BrainLMM: A Label-Free Framework for Mapping Multi-Semantic Representation in the Human Visual Cortex

Tan Gao[1], Mufan Xue[2], Haofang Zheng[3], Shuo Lv[4], Jia Xu[1], Dabin Sheng[1], Ziming Mao[1], Xinyu Wu[4], Andrew Luo[5], Guoyuan Yang[2,4*]

[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[2]School of Interdisciplinary Science, Beijing Institute of Technology, Beijing, China
[3]School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China
[4]School of Medical Technology, Beijing Institute of Technology, Beijing, China
[5]Institute of Data Science, University of Hong Kong, Hong Kong, China
yanggy@bit.edu.cn

Abstract

Previous studies leveraging artificial neural networks have been used to investigate the semantic coding within human visual cortex. However, building an interpretable label-free framework that can effectively map brain responses to multiple coexisting semantic concepts remains largely unexplored. Here, we propose BrainLMM, a label-free framework for multi-semantic mapping of voxel responses by combining diverse vision encoders with the Describe-and-Dissect strategy, enabling a hypothesis-free analysis of the human high-level visual cortex. First, we construct voxel-wise encoding models leveraging diverse vision encoders to predict visual cortical responses to natural scene images. Then, we use BrainLMM to map individual brain voxels to multiple semantics without requiring any predefined labels. To evaluate the effectiveness of our method, we compute Pearson correlation coefficients to compare the multi-semantic mappings produced by BrainLMM and CLIP-MSM with ground-truth voxel responses within selective cortical areas. Our findings indicate that BrainLMM achieves more accurate predictions of visual responses compared to CLIP-MSM. Finally, to demonstrate the multi-semantic mapping capability of our method, we project multiple representative semantic concepts onto the cortical surface for visualization. Our method enables the discovery of voxels that exhibit strong activation in response to previously undefined semantic concepts across two independent datasets: the Natural Scenes Dataset (NSD) and the Natural Object Dataset (NOD).

Code — https://github.com/BIT-YangLab/BrainLMM

Introduction

The human visual cortex is capable of transforming high-dimensional sensory input into structured semantic representations, forming category-selective activation patterns in distributed cortical regions (Grill-Spector and Weiner 2014; Bao et al. 2020). A substantial body of evidence from neuropsychological case studies and intracranial electrophysiology converges to demonstrate that higher-order visual regions are organized into specialized cortical areas, each selectively engaged in processing semantic categories such as faces, places, bodies, words, and food (Puce et al. 1996; Kanwisher, McDermott, and Chun 1997; Epstein and Kanwisher 1998; Maguire 2001; Grill-Spector 2003; Pennock et al. 2023; Jain et al. 2023). Large-scale brain imaging datasets (Chang et al. 2019; Allen et al. 2022) have driven significant advancements in encoding models for the ventral visual pathway (Qiao et al. 2021), including goal-driven and data-driven approaches (Cadena et al. 2019; Xiao et al. 2022). Recently, the rise of explainable neural network techniques, such as Network Dissection (Bau et al. 2017), Compositional Explanations (Mu and Andreas 2020), and CLIP Dissection (Oikarinen and Weng 2023), has enhanced the interpretability of human visual cortical encoding models. These techniques have broadened their use in hypothesis-free analyses, enabling the exploration of tuning features for ecologically significant intermediate properties (Sarch et al. 2023).

Prior study has demonstrated the parallelism between neural networks and the brain in information processing (Wehbe et al. 2014). In addition, research on modeling the IT cortex shows neural networks can be used to model the human ventral visual stream (Yamins et al. 2014). Recent hypothesis-free encoding models have revealed category-selective responses (e.g., food, person) in the human ventral visual cortex, mitigating biases introduced by predefined labels (Khosla and Wehbe 2022; Xue et al. 2024). Using vision–language embeddings, neural networks can directly predict brain responses to natural scenes with improved accuracy (Wang et al. 2023). Recent interpretable frameworks, such as BrainSCUBA (Luo et al. 2024) and CLIP-MSM (Yang et al. 2025), leverage CLIP-based voxel-wise models and semantic dissection to map brain voxels to multiple concepts, solving the problem of multi-semantic concepts corresponding to a single voxel (Tarr and Gauthier 2000; Doshi and Konkle 2023).

However, existing interpretable encoding models typically fall into one of two categories: those that enable multi-semantic representations but rely on predefined label sets, and those that support label-free analysis yet associate each voxel with only a single dominant concept. Thus, they cannot capture multiple neural selectivities without category-level priors. To overcome these limitations, we introduce

*Corresponding author: Guoyuan Yang

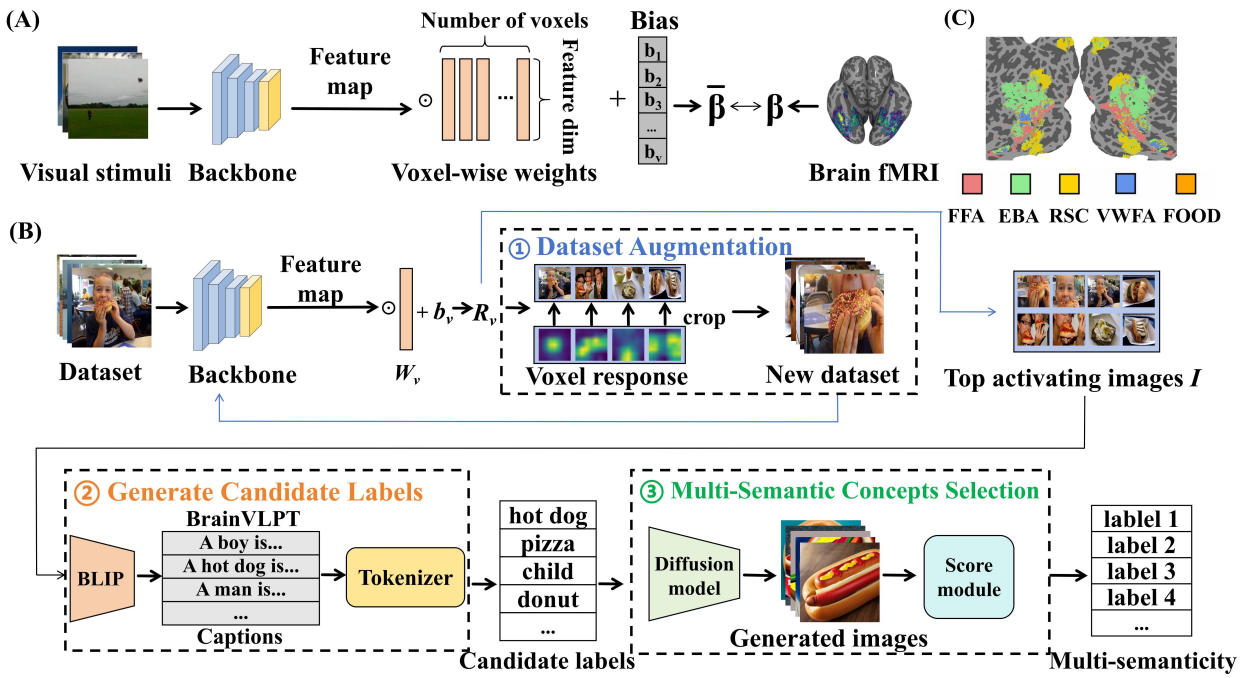

Figure 1: Architecture of BrainLMM. (A) For each voxel, an individualized ridge regression model is instantiated to capture its unique response profile. All voxels share a common backbone network based on ViT or CNN. The learned voxel-wise weights are applied to the output features of the backbone to predict each voxel's response to a given input image. The models are optimized by minimizing the mean squared error (MSE) between the predicted and actual voxel responses for both the NSD and NOD datasets. (B) The images are first processed by a pretrained convolutional backbone to extract feature maps from a specific intermediate layer. For each voxel, a linear transformation parameterized by voxel-specific weights W_v and bias b_v is applied to these features, producing a predicted activation that reflects the voxel's response to the input image. Guided by this voxel-wise activation, targeted data augmentation is performed to generate a diverse set of semantically relevant images. These augmented samples are then re-evaluated through the same voxel pathway (indicated by blue arrows), and the top - K images that elicit the strongest predicted responses are selected. For these images, we leverage BrainVLPT (Visual Linguistic Projection to Tokens) module to generate a pool of candidate semantic labels. To model the inherent semantic diversity of voxel responses, we employ a diffusion-based generative module in combination with a scoring network, which jointly selects multiple representative labels. This enables fully label-free discovery of multi-semantic mappings grounded in brain activity. (C) Visualization of selective ROIs along the ventral visual pathway for encoding models of subject S5 in the NSD.

BrainLMM: a label-free framework for multi-semantic mapping of voxel responses, designed for hypothesis-free analysis of the human high-level visual cortex. We begin by constructing voxel-wise encoding models that use CLIP image embeddings to predict brain responses to natural scene images from the NSD (Allen et al. 2022) and NOD (Gong et al. 2023). Then, we improve upon Describe-and-Dissect (DnD) (Bai et al. 2025), a training-free and label-free interpretability framework, and apply it to generate semantic labels for individual voxels. Our improved variant retains DnD's ability to produce open-vocabulary, natural language descriptions without relying on predefined semantic categories. We further compare the performance of BrainLMM with CLIP-MSM by assessing the similarity between the voxel-wise weights obtained from the training process and the labels of semantic mapping generated by the two explainable methods. To enable voxel-wise multi-semantic mapping, we normalize the semantic mapping scores to softly optimize the alignment between image-derived categorical labels and predicted brain responses. We validate this approach through a hypothesis-free brain activation analysis, quantifying the alignment between BrainLMM-reconstructed voxel selectivities and ground-truth semantic activations. We further confirm its reproducibility on two widely used fMRI datasets, NSD and NOD. In summary, our main contributions are as follows:

- We improve and apply DnD to fMRI response-optimized encoding models for the first time, enabling the identification of semantic concepts in images that elicit the strongest voxel-level selectivity.
- We propose BrainLMM, a framework that enables voxel-wise multi-semantic mapping in the human high-level visual cortex, and employ it to investigate cortical semantic selection without relying on predefined labels.
- We validated BrainLMM on two large-scale, high-quality fMRI datasets of natural scene images, including the NSD and NOD.

Related Work

Voxel-wise Encoding Models Based on CLIP

CLIP (Contrastive Language-Image Pretraining) is an efficient model designed to learn image representations directly from the raw text descriptions associated with images, providing a broader and more flexible source of supervision (Radford et al. 2021). By leveraging large-scale natural language supervision, CLIP enables image models to learn directly from massive amounts of web text. The CLIP model notably improves the accuracy of behavioral image predictions, suggesting that language plays a key role in shaping how the human mind interprets visual information in computer vision tasks (Conwell et al. 2023). In the context of human high-level visual cortex, CLIP has demonstrated strong effectiveness in predicting voxel-wise brain responses to natural scenes, as shown in the NSD dataset (Wang et al. 2023). Several research efforts have employed diffusion models to reconstruct both images and text from fMRI data, leading to enhanced semantic precision and deeper biological understanding (Luo et al. 2023; Ferrante et al. 2023; Takagi and Nishimoto 2023). CLIP-MSM focuses on mapping multiple semantic categories to individual voxels in the high-level visual cortex (Yang et al. 2025), combining CLIP-based encoding models with CLIP Dissection to identify overlapping semantic representations without being constrained to a single label per voxel.

Describe-and-Dissect

DnD is a label-free and training-free framework for interpreting deep neural networks (Bai et al. 2025). It uses large pre-trained vision-language models like BLIP (Li et al. 2022) to generate natural language descriptions for hidden neurons without relying on predefined concepts or annotated datasets. As a result, it can work with the unlabeled datasets (Krizhevsky, Hinton et al. 2009). DnD works by identifying the image patches that most strongly activate each neuron, matching them with semantic concepts via multimodal embeddings, and generating descriptions using a language model. Since it requires no retraining, DnD is computationally efficient and readily applicable across different models and datasets. Experiments have shown that DnD consistently outperforms earlier interpretability methods in both the quality and informativeness of neuron descriptions. Compared to traditional baselines, it is more than twice as likely to be selected as the best explanation for a given neuron. Overall, DnD offers a powerful and general-purpose solution for probing the functional roles of internal units in deep visual models.

Methods

First, we begin by introducing the datasets and defining the Regions of Interest (ROIs). Then, we detail the parameterization and training of our voxel-wise encoding models (Naselaris et al. 2011), which predict brain responses from image embeddings (Fig. 1A). Finally, we present how our BrainLMM framework captures multi-semantic representation of each voxel without relying on predefined category labels (Fig. 1B).

Regions of Interests

In the case of the NSD dataset, we used five regions: the fusiform face area (FFA) for face selectivity (Kanwisher, McDermott, and Chun 2002), the extrastriate body area (EBA) for body selectivity (Downing et al. 2001), the retrosplenial cortex (RSC) for place selectivity (Brodmann 1909), and the visual word form area (VWFA) for word selectivity (Cohen et al. 2000). These ROIs were defined using independent category localizer data (Allen et al. 2022), with a t-value threshold of $t > 0$. Additionally, we identified the food-selective region (FOOD) following the method outlined in (Jain et al. 2023). We merged these five areas to create a final selective ROI. For the NOD dataset, we used the official ROIs corresponding to FFA, RSC, EBA, and VWFA, with a more stringent threshold of $t > 2.3$. Similarly, the final selective ROI was obtained by combining these four regions. To visualize the results, we used Pycortex for rendering in the native cortical space (Gao et al. 2015), as demonstrated for subject S5 in Figure 1C.

BrainLMM Framework

Voxel-wise Encoding Models We employed OpenAI's CLIP models with ViT-B/32 and ResNet50 backbones for image feature extraction, as these models have been demonstrated to align well with the cortical hierarchy of the human brain (Millet et al. 2022). Image embeddings derived from $\text{ResNet50}_{\text{CLIP}}$ and $\text{ViT-B/32}_{\text{CLIP}}$ were used to predict voxel-wise brain responses, with feature dimensionalities of 1024 and 512, respectively. We also utilized ImageNet-pretrained (Deng et al. 2009) ResNet50 and AlexNet for image feature extraction. Features from the average pooling layer of $\text{ResNet50}_{\text{ImageNet}}$ and $\text{AlexNet}_{\text{ImageNet}}$ were used to predict brain responses, with feature dimensionalities of 2048 and 9216. We used ridge regression models implemented in PyTorch to predict the averaged voxel-wise fMRI response to individual images (Paszke et al. 2019). For both the NSD and NOD datasets, images were split into training and test sets using an 85:15 ratio. Regularization parameters were logarithmically spaced from 10^{-8} to 10^{10}. Model performance was quantified using the coefficient of determination (R^2). Finally, we assessed statistical significance by performing bootstrap testing on 2,000 resampled test sets, followed by false discovery rate (FDR) correction (Benjamini and Hochberg 1995).

Input For each voxel, the model predicts its response to all images, using as input the same stimulus images presented during the fMRI experiments in both the NSD and NOD datasets.

Algorithm BrainLMM involves three key steps (Fig. 1B).

1. Dataset Augmentation. For a given input image, the activation map of voxel v is obtained by passing the image through the model. We define the regions in the image with high activation values as the f (foreground) and those with low values as the b (background). To determine the optimal threshold for separating the foreground and background, we apply Otsu's method (Otsu et al. 1975), which automatically selects the threshold that maximizes the inter-class variance

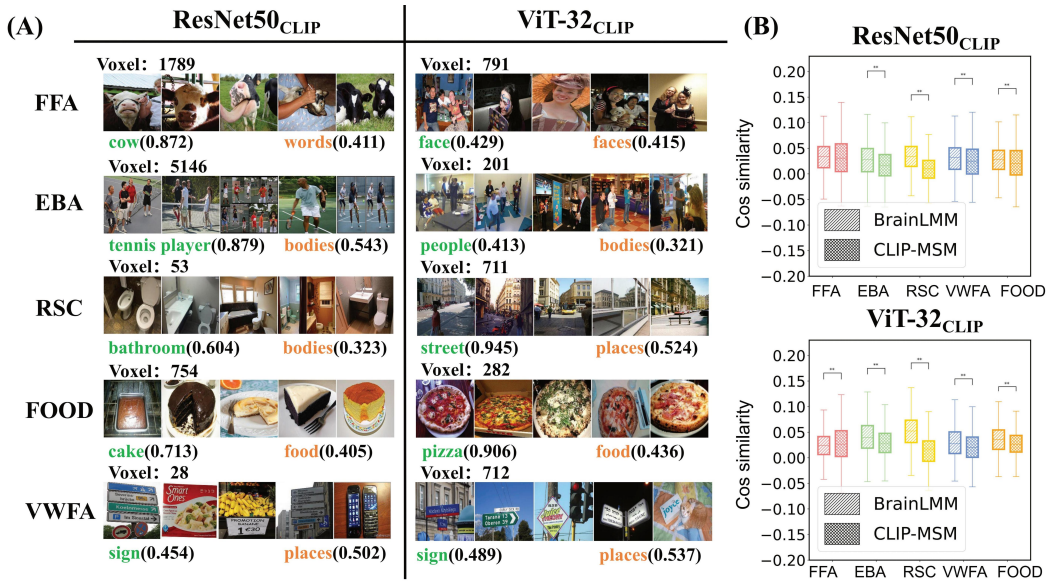

Figure 2: Performance evaluation of BrainLMM and CLIP-MSM for semantic map capture. We evaluate the performance of both methods by mapping the label of each voxel to its corresponding text embedding using the text encoder. Both ResNet50$_{\text{CLIP}}$ and ViT-32$_{\text{CLIP}}$ utilize their respective text encoders, in line with the original CLIP implementation. To compare the methods, we calculate the cosine similarity between two elements: the text embeddings derived from the labels produced by the two explainable methods for all voxels within the selected regions, and the voxel-specific weights obtained during the training process. (A) We visualize the labels and corresponding images for the same voxel across different selective regions, as generated by the two explainable methods. Labels rendered in green denote results produced by BrainLMM, whereas those rendered in orange denote results derived from CLIP-MSM. (B) A higher cosine similarity between the text embeddings and voxel-wise weights signifies a stronger semantic alignment with the selective regions. As demonstrated, BrainLMM provides a more precise semantic mapping compared to CLIP-MSM ($^{*}P < 0.05$, $^{**}P < 0.001$, paired t-test).

between the foreground and background pixels. We set P_b and P_f represent the proportions of background and foreground pixels, and r_b and r_f are their respective mean intensities. In the thresholding strategy used in BrainLMM, P_f corresponds to the fraction of pixels in the activation map that exceed a candidate threshold λ, while P_b denotes the fraction of pixels below λ. The value of λ that yields the maximum q^2 is selected as the activation threshold. q^2 is computed as:

$$q^2 = P_b P_f (r_b - r_f)^2 \tag{1}$$

We then use λ as the threshold to binarize the activation map. For the binary masked activation map, we utilize OpenCV's contour detection algorithm to highlight the contours of the salient regions. Then, we compress the traced contour segments into four end points, each representing a vertex of the bounding box for the salient region. Finally, we overlay the bounding boxes on the original stimulus images and crop them to construct the final augmented dataset.

2. Generate Candidate Labels. To identify the most highly activating images for a given voxel v, we select the top-N images from the new dataset, where each image m_i is chosen based on the highest values of $g(A_v(m_i))$ forming a set denoted as M. Here, $A_v(m_i)$ denotes the activation map of voxel v in response to input m_i, and g is defined as the spatial mean. To generate candidate labels for each voxel, we introduce the BrainVLPT (Visual Linguistic Projection to Tokens) module. This module combines a BLIP image-to-text model with a text tokenizer to produce labels. For each image m_i, we pass it through the BLIP model to generate a caption. All the captions of M are then tokenized, and noun tokens are extracted as the candidate labels.As a result, each voxel v is associated with a set of candidate labels L, where L contains n labels. The number n corresponds to the total number of distinct nouns extracted from all the captions generated for the images in set M.

3. Multi-Semantic Concepts Selection. For each label $l_i \in L$, t synthetic images are generated using a diffusion model. We denote the resulting image set for each label as M_i, where $\mid M_i \mid = t$. For each voxel, the entire new dataset $M_{all} = \sum_{j=1}^{n} M_i = \{m_1, \ldots, m_{n \cdot t}\}$, which represents the entire set of generated images. Then, We feed the synthetic dataset M_{all} back into the target model to rank the images based on the voxel response. Specifically, the scalar response of target voxel v for each image $m_i \in M_{all}$ is computed as $g(A_v(m_i))$. The resulting set of scalar responses is denoted as $G_v = g(A_v(m_1)), ..., g(A_v(m_{n \cdot t}))$. To record the rank of each image m_i in M_{all}, we set a function $\mathrm{R}(y; Y)$ to return the rank of an element y within the set Y. For each label l_i , the ranks of its corresponding generated images are recorded in $T_i = \{\mathrm{R}(g(A_v(m)); G_v), \forall m \in M_i\}$. The elements in T_i are then sorted in ascending order by their ranks with the corresponding numerical values decreasing accord-

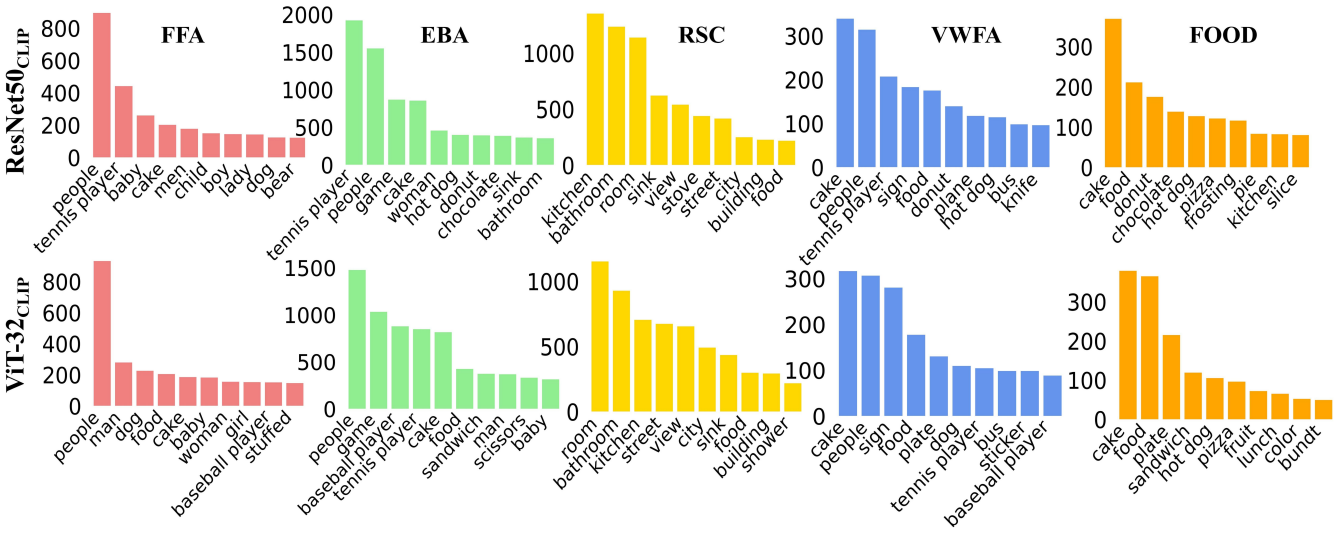

Figure 3: BrainLMM facilitates the fine-grained dissection of high-level visual categories within selective regions. For subject S5 in the NSD dataset, we present voxel-wise semantic mappings captured by BrainLMM across selective regions for two encoding models. Within each region of interest, the top 10 most frequent semantic labels are identified based on voxel-level mappings. The X-axis denotes these selected semantic labels, while the Y-axis indicates the number of voxels associated with each label. Notably, while "cake" emerged as the top-ranked label in the VWFA region, the corresponding images typically featured cakes decorated with textual patterns. The results for other models and subjects are presented in the Appendix.

ingly, such that T_{i1} corresponds to the lowest-ranking (i.e., weakest-activating) image. The concept score for the label l_i is computed as the mean of the squared ranks of the top-a images with the lowest ranks in T_i, which helps reduce sensitivity to poorly generated samples. The scoring function $V(T_i)$ assigns a semantic relevance score to each candidate concept based on the rank distribution of its associated generated images. For example, T_i = [8, 7, 6, 1], where T_{i1} (rank 8) denotes the weakest activation. The concept score for l_i is obtained by averaging the squared ranks of the top-a images with the lowest ranks in T_i(e.g., [7, 6, 1]), as defined in Equation (2).

$$V(T_i) = -\frac{1}{a}\sum_{j=1}^{a} T_{ij}^2 \tag{2}$$

Experiments

In this section, we train encoding models separately using the NSD and NOD datasets to predict brain responses for each subject. For the NSD dataset, all stimulus images were used as input to all models ($\text{ResNet50}_{\text{CLIP}}$, $\text{ViT-32}_{\text{CLIP}}$, $\text{ResNet50}_{\text{ImageNet}}$, and $\text{AlexNet}_{\text{ImageNet}}$). For the NOD dataset, only images actually viewed by participants are used as input to $\text{ResNet50}_{\text{CLIP}}$ and $\text{ViT-32}_{\text{CLIP}}$. We apply the proposed BrainLMM framework to these encoding models to automatically inspect the functional selectivity of individual brain voxels. Then we compare the multi-semantic mapping accuracy of BrainLMM and CLIP-MSM. Finally, we validate the effectiveness of BrainLMM by quantifying the correspondence between its reconstructed brain responses and the ground-truth activations for the semantic categories of faces, bodies, places, food, and words.

Large-scale fMRI Datasets

The NSD dataset (Allen et al. 2022) provides high-density fMRI recordings from eight participants (six female, aged 19-32). We used the *betas_fithrf_GLMdenoise_RR* estimates for beta value preparation. Cortical surface reconstructions were performed with FreeSurfer (Dale, Fischl, and Sereno 1999). We calculated the z-scores of the beta values across runs and averaged them over up to three repetitions for each image, resulting in one fMRI response per voxel per image. The visual stimuli consisted of square-cropped, resized images from the COCO dataset (Lin et al. 2014), each subtending a visual angle of $8.4 \times 8.4°$. We also employed the NOD dataset (Gong et al. 2023) for validation, which contains fMRI responses to 57,120 images collected from 30 participants scanned scanned on a 3T MRI scanner. We selected a subset of nine participants with high fMRI quality (five female, aged 19-26) who each viewed 4,000 unique ImageNet images along with 120 shared COCO images. The analysis used preprocessed surface-based data from *ciftify* (Dickie et al. 2019) to ensure data quality.

BrainLMM Improves Multi-Semantic Mapping

We conducted a comparative analysis between BrainLMM and CLIP-MSM. After model dissection, we visualized the labels and corresponding images for the same voxel across each selective region, as generated by both explainable methods (Fig. 2A). For both methods, the label with the highest semantic mapping score was selected, and the scores were normalized using the softmax function. The normalized score is shown in parentheses alongside each label. Our results indicate that, BrainLMM enables a richer and more fine-grained semantic mapping of individual voxels.

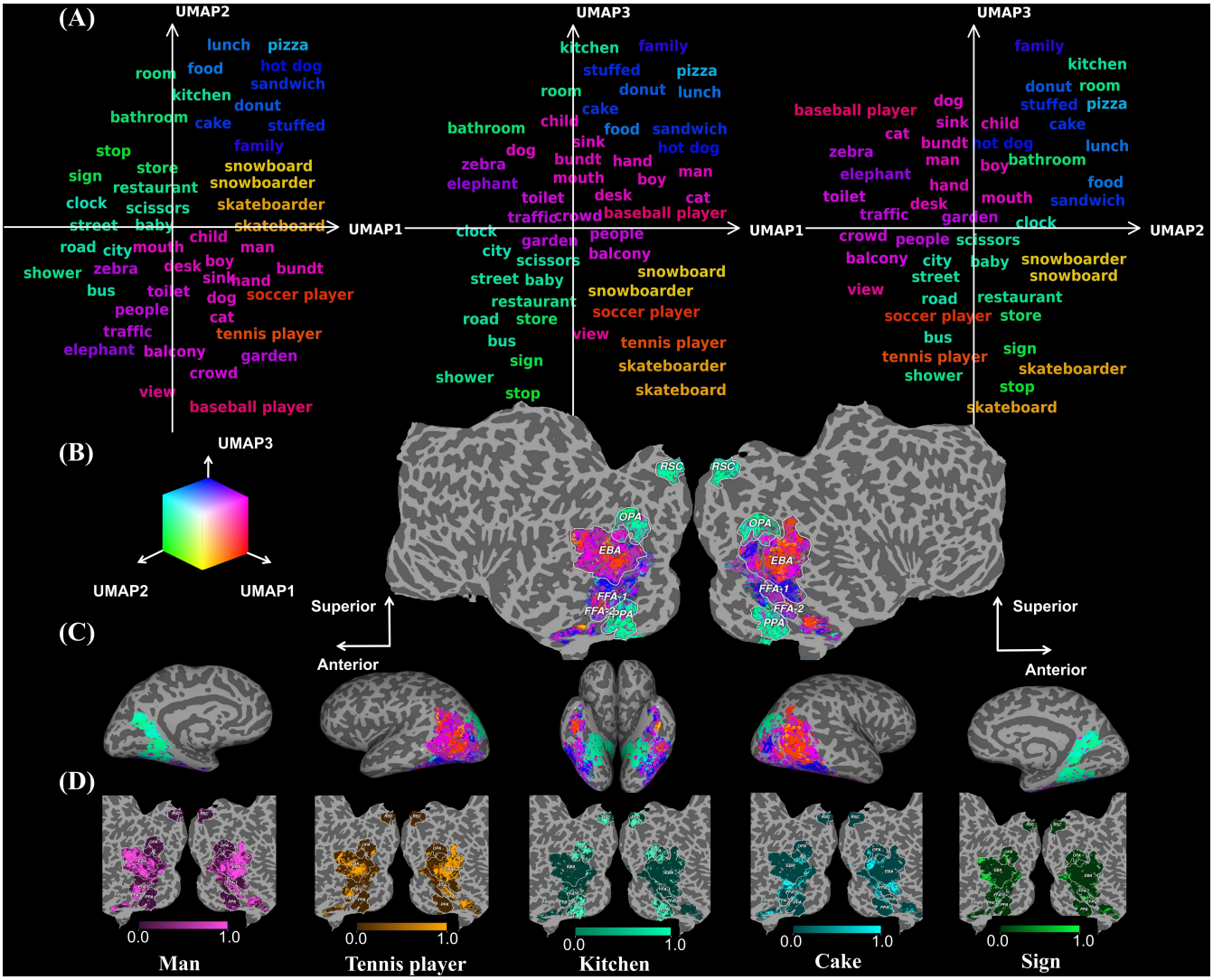

Figure 4: Voxel-wise visualization results on the cortical surface using BrainLMM. (A) Voxel-level dissection results for subject S5 in the NSD dataset, using ResNet50$_{\text{CLIP}}$ as the visual backbone. Image embeddings for all stimulus images are computed using the CLIP image encoder, followed by a UMAP projection into three dimensions. For each voxel, the highest-scoring label from its dissection result is embedded using the CLIP text encoder, projected into the same UMAP space, and normalized to generate RGB color values. (B) Flatmap of subject S5 with labeled ROIs. (C) Inflated cortical view showing the projection of semantic labels for S5. (D) Several representative semantic concepts are projected onto the flattened cortical surface. The color of each voxel reflects the similarity between its representation and the given concept, with intensity encoded by the color map. The visualization style is adapted from (Luo et al. 2024).

To conduct a more detailed quantitative comparison between the two methods, we focused on their ability to characterize semantic mappings across all voxels within each selective region. We computed the cosine similarity between the text embeddings of labels generated by the two explainable methods and the voxel-specific weights obtained from the CLIP models. Our analysis reveals that, for subject S5 in the NSD dataset, BrainLMM achieves higher cosine similarity than CLIP-MSM across most selective cortical regions (Fig. 2B), indicating more accurate semantic alignment at the voxel level. In the FFA region, the median cosine similarity between BrainLMM and CLIP-MSM is comparable. Although BrainLMM shows a slightly lower maximum, it achieves a substantially higher minimum, suggesting more stable performance across voxels. In the EBA, RSC, FOOD, and VWFA regions, BrainLMM consistently outperforms CLIP-MSM, with both higher median and overall similarity scores. Results from additional NSD subjects (S1-S8) and NOD participants (S1-S9) are provided in the Appendix.

Label-free Mapping with BrainLMM

For each voxel, we selected the top-ranked label from its candidate label set. We then aggregated these labels across all voxels within each category-selective region and identified the 10 most frequently occurring nouns (Fig. 3). Here, we present the results for subject S5 in the NSD dataset for the FFA, EBA, RSC, FOOD, and VWFA, using ResNet50$_{\text{CLIP}}$ and ViT-32$_{\text{CLIP}}$ models. The categories that most strongly activated in these brain regions show strong correspondence with those identified in prior studies (Stigliani, Weiner, and Grill-Spector 2015; Gauthier, Behrmann, and Tarr 1999; Khosla and Wehbe 2022; Dilks et al. 2013; Khosla et al. 2022), underscoring the validity of our data-driven approach. Additional results are provided in the Appendix.

Cortical Visualization of BrainLMM Results

To map the results onto the cortical surface, we first generate image embeddings for all images using the CLIP image encoder. We then utilize UMAP to perform dimensionality reduction on these embeddings. For each voxel, we input its associated labels into the CLIP text encoder to generate text embeddings, which are subsequently projected into the pre-trained UMAP space and reduced to three dimensions. The results are normalized and interpreted as RGB values, which are used to color the labels of each voxel. Finally, the 50 most common nouns are displayed on the axes (Fig. 4A), which is based on the dissection results of S5 in the NSD dataset, using ResNet50$_{\text{CLIP}}$ as the visual backbone. To further illustrate the multi-semantic mapping captured by our model, we project the highest-scoring label for each voxel onto the flattened cortical surface (Fig. 4B), and also present the corresponding inflated view (Fig. 4C). To demonstrate the multi-semantic mapping capability of our method, we project some representative semantic concepts onto the cortical surface, including tennis player, man, kitchen, cake, and sign (Fig. 4D). The color of each voxel reflects the similarity score between its representation and the given semantic concept, as indicated by the color map. As shown in the visualizations, voxels located in known category-selective areas, such as EBA (bodies), FFA (faces), RSC (places), VWFA (words), and the food-selective region adjacent to FFA, exhibit selectivity profiles aligned with the corresponding semantic concepts. Additional results for other subjects and models are provided in the Appendix.

BrainLMM Improves Brain Alignment

To further evaluate the alignment between our results and actual brain activity, we computed the Pearson correlation coefficient to assess the relationship between the multi-semantic mappings across concepts generated by BrainLMM and the ground-truth brain responses. Subsequently, we performed a statistical analysis comparing the correlation coefficients derived from BrainLMM and CLIP-MSM with respect to ground-truth responses. Our results show that BrainLMM achieves significantly higher accuracy in predicting visual responses than CLIP-MSM across the concepts displayed on the X-axis, using voxel-wise encoding models based on ResNet50$_{\text{CLIP}}$ (Fig. 5A) and ViT-32$_{\text{CLIP}}$

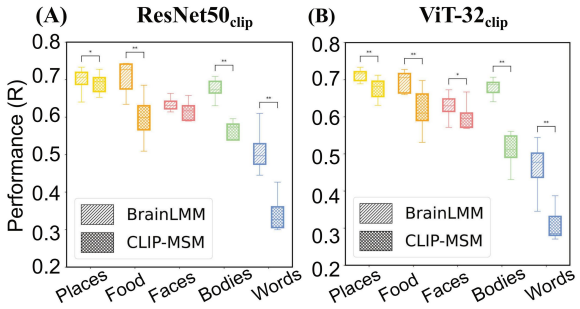

Figure 5: Comparison of multi-semantic mapping between BrainLMM and CLIP-MSM. Pearson correlation coefficients quantifying the alignment between multi-semantic mapping derived from both BrainLMM and CLIP-MSM and the ground-truth voxel responses across selective cortical regions. (A) Using ResNet50$_{\text{CLIP}}$ as the visual encoding backbone, BrainLMM achieves significantly higher prediction accuracy than CLIP-MSM for visual responses associated with places, food, faces, bodies, and words. (B) Using ViT-32$_{\text{CLIP}}$ as the visual encoding backbone, BrainLMM also outperforms CLIP-MSM in predicting responses for places, food, bodies, and words ($^{*}P < 0.05$, $^{**}P < 0.001$, paired *t*-test).

(Fig. 5B). These findings confirm the neuroscientific plausibility of the BrainLMM framework and underscore its ability to model the overlapping, multi-concept selectivity observed in the high-level visual cortex. Overall, BrainLMM effectively captures broad multi-semantic mappings across the high-level visual cortex without relying on any predefined label sets. Additional results are provided in the Appendix.

Conclusion

We introduce BrainLMM, a label-free framework for multi-semantic mapping of voxel responses that supports hypothesis-free analysis in the human high-level visual cortex. By combining deep neural networks with an enhanced DnD method, BrainLMM enables more detailed investigation of voxel-level concept selectivity across a wide range of natural scene categories. Compared to CLIP-MSM, BrainLMM demonstrates superior performance in interpretability evaluations. To assess mapping precision, we compare the brain activations reconstructed by BrainLMM with ground-truth responses for semantic categories. The results show that BrainLMM achieves closer alignment with actual brain activations than CLIP-MSM. To illustrate BrainLMM's multi-semantic mapping capability, we project several representative semantic concepts onto the flattened cortical surface. BrainLMM enables the identification of voxels that respond strongly to previously undefined semantic concepts. We further validate the robustness of BrainLMM using the NSD and NOD. Although this study focuses on the high-level visual cortex, BrainLMM holds promise for future investigations into finer-grained or more abstract representations across the brain.

Acknowledgments

This work was supported by the National Science and Technology Innovation 2030 Program (grant number 2021ZD0200506); the National Natural Science Foundation of China (grants number 82302175 and 62336002). We thank National Center for Protein Sciences at Peking University in Beijing, China, for assistance with data analysis. This work was partially supported by Beijing Institute of Technology Kunpeng&Ascend Center of Cultivation.

References

Allen, E. J.; St-Yves, G.; Wu, Y.; Breedlove, J. L.; Prince, J. S.; Dowdle, L. T.; Nau, M.; Caron, B.; Pestilli, F.; Charest, I.; et al. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1): 116–126.

Bai, N.; Iyer, R. A.; Oikarinen, T.; Kulkarni, A. R.; and Weng, T.-W. 2025. Interpreting Neurons in Deep Vision Networks with Language Models. *Transactions on Machine Learning Research*.

Bao, P.; She, L.; McGill, M.; and Tsao, D. Y. 2020. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814): 103–108.

Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6541–6549.

Benjamini, Y.; and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1): 289–300.

Brodmann, K. 1909. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Barth.

Cadena, S. A.; Denfield, G. H.; Walker, E. Y.; Gatys, L. A.; Tolias, A. S.; Bethge, M.; and Ecker, A. S. 2019. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS computational biology*, 15(4): e1006897.

Chang, N.; Pyles, J. A.; Marcus, A.; Gupta, A.; Tarr, M. J.; and Aminoff, E. M. 2019. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific data*, 6(1): 49.

Cohen, L.; Dehaene, S.; Naccache, L.; Lehéricy, S.; Dehaene-Lambertz, G.; Hénaff, M.-A.; and Michel, F. 2000. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2): 291–307.

Conwell, C.; Prince, J. S.; Hamblin, C. J.; and Alvarez, G. A. 2023. Controlled assessment of CLIP-style language-aligned vision models in prediction of brain & behavioral data. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Dale, A. M.; Fischl, B.; and Sereno, M. I. 1999. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*, 9(2): 179–194.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dickie, E. W.; Anticevic, A.; Smith, D. E.; Coalson, T. S.; Manogaran, M.; Calarco, N.; Viviano, J. D.; Glasser, M. F.; Van Essen, D. C.; and Voineskos, A. N. 2019. Ciftify: A framework for surface-based analysis of legacy MR acquisitions. *Neuroimage*, 197: 818–826.

Dilks, D. D.; Julian, J. B.; Paunov, A. M.; and Kanwisher, N. 2013. The occipital place area is causally and selectively involved in scene perception. *Journal of neuroscience*, 33(4): 1331–1336.

Doshi, F. R.; and Konkle, T. 2023. Cortical topographic motifs emerge in a self-organized map of object space. *Science Advances*, 9(25): eade8187.

Downing, P. E.; Jiang, Y.; Shuman, M.; and Kanwisher, N. 2001. A cortical area selective for visual processing of the human body. *Science*, 293(5539): 2470–2473.

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

Ferrante, M.; Boccato, T.; Ozcelik, F.; VanRullen, R.; and Toschi, N. 2023. Multimodal decoding of human brain activity into images and text. In *UniReps: the First Workshop on Unifying Representations in Neural Models*.

Gao, J. S.; Huth, A. G.; Lescroart, M. D.; and Gallant, J. L. 2015. Pycortex: an interactive surface visualizer for fMRI. *Frontiers in neuroinformatics*, 9: 23.

Gauthier, I.; Behrmann, M.; and Tarr, M. J. 1999. Can face recognition really be dissociated from object recognition? *Journal of cognitive neuroscience*, 11(4): 349–370.

Gong, Z.; Zhou, M.; Dai, Y.; Wen, Y.; Liu, Y.; and Zhen, Z. 2023. A large-scale fMRI dataset for the visual processing of naturalistic scenes. *Scientific Data*, 10(1): 559.

Grill-Spector, K. 2003. The neural basis of object perception. *Current opinion in neurobiology*, 13(2): 159–166.

Grill-Spector, K.; and Weiner, K. S. 2014. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8): 536–548.

Jain, N.; Wang, A.; Henderson, M. M.; Lin, R.; Prince, J. S.; Tarr, M. J.; and Wehbe, L. 2023. Selectivity for food in human ventral visual cortex. *Communications Biology*, 6(1): 175.

Kanwisher, N.; McDermott, J.; and Chun, M. M. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11): 4302–4311.

Kanwisher, N.; McDermott, J.; and Chun, M. M. 2002. Specialized for Face Perception. *Foundations in Social Neuroscience*, 259.

Khosla, M.; Jamison, K.; Kuceyeski, A.; and Sabuncu, M. 2022. Characterizing the ventral visual stream with response-optimized neural encoding models. *Advances in Neural Information Processing Systems*, 35: 9389–9402.

Khosla, M.; and Wehbe, L. 2022. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*, 2022–03.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.(2009).

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Luo, A.; Henderson, M.; Wehbe, L.; and Tarr, M. 2023. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *Advances in Neural Information Processing Systems*, 36: 75740–75781.

Luo, A.; Henderson, M. M.; Tarr, M. J.; and Wehbe, L. 2024. BrainSCUBA: Fine-Grained Natural Language Captions of Visual Cortex Selectivity. In *The Twelfth International Conference on Learning Representations*.

Maguire, E. 2001. The retrosplenial contribution to human navigation: a review of lesion and neuroimaging findings. *Scandinavian journal of psychology*, 42(3): 225–238.

Millet, J.; Caucheteux, C.; Boubenec, Y.; Gramfort, A.; Dunbar, E.; Pallier, C.; King, J.-R.; et al. 2022. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 35: 33428–33443.

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

Naselaris, T.; Kay, K. N.; Nishimoto, S.; and Gallant, J. L. 2011. Encoding and decoding in fMRI. *Neuroimage*, 56(2): 400–410.

Oikarinen, T.; and Weng, T.-W. 2023. CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks. *International Conference on Learning Representations*.

Otsu, N.; et al. 1975. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296): 23–27.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pennock, I. M.; Racey, C.; Allen, E. J.; Wu, Y.; Naselaris, T.; Kay, K. N.; Franklin, A.; and Bosten, J. M. 2023. Color-biased regions in the ventral visual pathway are food selective. *Current Biology*, 33(1): 134–146.

Puce, A.; Allison, T.; Asgari, M.; Gore, J. C.; and McCarthy, G. 1996. Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *Journal of neuroscience*, 16(16): 5205–5215.

Qiao, K.; Zhang, C.; Chen, J.; Wang, L.; Tong, L.; and Yan, B. 2021. Effective and efficient roi-wise visual encoding using an end-to-end cnn regression model and selective optimization. In *International Workshop on Human Brain and Artificial Intelligence*, 72–86. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Sarch, G. H.; Tarr, M. J.; Fragkiadaki, K.; and Wehbe, L. 2023. Brain dissection: fMRI-trained networks reveal spatial selectivity in the processing of natural images. *bioRxiv*, 2023–05.

Stigliani, A.; Weiner, K. S.; and Grill-Spector, K. 2015. Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, 35(36): 12412–12424.

Takagi, Y.; and Nishimoto, S. 2023. 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*, 14453–14463.

Tarr, M. J.; and Gauthier, I. 2000. FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature neuroscience*, 3(8): 764–769.

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

Wehbe, L.; Vaswani, A.; Knight, K.; and Mitchell, T. 2014. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 233–243.

Xiao, W.; Li, J.; Zhang, C.; Wang, L.; Chen, P.; Yu, Z.; Tong, L.; and Yan, B. 2022. High-level visual encoding model framework with hierarchical ventral stream-optimized neural networks. *Brain Sciences*, 12(8): 1101.

Xue, M.; Wu, X.; Li, J.; Li, X.; and Yang, G. 2024. A convolutional neural network interpretable framework for human ventral visual pathway representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6413–6421.

Yamins, D. L.; Hong, H.; Cadieu, C. F.; Solomon, E. A.; Seibert, D.; and DiCarlo, J. J. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23): 8619–8624.

Yang, G.; Xue, M.; Mao, Z.; Zheng, H.; Xu, J.; Sheng, D.; Sun, R.; Yang, R.; and Li, X. 2025. CLIP-MSM: A Multi-Semantic Mapping Brain Representation for Human High-Level Visual Cortex. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9184–9192.