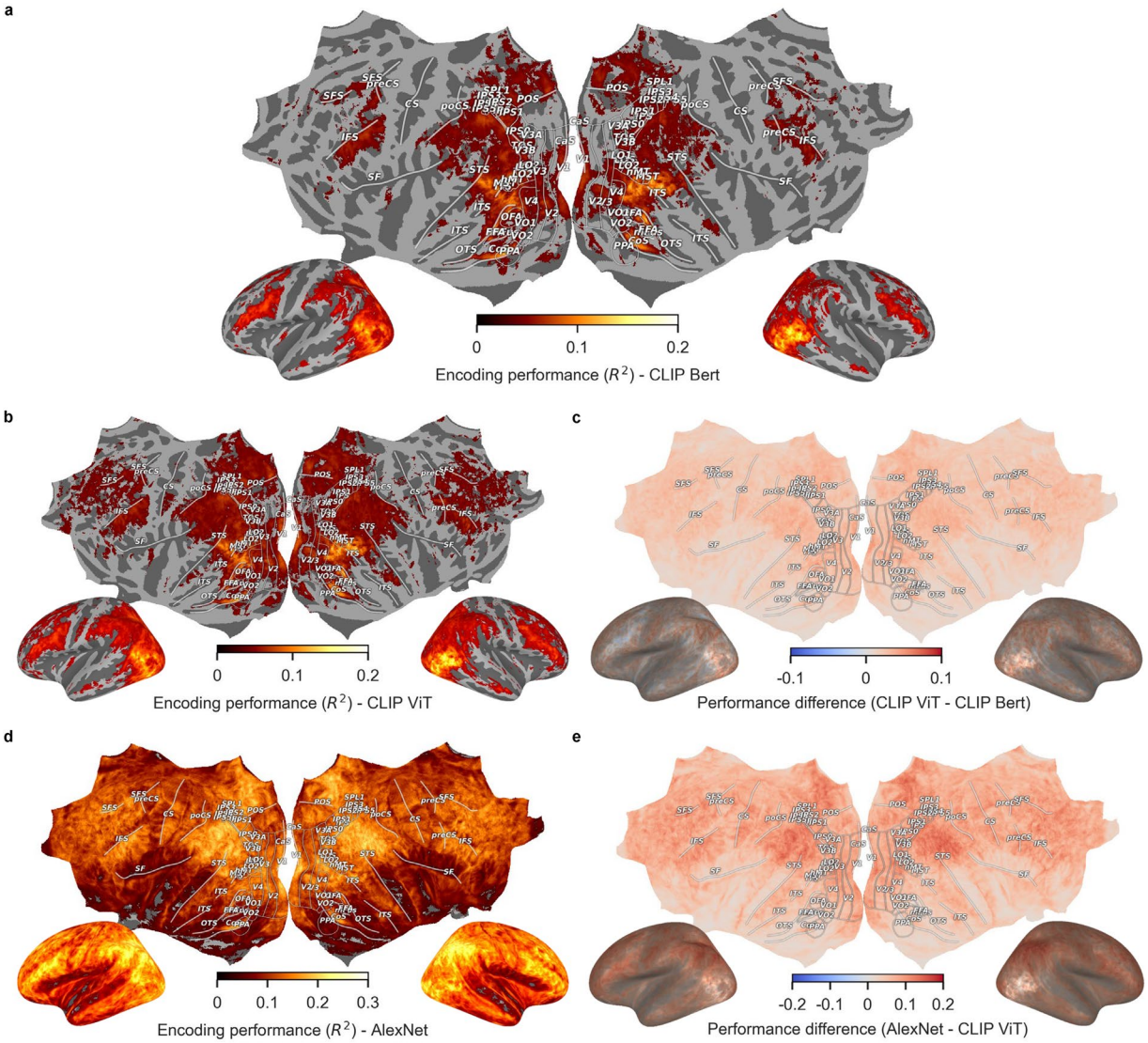


Fig. 9 Group-level encoding performance (R^2). (**a**) Group-averaged encoding results based on CLIP-Bert across 8 participants, visualized on the fsaverage surface. Only vertices with $R^2 > 0.04$ are displayed. Strong encoding performance is observed in high-level visual areas, particularly in the inferior temporal cortex, ventral temporal lobe, and regions near Broca's area. (**b**) Group-averaged encoding results based on CLIP ViT across 8 participants, visualized on the fsaverage surface. Only vertices with $R^2 > 0.04$ are displayed. (**c**) Group-averaged encoding performance between CLIP ViT and CLIP Bert. All vertices are displayed. Slight performance improvements are observed in frontal and dorsal parietal regions. (**d**) Group-averaged encoding results based on AlexNet across 8 participants, visualized on the fsaverage surface. Only vertices with $R^2 > 0.04$ are displayed. (**e**) Group-averaged encoding performance between AlexNet and CLIP ViT. All vertices are displayed. Marked performance improvements are observed in frontal and dorsal parietal regions.

Data availability

The whole dataset can be accessed from ScienceDB https://doi.org/10.57760/sciencedb.27580.

Code availability

All code used for preprocessing the data and technical validation will be made publicly available on https://github.com/lishurui0612/caption_scene_dataset. The repository includes scripts for preprocessing raw fMRI data, preprocessing eye-tracking data, computing temporal signal-to-noise ratio (tSNR), estimating noise ceilings, running general linear models (GLMs), and training encoding models.

Received: 27 November 2025; Accepted: 13 April 2026;
Published online: 18 April 2026

References

1. Lalanne, C. & Lorenceau, J. Crossmodal integration for perception and action. *J. Physiol.-Paris* **98**, 265–279 (2004).
2. Macaluso, E. & Driver, J. Multisensory spatial interactions: a window onto functional integration in the human brain. *Trends Neurosci.* **28**, 264–271 (2005).
3. Mesulam, M. M. From sensation to cognition. *Brain* **121**, 1013–1052 (1998).
4. Popham, S. F. *et al*. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nat. Neurosci.* **24**, 1628–1636 (2021).
5. Dong, D. T. & Toneva, M. Vision-Language Integration in Multimodal Video Transformers (Partially) Aligns with the Brain. Preprint at https://doi.org/10.48550/arXiv.2311.07766 (2023).
6. Anzellotti, S. & Caramazza, A. Multimodal representations of person identity individuated with fMRI. *Cortex* **89**, 85–97 (2017).
7. Visser, M., Jefferies, E., Embleton, K. V. & Lambon Ralph, M. A. Both the Middle Temporal Gyrus and the Ventral Anterior Temporal Area Are Crucial for Multimodal Semantic Processing: Distortion-corrected fMRI Evidence for a Double Gradient of Information Convergence in the Temporal Lobes. *J. Cogn. Neurosci.* **24**, 1766–1778 (2012).
8. Ralph, M. A. L., Jefferies, E., Patterson, K. & Rogers, T. T. The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* **18**, 42–55 (2017).
9. 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.* 1–12 https://doi.org/10.1038/s42256-023-00753-y (2023).
10. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
11. Chen, P. *et al*. Do Self-Supervised Speech and Language Models Extract Similar Representations as Human Brain? in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2225–2229. https://doi.org/10.1109/ICASSP48485.2024.10446334 (2024).
12. Li, Y. *et al*. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nat. Neurosci.* **26**, 2213–2225 (2023).
13. Du, C., Fu, K., Li, J. & He, H. Decoding Visual Neural Representations by Multimodal Learning of Brain-Visual-Linguistic Features. *IEEE Trans. Pattern Anal. Mach. Intell.* 1–17 https://doi.org/10.1109/TPAMI.2023.3263181 (2023).
14. Yang, H., Gee, J. & Shi, J. Memory Encoding Model. Preprint at https://doi.org/10.48550/arXiv.2308.01175 (2023).
15. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis With Latent Diffusion Models. in 10684–10695 (2022).
16. Ma, Y. *et al*. BrainCLIP: Brain Representation via CLIP for Generic Natural Visual Stimulus Decoding. *IEEE Trans. Med. Imaging.* **44**, 3962–3972 (2025).
17. Tang, J., LeBel, A., Jain, S. & Huth, A. G. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nat. Neurosci.* **26**, 858–866 (2023).
18. Caselles-Dupré, H. *et al*. Mind-to-Image: Projecting Visual Mental Imagination of the Brain from fMRI. In *ICML 2024 Workshop on AI for Science* (2024).
19. Li, Y., Yang, H. & Gu, S. Enhancing neural encoding models for naturalistic perception with a multi-level integration of deep neural networks and cortical networks. *Sci. Bull.* **69**, 1738–1747 (2024).
20. Eickenberg, M., Gramfort, A., Varoquaux, G. & Thirion, B. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage* **152**, 184–194 (2017).
21. Schrimpf, M. *et al*. Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron.* **108**, 413–423 (2020).
22. Güçlü, U. & van Gerven, M. A. J. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J. Neurosci.* **35**, 10005–10014 (2015).
23. Kriegeskorte, N. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).
24. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Comput. Biol.* **10**, e1003915 (2014).
25. Yamins, D. L. K. *et al*. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* **111**, 8619–8624 (2014).
26. Allen, E. J. *et al*. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).
27. Chang, N. *et al*. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Sci. Data* **6**, 49 (2019).
28. 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).
29. Gong, Z. *et al*. A large-scale fMRI dataset for the visual processing of naturalistic scenes. *Sci. Data* **10**, 559 (2023).
30. Zhang, G. *et al*. A large-scale MEG and EEG dataset for object recognition in naturalistic scenes. *Sci. Data* **12**, 857 (2025).
31. Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron* **76**, 1210–1224 (2012).
32. Glasser, M. F. *et al*. A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
33. Fedorenko, E., Ivanova, A. A. & Regev, T. I. The language network as a natural kind within the broader landscape of the human brain. *Nat. Rev. Neurosci.* **25**, 289–312 (2024).
34. Guillot, A. *et al*. Functional neuroanatomical networks associated with expertise in motor imagery. *NeuroImage* **41**, 1471–1483 (2008).
35. Barch, D. M. *et al*. Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage* **80**, 169–189 (2013).
36. Ayzenberg, V. & Behrmann, M. The Dorsal Visual Pathway Represents Object-Centered Spatial Relations for Object Recognition. *J. Neurosci.* **42**, 4693–4710 (2022).
37. Genç, E., Schölvinck, M. L., Bergmann, J., Singer, W. & Kohler, A. Functional Connectivity Patterns of Visual Cortex Reflect its Anatomical Organization. *Cereb. Cortex* **26**, 3719–3731 (2016).
38. Nastase, S. A. *et al*. The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Sci. Data* **8**, 250 (2021).
39. D'Esposito, M., Postle, B. R. & Rypma, B. Prefrontal cortical contributions to working memory: evidence from event-related fMRI studies. *Exp. Brain Res.* **133**, 3–11 (2000).
40. Miller, E. *et al*. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci. 16, 5154-5167. J. Neurosci. Off. J. Soc. Neurosci.* **16**, 5154–67 (1996).
41. Emch, M., von Bastian, C. C. & Koch, K. Neural Correlates of Verbal Working Memory: An fMRI Meta-Analysis. *Front. Hum. Neurosci.* **13**, (2019).
42. Pessoa, L., Gutierrez, E., Bandettini, P. A. & Ungerleider, L. G. Neural Correlates of Visual Working Memory: fMRI Amplitude Predicts Task Performance. *Neuron* **35**, 975–987 (2002).
43. Nichols, E. A., Kao, Y.-C., Verfaellie, M. & Gabrieli, J. D. E. Working memory and long-term memory for faces: Evidence from fMRI and global amnesia for involvement of the medial temporal lobes. https://doi.org/10.1002/hipo.20190.

44. Sisakhti, M., Sachdev, P. S. & Batouli, S. A. H. The Effect of Cognitive Load on the Retrieval of Long-Term Memory: An fMRI Study. *Front. Hum. Neurosci.* **15**, (2021).
45. Wittmann, B. C. *et al*. Reward-Related fMRI Activation of Dopaminergic Midbrain Is Associated with Enhanced Hippocampus-Dependent Long-Term Memory Formation. *Neuron* **45**, 459–467 (2005).
46. Gifford, A. T. *et al*. The Algonauts Project 2025 Challenge: How the Human Brain Makes Sense of Multimodal Movies. Preprint at https://doi.org/10.48550/arXiv.2501.00504 (2025).
47. LeBel, A. *et al*. A natural language fMRI dataset for voxelwise encoding models. *Sci. Data* **10**, 555 (2023).
48. Li, X. *et al*. COCO-CN for Cross-Lingual Image Tagging, Captioning, and Retrieval. *IEEE Trans. Multimed.* **21**, 2347–2360 (2019).
49. Lin, T.-Y. *et al*. Microsoft COCO: Common Objects in Context. in *Computer Vision – ECCV 2014* (eds Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T.) 740–755. https://doi.org/10.1007/978-3-319-10602-1_48 (Springer International Publishing, Cham, 2014).
50. Nikolaus, M. *et al*. Modality-Agnostic Decoding of Vision and Language from fMRI. *eLife* **14**, (2025).
51. Benson, N. C. *et al*. The Human Connectome Project 7 Tesla retinotopy dataset: Description and population receptive field analysis. *J. Vis.* **18**, 23 (2018).
52. Stigliani, A., Weiner, K. S. & Grill-Spector, K. Temporal Processing Capacity in High-Level Visual Cortex Is Domain Specific. *J. Neurosci.* **35**, 12412–12424 (2015).
53. Kleiner, M. *et al*. What's new in psychtoolbox-3. *Perception* **36**, 1–16 (2007).
54. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W. & Smith, S. M. F. S. L. *NeuroImage* **62**, 782–790 (2012).
55. Fischl, B. FreeSurfer. *NeuroImage* **62**, 774–781 (2012).
56. Prince, J. S. *et al*. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife* **11**, e77599 (2022).
57. Li, S., Jin, Z., Gu, S., Zhang, R.-Y. & Li, Y. CSD: Caption Scene Dataset. https://doi.org/10.57760/sciencedb.27580 (2025).
58. Murphy, K., Bodurka, J. & Bandettini, P. A. How long to scan? The relationship between fMRI temporal signal to noise ratio and necessary scan duration. *NeuroImage* **34**, 565–574 (2007).
59. Power, J. D. *et al*. Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* **84**, 320–341 (2014).
60. Nili, H. *et al*. A Toolbox for Representational Similarity Analysis. *PLOS Comput. Biol.* **10**, e1003553 (2014).
61. Kriegeskorte, N. *et al*. Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron* **60**, 1126–1141 (2008).
62. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* **17**, 401–412 (2013).
63. Lahner, B. *et al*. Modeling short visual events through the BOLD moments video fMRI dataset and metadata. *Nat. Commun.* **15**, 6241 (2024).
64. Radford, A. *et al*. Learning Transferable Visual Models From Natural Language Supervision. in *Proceedings of the 38th International Conference on Machine Learning* 8748–8763 (PMLR, 2021).

Acknowledgements

The authors thank the participants for their contributions to this work. The authors thank Hua Jiang for helping with the neuroimaging data collection. This work is supported by the National Science and Technology Major Project of China (2025ZD0217000, Y.L.), National Natural Science Foundation of China (32371154, Y.L.; 32441102, R.-Y.Z), Shanghai Rising-Star Program (24QA2705500, Y.L.), and the Lin Gang Laboratory (LG-GG-202402-06 and LGL-1987-18, Y.L.). The computations in this research are supported by the HPC Platform of ShanghaiTech University.

Author contributions

Conceptualization: Y.L., S.G., and R.-Y.Z.; Methodology: S.L., Z.J., S.G., R.-Y.Z. and Y.L.; Software: S.L.; Investigation: S.L., Z.J., S.G., R.-Y.Z., and Y.L.; Data Curation: S.L., Z.J. and R.-Y.Z.; Formal analysis: S.L., Z.J. and Y.L.; Resources: Y.L.; Writing – Original Draft: S.L. and Z.J.; Writing – Review & Editing: S.L., Z.J. and Y.L.; Supervision: Y.L.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.G., R.-Y.Z. or Y.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026