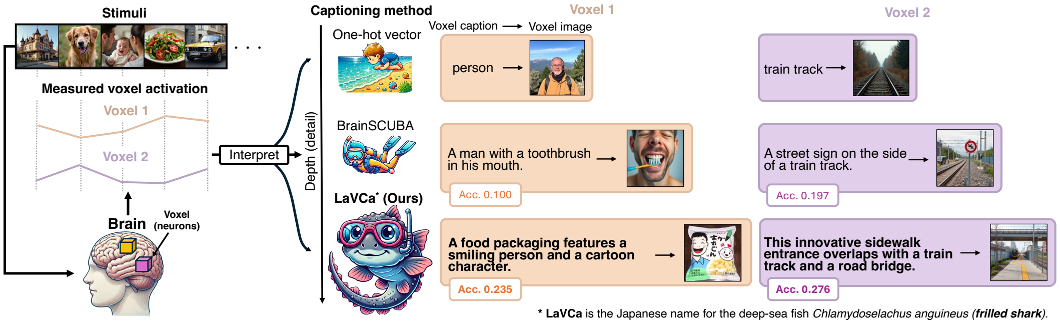
LaVCa: LLM-assisted Visual Cortex Captioning

Anonymous authors
Paper under double-blind review


Stimuli
. . .
Measured voxel activation
Voxel 1
Voxel 2
Brain
Voxel (neurons)
Interpret
Captioning method
Depth (detail)
One-hot vector
BrainSCUBA
LaVCa* (Ours)
Voxel caption → Voxel image
Voxel 1
Voxel 2
person →
train track →
A man with a toothbrush in his mouth. →
Acc. 0.100
A street sign on the side of a train track. →
Acc. 0.197
A food packaging features a smiling person and a cartoon character. →
Acc. 0.235
This innovative sidewalk entrance overlaps with a train track and a road bridge. →
Acc. 0.276

* **LaVCa** is the Japanese name for the deep-sea fish *Chlamydoselachus anguineus* (***frilled shark***).

Figure 1: Illustration of our paper. Our proposed method, LaVCa, produces natural language captions that provide a fine-grained description of voxel selectivity (representation) and more accurately capture the characteristics of brain activity in the visual cortex, outperforming conventional approaches.

Abstract

Understanding the properties of neural populations (or voxels) in the human brain can advance our comprehension of human perceptual and cognitive processing capabilities and contribute to developing brain-inspired computer models. Recent encoding models using deep neural networks (DNNs) have successfully predicted voxel-wise activity. However, interpreting the properties that explain voxel responses remains challenging because of the black-box nature of DNNs. As a solution, we propose LLM-assisted Visual Cortex Captioning (LaVCa), a data-driven approach that leverages large language models (LLMs) to generate natural-language captions for images to which voxels are selective. By applying LaVCa for image-evoked brain activity, we demonstrate that LaVCa generates captions that describe voxel selectivity more accurately than the previous approaches. The captions generated by LaVCa quantitatively capture more detailed properties than the existing method at both the inter-voxel and intra-voxel levels. Furthermore, we find richer representational content within cortical regions that prior neuroimaging studies have deemed selective for simpler categories. These findings offer profound insights into human visual representations by assigning detailed captions throughout the visual cortex while highlighting the potential of LLM-based methods in understanding brain representations.

1 Introduction

A primary goal of computer vision is to build systems capable of processing and understanding the complex visual world in a manner akin to human perception. Studying how the human brain—with its advanced visual functions—forms its visual representations deepens our understanding of the brain's visual network and holds promise for developing next-generation computer vision models.

Over the past decade, *encoding models* have become the standard tool for this endeavour (Kay et al., 2008; Nishimoto et al., 2011; Naselaris et al., 2011). Early work employed handcrafted, low-level filters or one-hot semantic labels, yielding interpretable—but coarse—descriptions of voxel-level (the spatial measurement unit of fMRI) selectivity. Modern approaches substitute deep neural-network

(DNN) features, which dramatically raise prediction accuracy (Güçlü & Van Gerven, 2015; Schrimpf et al., 2021; Takagi & Nishimoto, 2023). Yet the very richness that makes DNNs powerful also renders them opaque: it remains difficult to explain why a given voxel activates, especially at the single-voxel level where group-averaged semantic axes (Huth et al., 2016; Lescroart & Gallant, 2019) are too blunt.

In this study, we address the difficulty of voxel-level interpretation with a new method called LLM-assisted Visual Cortex Captioning (**LaVCa**), which generates data-driven captions for individual voxels (Figure 1). LaVCa proceeds in four steps: (1) building voxel-wise encoding models for brain activity evoked by images, (2) identifying the optimal images for each voxel's encoding model using an augmented image dataset, (3) generating captions for these optimal images, and (4) creating concise summaries from those captions. By leveraging large language models (LLMs) with access to a vast, open-ended vocabulary, LaVCa generates diverse inter-voxel captions. Moreover, generating captions from multiple keywords enables us to capture diverse intra-voxel properties.

Our contributions are as follows:

1. We propose LaVCa, which leverages LLMs to generate natural language captions of voxel-level visual selectivity. By adopting a multi-stage design that decomposes the captioning process into interpretable steps, LaVCa enhances interpretability compared to prior work while preserving descriptive richness.
2. We demonstrate that LaVCa produces more accurate captions than the earlier method BrainSCUBA (Luo et al., 2023) and better characterizes voxel-wise visual selectivity through brain activity prediction.
3. We also demonstrate that LaVCa can generate highly interpretable and accurate captions without sacrificing information from the optimal images (Figure 2).
4. The captions generated by LaVCa quantitatively capture more detailed properties than the existing method at both the inter-voxel and intra-voxel levels.
5. More detailed analysis of the voxel-specific properties generated by LaVCa reveals richer representational content within ROIs that prior neuroimaging studies have deemed selective for simpler categories.

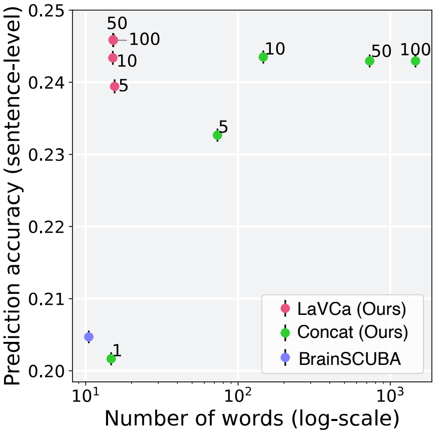


Figure 2: The relationship between brain activity prediction accuracy and voxel caption length (number of words) for a single subject (subj01). Numbers next to each point denote the number of optimal image captions employed by LaVCa for summarization and by Concat for direct concatenation.

2 Related Work

Two complementary approaches frame modern fMRI research: *encoding* and *decoding* models (Naselaris et al., 2011). Encoding models aggregate activity across many *stimuli* for each voxel to pinpoint the features that best explain its responses. Decoding models reverse the mapping, pooling activity across many *voxels* to reconstruct a participant's moment-to-moment percepts—ranging from continuous speech (Tang et al., 2023) to images (Takagi & Nishimoto, 2023). Relatedly, recent brain-to-text decoding studies align brain activity with models to generate natural-language descriptions directly from neural signals (Chen et al., 2025a;b). Although both lines of work exploit powerful deep models, they address distinct questions: decoders ask "*What was the observer perceiving?*," whereas encoders ask "*What information does this voxel represent?*." The present study targets the encoding side, judging captions by how accurately they predict voxel responses rather than how faithfully they reproduce the original stimulus.

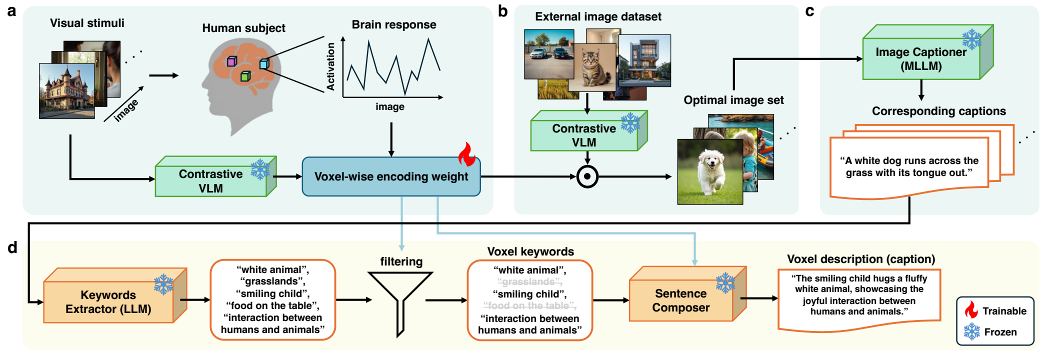

a
Visual stimuli
image
Human subject
Brain response
Activation
image
Contrastive VLM
Voxel-wise encoding weight
b
External image dataset
Contrastive VLM
Optimal image set
c
Image Captioner (MLLM)
Corresponding captions
"A white dog runs across the grass with its tongue out."
d
Keywords Extractor (LLM)
"white animal", "grasslands", "smiling child", "food on the table", "interaction between humans and animals"
filtering
Voxel keywords
"white animal", ~~"grasslands",~~ "smiling child", ~~"food on the table",~~ "interaction between humans and animals"
Sentence Composer
Voxel description (caption)
"The smiling child hugs a fluffy white animal, showcasing the joyful interaction between humans and animals."
Trainable
Frozen

Figure 3: Architecture of LaVCa. **a** We construct a voxel-wise encoding model for a human subject's brain activity data (measured using fMRI) while viewing images, using latent representations from a contrastive vision–language model (VLM). The encoding weight is obtained through ridge regression. **b** We identify the optimal images for a given voxel by calculating the inner product between the contrastive VLM embeddings of external image datasets and the voxel's trained encoding weight, selecting the top-N images (the "optimal image set") that produce the highest predicted activation. **c** Next, we use a multimodal large language model (MLLM) to generate captions for each optimal image set, allowing an LLM to interpret them. **d** Finally, we prompt an LLM to extract keywords from the captions, filter these keywords, and feed them into a "Sentence Composer," producing a concise voxel caption.

Early encoding work used handcrafted low-level filters or one-hot semantic labels, enabling straightforward but coarse voxel interpretation (Kay et al., 2008; Nishimoto et al., 2011; Naselaris et al., 2011; Huth et al., 2012). Swapping these features for deep-neural-network (DNN) embeddings dramatically improves prediction accuracy (Güçlü & Van Gerven, 2015; Schrimpf et al., 2021; Takagi & Nishimoto, 2023; Antonello et al., 2024), yet the high-dimensional representations make individual voxels hard to explain. Population-level remedies project many voxels onto a few semantic axes (Huth et al., 2016; Lescroart & Gallant, 2019; Nakagi et al., 2024), but sacrifice single-voxel nuance.

To obtain finer, voxel-specific explanations, data-driven text-generation approaches such as BrainSCUBA (Luo et al., 2023) and SASC (Singh et al., 2023) have been proposed. BrainSCUBA is an end-to-end method that uses an existing image captioning model to produce voxel-wise captions for the visual cortex, whereas SASC uses an LLM to merge multiple short phrases—those with the highest predicted voxel activations—into a single, data-driven caption, thus describing the semantic properties of voxels. However, their reliance on a single captioning model (BrainSCUBA) or on very short n-gram phrases (SASC) limits lexical richness and adaptability.

By (i) decoupling image selection from caption generation, (ii) using LLM-based keyword extraction followed by lightweight sentence composition (iii) allowing any vision-language backbone, and (iv) working with any LLM that has strong language skills without task-specific fine-tuning, **LaVCa retains the high predictive power of brain activity while yielding richer and more controllable voxel-level descriptions than prior work**.

3 Methods

3.1 fMRI dataset

This study uses the Natural Scenes Dataset (NSD) (Allen et al., 2022) following the same experimental conditions as in BrainSCUBA. The NSD consists of data collected over 30 to 40 sessions using a 7 Tesla fMRI scanner, with each participant viewing 10,000 images, repeated three times. We analyze data from the four participants (Subject 01, Subject 02, Subject 05, and Subject 07) who completed all imaging sessions. The images and captions used in NSD are drawn from MS COCO and resized to 224×224 pixels to align with the input requirements of the vision models used. We average the brain activity data for each subject across repeated trials of the same image to improve the signal-to-noise ratio. Up to 9,000 images per subject are used as training data, and the remaining 1,000 images are reserved for testing. We use the preprocessed scans with a resolution of $1.8\,\mathrm{mm}$ provided by NSD for the functional data. We use single-trial beta weights estimated via a generalized linear model within ROIs. Moreover, we standardize the response of each voxel to have a mean of zero and a variance of one within each session. We use the ROIs provided by NSD, which include early and higher-level (ventral) visual areas and face, place, body, and word-selective regions.

3.2 LLM-ASSISTED VISUAL CORTEX CAPTIONING (LAVCA)

We propose a method, **LaVCa (LLM-assisted Visual Cortex Captioning)**, to automatically generate data-driven natural language captions that characterize each voxel's selectivity in the visual cortex. LaVCa consists of four stages (Figure 3):

1. Construct voxel-wise encoding models for each subject while they view natural images.
2. Identify the optimal image set by finding the top-N images that most strongly activate each voxel (according to the trained encoding models).
3. Generate captions for these optimal images using a multimodal large language model (MLLM) for summarization by an LLM in the next step.
4. Derive concise voxel captions by extracting and filtering keywords from the image captions, then feeding these keywords into a "Sentence Composer."

We describe the core pipeline here; ablations are detailed in Appendix A.3.

3.2.1 ENCODING MODEL CONSTRUCTION

First, we construct voxel-wise encoding models to predict each voxel's activity in response to natural images (Figure 3a). To obtain high-level feature representations of visual stimuli that can be linked to neural responses, we use embeddings from a contrastive vision–language model (VLM; e.g., CLIP (Radford et al., 2021)). Specifically, for comparability with BrainSCUBA, we adopt the projection layer embedding of CLIP's vision branch and use the same pretrained checkpoint as reported in that work (see Appendix A.2.3). We also re-implemented BrainSCUBA in-house; note that our implementation differs in the dataset used for the projection step and in the training approach for the encoding model (Appendix A.2.4 for details).

For each image stimulus i, we extract its CLIP-Vision projection-layer embedding $\mathbf{x}_i \in \mathbb{R}^d$, L2-normalize it to unit norm, and pair it with the measured responses across all voxels $\mathbf{y}_i \in \mathbb{R}^v$. The encoding model assumes a linear relationship

$$\mathbf{y}_i = \mathbf{W}\,\mathbf{x}_i + \boldsymbol{\varepsilon}_i,$$

where $\mathbf{W} \in \mathbb{R}^{v \times d}$ represents the voxel-wise encoding parameters and $\boldsymbol{\varepsilon}_i$ captures residual noise. We estimate $\mathbf{W}$ using ridge regression on the NSD training set.

3.2.2 EXPLORATION OF OPTIMAL IMAGE SETS FOR VOXELS

Next, we identify the optimal image set for each voxel (Figure 3b). We compute the inner product between the voxel's encoding weight and CLIP-Vision latent representations from a large-scale external dataset (distinct from NSD) to obtain predicted voxel responses for each image. We then select the top-N images that generate the highest predicted activation. This process is equivalent to calculating the predicted responses of each voxel for every image. This study uses approximately 1.7 million images from OpenImages-v6 (Kuznetsova et al., 2020)

3.2.3 CAPTIONING OPTIMAL IMAGE SETS WITH MLLM

To enable an LLM to interpret each voxel's optimal image set, we first generate captions for these image sets using an MLLM. We use MiniCPM-V (Yao et al., 2024) with the prompt *"Describe the image briefly."* For our accuracy evaluation, we also form a simple baseline by concatenating the top-N captions from the optimal image set.

3.2.4 Generating Voxel Captions

Finally, we generate interpretable voxel captions from the image captions. First, we use an LLM to extract common keywords across the captions within each voxel's optimal image set (Figure 3d). Following the in-context learning prompt approach from (Dunlap et al., 2024), we extract multiple keywords from the caption sets using an LLM (A6 for the prompt). We use *gpt-4o* (gpt-4o-2024–08–06 in the OpenAI API) as the LLM. To remove irrelevant or noisy keywords, we compute the cosine similarity between each keyword's embedding from CLIP' text branch (prompted as *"A photo of {keyword}."*) and the encoding weight for that voxel, then apply a softmax threshold to retain only sufficiently relevant keywords. Hereafter, we refer to CLIP's text branch as "CLIP-Text". Next, we transform these filtered keywords into a sentence-level caption using the "Sentence Composer" from MeaCap (Zeng et al., 2024), initially designed to generate image captions from keyword sets. MeaCap can generate a caption by inputting the target image's keywords into the Sentence Composer while referencing similarities to the image features. In this study, we replace image features with encoding weights so that the model composes a coherent sentence from the voxel-specific keywords (for details, see Section A.2.1).

3.3 Caption Evaluation

3.3.1 Brain Activity Prediction at Sentence Level

A voxel caption that truly reflects a voxel's selectivity should be more similar to the caption of an NSD image that strongly activates that voxel, and less similar to captions of images that do not. We therefore predict voxel-wise brain activity from sentence similarity to evaluate how accurately each caption captures voxel selectivity (Figure A7a). Importantly, this procedure differs conceptually from decoding, which predicts a caption for every stimulus. Following (Singh et al., 2023), we:

1. Use a pretrained Sentence-BERT to compute text embeddings for each voxel caption and each NSD image caption.
2. Compute the cosine similarity between the voxel caption embedding and each NSD image caption embedding.
3. Treat this similarity value as the predicted activity for that voxel on that image.

For each voxel v, we then calculate the Spearman's rank correlation between the vector of predicted sentence-level similarities and the measured activity; this correlation coefficient is regarded as the prediction accuracy for that voxel.

For statistical significance, we use a permutation test to assess voxel-wise prediction accuracy. Multiple comparisons are corrected using the Benjamini–Hochberg false discovery rate procedure ($\alpha = 0.05$). Detailed procedures are provided in Appendix A.2.2.

3.3.2 Brain Activity Prediction at Image Level

Because sentence-based evaluation can be influenced by non-visual linguistic features (e.g., sentence length, clarity of phrasing, or stylistic variation), we also assess voxel selectivity using *image* similarity (Figure A7b). We use FLUX.1-schnell to create a *voxel image* and then compute vision embeddings (via CLIP-Vision) for both the generated voxel image and each NSD trial image. We obtain an image-level metric of predicted brain activity by comparing these embeddings, focusing purely on visual content. Crucially, this procedure is not image reconstruction in the decoding sense; it characterises voxel selectivity through images rather than attempting to recreate the stimuli themselves.

4 Results

4.1 Voxel Activity Prediction

We examine whether LaVCa can generate concise and interpretable voxel captions without losing critical information in each voxel's optimal image set. We compare two approaches from the perspective of interpretability by varying the number of optimal images used by LaVCa (Top-N) and

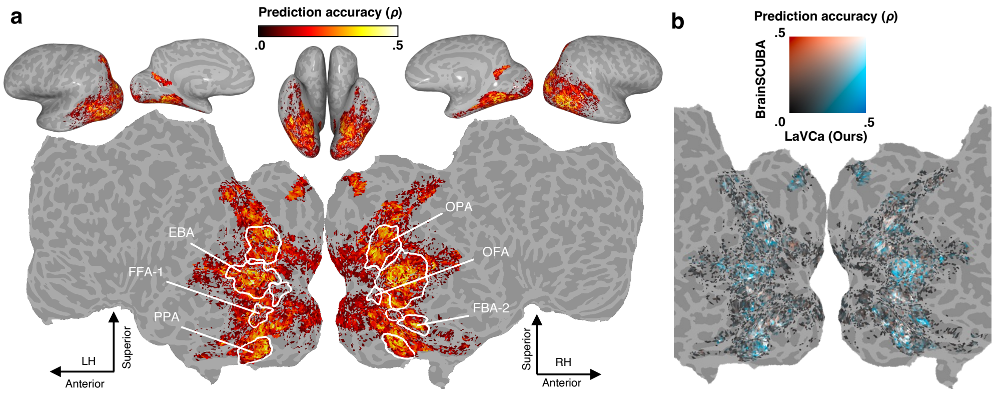


Figure 4: Mapping of brain activity prediction accuracy (subj01). **a** The sentence-level prediction performance is projected onto inflated cortical surfaces (top: lateral, medial, and dorsal views) and flattened cortical surfaces (bottom, with the occipital areas at the center) for both hemispheres. Voxels with significant prediction performance are color-coded (all colored voxels $P < 0.05$, FDR corrected). The white outlines indicate the ROIs that are among the top two in terms of the total voxel count across subjects for each semantic category—Body (Extra Striate Body Area; EBA, and Fusiform Body Area; FBA-2), Face (Fusiform Face Area; FFA-1, and Occipital Face Area; OFA), and Places (Parahippocampal Place Area; PPA, and Occipital Place Area; OPA). Word areas are shown in Figure A8. **b** A comparison of sentence-level prediction performance between our method, LaVCa, and the existing method, BrainSCUBA on the flattened cortical surface.

by simply concatenating the captions of the optimal images (Concat-N). Figure 2 plots prediction accuracy against the average caption length on the horizontal axis, highlighting the trade-off between accuracy and interpretability. Concat-N achieves better accuracy as N increases (up to $N = 10$) but at the cost of a much longer caption, which can reduce interpretability. In contrast, LaVCa merges information across the optimal image set into a concise summary, retaining interpretability even as N grows and reaching accuracy comparable to Concat-N (see Figure A10b). Results for all participants are provided in Figure A10a.

Next, we determine whether the generated captions accurately capture the properties of voxels in the visual cortex. To this end, we map sentence-level prediction accuracy onto both inflated and flattened cortical surfaces (Figure 4a). These maps illustrate that LaVCa captions significantly predict voxel activity throughout the visual cortex ($P < 0.05$, FDR-corrected). Results for all subjects at the sentence and image levels are presented in Figures A8 and A9.

Finally, we compare two configurations of LaVCa—its default five-keyword version with the Sentence Composer and a simplified single-keyword variant without the Sentence Composer—against the existing method BrainSCUBA and a shuffled variant (LaVCa captions shuffled across voxels) at both the sentence and image levels, focusing on the top 5,000 voxels with the highest accuracy on the training data (Table 1). Our proposed method, LaVCa, outperforms BrainSCUBA and the single-keyword variant ($P < 0.05$, paired t-test). This finding suggests that using multiple keywords and composing them into a coherent sentence provides a more accurate explanation of voxel selectivity. Importantly, LaVCa outperforms BrainSCUBA at the *image*-level, indicating that the improvement is not merely due to better handling of non-visual linguistic features (e.g., sentence length or phrasing), but reflects a genuinely enhanced characterization of visual selectivity. Furthermore, LaVCa achieves far higher accuracy than the shuffled condition. Results for the top 1,000, 3,000, and 10,000 voxels appear in Table A8 and A9. After visualizing sentence-level prediction accuracy across the cortex, we find that LaVCa exceeds BrainSCUBA's performance throughout the visual cortex (Figure 4b). See Figure A8 for the results of all subjects.

4.2 Lexical and Semantic Diversity Analysis

We next assess how effectively LaVCa captions capture both lexical and semantic diversity across voxels, focusing first on *inter-voxel* diversity (Table A10, left). For this quantitative evaluation, we

Table 1: Comparison of brain activity prediction accuracy at the sentence and image levels. For each subject, the mean and standard deviation of accuracy on the test data are displayed for the top 5,000 voxels with the highest accuracy on the train data.

			Sentence level			
Model	# keywords	Sentence Composer	subj01	subj02	subj05	subj07
Shuffled	–	–	0.007 ± 0.199	0.058 ± 0.223	0.068 ± 0.243	0.009 ± 0.175
BrainSCUBA	–	–	0.207 ± 0.062	0.251 ± 0.071	0.264 ± 0.084	0.182 ± 0.065
LaVCa (Ours)	1	✗	0.205±0.068	0.250±0.075	0.272±0.086	0.186±0.072
LaVCa (Ours)	5	✓	**0.246 ± 0.066**	**0.287 ± 0.075**	**0.306 ± 0.084**	**0.218 ± 0.073**

			Image level			
Model	# keywords	Sentence Composer	subj01	subj02	subj05	subj07
Shuffled	–	–	0.017 ± 0.163	0.052 ± 0.185	0.066 ± 0.204	0.009 ± 0.149
BrainSCUBA	–	–	0.188 ± 0.067	0.226 ± 0.070	0.250 ± 0.078	0.169 ± 0.069
LaVCa (Ours)	1	✗	0.182 ± 0.063	0.221 ± 0.066	0.252 ± 0.077	0.158 ± 0.064
LaVCa (Ours)	5	✓	**0.213 ± 0.072**	**0.250 ± 0.070**	**0.273 ± 0.079**	**0.187 ± 0.073**

use three metrics: (1) the total vocabulary size (excluding stop-words) across all voxel captions (Lexical); (2) the average variance across each dimension of the CLIP-Text embedding computed on all voxel captions (Semantic); and (3) the number of principal components (PCs) required to capture 90% of the variance of CLIP-Text embedding across captions in a principal component analysis (PCA; Semantic).

First, we evaluate the diversity of LaVCa captions compared with the existing method, BrainSCUBA. When averaged across subjects, LaVCa markedly outperforms BrainSCUBA in both lexical (16,922 vs. 3,193 in vocab. size) and semantic (0.0642 vs. 0.0588 in variance of embeddings; 219 vs. 127 in PCs required for 90% variance explained) diversity. These findings confirm that our open-ended LLM–based approach can produce richer word usage and more meaningful captions across inter-voxel comparisons.

We evaluate the diversity of LaVCa captions compared with more detailed captions. BrainSCUBA leverages ClipCap (Mokady et al., 2021), a model that produces relatively simple image captions. We use the top-1 captions generated by the MLLM on the optimal image sets (equivalent to the case where $N = 1$ in Concat-N) to compare the diversity of LaVCa with more detailed captions. When averaged across subjects, Top-1 (13,959 vocab. size, 0.0638 avg. variance, 210 PCs) exhibits both a vocabulary range and semantic diversity close to LaVCa. However, LaVCa achieves a higher prediction accuracy (0.264 vs. 0.224), indicating that LaVCa can preserve robust brain activity prediction performance while enhancing the diversity of generated captions.

Next, we evaluate diversity from an *intra-voxel* perspective by comparing captions generated by three models in both lexical and semantic dimensions (Table A10, right). We use three metrics: (1) the vocabulary size of each voxel's caption (Lexical), (2) the average sentence length in each voxel's caption (Lexical), and (3) the average variance across all dimensions of Word2Vec embeddings of each caption's words (excluding stop-words) (Semantic). When averaged across subjects, LaVCa markedly outperforms BrainSCUBA in both lexical (11.4 vs. 6.09 in vocab. size) and semantic (11.9 vs. 6.19 in avg. length; 0.0199 vs. 0.0160 in variance of semantic embeddings) diversity. This improvement suggests that LaVCa more precisely captures the fine-grained intra-voxel characteristics.

For examples of voxel captions and images from various OFA and PPA voxels—along with their corresponding quantitative metrics—compared across three models (LaVCa, BrainSCUBA, Top-1), see Figures A14, A15, A16, and A17.

4.3 ROI-LEVEL DIVERSITY ANALYSIS

Our results thus far demonstrate that LaVCa produces more accurate voxel captions than BrainSCUBA and better captures both inter- and intra-voxel diversity. We next ask whether LaVCa can reveal richer representational content inside ROIs that earlier neuroimaging studies have largely described as selective for simpler categories—for example, faces in the OFA or places in the PPA (Gauthier et al., 2000; Haxby et al., 2000; Epstein & Kanwisher, 1998). We conduct a qualitative and quantitative

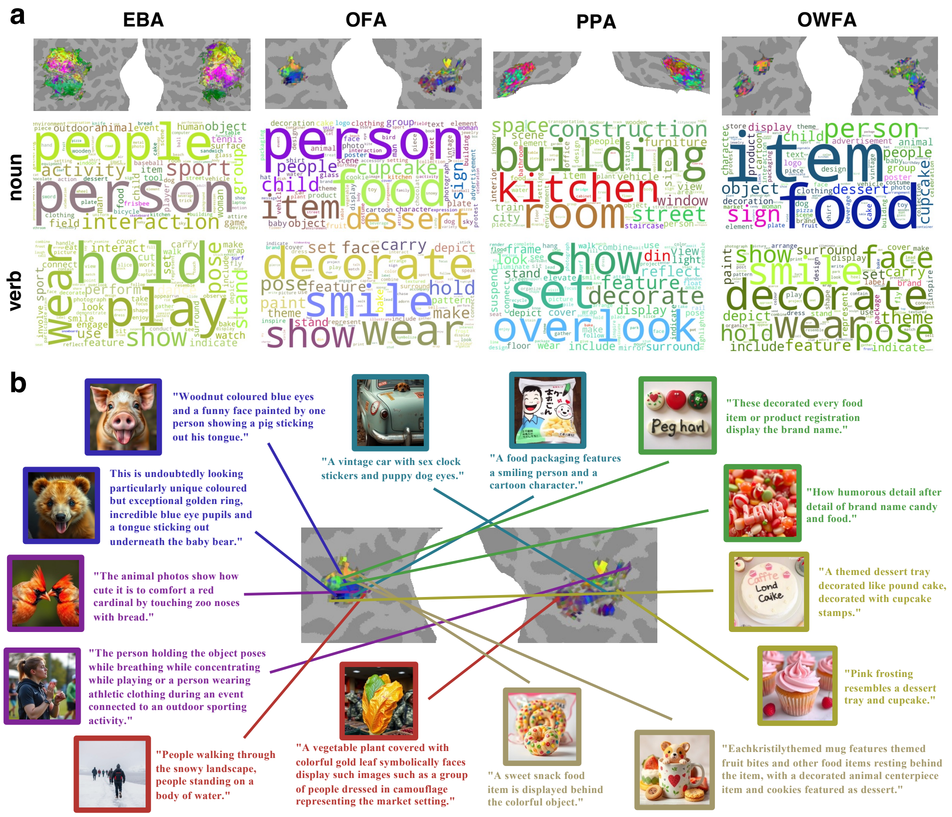


Figure 5: Interpretation of LaVCa captions (subj02). **a** UMAP projection of caption text across four ROIs (EBA, OFA, PPA, OWFA), visualized on a flatmap (top). A word cloud of the 100 most frequent **nouns** in these captions (middle), colored by location in the UMAP space. A word cloud of the 100 most frequent **verbs** (bottom). **b** Visualization of the top two captions (by accuracy) for eight clusters on the flatmap in OFA. The images generated for each caption appear to the left or above the text. Voxels are connected to their corresponding captions and images by lines. The color of each caption and image border reflects the average UMAP color of all voxels in the cluster.

evaluation using LaVCa's captions and generated images to analyze diversity that exists beyond the known selectivity in the ROI.

Qualitative Assessment. We explore the semantic diversity of LaVCa captions across four ROIs (EBA, OFA, PPA, OWFA) by applying UMAP to their CLIP-Text embeddings and visualizing the resulting distributions on a flatmap (Figure 5a, top). In each ROI, we observe a broad spectrum of UMAP colors, indicating multiple meaningful clusters within regions known for distinct category-selective responses. The presence of this broad spectrum is consistent across participants (Figure A18, A19).

Across ROIs, we observe diverse nouns and verbs that not only align with prior selectivity profiles but also reveal richer, voxel-level selectivity for object categories and actions. Both EBA and OFA frequently include common nouns such as people" and person," and the verb distributions highlight ROI-specific action tendencies: EBA is enriched for body-related actions (e.g., hold"), whereas OFA is enriched for face-related actions (e.g., smile"). These patterns are consistent across participants (Figure A18, A19).

Finally, to highlight how each caption and its corresponding voxel image relate to specific colors in semantic space, we project them onto a flatmap (Figure 5b). We divide the samples into eight clusters by labeling each of the three UMAP dimensions as “High” ($\geq 2/3$) or “Low” ($\leq 1/3$). From each

Table 2: Average prediction accuracy with standard error across subjects when captions within each ROI are shuffled (Shuffled) versus used as is (Original).

	Body areas		**Face areas**		**Place areas**		**Word areas**	
Model	EBA	FBA-2	OFA	FFA-1	OPA	PPA	OWFA	VWFA-1
Shuffled	0.018±0.008	0.018±0.005	0.028±0.004	0.016±0.003	0.116±0.024	0.151±0.028	0.025±0.005	0.034±0.009
Original	**0.157±0.005**	**0.125±0.010**	**0.095±0.009**	**0.111±0.003**	**0.200±0.022**	**0.213±0.027**	**0.084±0.013**	**0.158±0.007**

cluster, we pick the two voxels with the highest prediction accuracy (or one if only one qualifies, or none if none qualify) and illustrate their captions and generated images.

In OFA, some captions are related to faces (e.g., “face,” “person,” “animal”), while particular voxels encoded more fine-grained features such as “eye,” “tongue,” or “smiling,” and other voxels encoded information like “animal,” “bear,” or “cardinal.” Thus, even within this ROI, there appears to be substantial functional differentiation among inter-voxel that extends beyond a generic “face” category.

Moreover, we observe *intra-voxel* diversity, where a single caption incorporates multiple ideas (e.g., *“A food packaging features a smiling person and a cartoon character”*), suggesting that individual voxels can simultaneously encode several distinct concepts. These findings highlight the fine-grained functional specialization across inter-voxel within the ROI and the diverse nature of intra-voxel encoding beyond singular concepts.

The results for all participants, visualizing the top two captions for each cluster directly in the UMAP space, can be found in Figures A20, A21, A22, and A23.

Quantitative Assessment. We next determine how many distinct captions appear in each ROI by comparing the sentence-level prediction accuracy of each ROI when captions are maintained in their original form versus shuffled within the ROI (Table 2). For each category (body, face, place, and word area), we select two ROIs with the largest total voxel count across all subjects, resulting in eight ROIs in total. In all ROIs, shuffling reduces prediction accuracy significantly. For example, in the OFA, accuracy drops from 0.0945 (Original) to 0.0280 (Shuffled), a 3.3-fold decrease; in the PPA, accuracy falls from 0.213 (Original) to 0.151 (Shuffled), a 1.4-fold decrease. Thus, even in regions traditionally linked to particular concepts, voxels exhibit a range of distinct selectivities. Furthermore, the average caption similarity between the same ROIs of different subjects is relatively high at 0.227, compared to 0.171 between different ROIs of different subjects, indicating that such diversity is reproducible across subjects (Figure A11).

Next, we quantify how many different semantic concepts a single voxel can encode (i.e., its degree of multi-concept selectivity). We perform the following analysis: (1) extract every unique noun from all voxel captions within the ROI; (2) obtain CLIP-Text embeddings for each noun using the prompt “A photo of $\{\text{word}\}$.” and cluster them with k-means ($k = 6$); (3) for each voxel, count how many of its nouns fall into different clusters. Across all ROIs, we find that most voxels are associated with multiple clusters, indicating multi-concept selectivity (Table A12). Thus, even within ROIs whose vocabulary is relatively coherent, individual voxels can encode several distinct concepts. Furthermore, by aggregating the nouns used in this clustering analysis from all subjects and examining the extent to which each subject’s voxels belong to the subject-shared clusters, we evaluate the cross-subject reproducibility of ROI diversity (Figure A12). In both the OFA and PPA, voxels from all subjects populate the same clusters, suggesting that such diversity is, to some extent, consistent across individuals.

5 Discussion & Conclusions

In this study, we introduce a novel method called LaVCa, which leverages LLMs to produce data-driven, natural-language descriptions of voxel selectivity in the human visual cortex. The voxel captions generated by LaVCa exhibit higher accuracy and greater semantic diversity than those generated by the existing approach, BrainSCUBA. We attribute this improvement to our mechanism for integrating multiple keywords extracted by advanced LLMs, which enables a more comprehensive capture of the diverse selectivity patterns across voxels. Furthermore, LaVCa uncovers richer representational content within ROIs that earlier neuroimaging studies had characterized as selective for simpler categories. By revealing that even “category-selective” areas such as the OFA and PPA encode a broader spectrum of concepts, our findings challenge long-standing assumptions about functional specialization in the visual cortex. See Sections A.6 and A.7 for the Limitation and Impact Statement.

Ethics Statement

This study did not involve the collection of any new neural recording data. Instead, we relied exclusively on the Natural Scenes Dataset (NSD), which is openly accessible to the research community. The dataset can be obtained from `https://naturalscenesdataset.org/` , subject to their terms of use.

We conducted all analyses on this publicly released dataset and did not handle any personally identifiable information. Based on the nature of the data and the scope of our methods, we do not anticipate harmful applications of this work.

Reproducibility Statement

We have made substantial efforts to ensure the reproducibility of our results. Details of the LaVCa pipeline, including model architecture, training procedure, and evaluation metrics, are described in Section 3. Additional implementation details, hyperparameter settings, and preprocessing steps for the Natural Scenes Dataset (NSD) are provided in the Appendix. Furthermore, we include anonymized source code and scripts as part of the Supplementary Material to facilitate reproduction of our experiments.

LLM Usage

In accordance with the ICLR policy on the use of large language models (LLMs), we report that LLMs were employed exclusively for language-related assistance. Specifically, we used LLMs to aid in the translation of text into English and to polish the grammar and style of the manuscript. All research ideas, experimental design, data analysis, and scientific interpretations were conceived and conducted entirely by the authors.

The use of LLMs did not contribute to the formulation of research questions, methodology, or conclusions. The authors take full responsibility for the final content of the paper, including all text that was assisted by LLM-based tools.

References

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36, 2024.

Nicholas Bai, Rahul A Iyer, Tuomas Oikarinen, and Tsui-Wei Weng. Describe-and-dissect: Interpreting neurons in vision networks with language models. *arXiv preprint arXiv:2403.13771*, 2024.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Kirill Bykov, Laura Kopf, Shinichi Nakajima, Marius Kloft, and Marina Höhne. Labeling neural representations with inverse recognition. *Advances in Neural Information Processing Systems*, 36, 2024.

Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):49, 2019.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Bridging the gap between brain and machine in interpreting visual semantics: Towards self-adaptive brain-to-text decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21938–21948, 2025a.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Mindgpt: Interpreting what you see with non-invasive brain recordings. *IEEE Transactions on Image Processing*, 2025b.

Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24199–24208, 2024.

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.

Isabel Gauthier, Michael J Tarr, Jill Moylan, Pawel Skudlarski, John C Gore, and Adam W Anderson. The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of cognitive neuroscience*, 12(3):495–504, 2000.

Umut Güçlü and Marcel AJ Van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

James V Haxby, Elizabeth A Hoffman, and M Ida Gobbini. The distributed human neural system for face perception. *Trends in cognitive sciences*, 4(6):223–233, 2000.

Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*, 2021.

Nhat Hoang-Xuan, Minh Vu, and My T Thai. Llm-assisted concept discovery: Automatically identifying and explaining neuron functions. *arXiv preprint arXiv:2406.08572*, 2024.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, 2016.

Neha Kalibhat, Shweta Bhardwaj, C Bayan Bruss, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. Identifying interpretable subspaces in image representations. In *International Conference on Machine Learning*, pp. 15623–15638. PMLR, 2023.

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

Tom Dupré La Tour, Michael Eickenberg, Anwar O Nunez-Elizalde, and Jack L Gallant. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728, 2022.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G Huth. A natural language fmri dataset for voxelwise encoding models. *Scientific Data*, 10(1):555, 2023.

Mark D Lescroart and Jack L Gallant. Human scene-selective areas represent 3d configurations of surfaces. *Neuron*, 101(1):178–192, 2019.

Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. Factual: A benchmark for faithful and consistent textual scene graph parsing. *arXiv preprint arXiv:2305.17497*, 2023.

Andrew F Luo, Margaret M Henderson, Michael J Tarr, and Leila Wehbe. Brainscuba: Fine-grained natural language captions of visual cortex selectivity. *arXiv preprint arXiv:2310.04420*, 2023.

Andrew F Luo, Jacob Yeung, Rushikesh Zawar, Shaurya Dewan, Margaret M Henderson, Leila Wehbe, and Michael J Tarr. Brain mapping with dense features: Grounding cortical semantic selectivity in natural images with vision transformers. *arXiv preprint arXiv:2410.05266*, 2024.

Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

Jesse Mu and Jacob Andreas. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163, 2020.

Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Q Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. Unveiling multi-level and multi-modal semantic representations in the human brain using large language models. *bioRxiv*, pp. 2024–02, 2024.

Tomoya Nakai and Shinji Nishimoto. Quantitative models reveal the organization of diverse cognitive functions in the brain. *Nature communications*, 11(1):1142, 2020.

Tomoya Nakai and Shinji Nishimoto. Representations and decodability of diverse cognitive functions are preserved across the human cortex, cerebellum, and subcortex. *Communications Biology*, 5(1): 1245, 2022.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. *arXiv preprint arXiv:2204.10965*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Chandan Singh, Aliyah R. Hsu, Richard Antonello, Shailee Jain, Alexander G. Huth, Bin Yu, and Jianfeng Gao. Explaining black box text modules in natural language with language models, 2023. URL https://arxiv.org/abs/2305.09863.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14453–14463, 2023.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Tung-Yu Wu, Yu-Xiang Lin, and Tsui-Wei Weng. And: Audio network dissection for interpreting deep acoustic models. *arXiv preprint arXiv:2406.16990*, 2024.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. *arXiv preprint arXiv:2505.05071*, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint 2408.01800*, 2024.

Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. Meacap: Memory-augmented zero-shot image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14100–14110, 2024.

A Appendix

A.1 Full Related Work

A.1.1 Interpreting the representations of the brain's neurons.

Encoding models have long been used in neuroscience to interpret neural representations within the brain (Kay et al., 2008; Nishimoto et al., 2011; Naselaris et al., 2011; Huth et al., 2012). These studies used interpretable features, such as low-level visual attributes, or high-level semantic features, such as one-hot encoding of words, for straightforward voxel-wise interpretation.

Recent approaches use features derived from DNNs and have demonstrated higher explanatory power for brain activity than those using simpler, more interpretable features (Güçlü & Van Gerven, 2015; Schrimpf et al., 2021; Takagi & Nishimoto, 2023; Antonello et al., 2024). However, the interpretability of these DNN-based encoding models remains challenging, leading to the development of methods that condense the entire set of voxels into a small number of universal and interpretable axes (Huth et al., 2016; Lescroart & Gallant, 2019; Nakagi et al., 2024).

Recent approaches propose data-driven methods to describe the properties of individual brain voxels using natural language (Luo et al., 2023; Singh et al., 2023) when analyzing brain representations at a finer, voxel-wise level. BrainSCUBA (Luo et al., 2023) is an end-to-end method that uses an existing image captioning model, which provides voxel-wise captions of the visual cortex in a data-driven manner. BrainSCUBA projects each voxel‘s encoding weight onto the image feature space via dot-product attention, identifies regions of highest similarity, and then uses a text decoder to generate captions describing the images to which the voxel is most selective. This approach provides a data-driven natural-language description of voxel selectivity without additional training. Similarly, SASC (Singh et al., 2023) uses fMRI data collected during speech listening (LeBel et al., 2023) to identify the short phrases that most strongly activate each voxel. It then uses an LLM to combine these short phrases into a single, data-driven caption describing each voxel's semantic properties.

Our proposed method also generates data-driven voxel captions but differs in several ways. First, BrainSCUBA is constrained to pre-existing, end-to-end image captioning models. In contrast, our approach divides the task into (i) identifying an optimal set of images and (ii) converting these images into a caption, allowing us to use any vision model aligned with language and any LLM with advanced language capabilities without requiring specialized fine-tuning. Furthermore, although SASC uses an LLM to create voxel captions, it primarily synthesizes short, low-information phrases (e.g., trigrams), producing only simple keyword-based captions. In contrast, our method summarizes more diverse and informative text and then uses these extracted keywords to compose a complete sentence, capturing a richer range of voxel-level properties.

A.1.2 Interpreting the Representations of Artificial Neurons in DNNs

Interpreting artificial neurons is a key challenge in understanding how DNNs process information. We can potentially examine human neural representations at a finer granularity by applying the data-driven and highly accurate interpretation methods developed for artificial neurons to analyze human brain voxels.

Numerous studies have aimed to associate artificial neurons with human-interpretable concepts (Bau et al., 2017; Mu & Andreas, 2020; Oikarinen & Weng, 2022; Kalibhat et al., 2023; Bykov et al., 2024). These methods link neurons to textual concepts by comparing neuron output feature maps with outputs from segmentation models. However, these approaches are constrained by predefined concept sets or limited to the dataset's words and phrases. MILAN (Hernandez et al., 2021) introduced a generative approach, enabling adaptation to different domains and tasks, but it requires annotated data, which poses challenges for scalable applications.

LLMs permit open-ended descriptions of artificial neurons without additional model training (Singh et al., 2023; Bai et al., 2024; Wu et al., 2024; Hoang-Xuan et al., 2024). Analogous to these methods, our study also leverages LLMs to generate open-ended concepts for **brain** neurons rather than artificial neurons, seeking flexible and diverse interpretations that do not depend on predefined vocabularies.

A.2 Implementation Details

A.2.1 Generating Voxel Captions

In this study, we leverage the “Sentence Composer” proposed in the image captioning model MeaCap (Zeng et al., 2024)—referred to as the “keywords-to-sentence LM” in the original paper—to generate sentence-level captions from keywords.

Notation

- $K = \{k_1, \dots, k_m\}$: keyword set extracted by an LLM.
- $W \in \mathbb{R}^d$: voxel-wise encoding weight.
- $\{\tilde{c}_1, \dots, \tilde{c}_k\}$: top-$k$ captions of the voxel’s optimal images.
- $\phi_T(\cdot)$: CLIP-Text embedding operator.
- $\mathrm{sim}(\mathbf{u}, \mathbf{v}) = \dfrac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}$: cosine similarity.

Iterative Decoding Procedure We begin with an *initial draft caption*

$$\mathbf{c}^{(0)} = [k_1 \; k_2 \; \dots \; k_m],$$

obtained by concatenating the keywords K in their given order (separated by spaces). Starting from this seed, CBART iteratively refines the caption through the following steps:

1. **Action prediction.** For each position j in the current draft $\mathbf{c}^{(t)}$, the encoder assigns one of the actions Copy, Replace, or Insert.
2. **Candidate generation.** At positions marked for replacement or insertion, the decoder proposes the top-n lexical candidates $\mathcal{W}_j = \{w_{j,1}, \dots, w_{j,n}\}$ ranked by the token likelihood $P_\theta(w \mid \mathbf{c}^{(t)}_{<j})$.
   - *Fluency* ($\log P_\theta$) ensures linguistic naturalness.
   - *Image relevance* grounds the sentence in visual evidence from the voxel’s optimal images.
   - *Voxel relevance* ($\mathrm{sim}(\phi_T(w), W)$) links the caption to the voxel’s representation.

   Each candidate $w \in \mathcal{W}_j$ is scored by

   $$S(w) = \lambda_1 \log P_\theta\big(w \mid \mathbf{c}^{(t)}_{<j}\big) + \lambda_2 \frac{1}{k} \sum_{i=1}^{k} \mathrm{sim}\big(\phi_T(w), \phi_T(\tilde{c}_i)\big) + \lambda_3 \, \mathrm{sim}\big(\phi_T(w), W\big), \quad \text{(A1)}$$

   where we set $(\lambda_1, \lambda_2, \lambda_3) = (0.2,\ 0.2,\ 1.2)$.
3. **Token selection and refinement.** The token with the highest $S(w)$ replaces or is inserted at position j, yielding the updated draft $\mathbf{c}^{(t+1)}$. The loop repeats until every position is predicted as Copy, producing the final caption $\hat{\mathbf{c}}$.

Rationale

- *Fluency* $\big(\log P_\theta\big)$ encourages linguistic naturalness.
- *Image relevance* $\big(\frac{1}{k} \sum_{i=1}^{k} \mathrm{sim}\big(\phi_T(w), \phi_T(\tilde{c}_i)\big)\big)$ grounds the sentence in visual evidence drawn from the voxel’s optimal images.
- *Voxel relevance* $\big(\mathrm{sim}\big(\phi_T(w), W\big)\big)$ ties the caption to the voxel’s learned representation.

By jointly optimizing the score in equation A1, the method transforms discrete keyword sets into a coherent sentence that is *linguistically natural*, *visually grounded*, and *specifically aligned* with the voxel’s weights W.

Table A1: Pretrained checkpoints used in our experiments.

Category	Model	Repository (Hugging Face)
Contrastive VLM	CLIP	`openai/clip-vit-base-patch32`
	SigLIP2	`google/siglip2-base-patch16-224`
	FG-CLIP	`qihoo360/fg-clip-base`
MLLM	MiniCPM-Llama3-V2.5	`openbmb/MiniCPM-Llama3-V-2_5`
	BLIP	`Salesforce/blip-image-captioning-base`
LLM	gpt-4o	*N/A (OpenAI API)*
	Llama 3.1-70B	`meta-llama/Llama-3.1-70B-Instruct`
Sentence Similarity Model	Sentence-BERT	`sentence-transformers/all-MiniLM-L6-v2`
	MPNet	`sentence-transformers/all-mpnet-base-v2`
Text-to-Image Model	FLUX.1-schnell	`black-forest-labs/FLUX.1-schnell`

A.2.2 Statistical Testing

For each voxel v, the observed prediction accuracy is quantified as Spearman's rank correlation ρ_v^{obs}. To realize the null hypothesis of no association, the activity vector is randomly permuted B times, yielding surrogate correlations $\{\rho_{v,b}^{\text{null}}\}_{b=1}^{B}$. Pooling across all N voxels produce a global null distribution. The one-tailed p-value is computed as

$$p_v = \frac{\#\{\rho_{v,b}^{\text{null}} \geq \rho_v^{\text{obs}}\} + 1}{B \times N + 1}.$$

We control for multiple comparisons using the Benjamini–Hochberg false-discovery-rate (FDR) procedure ($\alpha = 0.05$); voxels with $q < 0.05$ are declared significant. In our experiments, we set $B = 1000$.

A.2.3 Pretrained Checkpoints

We rely on a variety of pretrained models for different components of our pipeline. Most of the models are publicly available checkpoints hosted on Hugging Face, including contrastive vision–language models (CLIP, SigLIP2, FG-CLIP), multimodal LLMs (MiniCPM-Llama3-V2.5, BLIP), LLM (Llama 3.1-70B), sentence similarity models (Sentence-BERT, MPNet), and a text-to-image model (FLUX.1-schnell). For keyword extraction, we use gpt-4o, which is not available as a checkpoint but is accessed via the OpenAI API. A complete list of all models and their repositories is summarized in Table A1.

A.2.4 BrainSCUBA

At the outset of our project in January 2025, the original BrainSCUBA codebase had not yet been released, so we implemented the method ourselves for this study. In BrainSCUBA, the encoding weights (linear layers) are learned using gradient descent. In our implementation, consistent with our proposed method, we trained the encoding weights using L2-regularized linear regression from the *himalaya* library package[1] (La Tour et al., 2022).

Moreover, BrainSCUBA projects each voxel‘s encoding weight into image space using a dataset of 2 million images, combining OpenImages (Kuznetsova et al., 2020) and LAION-A v2 (6+ subset) (Schuhmann et al., 2022). However, the specific images selected from each dataset are not disclosed. We ensure a fair and dataset-independent comparison by relying solely on the 1.7 million images from OpenImages (the same dataset used by our proposed method, LaVCa). We leverage the training set of the subset that is accompanied by bounding boxes, object segmentations, visual relationships, and localized narratives.

For other hyperparameters, we tested temperature values of 1.0, $1/10$, $1/100$, $1/150$ (the value used in the BrainSCUBA paper), and $1/500$ for the softmax projection (Figure A1). We used beam search with a beam width of 5 to generate the text decoder's caption as described in the BrainSCUBA paper.

[1] https://github.com/gallantlab/himalaya

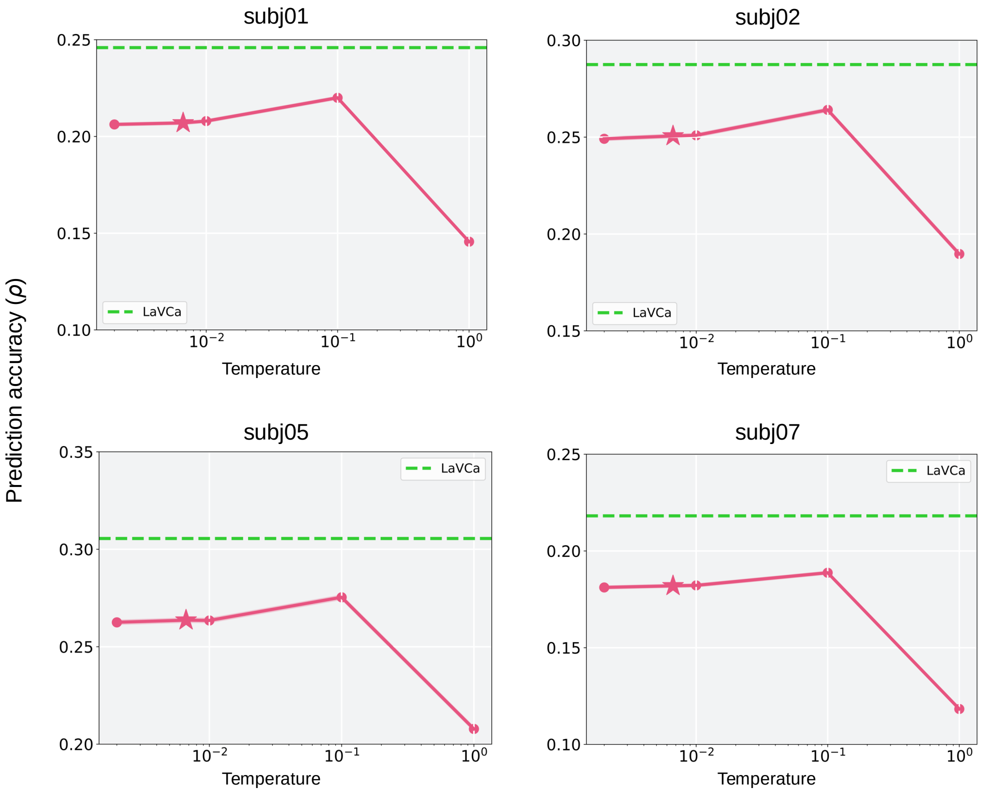


Figure A1: The change in accuracy of BrainSCUBA with respect to temperature. The error bars represent standard error. Moreover, the star markers on the plot indicate the point where the temperature is set to 1/150, as adopted in the original BrainSCUBA paper. The green line represents the average value of LaVCa.

A.2.5 Used Compute Resources

All experiments are conducted on a single Lambda Labs cloud instance equipped with eight NVIDIA A100-SXM4 GPUs (40 GB each). The host system features a dual-socket AMD EPYC 7542 processor, providing 124 logical CPU cores and 512 GB of DDR4 RAM.

In this setting, voxel-wise Ridge-regression training for one subject finishes in ~7 s and occupies 5.4 GB of GPU memory. Loading the 1.7 M candidate images for optimal-image search, performed on the CPU, takes ~1,490 s per subject. The subsequent per-voxel pipeline—optimal-image search, multimodal-LLM captioning, keyword extraction, and SentenceComposer generation—runs in ~15 s per voxel with a peak GPU footprint of 5.0 GB. Processing 20,000 voxels for a single subject therefore requires ~83 h end-to-end, which remains practical for offline analyses in systems neuroscience.

Table A2: Hyper-parameter settings for gradient-descent training.

Parameter	Value
Optimizer	AdamW ($\beta_1 = 0.9,\ \beta_2 = 0.999$)
Batch size	64
Initial learning rate (η_0)	3×10^{-4}
Final learning rate (η_T)	1.5×10^{-4}
Scheduler	Exponential decay
Decay factor (γ)	$(\eta_T/\eta_0)^{1/50} = 0.87$
Weight decay	2×10^{-2}
Maximum epochs	50
Early-stopping patience	5 epochs

Table A3: Comparison of prediction accuracy for the encoding model. For each subject, the mean ± standard error on the test set is reported for the top 5,000 voxels that achieve the highest accuracy on the train data.

Methods	ridge	gradient	layers	voxel-wise	shared-weights	subj01	subj02	subj05	subj07
Linear	✓	–	1	✓	–	0.501±0.002	0.524±0.001	0.570±0.001	0.421±0.001
	✓	–	1	–	✓	0.500±0.002	0.523±0.001	0.567±0.001	0.420±0.001
	–	✓	1	–	✓	0.484±0.002	0.512±0.002	0.563±0.001	0.405±0.002
Nonlinear	–	✓	2	–	✓	0.492±0.002	0.516±0.001	0.565±0.001	0.411±0.002
	–	✓	3	–	✓	0.491±0.002	0.518±0.001	0.563±0.001	0.411±0.002

A.3 Ablation Study

A.3.1 Encoding Model

We conduct an ablation study to evaluate how hyper-parameter choices affect prediction accuracy for the encoding model (Table A3). Here, "accuracy" denotes the correlation coefficient obtained when predicting brain activity with the encoding model, rather than a voxel-caption–based metric).

Voxel-wise versus shared weights. We compare two variants of ridge regression: a *voxel-wise* version that learns an individual weight vector for each voxel and a *shared-weights* version that learns a single weight vector common to all voxels. The regularisation coefficient λ is sampled at 25 logarithmically spaced points from 10^{-4} to 10^{20}, and the optimal value is selected via five-fold cross-validation. This comparison reveals almost no difference in accuracy between the two variants.

Ridge regression versus gradient descent. We next contrast linear layers trained with ridge regression against those trained with gradient descent. The hyper-parameter settings for gradient-descent training are summarised in Table A2. During optimisation, the final 10 % of the training data are held out for validation, and early stopping is triggered if validation accuracy fails to improve for five consecutive epochs. To ensure a fair comparison, ridge regression is fitted using only the first 90 % of the training data, so that both methods see an identical amount of data. Under the *shared-weights* setting, ridge regression outperforms gradient descent, indicating that an analytically derived linear solution is more effective for training the encoding model.

Linear versus non-linear models. Finally, we compare linear and non-linear architectures. For the non-linear networks, the hidden-layer width is set equal to the input dimensionality (512, corresponding to the CLIP-Vision feature size). Introducing non-linearities yields higher accuracy than the single-layer linear model trained with gradient descent; however, increasing the depth from two to three layers offers no further benefit. Moreover, none of the non-linear networks surpass the performance of ridge regression. These findings indicate that, although non-linear models trained with gradient descent can exceed their linear counterparts, they still fall short of the accuracy achieved by ridge regression.

Table A4: Sentence-level prediction accuracy for β-based versus encoding-model–based models. Values denote mean ± standard deviation across test images for the top 5,000 voxels per subject, selected based on their training-set prediction accuracy.

Condition	subj01	subj02	subj05	subj07
w/o Encoding model	0.201 ± 0.082	0.248 ± 0.086	0.276 ± 0.096	0.094 ± 0.054
w/ Encoding model	0.246 ± 0.066	0.287 ± 0.075	0.306 ± 0.084	0.218 ± 0.073

A.3.2 Caption Generation

We investigate how various hyper-parameter settings influence the accuracy of the voxel captions (Tables A4, A5). Unless otherwise specified, we adopt the primary-analysis configuration: CLIP-Vision as the default contrastive VLM for feature extraction, 50 optimal images per voxel, MiniCPM-V as the MLLM for captioning the optimal images, five extracted keywords, *gpt-4o* as the keyword extraction model, exemplar-based prompting as the prompting strategy, and use of the Sentence Composer.

Effect of the Encoding Model Our goal in LaVCa is to characterise voxel selectivity at the level of DNN-derived visual features, rather than to interpret voxels directly from raw β-values. Nevertheless, to verify that the encoding model indeed improves interpretability—by suppressing noise and non-visual components in the β-responses and enabling better generalisation to novel stimuli—we performed an additional control analysis using a purely β-driven baseline.

To directly assess whether the encoding model is necessary for identifying optimal images, we conducted a control analysis in which, for each voxel, we selected the NSD images that elicited the strongest raw β-values and generated voxel captions from their MS-COCO captions. Importantly, in this baseline the encoding-model weights used in the later stages—keyword filtering and the Sentence Composer—are replaced with a uniform weighting scheme by setting all weights to 1. This ensures that the entire pipeline operates without any feature-level information from the encoding model, providing an interpretation grounded solely in the measured β-values.

This β-based approach produced substantially lower predictive accuracy compared to LaVCa (Table A4). Because this baseline relies solely on raw β-values, it is highly sensitive to measurement noise and non-visual components of brain activity, resulting in captions that generalised poorly to held-out images.

Contrastive VLM comparison. We next compare different contrastive vision–language models (VLMs) used to extract latent features for voxel-wise encoding. Specifically, we evaluate SigLIP2 (Tschannen et al., 2025), FG-CLIP (Xie et al., 2025), and both the text and vision branches of CLIP (Radford et al., 2021). For CLIP-Text, we obtain projection-layer embeddings from the COCO captions that were pre-assigned to the NSD image stimuli. Overall, CLIP-based representations (CLIP-Text and CLIP-Vision) achieve the highest accuracies. FG-CLIP performs comparably to CLIP-Vision, while SigLIP2 does not surpass CLIP in our setting—despite reports in the original SigLIP work that it often outperforms CLIP on benchmark tasks—yet its inclusion demonstrates that LaVCa generalises well across diverse VLM backbones.

Number of optimal images. We vary the number of optimal images used for keyword extraction from 5 to 10, 50, and 100. Increasing the number up to 50 improves accuracy, presumably because relying only on top-ranked images can omit useful second- and third-ranked keywords. Using more images therefore captures a broader range of selective concepts. However, once the number of optimal images reaches 100, the improvement plateaus, likely because additional concepts can no longer be adequately represented with only five keywords. These observations suggest that increasing the number of keywords, rather than merely adding more images, may further enhance accuracy.

Multimodal LLM comparison. We compare two multimodal LLMs for captioning the optimal images: the state-of-the-art MiniCPM-V and the lighter, less accurate BLIP. MiniCPM-V slightly outperforms BLIP, indicating that a more capable MLLM can further improve LaVCa's voxel-caption

Table A5: Comparison of sentence-level brain activity prediction accuracy using different hyperparameters. Accuracy is reported as the mean ± standard deviation for the top 5,000 voxels in the test data, selected by training-set accuracy.

Parameter	Setting	subj01	subj02	subj05	subj07
Contrastive VLM	SigLIP2	0.232±0.065	0.275±0.075	0.294±0.083	0.206±0.070
	FG-CLIP	0.244±0.070	0.285±0.076	0.306±0.086	0.214±0.074
	CLIP-Text	0.246±0.067	0.281±0.074	0.309±0.084	0.216±0.071
	CLIP-Vision	0.246±0.066	0.287±0.075	0.306±0.084	0.218±0.073
# optimal images	5	0.239±0.068	0.279±0.073	0.294±0.083	0.209±0.073
	10	0.243±0.068	0.281±0.072	0.297±0.083	0.212±0.074
	50	0.246±0.066	0.287±0.075	0.306±0.084	0.218±0.073
	100	0.246±0.067	0.285±0.074	0.301±0.083	0.215±0.072
Multimodal LLM	MiniCPM-V	0.246±0.066	0.287±0.075	0.306±0.084	0.218±0.073
	BLIP	0.242±0.068	0.285±0.075	0.302±0.084	0.215±0.072
# keywords	1	0.237±0.066	0.274±0.073	0.295±0.085	0.207±0.072
	5	0.246±0.066	0.287±0.075	0.306±0.084	0.218±0.073
	10	0.241±0.067	0.279±0.074	0.296±0.084	0.212±0.074
Extraction model	TextGraphParser	0.242±0.067	0.276±0.073	0.296±0.084	0.205±0.071
	Llama3.1-70B	0.238±0.067	0.281±0.075	0.298±0.085	0.214±0.073
	gpt-4o	0.246±0.066	0.287±0.075	0.306±0.084	0.218±0.073
Prompt	Hidden CoT	0.239±0.064	0.283±0.074	0.298±0.083	0.214±0.072
	Exemplar-based	0.246±0.066	0.287±0.075	0.306±0.084	0.218±0.073
Sentence Composer	✓	0.246±0.066	0.287±0.075	0.306±0.084	0.218±0.073
	✗	0.230±0.066	0.279±0.078	0.296±0.087	0.201±0.070

accuracy. Conversely, the modest gap between BLIP and MiniCPM-V suggests that our approach generalises well even with simpler captioning models.

Number of extracted keywords. With the number of optimal images fixed at 50, we vary the number of extracted keywords among 1, 5, and 10. Increasing the output concepts from one to five boosts accuracy, whereas extending the list to ten decreases accuracy—likely because irrelevant or noisy concepts are introduced. The improvement at five keywords indicates that voxels encode multiple concepts, but extracting too many can introduce noise. Thus, expanding the image set rather than the keyword count may be a more effective strategy for capturing additional informative concepts.

Keyword-extraction model comparison. We evaluate three models for extracting keywords from the optimal image set: *gpt-4o*, an 8-bit-quantised Llama 3.1-70B-Instruct, and the TextGraphParser Li et al. (2023) employed in MeaCap. *gpt-4o* surpasses TextGraphParser, showing that an open-ended LLM makes concept extraction more effective than simply pulling words from captions. It also exceeds Llama 3.1-70B-8bit, demonstrating that stronger LLMs can further raise accuracy. These results imply that LaVCa's interpretability will improve in tandem with future advances in LLM capability.

Prompt comparison. We compare two prompting strategies for summarising the captions associated with each voxel's optimal image set: *Exemplar-based Prompting* and a more structured *Hidden CoT Prompting* formulation. Exemplar-based Prompting follows the classical in-context learning paradigm (Brown et al., 2020; Dunlap et al., 2024), where multiple positive and negative examples constrain the output format and guide the model toward producing concept-like captions. In contrast, the structured prompt incorporates an expert role assignment and an internal ("hidden") chain-of-thought instruction (Wei et al., 2022), encouraging the model to silently derive recurrent patterns before producing the final set of concepts.

Across subjects, both prompting strategies performed similarly, indicating that LaVCa is robust to prompt specification. Nonetheless, Exemplar-based Prompting consistently achieved slightly higher

Table A6: Comparison of sentence similarity models for accuracy evaluation. Accuracy is reported as the mean ± standard deviation for the top 5,000 voxels in the test data, selected by training-set accuracy.

Model	subj01	subj02	subj05	subj07
MPNet	0.245±0.067	0.287±0.075	0.305±0.088	0.216±0.069
Sentence-BERT	0.246±0.066	0.287±0.075	0.306±0.084	0.218±0.073

accuracies for all subjects, suggesting that explicit examples remain an effective mechanism for stabilising the output structure in this caption-concept abstraction task.

Effect of the Sentence Composer. Finally, we assess the Sentence Composer by comparing results with and without it. Incorporating the Sentence Composer yields higher accuracy than relying on keywords alone, suggesting that contextual information beyond isolated concepts enables a more fine-grained interpretation of voxel properties.

A.3.3 Sentence similarity model for evaluation

We also examine how the choice of sentence similarity model used for evaluation affects voxel-caption accuracy (Table A6). Specifically, we compare MPNet (Song et al., 2020) and Sentence-BERT (Reimers & Gurevych, 2019), both widely used models for computing sentence embeddings. Overall, the two models yield nearly identical accuracies across all subjects, with Sentence-BERT performing marginally better. This consistency suggests that LaVCa's evaluation results are robust to the particular choice of sentence similarity model, and that the observed improvements are not dependent on model-specific idiosyncrasies.

A.4 Data-size Sensitivity

In this appendix, we examine how LaVCa's caption-prediction accuracy depends on the amount and structure of the stimulus data used within the NSD dataset. We vary (i) the number of stimulus images supplied to LaVCa and (ii) the number of stimulus-image categories used as part of the method's categorization stage.

A.4.1 Sample-size Sensitivity

To evaluate how the size of the NSD dataset influences LaVCa's performance, we vary the number of stimulus images provided to the method across 1, 10, 100, 500, 1000, 5000, and approximately 9000 (the full NSD stimulus set). Figure A2 summarizes how prediction accuracy changes across these sample sizes.

The number of stimulus-image categories is fixed at 80, based on the COCO object categories, except when fewer than 80 images are used, in which case the number of categories matches the sample size.

LaVCa's accuracy shows a clear dependence on the number of available stimulus images. With only a small number of images (such as 1 or 10), accuracy decreases substantially because the limited variety of stimuli restricts the method's ability to infer stable relationships between fMRI responses and visual content. As more images are incorporated, accuracy improves rapidly and stabilizes once approximately 500 to 1000 stimulus images are available. Increasing the stimulus dataset beyond this range yields only modest gains, and the performance with the full ~9000-image set is only slightly higher than that achieved with 5000 images. These results indicate that roughly 500–1000 stimulus images are sufficient for LaVCa to produce reliable voxel-level captions, without requiring the full scale of the NSD stimulus set.

A.4.2 Category-size Sensitivity

We next analyze how the number of stimulus-image categories provided to LaVCa affects caption-prediction accuracy. Figure A3 provides a summary of this analysis. The number of categories is

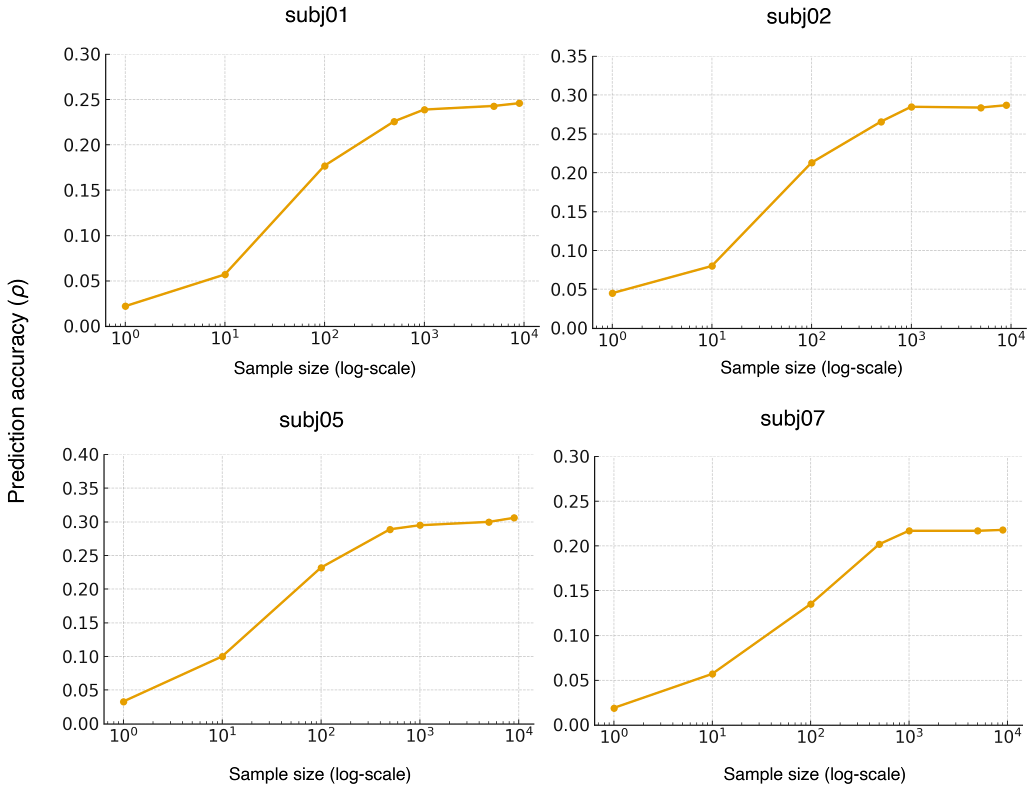


Figure A2: Effect of sample size on test prediction accuracy. Mean prediction accuracy on the test data for the top 5000 voxels selected using training-data performance, plotted across different sample sizes (x-axis in log scale).

varied across 2, 10, 40, and 80 only during the training stage, while the test evaluation uses the full set of stimulus images.

LaVCa's accuracy increases as the categories become more fine-grained. Extremely coarse category sets, such as those with 2 or 10 categories, compress variability across the stimulus set and lead to reduced accuracy. Using 40 categories captures substantially more structure within the stimulus data and already approaches the performance of the full 80-category condition. The 80-category condition performs best, particularly in higher-level visual areas where finer distinctions between visual stimulus types are beneficial.

A.5 Additional fMRI Dataset: BOLD5000

A.5.1 Dataset and preprocessing

To assess whether LaVCa generalises beyond NSD, we additionally evaluate our method on the BOLD5000 dataset, which contains fMRI responses from four subjects viewing approximately 5,000 natural images (Chang et al., 2019). Following our NSD pipeline, we first compute single-trial response patterns using GLMsingle β-estimates for each session and concatenate them along the time dimension to obtain a 4D volume for each subject.

For our analyses, we restrict all experiments to the functional masks provided in BOLD5000, consisting of EarlyVis (early visual cortex), LOC (lateral occipital complex), OPA (occipital place area), PPA (parahippocampal place area), and RSC (retrosplenial cortex). Across these regions, the total number of voxels amounts to approximately 2,500 per subject. Within each subject, we vectorise the GLMsingle β-estimates into trial-wise response matrices and standardise each voxel by z-scoring

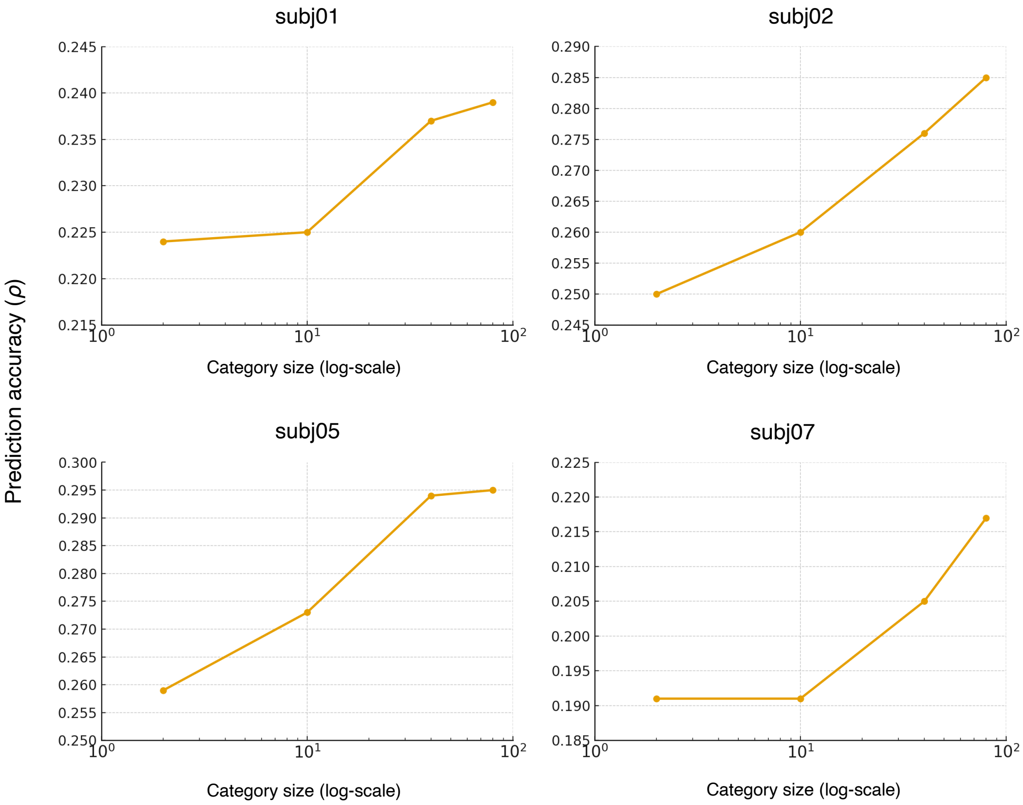


Figure A3: Effect of category size on test prediction accuracy. Mean prediction accuracy on the test data for the top 5000 voxels selected using training-data performance, shown for different category sizes.

across all trials. Following the same procedure as in NSD, we first split the trials into 90% for training and validation and 10% for testing.

We then apply LaVCa to BOLD5000 using exactly the same configuration as in the main NSD experiments, without any dataset-specific tuning.

Because LaVCa evaluates voxel captions by comparing them with the ground-truth captions of the held-out stimuli, textual captions for the test images are required. In BOLD5000, the only subset of stimuli accompanied by human-written captions is the set of COCO images; therefore, test-set accuracy is computed exclusively on COCO trials using their official COCO captions and the corresponding β values. Importantly, the encoding model is trained on all available BOLD5000 trials, and only the evaluation stage is restricted to COCO trials.

A.5.2 Results on BOLD5000

We apply LaVCa to the BOLD5000 dataset and compute voxel-caption accuracy on the COCO-captioned trials in the test split. Table A7 summarises the caption-prediction accuracy for the top 100, top 500, and top 1000 voxels in each subject, selected based on their training-set performance.

Across all voxels, LaVCa achieved significant caption-prediction accuracy in roughly 40% of voxels.

Although LaVCa's performance on BOLD5000 is lower than on NSD, this reduction closely parallels the decrease in encoding-model accuracy between the two datasets. Specifically, the average encoding-model performance drops from 0.504 in NSD to 0.302 in BOLD5000—an approximately 40% reduction. The proportional decline in both encoding and caption-prediction accuracy suggests that the diminished performance reflects inherent differences in dataset quality (e.g., shorter scan durations and lower SNR in BOLD5000), rather than overfitting of LaVCa to NSD.

Table A7: Comparison of sentence-level caption prediction accuracy on BOLD5000 (subj01–subj04). "Top-N voxels" refers to the voxels with top-N prediction performance in the training data. Values are mean ± standard deviation on the test data.

Top-N voxels	**subj01**	**subj02**	**subj03**	**subj04**
Top-100	0.2655 ± 0.0609	0.1927 ± 0.0656	0.1525 ± 0.0540	0.2117 ± 0.0576
Top-500	0.1652 ± 0.0780	0.1138 ± 0.0737	0.1185 ± 0.0560	0.1523 ± 0.0668
Top-1000	0.1286 ± 0.0769	0.0776 ± 0.0749	0.1006 ± 0.0575	0.1221 ± 0.0743

Taken together, these results demonstrate that LaVCa generalises to a dataset with different subjects and stimulus sets.

To assess the reproducibility of voxel-level semantic organization across datasets and subject groups, we conducted a WordCloud-based analysis analogous to that used in the NSD experiments (Figure A4).

Across ROIs, the BOLD5000 results exhibit semantic patterns that closely parallel those observed in NSD despite differences in subjects, scanner parameters, and stimulus sets. In BOLD5000 LOC, which is broadly associated with object- and body-related processing, we observed frequent person-related nouns such as *person*, reproducing the person-related cluster identified in NSD's EBA. Verb distributions show a similar correspondence: BOLD5000 LOC contained action verbs such as *wear* and *hold*, mirroring the body- and interaction-related action clusters prominent in NSD's EBA and OFA.

Scene-selective regions demonstrated an even stronger cross-dataset alignment. BOLD5000 OPA, PPA, and RSC all exhibited WordClouds enriched with place-related nouns including *room*, *kitchen*. These terms match the dominant scene-related cluster observed in NSD PPA, indicating that the underlying semantic structure of these ROIs is stable across datasets.

Taken together, these results show that the semantic clusters identified by LaVCa are not specific to a particular dataset. Instead, similar diversity patterns emerge across both NSD and BOLD5000: object- and body-related concepts appear in lateral occipito-temporal regions, and place-related concepts are consistently represented in medial scene-selective regions. This cross-dataset reproducibility directly addresses the reviewer's question regarding the stability of semantic clusters across large populations and distinct image sets.

A.6 LIMITAION

Despite the overall improvement in brain activity prediction, we observe that face-selective regions do not achieve accuracy levels as high as those in other ROIs (Figure 2). One reason may be that our current approach uses a Multimodal LLM (MLLM) to produce relatively simple captions for optimal images, often omitting important local features (e.g., "eyes," "nose") and focusing on more global terms (e.g., "face," "person"). Consequently, the subsequent summarization step lacks access to these local details. Because our method relies on language descriptions, it has inherent limitations in capturing the fine-grained, local selectivity of these voxels. Incorporating recent techniques that visually interpret local voxel selectivity (Luo et al., 2024) could help address this gap.

Furthermore, while our current study describes voxel selectivity primarily in response to visual stimuli in the occipital cortex, there exist "multimodal voxels" in the brain that are simultaneously activated by auditory and linguistic information, and higher-order cognitive processes such as calculations, memory retrieval, and reasoning (Nakai & Nishimoto, 2020; 2022). Designing stimuli and experimental tasks encompassing diverse sensory inputs (e.g., auditory, textual) and cognitive challenges (e.g., recalling past events, performing reasoning tasks) is essential when interpreting such voxels. Because our approach uses LLM-based textual summarization, it can be adapted to represent a wide range of stimuli and cognitive states in text form, providing a unified framework for multimodal integration. Looking ahead, by jointly modeling images, semantic information, auditory features, and cognitive tasks, we anticipate capturing the brain's integrated representation of both sensory and higher-order cognitive functions with greater accuracy.

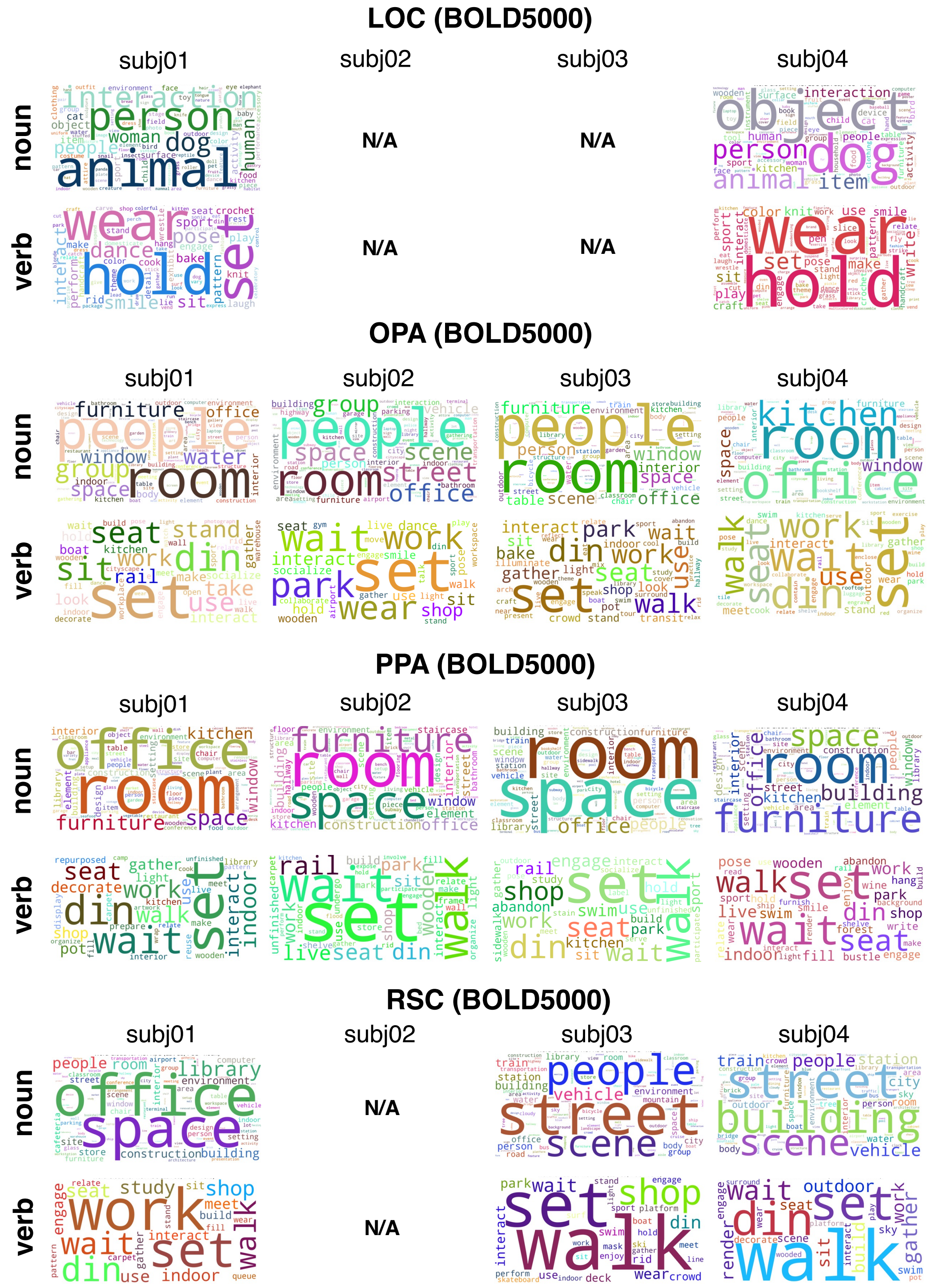


Figure A4: Word clouds depicting the 100 most frequent nouns and verbs appearing in voxel captions for the BOLD5000 dataset. Each panel shows the word cloud for a particular brain region and subject.

A.7 Impact Statement

We introduce a data-driven method that uses a large language model to generate natural language captions of voxel-level visual selectivity. Using the method detailed in this paper, we aim to provide a more fine-grained understanding of human visual function than previously achieved. We acknowledge that this human brain research could raise concerns regarding individual privacy. Although the present study examined relatively coarse-grained voxel-level data, we cannot dismiss the possibility that future advances in measurement and analysis techniques may enable the extraction of more detailed individual-specific information. In any case, obtaining explicit informed consent from participants remains crucial when collecting and using human brain activity data, as with the NSD dataset used in this study.

Prompt

The following are the result of captioning a group of images:

“Three people in a dark room with masks and lights.”
“A band is performing on stage with various instruments and lights.”
“A rock band is performing on a stage with a red Coca-Cola logo.”
“The image depicts a group of people singing together in a room.”
“A band is performing on stage with a singer holding a microphone and a guitarist playing his instrument.”
…

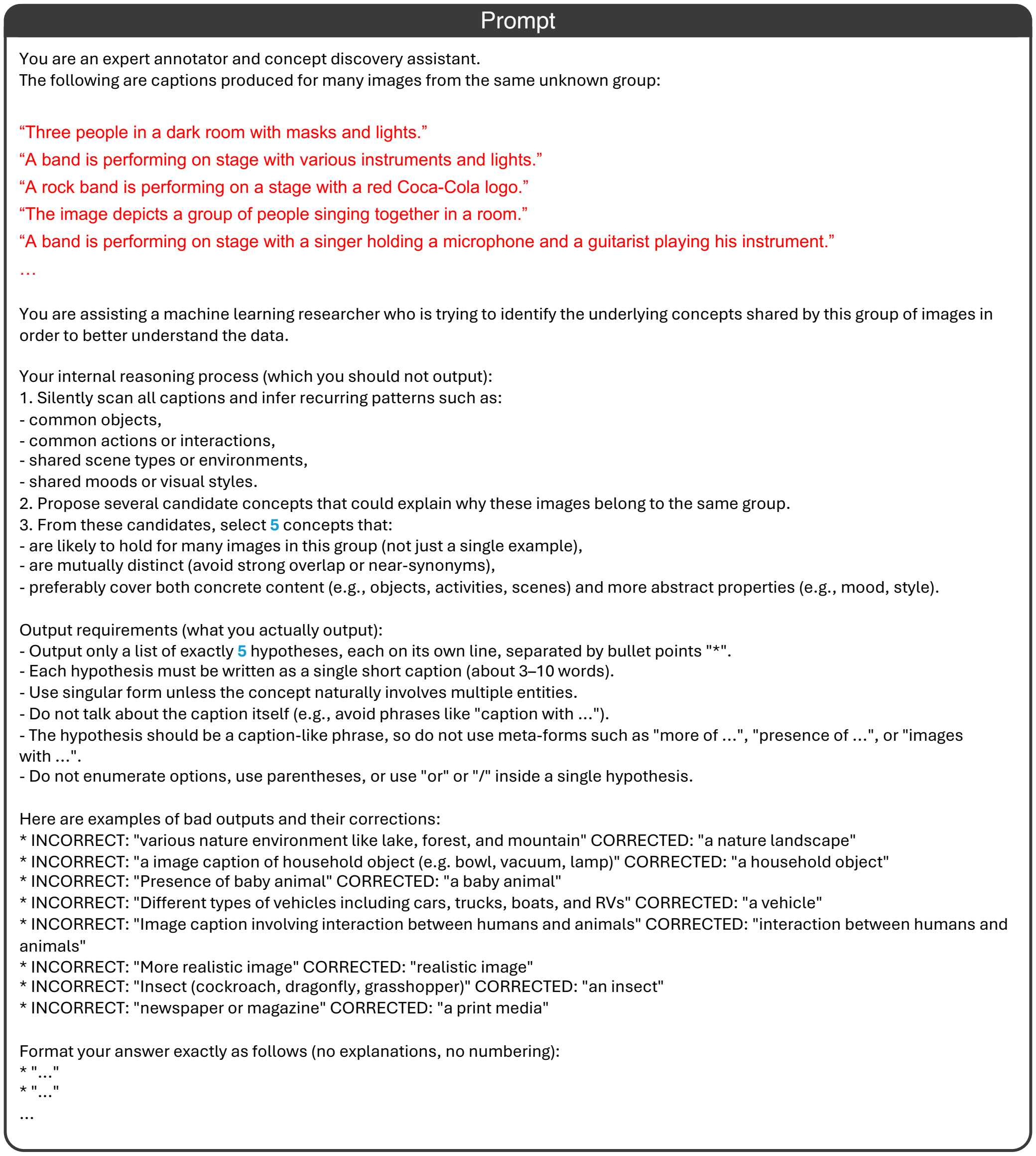
I am a machine learning researcher seeking to elucidate the concepts of this group in order to better understand my data.

Come up with 5 distinct concepts that are likely to be true for this group. Please write a list of captions separated by bullet points ("*"). For example:
* "a dog next to a horse"
* "a car in the rain"
* "low quality"
* "cars from a side view"
* "people in a intricate dress"
* "a joyful atmosphere"

Do not talk about the caption, e.g., "caption with one word" and do not list more than one concept. Also use singular form unless the concept naturally involves multiple objects.
The hypothesis should be a caption, so hypotheses like "more of ...", "presence of ...", "images with ..." are incorrect. Also do not enumerate possibilities within parentheses. Do not provide multiple options by using 'or' or '/' to maintain clarity. Here are examples of bad outputs and their corrections:
* INCORRECT: "various nature environment like lake, forest, and mountain" CORRECTED: "nature"
* INCORRECT: "a image caption of household object (e.g. bowl, vacuum, lamp)" CORRECTED: "a household object"
* INCORRECT: "Presence of baby animal" CORRECTED: "a baby animal"
* INCORRECT: "Different types of vehicles including cars, trucks, boats, and RVs" CORRECTED: "a vehicle"
* INCORRECT: "Image caption involving interaction between humans and animals" CORRECTED: "interaction between humans and animals"
* INCORRECT: "More realistic image" CORRECTED: "realistic image"
* INCORRECT: "Insect (cockroach, dragonfly, grasshopper)" CORRECTED: "a insect"
* INCORRECT: "newspaper or magazine" CORRECTED: "a print media"

Again, I want to identify the characteristics of this group. List properties that hold more often for the images in this group. Answer only with a list (separated by bullet points “*”). Your response:

Figure A5: The prompt used for summarizing the captions of optimal image groups with an LLM (Exemplar-based Prompting). The text in red represents the captions of the optimal image group, which depend on the target voxel and the number of optimal images used. The blue number specifies the required number of concepts, which was varied during the ablation study.



Prompt

You are an expert annotator and concept discovery assistant.
The following are captions produced for many images from the same unknown group:

“Three people in a dark room with masks and lights.”
“A band is performing on stage with various instruments and lights.”
“A rock band is performing on a stage with a red Coca-Cola logo.”
“The image depicts a group of people singing together in a room.”
“A band is performing on stage with a singer holding a microphone and a guitarist playing his instrument.”
…

You are assisting a machine learning researcher who is trying to identify the underlying concepts shared by this group of images in order to better understand the data.

Your internal reasoning process (which you should not output):
1. Silently scan all captions and infer recurring patterns such as:
- common objects,
- common actions or interactions,
- shared scene types or environments,
- shared moods or visual styles.
2. Propose several candidate concepts that could explain why these images belong to the same group.
3. From these candidates, select 5 concepts that:
- are likely to hold for many images in this group (not just a single example),
- are mutually distinct (avoid strong overlap or near-synonyms),
- preferably cover both concrete content (e.g., objects, activities, scenes) and more abstract properties (e.g., mood, style).

Output requirements (what you actually output):
- Output only a list of exactly 5 hypotheses, each on its own line, separated by bullet points "*".
- Each hypothesis must be written as a single short caption (about 3–10 words).
- Use singular form unless the concept naturally involves multiple entities.
- Do not talk about the caption itself (e.g., avoid phrases like "caption with ...").
- The hypothesis should be a caption-like phrase, so do not use meta-forms such as "more of ...", "presence of ...", or "images with ...".
- Do not enumerate options, use parentheses, or use "or" or "/" inside a single hypothesis.

Here are examples of bad outputs and their corrections:
* INCORRECT: "various nature environment like lake, forest, and mountain" CORRECTED: "a nature landscape"
* INCORRECT: "a image caption of household object (e.g. bowl, vacuum, lamp)" CORRECTED: "a household object"
* INCORRECT: "Presence of baby animal" CORRECTED: "a baby animal"
* INCORRECT: "Different types of vehicles including cars, trucks, boats, and RVs" CORRECTED: "a vehicle"
* INCORRECT: "Image caption involving interaction between humans and animals" CORRECTED: "interaction between humans and animals"
* INCORRECT: "More realistic image" CORRECTED: "realistic image"
* INCORRECT: "Insect (cockroach, dragonfly, grasshopper)" CORRECTED: "an insect"
* INCORRECT: "newspaper or magazine" CORRECTED: "a print media"

Format your answer exactly as follows (no explanations, no numbering):
* "..."
* "..."
...

Figure A6: The prompt used for summarizing the captions of optimal image groups with an LLM (Hidden CoT Prompting). The text in red represents the captions of the optimal image group, which depend on the target voxel and the number of optimal images used. The blue number specifies the required number of concepts, which was varied during the ablation study.

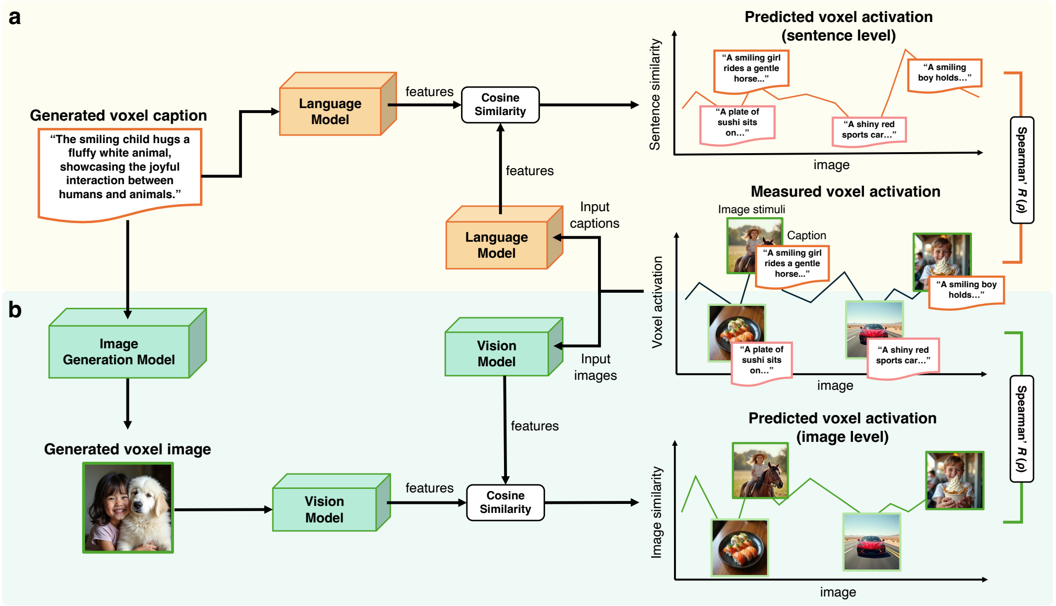

a
Generated voxel caption
"The smiling child hugs a fluffy white animal, showcasing the joyful interaction between humans and animals."
Language Model
features
Cosine Similarity
features
Language Model
Input captions
Predicted voxel activation (sentence level)
Sentence similarity
"A smiling girl rides a gentle horse..."
"A plate of sushi sits on..."
"A shiny red sports car..."
"A smiling boy holds..."
image
Measured voxel activation
Image stimuli
Caption
Voxel activation
"A smiling girl rides a gentle horse..."
"A smiling boy holds..."
"A plate of sushi sits on..."
"A shiny red sports car..."
image
Spearman' R (ρ)
b
Image Generation Model
Vision Model
Input images
features
Generated voxel image
Vision Model
features
Cosine Similarity
Predicted voxel activation (image level)
Image similarity
image
Spearman' R (ρ)

Figure A7: Overview of our evaluation methods. (a) We first obtain text embeddings for both the generated voxel captions (from the language model) and the NSD image captions. We then compute the cosine similarity between the voxel caption embeddings and each NSD caption embedding to derive a rough prediction of voxel activity. Finally, we evaluate text-level prediction performance by calculating Spearman's rank correlation coefficient between these predicted values and the actual voxel responses. (b) We generate voxel images by visualizing voxel captions with an image generation model and, using the same vision model, compute vision-based embeddings for both the generated voxel images and the NSD image stimuli. As in (a), we compute the cosine similarity between voxel-image embeddings and NSD image embeddings and use Spearman's rank correlation coefficient to evaluate image-level prediction performance.

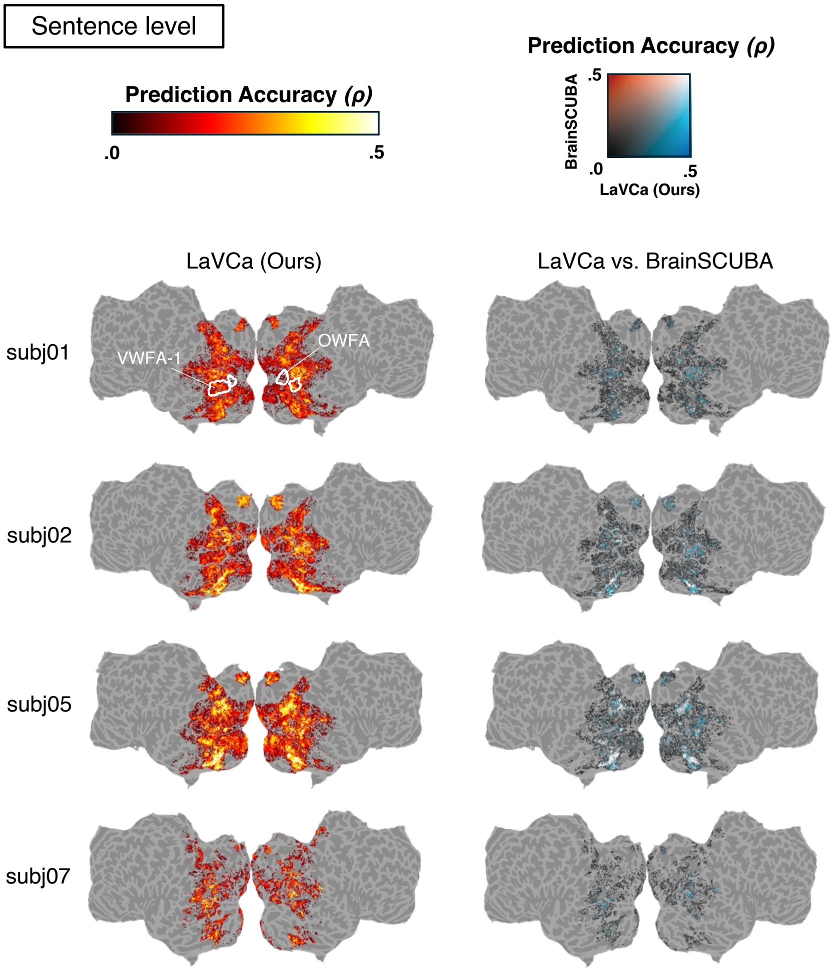


Figure A8: Mapping of brain activity prediction accuracy at the sentence level for LaVCa (left) and a comparison of brain activity prediction accuracy at the sentence level between LaVCa and BrainSCUBA (right) onto the flatmap for all subjects. The white outlines indicate Visual Word Form Area (VWFA-1) and Occipital Word Form Area (OWFA), which are ranked among the top two Words-category ROIs based on the mean number of voxels across subjects.

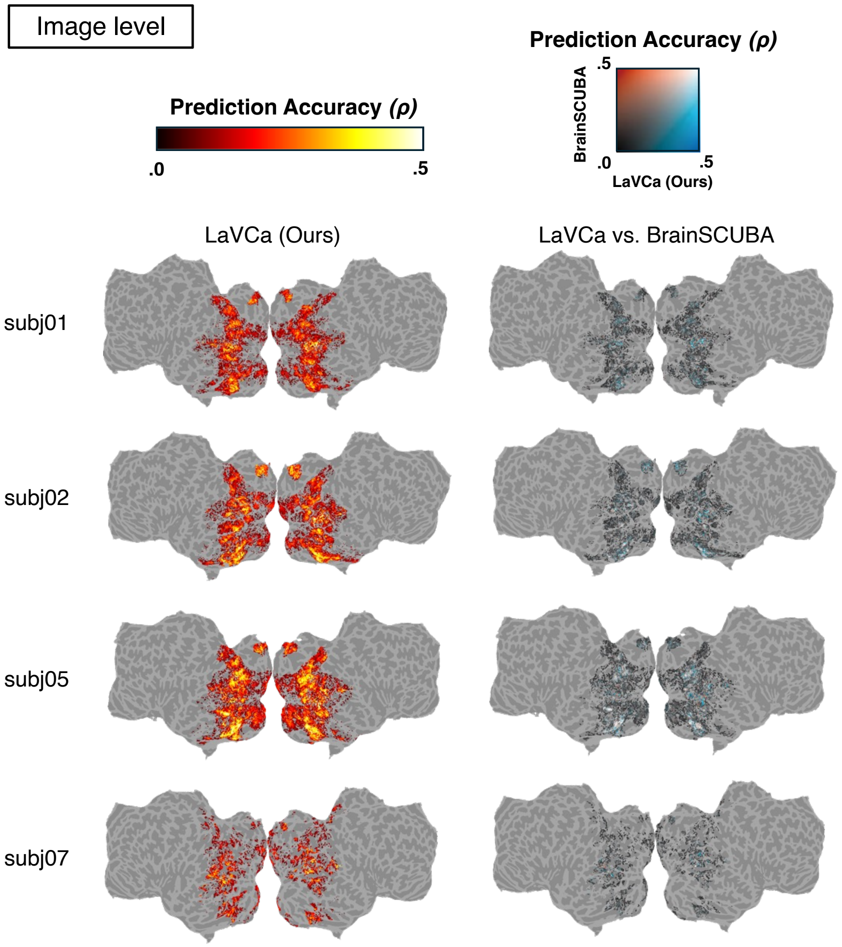


Figure A9: Mapping of brain activity prediction accuracy at the image level for LaVCa (left) and a comparison of brain activity prediction accuracy at the image level between LaVCa and BrainSCUBA (right) onto the flatmap for all subjects.

Table A8: Comparison of sentence-level brain activity prediction performance (all subjects). "Top-N voxels" refers to the voxels with top-N prediction performance in the training data. Each cell shows the mean ± standard deviation of prediction performance on the test data. Two additional columns indicate the hyper-parameter setting of LaVCa variants.

			Top1000 voxels			
Model	# keywords	Sentence Composer	subj01	subj02	subj05	subj07
Shuffled	–	–	-0.010 ± 0.281	0.094 ± 0.312	0.145 ± 0.331	0.013 ± 0.265
BrainSCUBA	–	–	0.291 ± 0.049	0.347 ± 0.062	0.378 ± 0.068	0.267 ± 0.055
LaVCa (Ours)	1	✗	0.300 ± 0.054	0.352 ± 0.057	0.393 ± 0.059	0.286 ± 0.056
LaVCa (Ours)	5	✓	**0.338 ± 0.051**	**0.392 ± 0.057**	**0.420 ± 0.061**	**0.320 ± 0.060**
			Top3000 voxels			
Model	# keywords	Sentence Composer	subj01	subj02	subj05	subj07
Shuffled	–	–	0.000 ± 0.228	0.059 ± 0.255	0.099 ± 0.274	0.004 ± 0.205
BrainSCUBA	–	–	0.237 ± 0.057	0.284 ± 0.067	0.305 ± 0.077	0.212 ± 0.061
LaVCa (Ours)	1	✗	0.240 ± 0.062	0.288 ± 0.068	0.317 ± 0.077	0.221 ± 0.067
LaVCa (Ours)	5	✓	**0.280 ± 0.059**	**0.325 ± 0.068**	**0.349 ± 0.075**	**0.253 ± 0.069**
			Top5000 voxels			
Model	# keywords	Sentence Composer	subj01	subj02	subj05	subj07
Shuffled	–	–	0.007 ± 0.199	0.058 ± 0.223	0.067 ± 0.243	0.009 ± 0.175
BrainSCUBA	–	–	0.207 ± 0.062	0.251 ± 0.071	0.264 ± 0.084	0.182 ± 0.065
LaVCa (Ours)	1	✗	0.205 ± 0.068	0.250 ± 0.075	0.272 ± 0.086	0.186 ± 0.072
LaVCa (Ours)	5	✓	**0.246 ± 0.066**	**0.287 ± 0.075**	**0.306 ± 0.084**	**0.218 ± 0.073**
			Top10000 voxels			
Model	# keywords	Sentence Composer	subj01	subj02	subj05	subj07
Shuffled	–	–	0.008 ± 0.157	0.039 ± 0.178	0.051 ± 0.192	0.012 ± 0.134
BrainSCUBA	–	–	0.159 ± 0.071	0.195 ± 0.081	0.199 ± 0.095	0.134 ± 0.072
LaVCa (Ours)	1	✗	0.154 ± 0.076	0.190 ± 0.086	0.199 ± 0.101	0.132 ± 0.080
LaVCa (Ours)	5	✓	**0.191 ± 0.077**	**0.227 ± 0.086**	**0.237 ± 0.098**	**0.163 ± 0.081**

Table A9: Comparison of image-level brain activity prediction performance (all subjects). "Top-*N* voxels" refers to the voxels with top-*N* prediction performance in the training data. Values are mean ± standard deviation on the test data.

Top1000 voxels

Model	# keywords	Sentence Composer	subj01	subj02	subj05	subj07
Shuffled	–	–	0.022 ± 0.235	0.048 ± 0.254	0.104 ± 0.273	0.036 ± 0.230
BrainSCUBA	–	–	0.278 ± 0.056	0.322 ± 0.052	0.357 ± 0.057	0.262 ± 0.061
LaVCa (Ours)	1	✗	0.267 ± 0.050	0.311 ± 0.047	0.355 ± 0.054	0.241 ± 0.052
LaVCa (Ours)	5	✓	**0.314 ± 0.059**	**0.347 ± 0.053**	**0.379 ± 0.054**	**0.289 ± 0.060**

Top3000 voxels

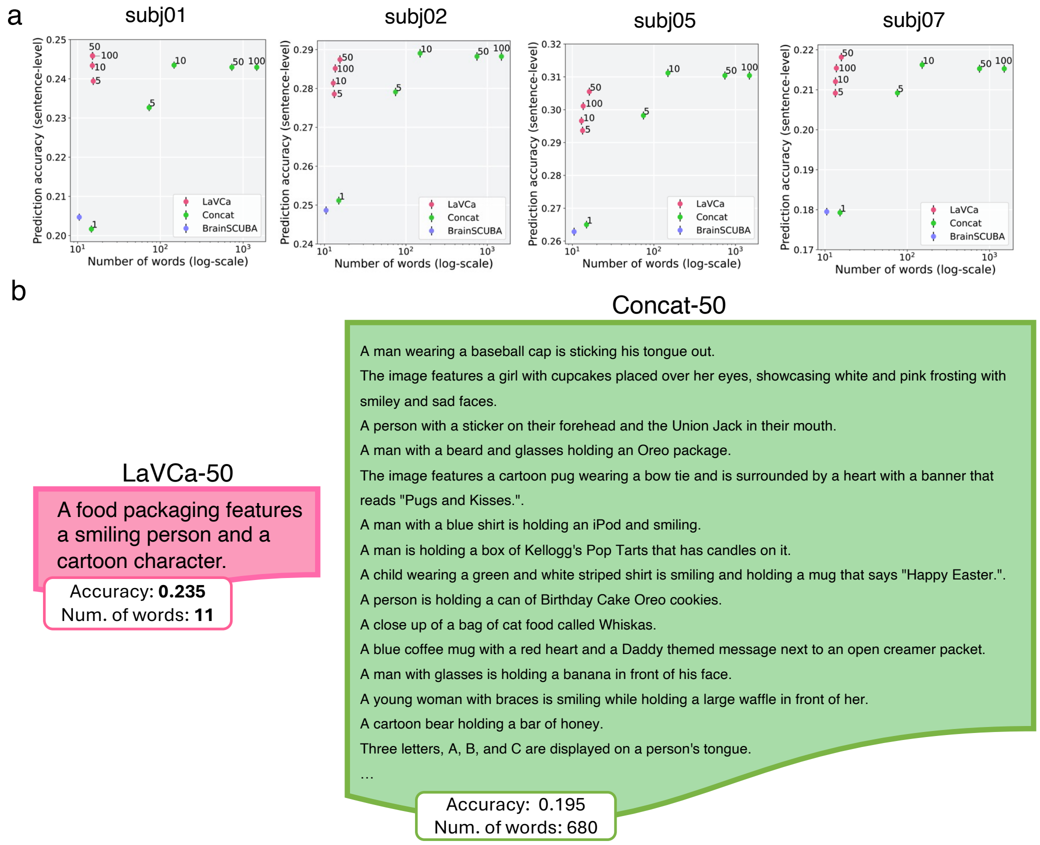
Model	# keywords	Sentence Composer	subj01	subj02	subj05	subj07
Shuffled	–	–	0.017 ± 0.187	0.059 ± 0.210	0.087 ± 0.228	0.007 ± 0.174
BrainSCUBA	–	–	0.220 ± 0.062	0.262 ± 0.063	0.291 ± 0.070	0.201 ± 0.067
LaVCa (Ours)	1	✗	0.213 ± 0.058	0.255 ± 0.058	0.292 ± 0.067	0.187 ± 0.061
LaVCa (Ours)	5	✓	**0.248 ± 0.066**	**0.286 ± 0.063**	**0.315 ± 0.068**	**0.221 ± 0.069**

Top5000 voxels

Model	# keywords	Sentence Composer	subj01	subj02	subj05	subj07
Shuffled	–	–	0.017 ± 0.163	0.052 ± 0.185	0.066 ± 0.204	0.009 ± 0.148
BrainSCUBA	–	–	0.188 ± 0.067	0.226 ± 0.070	0.250 ± 0.078	0.169 ± 0.069
LaVCa (Ours)	1	✗	0.182 ± 0.063	0.221 ± 0.066	0.252 ± 0.077	0.158 ± 0.064
LaVCa (Ours)	5	✓	**0.213 ± 0.072**	**0.249 ± 0.070**	**0.273 ± 0.079**	**0.187 ± 0.073**

Top10000 voxels

Model	# keywords	Sentence Composer	subj01	subj02	subj05	subj07
Shuffled	–	–	0.010 ± 0.128	0.034 ± 0.145	0.049 ± 0.159	0.006 ± 0.114
BrainSCUBA	–	–	0.139 ± 0.073	0.170 ± 0.081	0.188 ± 0.090	0.122 ± 0.073
LaVCa (Ours)	1	✗	0.134 ± 0.071	0.168 ± 0.076	0.187 ± 0.091	0.114 ± 0.069
LaVCa (Ours)	5	✓	**0.160 ± 0.078**	**0.191 ± 0.082**	**0.208 ± 0.092**	**0.138 ± 0.077**



A food packaging features a smiling person and a cartoon character.

Accuracy: **0.235**
Num. of words: **11**

A man wearing a baseball cap is sticking his tongue out.
The image features a girl with cupcakes placed over her eyes, showcasing white and pink frosting with smiley and sad faces.
A person with a sticker on their forehead and the Union Jack in their mouth.
A man with a beard and glasses holding an Oreo package.
The image features a cartoon pug wearing a bow tie and is surrounded by a heart with a banner that reads "Pugs and Kisses.".
A man with a blue shirt is holding an iPod and smiling.
A man is holding a box of Kellogg's Pop Tarts that has candles on it.
A child wearing a green and white striped shirt is smiling and holding a mug that says "Happy Easter.".
A person is holding a can of Birthday Cake Oreo cookies.
A close up of a bag of cat food called Whiskas.
A blue coffee mug with a red heart and a Daddy themed message next to an open creamer packet.
A man with glasses is holding a banana in front of his face.
A young woman with braces is smiling while holding a large waffle in front of her.
A cartoon bear holding a bar of honey.
Three letters, A, B, and C are displayed on a person's tongue.
…

Accuracy: 0.195
Num. of words: 680

Figure A10: **a** Relationship between voxel caption prediction performance and word count (all subjects). The color of the plot corresponds to the lineage of each model. The numbers associated with LaVCa indicate the number of optimal images used for summarization, while the numbers associated with Concat represent the number of captions for concatenated optimal images. Error bars indicate the standard error. **b** Comparison of actual voxel captions between Concat-50 and LaVCa-50. Only a portion of the captions is depicted for Concat-50.

Table A10: Evaluation of the diversity of three models. PCs (90% Var) means the number of principal components required to explain 90% variance of the text embeddings. For intra-voxel comparisons, the mean ± standard deviation across subjects is presented. The inter-subject average (Average) is presented as the mean ± standard error.

			Inter-voxel			Intra-voxel		
			Lexical	Semantic		Lexical		Semantic
Subject	Model	Acc.	Vocab. size	Variance	PCs (90% Var)	Vocab. size	Length	Variance
subj01	BrainSCUBA	0.207±0.062	3400	0.0591	99	6.21±1.27	6.32±1.46	0.0163±0.0025
	Top-1 (Ours)	0.202±0.064	15384	0.0640	210	9.65±3.59	10.0±4.24	0.0194±0.0027
	LaVCa (Ours)	0.246±0.066	16477	0.0639	218	11.0±2.89	11.5±3.19	0.0198±0.0025
subj02	BrainSCUBA	0.251±0.071	3287	0.0591	133	6.17±1.32	6.27±1.51	0.0162±0.0026
	Top-1 (Ours)	0.251±0.070	14135	0.0632	206	9.99±3.65	10.5±4.28	0.0195±0.0027
	LaVCa (Ours)	0.287±0.075	17242	0.0639	218	11.3±3.64	11.8±3.93	0.0198±0.0027
subj05	BrainSCUBA	0.263±0.084	3043	0.0583	127	6.18±1.37	6.26±1.52	0.0159±0.0027
	Top-1 (Ours)	0.265±0.081	13485	0.0631	206	9.99±3.68	10.4±4.34	0.0195±0.0028
	LaVCa (Ours)	0.306±0.084	17459	0.0644	218	11.8±3.88	12.2±4.14	0.0199±0.0027
subj07	BrainSCUBA	0.182±0.065	3042	0.0587	131	6.23±1.30	6.36±1.47	0.0163±0.0026
	Top-1 (Ours)	0.179±0.066	12831	0.0632	203	10.1±3.51	10.6±4.14	0.0197±0.0026
	LaVCa (Ours)	0.218±0.073	16508	0.0646	222	11.6±3.76	12.0±4.02	0.0202±0.0026
Average	BrainSCUBA	0.226±0.019	3193±90	0.0588±0.0002	123±7.93	6.20±0.01	6.30±0.02	0.0162±0.0001
	Top-1 (Ours)	0.224±0.020	13959±545	0.0634±0.0002	206±1.44	9.93±0.10	10.4±0.132	0.0195±0.0001
	LaVCa (Ours)	0.264±0.020	16922±252	0.0642±0.0002	219±1.00	11.4±0.175	11.9±0.149	0.0199±0.0001

Table A11: The average prediction accuracy for each subject and the inter-subject average prediction accuracy when captions were shuffled within the ROI (Shuffled) and when they were used as-is (Original). For each subject, the average prediction accuracy ± standard deviation is depicted, while for the inter-subject average, the average prediction accuracy ± standard error is presented.

		Body areas		**Face areas**	
		EBA	FBA-2	OFA	FFA-1
subj01	Shuffled	0.035±0.147	0.014±0.128	0.031±0.067	0.017±0.113
	Original	**0.169±0.105**	**0.124±0.102**	**0.083±0.069**	**0.117±0.083**
subj02	Shuffled	0.010±0.144	0.026±0.109	0.036±0.071	0.024±0.097
	Original	**0.158±0.101**	**0.128±0.103**	**0.079±0.066**	**0.105±0.078**
subj05	Shuffled	-0.001±0.148	0.007±0.148	0.017±0.118	0.009±0.116
	Original	**0.152±0.111**	**0.149±0.114**	**0.120±0.100**	**0.112±0.089**
subj07	Shuffled	0.028±0.135	0.025±0.096	0.027±0.096	0.013±0.100
	Original	**0.149±0.104**	**0.099±0.099**	**0.097±0.099**	**0.108±0.090**
Average	Shuffled	0.018±0.008	0.018±0.005	0.028±0.004	0.016±0.003
	Original	**0.157±0.005**	**0.125±0.010**	**0.095±0.009**	**0.111±0.003**
		Place areas		**Word areas**	
		OPA	PPA	OWFA	VWFA-1
subj01	Shuffled	0.080±0.108	0.105±0.107	0.015±0.057	0.054±0.114
	Original	**0.163±0.093**	**0.172±0.099**	**0.055±0.048**	**0.147±0.088**
subj02	Shuffled	0.118±0.139	0.178±0.147	0.037±0.071	0.031±0.135
	Original	**0.204±0.114**	**0.243±0.139**	**0.085±0.066**	**0.150±0.099**
subj05	Shuffled	0.184±0.140	0.217±0.153	0.028±0.109	0.039±0.148
	Original	**0.260±0.124**	**0.275±0.149**	**0.118±0.105**	**0.177±0.108**
subj07	Shuffled	0.083±0.119	0.105±0.108	0.020±0.070	0.012±0.159
	Original	**0.175±0.096**	**0.163±0.106**	**0.079±0.069**	**0.157±0.112**
Average	Shuffled	0.116±0.024	0.151±0.028	0.025±0.005	0.034±0.009
	Original	**0.200±0.022**	**0.213±0.027**	**0.084±0.013**	**0.158±0.007**

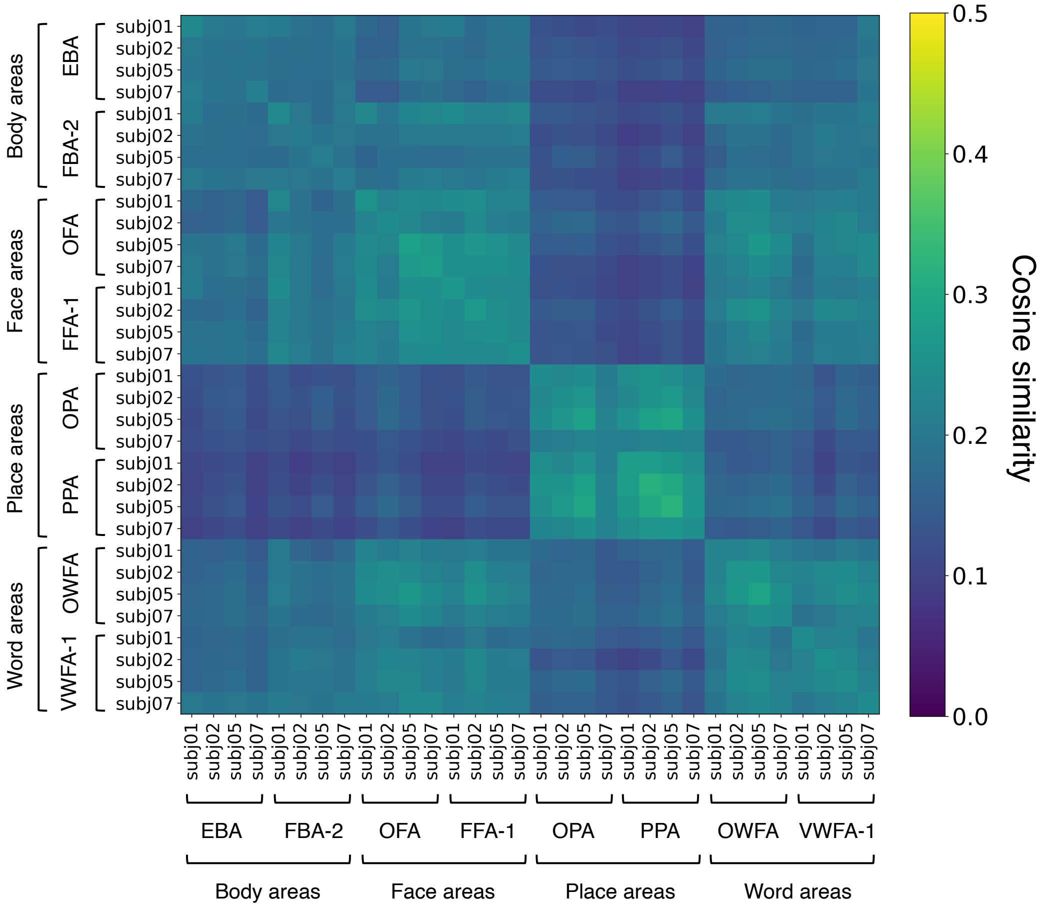


Figure A11: Cross-subject similarity of voxel captions in ROIs. Each cell shows the mean cosine similarity between sentence embeddings of all voxel captions in the two sets.

Table A12: Voxel counts for each ROI, categorized by the number of clusters to which each voxel belongs (*# assigned clusters*). Average rows show the mean ± standard error across subjects.

		Body areas		Face areas		Place areas		Word areas	
	# assigned clusters	EBA	FBA-2	OFA	FFA-1	OPA	PPA	OWFA	VWFA-1
subj01	1	40	11	10	14	14	14	4	6
	2	517	45	39	60	141	90	60	74
	3	1083	126	130	151	465	288	136	257
	4	904	147	112	165	579	369	165	263
	5	377	82	53	84	325	226	83	149
	6	50	19	11	10	87	46	16	23
subj02	1	46	13	4	5	20	16	3	2
	2	517	137	49	44	182	140	39	32
	3	1037	381	124	124	457	380	137	108
	4	1081	424	147	112	454	337	207	124
	5	624	219	95	52	216	116	106	64
	6	134	43	22	3	52	5	27	15
subj05	1	39	7	9	29	19	13	5	8
	2	607	72	102	108	178	171	62	55
	3	1449	182	257	160	416	444	134	139
	4	1446	181	248	111	450	379	147	157
	5	829	65	139	41	218	177	73	92
	6	214	1	26	3	50	37	17	35
subj07	1	34	11	4	5	34	8	1	6
	2	303	69	26	45	193	118	40	34
	3	1123	158	88	88	393	298	134	112
	4	1041	180	99	113	308	323	267	101
	5	492	116	84	72	141	143	155	67
	6	69	18	15	23	14	22	31	7
Average	1	40±2	10±1	7±2	13±6	22±4	13±2	3±1	6±1
	2	486±65	81±20	54±17	64±15	174±11	130±17	50±6	49±10
	3	1173±94	212±58	150±37	131±16	433±17	352±37	135±1	154±35
	4	1118±116	233±64	152±34	125±13	448±55	352±13	197±27	161±36
	5	580±97	120±35	93±18	62±10	225±38	166±24	104±18	93±20
	6	117±37	20±9	18±3	10±5	51±15	28±9	23±4	20±6

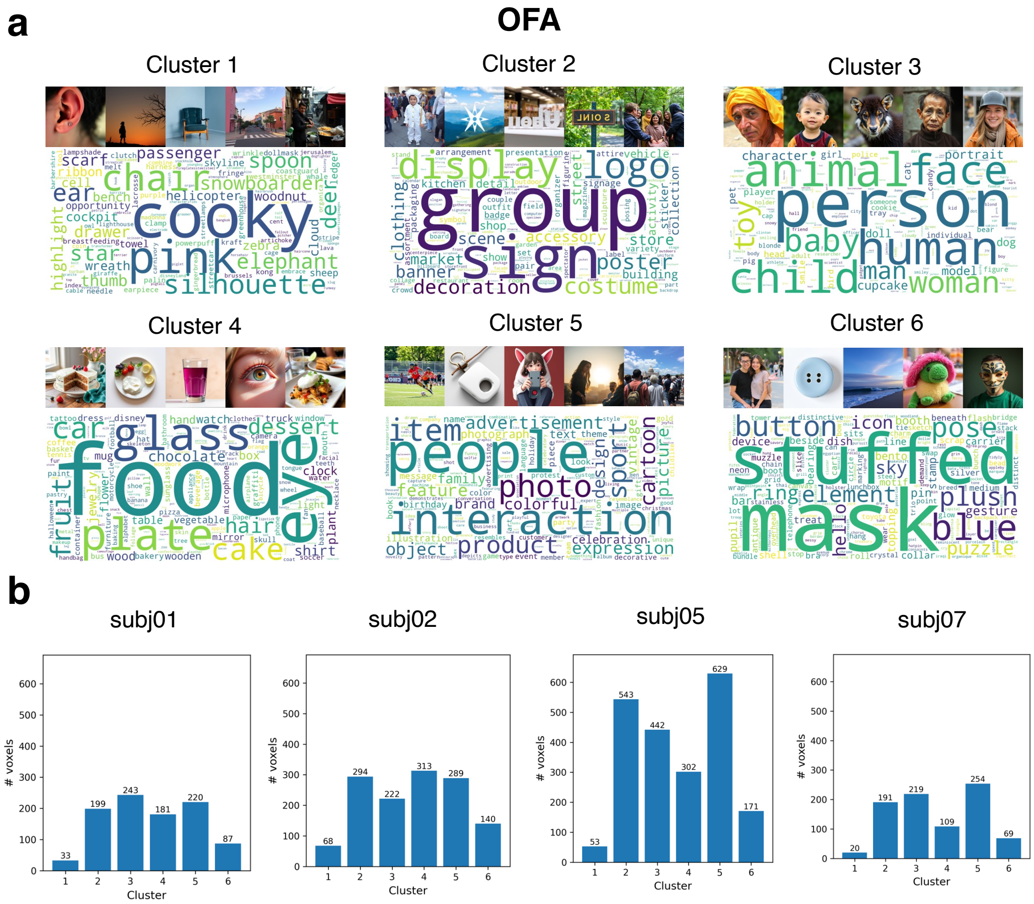


Figure A12: Subject-shared noun cluster analysis in the OFA. **a** Word clouds and generated images for the top-5 most frequent nouns in each subject-shared cluster. **b** Bar graphs showing the number of voxels assigned to each subject-shared cluster for individual subjects.

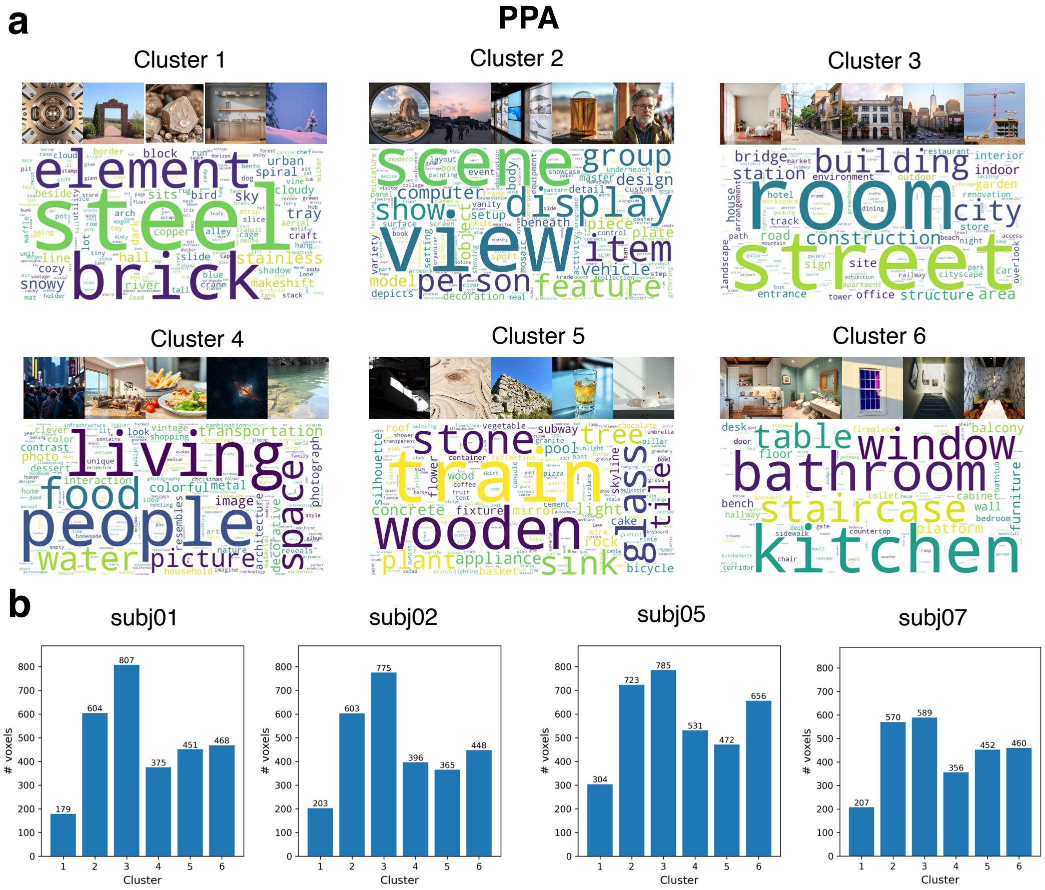


Figure A13: Subject-shared noun cluster analysis in the PPA. **a** Word clouds and generated images for the top-5 most frequent nouns in each subject-shared cluster. **b** Bar graphs showing the number of voxels assigned to each subject-shared cluster for individual subjects.

Examples of OFA (1/2)

Voxel index: 270546

Top49 most activated images

Model	Voxel caption	Voxel image	Acc. (Text-level)	Acc. (Image-level)	Vocab. size	Length	W2V variance
LaVCa (Ours)	Each food item in the organizer and container bears a colored themed decoration with unique lettering or name brand logo.		**0.1546**	**0.1034**	**14**	**14**	0.02239
Top-1 (Ours)	Six different flavors of Pringles chips are lined up side by side.		0.06317	0.08339	8	9	**0.02289**
BrainSCUBA	A group of four hot dogs sitting on top of a table.		0.08565	0.05684	8	14	0.01494

Voxel index: 296456

Top49 most activated images

Model	Voxel caption	Voxel image	Acc. (Text-level)	Acc. (Image-level)	Vocab. size	Length	W2V variance
LaVCa (Ours)	A child wearing a themed shirt, or someone else brandishing 50 smiling person holding a sign with the food network logo.		**0.1229**	**0.03068**	**17**	**17**	0.02000
Top-1 (Ours)	The storefront of PornTip Jewelry is brightly colored and has many items on display.		0.08345	0.02695	9	9	**0.02171**
BrainSCUBA	A store with a lot of signs on the wall.		0.06796	0.02664	5	5	0.01715

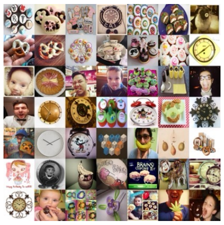
Voxel index: 296535

Top49 most activated images

Model	Voxel caption	Voxel image	Acc. (Text-level)	Acc. (Image-level)	Vocab. size	Length	W2V variance
LaVCa (Ours)	A food packaging features a smiling person and a cartoon character.		**0.2349**	**0.1999**	**8**	**8**	0.01908
Top-1 (Ours)	A man wearing a baseball cap is sticking his tongue out.		0.1068	0.1146	7	7	**0.01992**
BrainSCUBA	A man with a toothbrush in his mouth.		0.09994	0.1482	4	4	0.01745

Voxel index: 305512

Top49 most activated images



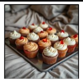
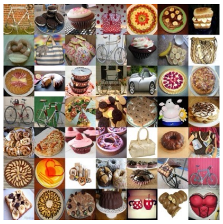
Model	Voxel caption	Voxel image	Acc. (Text-level)	Acc. (Image-level)	Vocab. size	Length	W2V variance
LaVCa (Ours)	Custom caricature logo illustration featuring the signature logo of puppy eyes complete with a clock face, child tooth and cute cupcake man.		**0.1643**	**0.1186**	**18**	**19**	**0.02306**
Top-1 (Ours)	The image shows a collection of cupcakes with various medical-themed decorations.		0.1239	0.1034	8	8	0.01930
BrainSCUBA	a close up of a tray of cupcakes on a bed		0.1084	0.1139	4	4	0.01847

Figure A14: Comparison of voxel captions and voxel images in the OFA voxels of subj02 (1/2).

Examples of OFA (2/2)

Voxel index: 305843

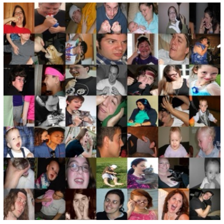
Top49 most activated images

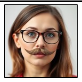


Model	Voxel caption	Voxel image	Acc. (Text-level)	Acc. (Image-level)	Vocab. size	Length	W2V variance
LaVCa (Ours)	Car boot handbag baked in good polka dot patterned bicycle dessert.		**0.06425**	**0.1166**	**11**	11	**0.02518**
Top-1 (Ours)	A white bike with black handlebars is leaning against a red brick wall.		0.01513	0.08413	9	9	0.01963
BrainSCUBA	A bicycle that is leaning against a wall.		0.008889	0.08090	8	**14**	0.01527

Voxel index: 296456

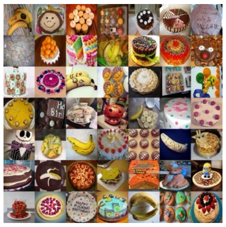
Top49 most activated images



Model	Voxel caption	Voxel image	Acc. (Text-level)	Acc. (Image-level)	Vocab. size	Length	W2V variance
LaVCa (Ours)	The person holding an animal mouth is a person wearing glasses.		**0.1225**	**0.07516**	**7**	**8**	0.01924
Top-1 (Ours)	A woman with a fake mustache and glasses.		0.04519	0.04770	5	5	**0.02047**
BrainSCUBA	a close up of a person brushing her teeth		0.07279	0.03866	4	4	0.01651

Voxel index: 278988

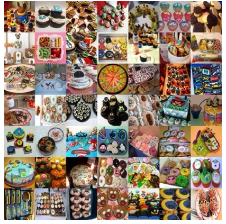
Top49 most activated images



Model	Voxel caption	Voxel image	Acc. (Text-level)	Acc. (Image-level)	Vocab. size	Length	W2V variance
LaVCa (Ours)	This is a fruit cake with banana icing and a decorative design.		0.06029	0.07104	7	7	**0.01965**
Top-1 (Ours)	The image shows two bananas with faces drawn on them, placed on a blue and white striped cloth.		0.08880	**0.1013**	**13**	**13**	0.01781
BrainSCUBA	A banana with a smiley face drawn on it.		**0.1057**	0.09654	5	5	0.01722

Voxel index: 305475

Top49 most activated images



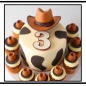
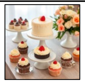
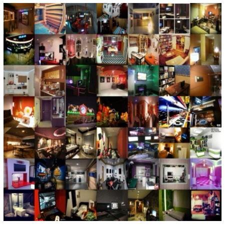
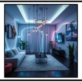
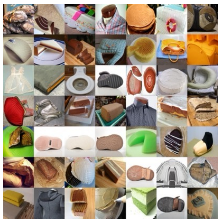
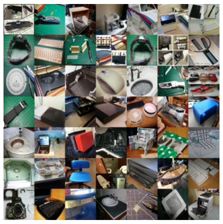
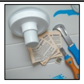
Model	Voxel caption	Voxel image	Acc. (Text-level)	Acc. (Image-level)	Vocab. size	Length	W2V variance
LaVCa (Ours)	Their food design and decoration scheme is paired with samples from holiday theme trucks.		0.1953	0.1583	10	10	**0.02306**
Top-1 (Ours)	The image features a cowboy-themed birthday cake with number 3 on it, surrounded by cupcakes with hat toppers.		0.2178	0.1673	**13**	**13**	0.02264
BrainSCUBA	A table topped with cakes and cup cakes.		**0.2531**	**0.2058**	5	6	0.01670

Figure A15: Comparison of voxel captions and voxel images in the OFA voxels of subj02 (2/2).

Examples of PPA (1/2)

Voxel index: 210730 Top49 most activated images	Model	Voxel caption	Voxel image	Acc. (Text-level)	Acc. (Image-level)	Vocab. size	Length	W2V variance
	**LaVCa (Ours)**	The digitally decorated room is a sign of the modern atmosphere.		**0.2527**	**0.1728**	**7**	7	**0.02056**
	Top-1 (Ours)	A restaurant called Fresco is next to another restaurant called Kabob.		0.1888	0.1307	7	**9**	0.01683
	BrainSCUBA	A view of a building with a lot of neon signs on it.		0.1576	0.0868	6	6	0.01636

Voxel index: 217592 Top49 most activated images	Model	Voxel caption	Voxel image	Acc. (Text-level)	Acc. (Image-level)	Vocab. size	Length	W2V variance
	**LaVCa (Ours)**	A shoe sole inside a cake depicting a chocolate mannequin.		**0.2022**	**0.1828**	**8**	**8**	**0.02049**
	Top-1 (Ours)	A pile of bread with a gold and black label on top.		0.1792	0.1155	7	7	0.01857
	BrainSCUBA	a close up of a stuffed animal on a bag		0.1340	0.1154	4	4	0.01653

Voxel index: 217647 Top49 most activated images	Model	Voxel caption	Voxel image	Acc. (Text-level)	Acc. (Image-level)	Vocab. size	Length	W2V variance
	**LaVCa (Ours)**	A bathroom fixture, newspaper slices and slices of some silver computer component, a wooden object, covered with the blue paint and pattern tool.		**0.2786**	**0.2029**	**16**	**19**	0.02198
	Top-1 (Ours)	The image showcases a pocket knife with various close-up views.		0.1775	0.1433	8	8	**0.02239**
	BrainSCUBA	a collage of photos with a green and black handle		0.2073	0.1223	5	5	0.01771

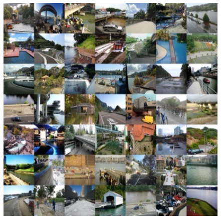
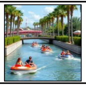
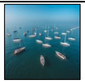
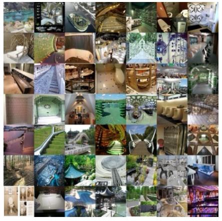
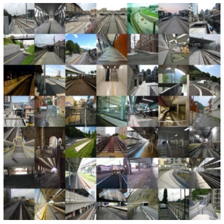
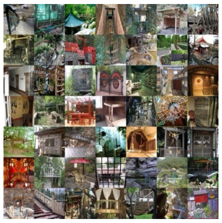
Voxel index: 217753 Top49 most activated images	Model	Voxel caption	Voxel image	Acc. (Text-level)	Acc. (Image-level)	Vocab. size	Length	W2V variance
	**LaVCa (Ours)**	This is the perfect urban setting, with a bridge deck and vehicle storage below.		**0.1454**	**0.1109**	9	9	0.01926
	Top-1 (Ours)	People are riding in small boats through a waterway at an amusement park with palm trees and a bridge in the background.		0.05586	0.03552	**12**	**12**	**0.02358**
	BrainSCUBA	a number of small boats in a body of water		0.06832	0.02745	5	5	0.01626

Figure A16: Comparison of voxel captions and voxel images in the PPA voxels of subj07 (1/2).

Examples of PPA (2/2)

Voxel index: 224691 **Top49 most activated images**	**Model**	**Voxel caption**	**Voxel image**	**Acc. (Text-level)**	**Acc. (Image-level)**	**Vocab. size**	**Length**	**W2V variance**
	**LaVCa (Ours)**	A spiral staircase, an aerial view of the bathroom and a modern interior.		**0.2733**	**0.2030**	9	9	**0.02184**
	Top-1 (Ours)	The image shows a bird's eye view of an artificial lake in the middle of a resort.		0.1005	0.1283	**11**	**11**	0.01993
	BrainSCUBA	An aerial view of a large pool of water.		0.04142	0.1185	6	6	0.01630

Voxel index: 239901 **Top49 most activated images**	**Model**	**Voxel caption**	**Voxel image**	**Acc. (Text-level)**	**Acc. (Image-level)**	**Vocab. size**	**Length**	**W2V variance**
	**LaVCa (Ours)**	The pedestrian infrastructure shows a train station and railway track.		**0.2759**	**0.4021**	8	8	0.01811
	Top-1 (Ours)	A railway track with a sign that says "Authorized personnel only."		0.1968	0.1826	**9**	**9**	**0.02029**
	BrainSCUBA	A street sign on the side of a train track.		0.1597	0.2468	6	6	0.01369

Voxel index: 239644 **Top49 most activated images**	**Model**	**Voxel caption**	**Voxel image**	**Acc. (Text-level)**	**Acc. (Image-level)**	**Vocab. size**	**Length**	**W2V variance**
	**LaVCa (Ours)**	The wooden structure, with asian architecture, mirrors the zoo gate.		0.1727	0.09343	9	10	**0.02312**
	Top-1 (Ours)	The image displays an outdoor Asian temple setting with various elements like a railing, steps, lanterns, and a roof.		0.1400	**0.1539**	**15**	**17**	0.02167
	BrainSCUBA	A row of wooden benches sitting on top of a walkway.		**0.1930**	0.1348	7	7	0.01691

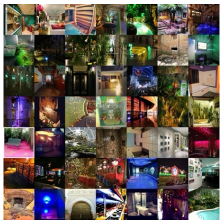
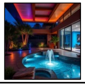
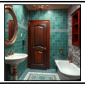
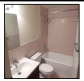
Voxel index: 203319 **Top49 most activated images**	**Model**	**Voxel caption**	**Voxel image**	**Acc. (Text-level)**	**Acc. (Image-level)**	**Vocab. size**	**Length**	**W2V variance**
	**LaVCa (Ours)**	The interior design highlighted a colorful nighttime scene and a water feature.		0.1364	0.1183	9	9	0.01651
	Top-1 (Ours)	An ornate bathroom with green tiles, a wooden door, and stone wall.		**0.2443**	0.1894	**10**	**11**	**0.01973**
	BrainSCUBA	A bathroom with a sink, mirror, toilet and bathtub.		0.2115	**0.1972**	7	8	0.01605

Figure A17: Comparison of voxel captions and voxel images in the PPA voxels of subj07 (2/2).

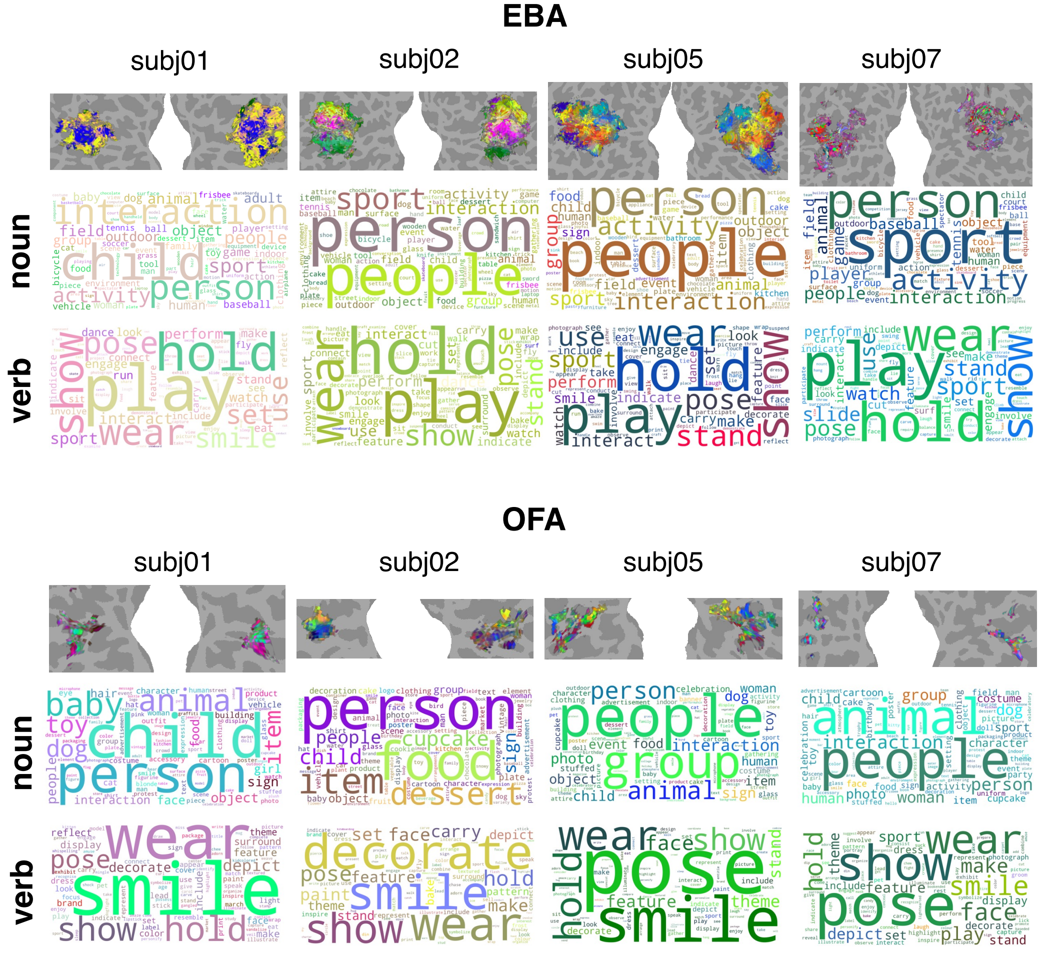


Figure A18: The UMAP projection of caption text across EBA and OFA for all subjects, visualized on a flatmap (top). A word cloud of the 100 most frequent **nouns** in these captions (middle), colored according to their location in the UMAP space. A word cloud of the 100 most frequent **verbs** (bottom).

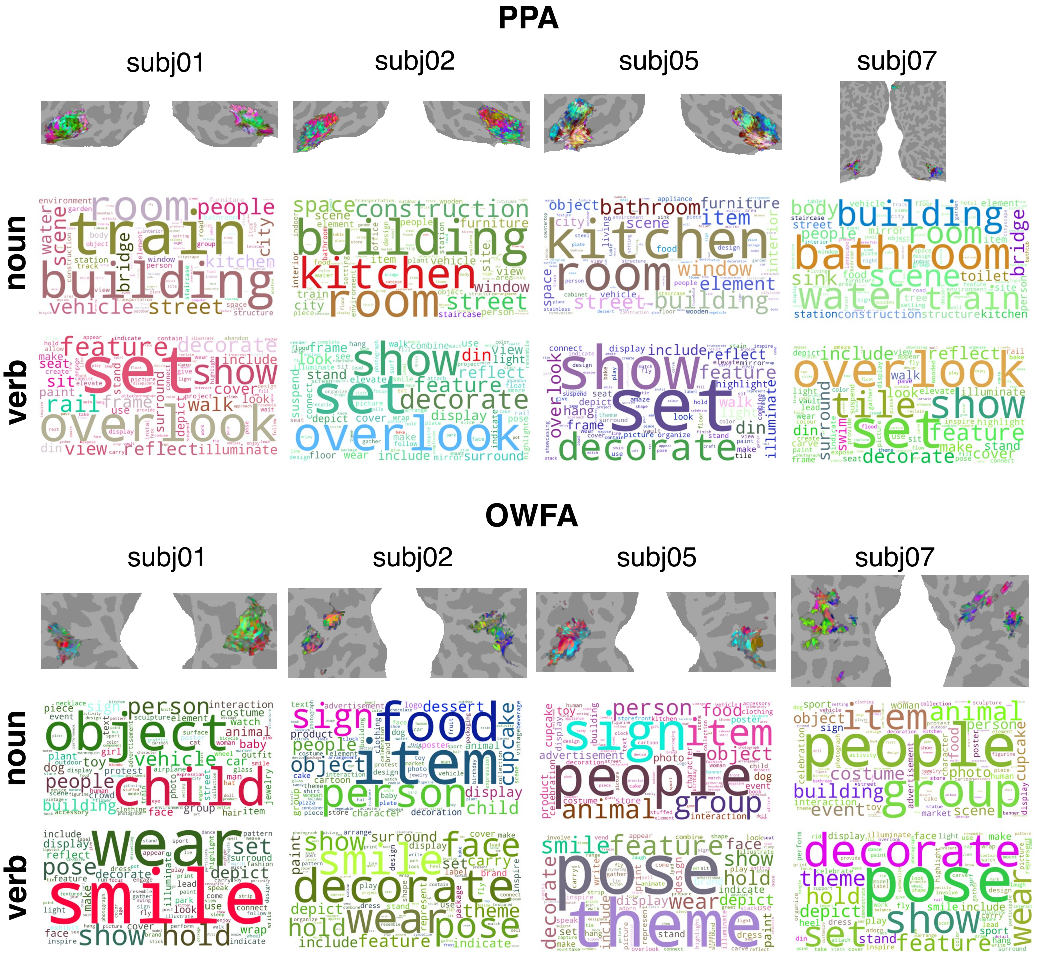


Figure A19: The UMAP projection of caption text across PPA and OWFA for all subjects, visualized on a flatmap (top). A word cloud of the 100 most frequent **nouns** in these captions (middle), colored according to their location in the UMAP space. A word cloud of the 100 most frequent **verbs** (bottom).

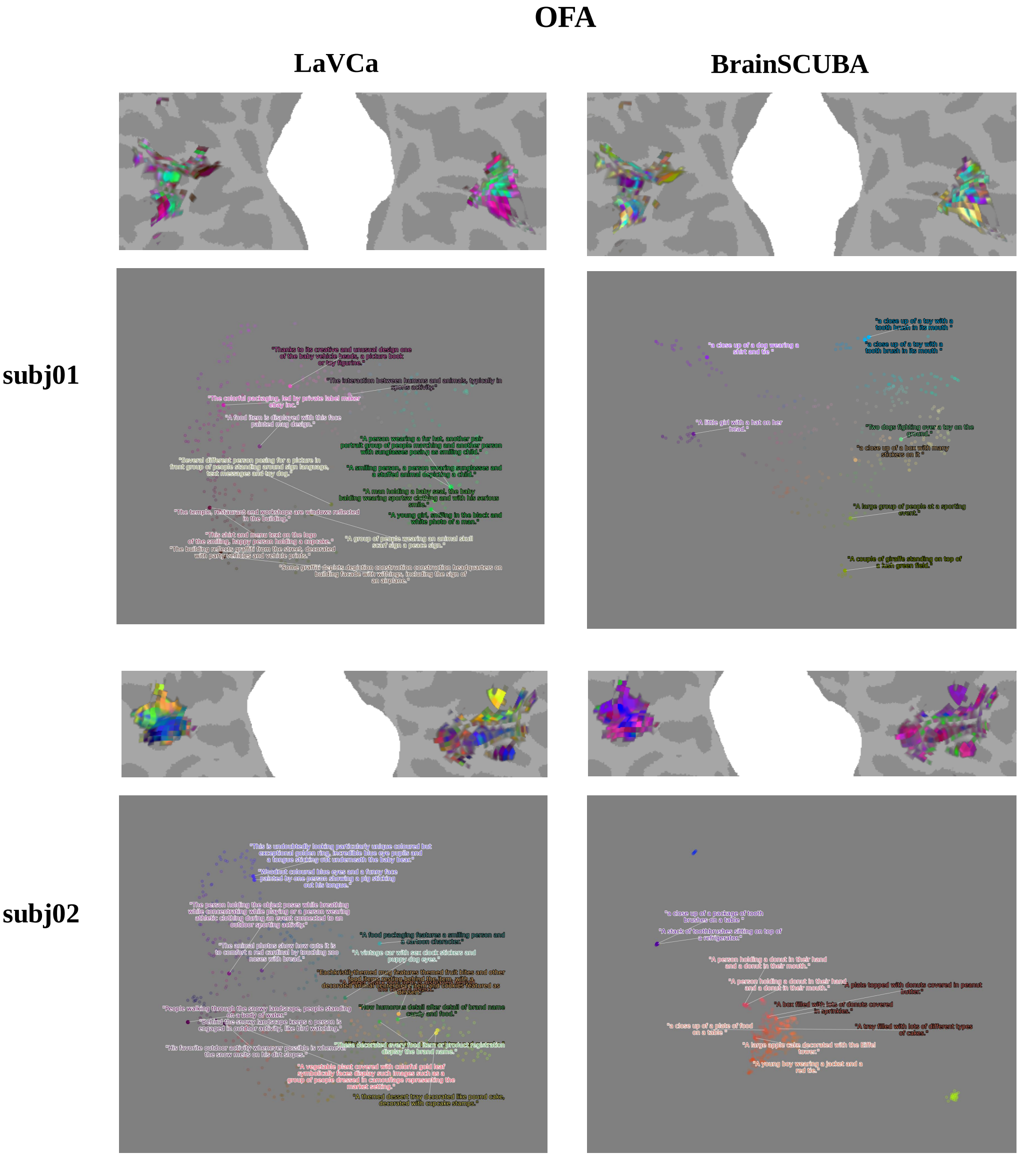


Figure A20: Visualization of OFA captions for subj01 and subj02. For each subject, the captions' UMAP representations were mapped onto a flatmap (top). The top 2 captions of each cluster in the UMAP space were visualized (bottom). The horizontal axis represents UMAP2, and the vertical axis represents UMAP2.

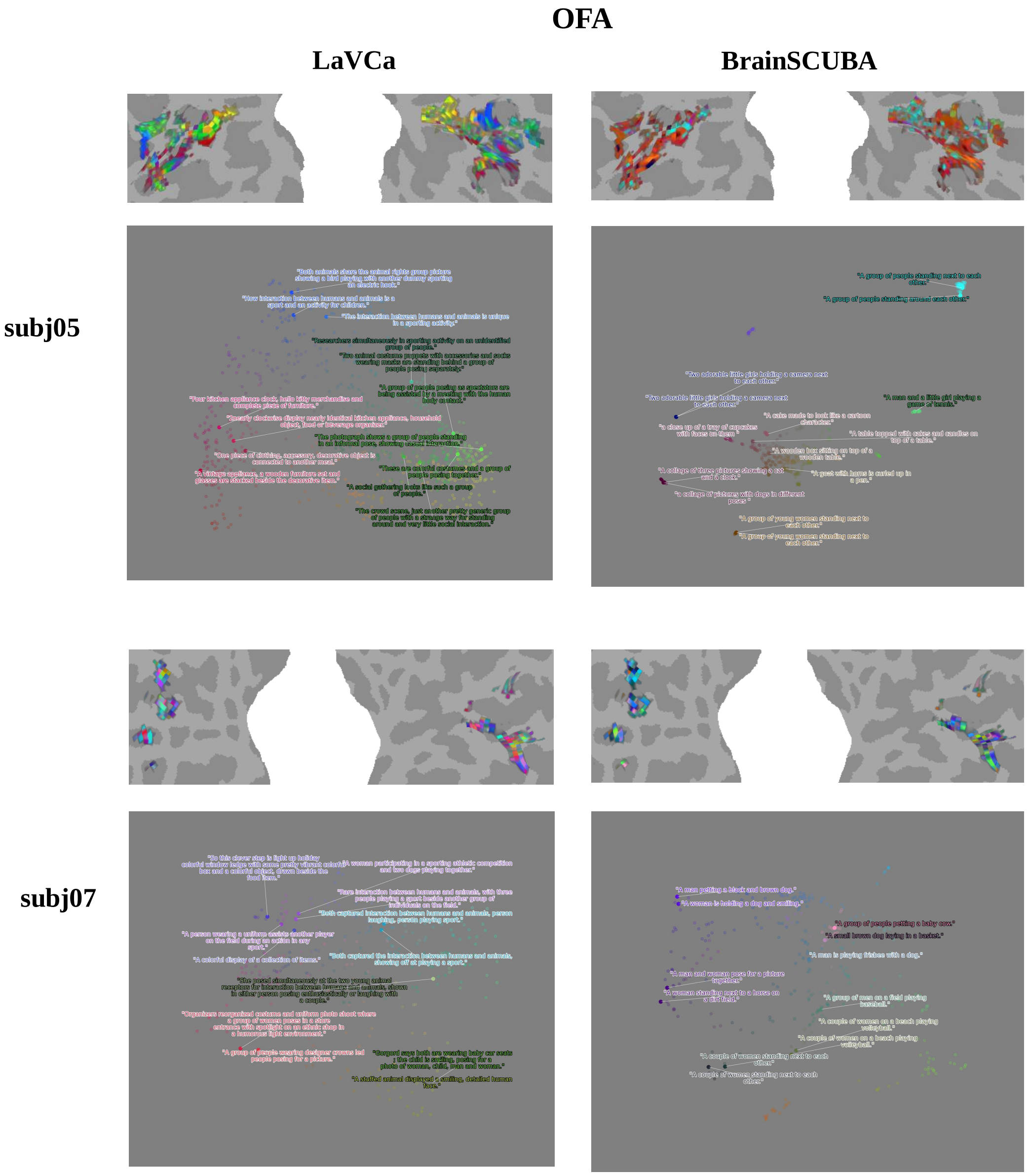


Figure A21: Visualization of OFA captions for subj05 and subj07. The captions' UMAP representations were mapped onto a flatmap (top) for each subject. The top 2 captions of each cluster in the UMAP space were visualized (bottom). The horizontal axis represents UMAP2, and the vertical axis represents UMAP2.

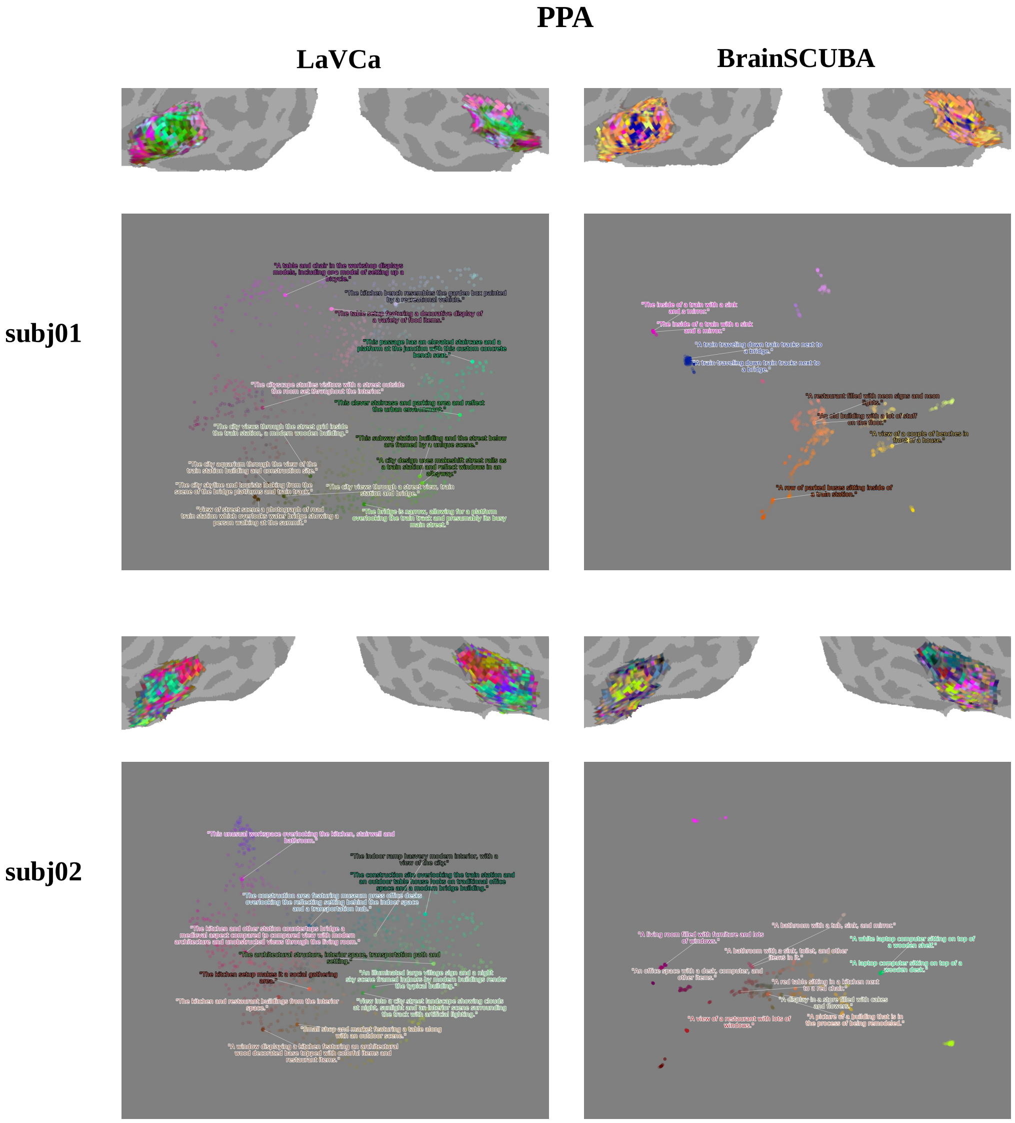


Figure A22: Visualization of PPA captions for subj01 and subj02. The captions' UMAP representations were mapped onto a flatmap (top) for each subject. The top 2 captions of each cluster in the UMAP space were visualized (bottom). The horizontal axis represents UMAP2, and the vertical axis represents UMAP2.

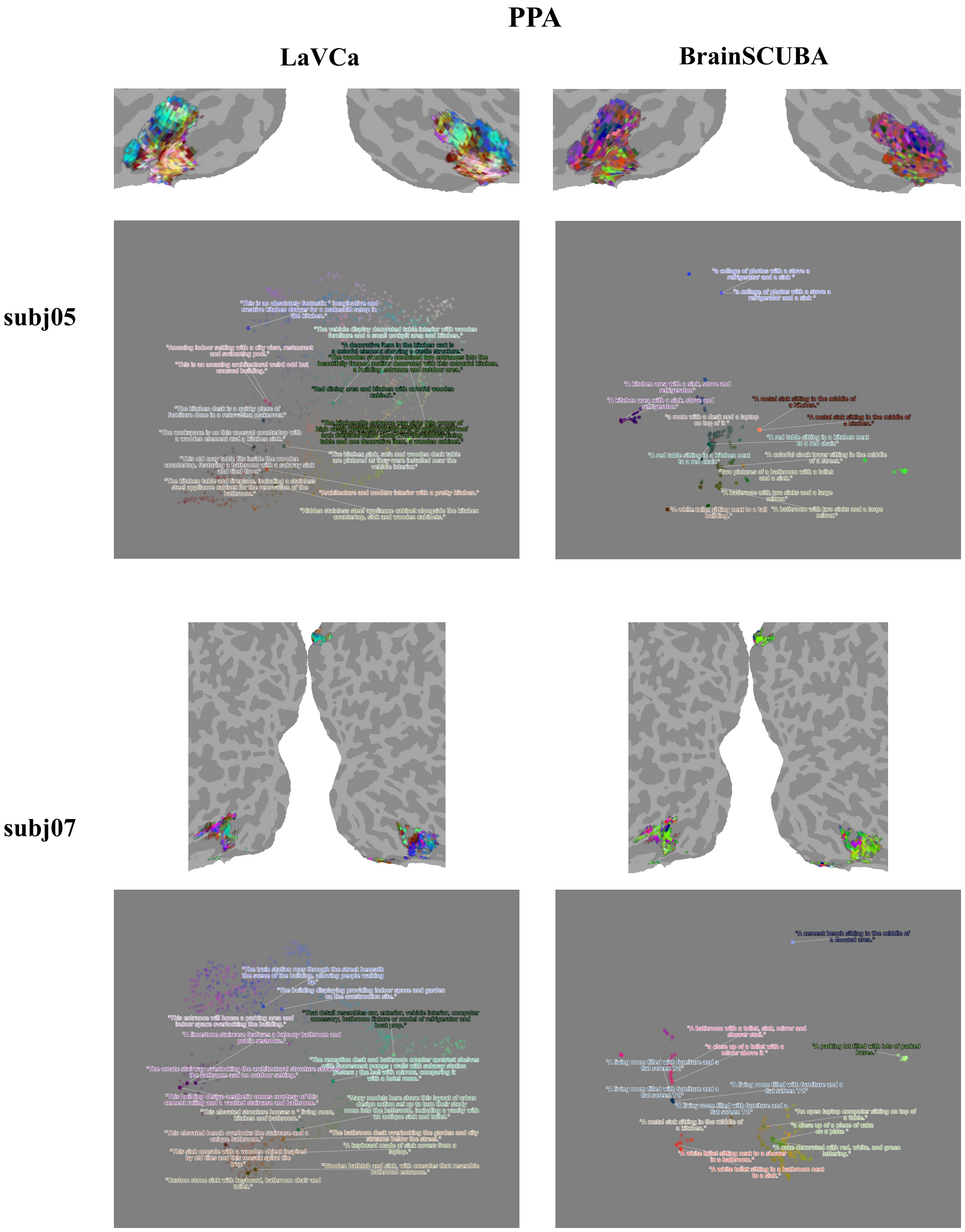


Figure A23: Visualization of PPA captions for subj05 and subj07. The captions' UMAP representations were mapped onto a flatmap (top) for each subject. The top 2 captions of each cluster in the UMAP space were visualized (bottom). The horizontal axis represents UMAP2, and the vertical axis represents UMAP2.