# CLIP-MSM: A Multi-Semantic Mapping Brain Representation for Human High-Level Visual Cortex

**Guoyuan Yang[1,2*], Mufan Xue[1], Ziming Mao[3], Haofang Zheng[4], Jia Xu[3], Dabin Sheng[3], Ruotian Sun[3], Ruoqi Yang[3], Xuesong Li[3]**

[1]Advanced Research Institute of Multidisciplinary Sciences, Beijing Institute of Technology, Beijing 100081, China
[2]School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China
[3]School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[4]School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing 100081, China
yanggy@bit.edu.cn

## Abstract

Prior work employing deep neural networks (DNNs) with explainable techniques has identified human visual cortical selective representation to specific categories. However, constructing high-performing encoding models that accurately capture brain responses to coexisting multi-semantics remains elusive. Here, we used CLIP models combined with CLIP Dissection to establish a multi-semantic mapping framework (CLIP-MSM) for hypothesis-free analysis in human high-level visual cortex. First, we utilize CLIP models to construct voxel-wise encoding models for predicting visual cortical responses to natural scene images. Then, we apply CLIP Dissection and normalize the semantic mapping score to achieve the mapping of single brain voxels to multiple semantics. Our findings indicate that CLIP Dissection applied to DNNs modeling the human high-level visual cortex demonstrates better interpretability accuracy compared to Network Dissection. In addition, to demonstrate how our method enables fine-grained discovery in hypothesis-free analysis, we quantify the accuracy between CLIP-MSM's reconstructed brain activation in response to categories of faces, bodies, places, words and food, and the ground truth of brain activation. We demonstrate that CLIP-MSM provides more accurate predictions of visual responses compared to CLIP Dissection. Our results have been validated using two large natural image datasets: the Natural Scenes Dataset (NSD) and the Natural Object Dataset (NOD).

**Code** — https://github.com/BIT-YangLab/CLIP-MSM

## Introduction

In neuroscience, visual perception is amazingly rapid, allowing humans to categorize visual scenes in just one-tenth of a second (Thorpe, Fize, and Marlot 1996). There is substantial evidence from neuropsychological patients, functional Magnetic Resonance Imaging (fMRI) studies and intracranial recordings indicating that the ventral visual pathway contains distinct regions in processing semantic categories such as faces, places, bodies, words, and food (Puce et al. 1996; Epstein and Kanwisher 1998; Maguire 2001; Kanwisher, McDermott, and Chun 2002; Grill-Spector 2003; Kanwisher and Yovel 2006; Pennock et al. 2023). However, it remains unclear whether all categories have been covered or if there are other, yet undiscovered categories. Several factors contribute to this uncertainty. Firstly, previous studies of the visual ventral pathway have examined only a limited number of stimulus categories, which may not fully encompass the stimulus space preferred by certain neural populations. Additionally, these studies have often been hypothesis-driven, limiting their ability to flexibly capture a wide range of neural features (Gauthier, Behrmann, and Tarr 1999).

Recently, leveraging the convolutional neural networks (CNNs) and combined with network explainable techniques, hypotheses-free approaches have been established to overcome biases toward existing theories and achieve the detection of semantic concepts like "person" or "words" (Khosla and Wehbe 2022; Xue et al. 2024). Through training neural networks to directly predict brain responses to images from natural scenes, encoding models with better model prediction performance can be established (Wang et al. 2023). The emergence of explainable neural network techniques, such as Network Dissection (Bau et al. 2017) and Compositional Explanations (Mu and Andreas 2020), has improved the interpretability of human visual cortical encoding models and enhanced their application in hypothesis-free analyses. These hypothesis-free analyses could even be used to explore the tuning features of ecologically important intermediate properties, including depth, surface normal, curvature, and object relations (Sarch et al. 2023). By further combining the human visual interpretable encoding model with the diffusion models, researchers show that the generated images could semantically align with the original images corresponding to targeted brain regions (Luo et al. 2024, 2023). However, these interpretable encoding models can only simultaneously correspond to each voxel of the human high-level visual cortex to a single category label with the highest explained score, ignoring the condition that multiple neural selectivities coexist within voxels. The study of single-neuron recordings has shown that single neurons can simultaneously encode multiple visual representations with invariant metric properties (Quiroga et al. 2005). Furthermore, other studies demonstrate that the representation of all objects of a category as a manifold in a high-dimensional space (Grill-Spector and Weiner 2014). In

*Corresponding author: Guoyuan Yang.

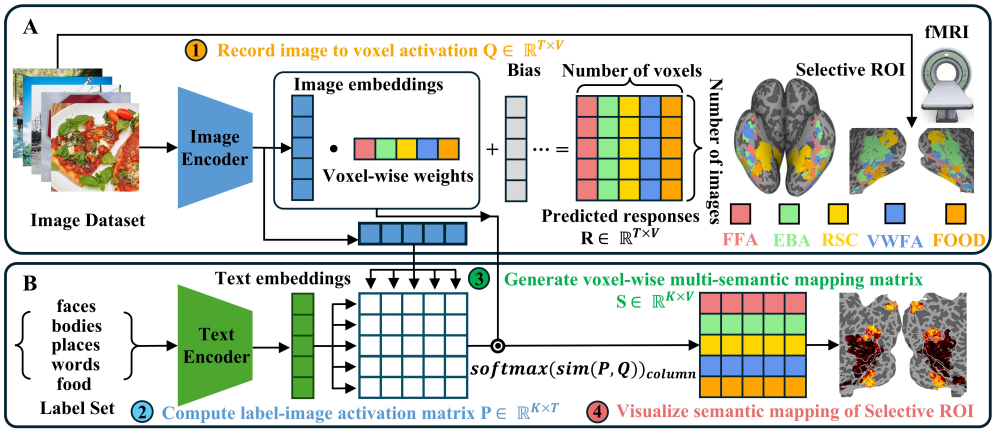


Figure 1: **CLIP-MSM Architecture.** A. Construction of voxel-wise encoding models for the high-level visual cortex. All voxels in the Selective ROI are characterized by image embeddings extracted from the final layer of the CLIP image encoder. The Selective ROI is visualized in native space for subject S5 from the NSD dataset (see Methods for details). B. Multi-semantic mapping. In CLIP Dissection, we used the label set which consists of faces, bodies, places, words, and food, because these categorical labels effectively characterize the selectivity of high-level visual cortical regions. All labels in the images are characterized by the text embeddings extracted from the CLIP text encoder. We compute the label-image activation matrix $\boldsymbol{P}$ using both the image embeddings and text embeddings of the CLIP model. The multi-semantic mapping matrix $\boldsymbol{S}$ was calculated by "$\boldsymbol{softmax}$" the similarity between voxel activation $\boldsymbol{Q}$ and label-image activation matrix $\boldsymbol{P}$. The similarity function used to generate $\boldsymbol{S}$ is SoftWPMI (see methods for more details). Finally, we visualize the third row of $\boldsymbol{S}$ as an example in the high-level visual cortex, reflecting the semantic mapping of Selective ROI to the label of "places".

cortical topographical space, these multiple functional representations are superimposed on the same cortical expanse (Bao et al. 2020; Doshi and Konkle 2023). Therefore, considering the spatial topological overlap of categorical representations in the human high-level visual cortex, there is an urgent need to establish an explainable framework for encoding models that correspond single voxels to multiple category attributes for hypothesis-free analysis.

To address these above limitations, we propose CLIP-MSM, a novel framework for hypothesis-free analysis in the human high-level visual cortex. We first built voxel-wise encoding models based on CLIP image embeddings to predict brain responses to images from the NSD (Allen et al. 2022) and NOD (Gong et al. 2023), which exhibit better performance than other models trained only with image/label pairs (ImageNet-pretrained ResNet50 and AlexNet) (Wang et al. 2023). Then, we leverage CLIP Dissection, a more accurate and flexible explainable AI technique, for enhancing neural network interpretability by automatically dissecting encoding models with unrestricted concepts without the need for any prior labeled data (Oikarinen and Weng 2022). We compared the performance of CLIP Dissection with Network Dissection by measuring the similarity between the labels of semantic mapping generated in the explainable methods and the voxel-wise weights obtained during training. Furthermore, we normalized the semantic mapping score to soft maximize the establishment of brain responses to categorical labels in images to achieve voxel-wise multi-semantic mapping. We verify this method by developing a hypothesis-free analysis of brain activation, in which we quantify the accuracy between ground truth of brain activation patterns in response to categories of faces, bodies, places, words, and food, and the CLIP-MSM's reconstructed brain selectivities. Our CLIP-MSM shows a higher similarity with the ground truth of brain activation compared to the CLIP Dissection of maximizing the semantic score. Finally, we validated the reproducibility of our CLIP-MSM on two widely used brain fMRI datasets, including NSD and NOD. In summary, our key contributions are as follows:

- We apply CLIP Dissection to fMRI response-optimized encoding models for the first time to identify the semantic concepts in images that elicit the most significant selectivity for each voxel at the voxel level.
- We develop the CLIP-MSM framework that can generate multi-semantic mapping in the human high-level visual cortex.
- We used CLIP-MSM to study the human visual cortex responses to natural scene images with categories of faces, bodies, places, words, and food.
- We validated CLIP-MSM on two unprecedented scales and quality of natural scene images fMRI datasets, including NSD and NOD.

## Related Work

### CLIP Model

CLIP stands for Contrastive Language-Image Pretraining, which is an efficient model for learning image representations directly from the raw text about images with a broader source of supervision (Radford et al. 2021). The core idea of CLIP is to address the limitations in other fields such as computer vision, which still pretraining standard models

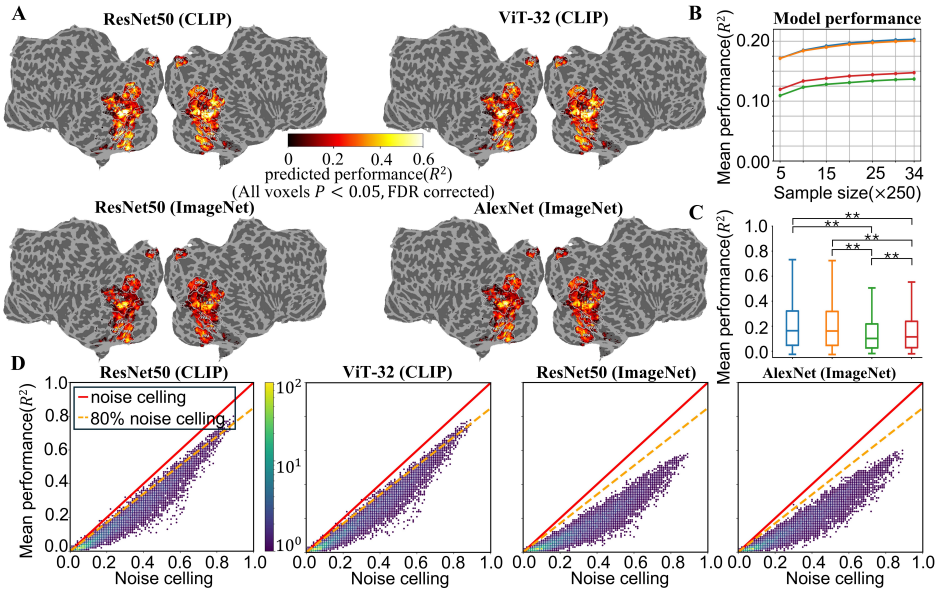


Figure 2: **Comparison of prediction performance across four encoding models.** A. Voxel-wise prediction performance ($R^2$) of ResNet50$_{\text{CLIP}}$, ViT-32$_{\text{CLIP}}$, ResNet50$_{\text{ImageNet}}$ and AlexNet$_{\text{ImageNet}}$ on a held-out test set of the NSD subject S5. We plot voxels that are predicted significantly higher than chance in the flatmap ($P < 0.05$, one-sided test, FDR-corrected) (Benjamini and Hochberg, 1995). B. Performance of all models assessed by varying the number of images used for model training. C. Prediction performance comparison across four models when training with all images ($^*P < 0.05$, $^{**}P < 0.001$, paired $t$-test). In Figure 2B and 2C, blue = ResNet50$_{\text{CLIP}}$, orange = ViT-32$_{\text{CLIP}}$, green = ResNet50$_{\text{ImageNet}}$, red = AlexNet$_{\text{ImageNet}}$. D. The noise ceiling is calculated as 100 times the square of the noise ceiling signal-to-noise ratio, divided by the square of the noise ceiling signal-to-noise ratio plus the reciprocal of the number of averaged trials. Two-dimensional histogram of model performance in $R^2$ against noise ceiling and 85% noise ceiling across all voxels in the Selective ROI. The density of voxels is shown in a log scale.

on crowd-labeled datasets, such as ImageNet (Deng et al. 2009). Specifically, CLIP through understanding the behaviors of image models trained with large-scale natural language supervision enables direct learning from a bulk of web text, which outperforms the best publicly available ImageNet model. In the human high-level visual cortex, CLIP is extraordinarily good at predicting voxel-wise responses when humans view scenes in the NSD (Wang et al. 2023). Furthermore, in conjunction with empirically controlled alternatives, the CLIP model strongly increased the predictivity of behavioral image assessments, demonstrating that language shapes the perceptual representation of the human mind in computer vision (Conwell et al. 2023).

Recent studies have expanded CLIP's application in neuroscience. For example, BrainSAIL uses dense semantic embeddings from CLIP to map neural selectivity in the human visual cortex, shedding light on how specific image regions drive voxel responses (Luo et al. 2024). Similarly, BrainSCUBA leverages CLIP embeddings to generate detailed captions for voxel-wise preferred stimuli, enhancing our understanding of cortical selectivity (Luo et al. 2023). BrainDiVE employs CLIP-driven diffusion models to synthesize brain-guided images, enabling data-driven exploration of cortical organization with high semantic specificity (Luo et al. 2024). Other studies have used diffusion models to reconstruct images and text from fMRI data, improving semantic accuracy and biological insights (Ferrante et al. 2023; Takagi and Nishimoto 2023).

## CLIP Dissection

CLIP Dissection is a flexible, automatic, and generalizable method for dissecting DNNs without requiring concept-labeled data (Oikarinen and Weng 2022). It is training-free and utilizes pretrained multi-modal models like CLIP to efficiently identify the functionality of individual units in a model. The method requires three inputs: a DNN to be probed, a set of probing images, and a set of concepts. Unlike Network Dissection, CLIP Dissection does not need concept labels, allowing it to work with unlabeled data, such as ImageNet (Deng et al. 2009), CIFAR-100 (Krizhevsky, Hinton et al. 2009), or other images from the internet. By using the CLIP image and text encoders to compute the concept activation matrix $\boldsymbol{P}$ and calculating the maximum similarity with the activation vector $\boldsymbol{q}$ of the model's neurons, CLIP Dissection identifies the corresponding label $\boldsymbol{l}$ for each neuron's strongest response. This framework significantly enhances the flexibility and efficiency of detecting neuron unit concepts.

# Methods

We first describe the dataset and the definition of Regions of Interest (ROIs). We then describe the parameterization

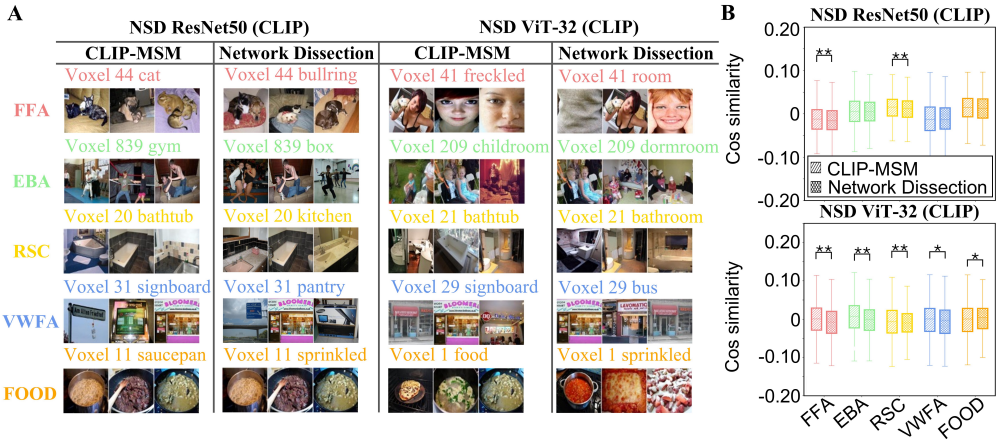


Figure 3: **Comparison of the performance of CLIP-MSM and Network Dissection for capturing semantic map.** To compare the performance of the two explainable methods, we input the label corresponding to each voxel into the text encoder to obtain the corresponding text embeddings. Here, BERT (Devlin et al., 2019) is used as the text encoder for ResNet50$_{\text{CLIP}}$, and transformer (Vaswani et al., 2017) is used as the text encoder for ViT-32$_{\text{CLIP}}$. Then, we calculate the cosine similarity between the text embeddings corresponding to labels generated by the two explainable methods for all voxels in each selective region and the voxel-wise weights obtained during training. A. We visualize labels and corresponding images for an identical voxel in each selective region generated by two explainable methods. B. The higher cosine similarity between text embedding and voxel-wise weights represents higher semantic mapping consistent with selective regions. As shown, we observed that CLIP-MSM captures more accurate semantic mapping than Network Dissection ($^{*}P < 0.05$, $^{**}P < 0.001$, paired $t$-test).

and training of our voxel-wise encoding models which go from images to brain responses. Finally, we describe how to capture the multi-semantic mapping of a given voxel for all stimuli that elicit its responses using our CLIP-MSM framework. The overall framework is illustrated in Figure 1.

## Datasets

All encoding models were trained using the NSD dataset (Allen et al. 2022), which includes high-density fMRI data from eight participants (six females, aged 19–32 years). We utilized the $betas\_fithrf\_GLMdenoise\_RR$ for beta value preparation. Cortical surface reconstructions were generated using FreeSurfer (Dale, Fischl, and Sereno 1999; Fischl, Sereno, and Dale 1999), and beta values were z-scored across runs and averaged across up to three repetitions per image, yielding one fMRI response per voxel per image. The stimuli were square-cropped, resized COCO images (Lin et al. 2014) subtending 8.4 × 8.4°.

For validation, we used the NOD dataset (Gong et al. 2023), comprising fMRI responses to 57,120 images from 30 participants scanned on a 3T scanner. We focused on nine participants (five females, aged 19–26 years) who viewed 4,000 unique ImageNet images and 120 shared COCO images. Preprocessed surface-based data from $ciftify$ (Dickie et al. 2019) was used, incorporating motion correction, field distortion correction, and spatial smoothing for data quality assurance.

## Regions of Interests

For the NSD dataset, we focused on five ROIs in the high-level visual cortex. The fusiform face area (FFA) for face selectivity (Kanwisher, McDermott, and Chun 2002), extrastriate body area (EBA) for body selectivity (Downing et al. 2001), retrosplenial cortex (RSC) for place selectivity (Brodmann 1909), visual word form area (VWFA) for word selectivity (Cohen et al. 2000) were defined using independent category localizer data (Allen et al. 2022) with a threshold of $t > 0$. The food-selective region (FOOD) was defined by (Jain et al. 2023). The final Selective ROI was derived by combining these five ROIs.

For the NOD dataset, we used the officially defined face-, place-, body-, and word-selective regions corresponding to FFA, RSC, EBA, and VWFA, with a stricter threshold of $t > 2.3$. The final Selective ROI was similarly obtained by combining these four ROIs.

For the NSD, brain visualizations were rendered in native space using Pycortex (Gao et al. 2015), as shown in Figure 1 for subject S5. For the NOD, visualizations were rendered in MNI152 standard space using Connectome Workbench (version 1.5.0) (Marcus et al. 2013).

## Voxel-wise Encoding Models for the Human High-level Visual Cortex

We employed the following models for feature extraction: (1) OpenAI CLIP models with ViT-32 and ResNet50 backbones; (2) ImageNet-pretrained ResNet50 and AlexNet. These models have been shown to align with the cortical hierarchy of the human brain (Millet et al. 2022; Xue et al. 2024; Conwell et al. 2024). For NSD, all stimulus images were used as input to all models, while for NOD, only images viewed by participants were input into ResNet50$_{\text{CLIP}}$, ViT-32$_{\text{CLIP}}$. Image embeddings from ResNet50$_{\text{CLIP}}$, ViT-32$_{\text{CLIP}}$, and features from the average

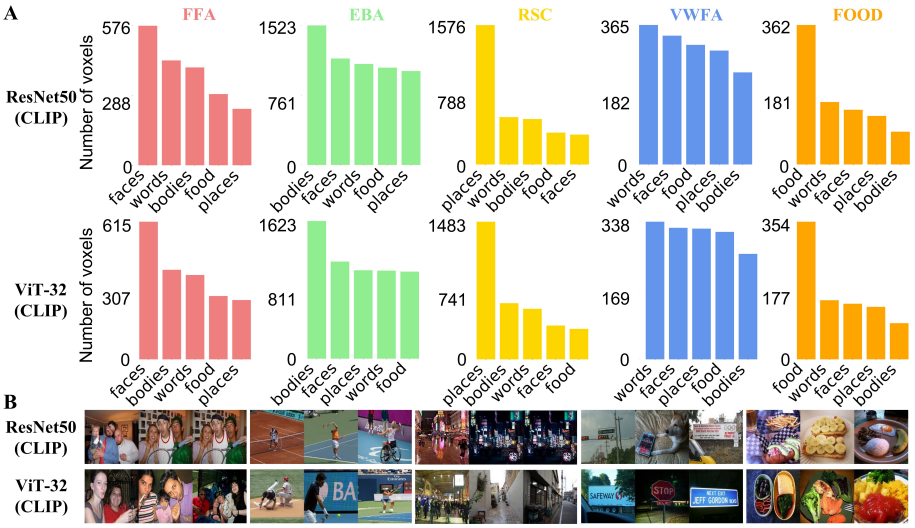


Figure 4: **CLIP-MSM enables precise dissection of high-level visual categories in the selective regions.** A. Single semantic mapping of subject S5 in the NSD for each encoding model across selective regions. The X-axis is the semantic label, and the Y-axis is the number of voxels mapped to each semantic label. B. Visualization of the top 3 images with the highest score within the concept that corresponds to the highest number of voxels in selective regions.

pooling layer of $\text{ResNet50}_{\text{ImageNet}}$ and $\text{AlexNet}_{\text{ImageNet}}$, were used to predict brain responses. Feature dimensions were as follows: $\text{ViT-32}_{\text{CLIP}}$ $(\mathbf{512})$, $\text{ResNet50}_{\text{CLIP}}$ $(\mathbf{1024})$, $\text{ResNet50}_{\text{ImageNet}}$ $(\mathbf{2048})$, and $\text{AlexNet}_{\text{ImageNet}}$ $(\mathbf{9216})$.

We implemented ridge regression models in PyTorch (Paszke et al. 2019) to predict the averaged fMRI response for each image per voxel. Images were divided into training and test sets with a ratio of $\mathbf{0.85 : 0.15}$ for both NSD and NOD. Regularization parameters, following (Wang et al. 2023), were spaced logarithmically from $\mathbf{10^{-8} \sim 10^{10}}$. Model performance was evaluated using the coefficient of determination ($\boldsymbol{R^2}$). Prediction significance was assessed via bootstrap testing with 2,000 resampled test sets and FDR-corrected $P$-value thresholds (Benjamini and Hochberg 1995).

## CLIP-MSM Framework

**Input & Output** We used stimulus images from the NSD and NOD datasets for the CLIP image encoder, which generates a predicted response matrix $\boldsymbol{R}$ for all voxels across images. The text encoder allows for hypothesis-free modeling by inputting any set of labels, a key advantage of CLIP Dissection over Network Dissection. The voxel-wise multi-semantic mapping matrix $\boldsymbol{S}$ reveals the hidden semantic mapping between voxels and the stimuli eliciting their strongest selectivity. To our knowledge, this is the first application of CLIP Dissection for interpreting encoding models in the human high-level visual cortex.

**Algorithm** CLIP-MSM involves three key steps (Fig. 1)
1. Record image to voxel activation $\boldsymbol{Q}$. The activation $\boldsymbol{Q} \in \mathbb{R}^{\boldsymbol{T}\times\boldsymbol{V}}$ is computed by the formula: $\boldsymbol{Q} = \boldsymbol{I} \cdot \boldsymbol{W}^{\boldsymbol{T}}$. The image embeddings $\boldsymbol{I} = [\boldsymbol{I_1}, \boldsymbol{I_2}, ..., \boldsymbol{I_t}] \in \mathbb{R}^{\boldsymbol{T}}$ are obtained from the image encoder $\boldsymbol{E_I}$ of the CLIP model and voxel-wise weights $\boldsymbol{W} = [\boldsymbol{W_1}, \boldsymbol{W_2}, ..., \boldsymbol{W_v}] \in \mathbb{R}^{\boldsymbol{V}}$ from ridge regression training. The predicted response matrix $\boldsymbol{R} \in \mathbb{R}^{\boldsymbol{T}\times\boldsymbol{V}}$ is obtained by the following formula:

$$\boldsymbol{R} = \boldsymbol{I} \cdot \boldsymbol{W}^{\boldsymbol{T}} + \boldsymbol{B} = \boldsymbol{Q} + \boldsymbol{B}$$

We define vector $\boldsymbol{b} = [\boldsymbol{b_1}, \boldsymbol{b_2}, ..., \boldsymbol{b_v}] \in \mathbb{R}^{\boldsymbol{V}}$ as the voxel-wise bias, $\boldsymbol{B} = [\boldsymbol{b}, \boldsymbol{b}, ..., \boldsymbol{b}]^{\boldsymbol{T}} \in \mathbb{R}^{\boldsymbol{T}\times\boldsymbol{V}}$ has the sample number of rows as $\boldsymbol{Q}$. The rows of $\boldsymbol{R}$ represent the number of voxels, and the columns represent the number of images.
2. Compute label-image activation matrix $\boldsymbol{P}$. The matrix $\boldsymbol{P} \in \mathbb{R}^{K\times T}$ is calculated by the inner product $\boldsymbol{P_{ij}} = \boldsymbol{T_i} \cdot \boldsymbol{I_j}$, where $\boldsymbol{T} = [\boldsymbol{T_1}, \boldsymbol{T_2}, ..., \boldsymbol{T_k}] \in \mathbb{R}^{\boldsymbol{K}}$ is the text embedding of the labels. The text encoder used in our study is a transformer model consistent with (Oikarinen and Weng 2022).
3. Generate voxel-wise multi-semantic-mapping matrix $\boldsymbol{S}$. The semantic-mapping score $\boldsymbol{S_{kv}}$ between voxel $v$ and label $k$ is defined as the similarity between $\boldsymbol{Q}_{:,\boldsymbol{v}}$ (voxel $v$'s responses to all images) and $\boldsymbol{P}_{\boldsymbol{k},:}$ (label $k$'s semantic activation): $\boldsymbol{S_{kv}} = \boldsymbol{sim}(\boldsymbol{P}_{\boldsymbol{k},:}, \boldsymbol{Q}_{:,\boldsymbol{v}})$. We use Soft Weighted Point-wise Mutual Information (SoftWPMI) as the similarity function for its high performance (Oikarinen and Weng 2022). The matrix $\boldsymbol{S} \in \mathbb{R}^{\boldsymbol{K}\times\boldsymbol{V}}$ is computed as:

$$\boldsymbol{S} = \boldsymbol{softmax}(\boldsymbol{sim}(\boldsymbol{P}, \boldsymbol{Q}))_{\boldsymbol{column}}$$

where $\boldsymbol{sim}(\boldsymbol{P}, \boldsymbol{Q})$ represents the semantic-mapping score calculated by SoftWPMI for every label to voxel pairs. The operation $\boldsymbol{softmax}(\cdot)_{\boldsymbol{column}}$ denotes column-wise normalization of the resulting matrix.

**Visualization** To visualize the multi-semantic mapping of the Selective ROI to specific labels, we use the voxel-wise matrix $\boldsymbol{S}$. The columns represent each voxel's selectivity for all labels, while the rows show the selectivity of all voxels for a specific label $\boldsymbol{k}$. We visualize $\boldsymbol{S}$ row by row for

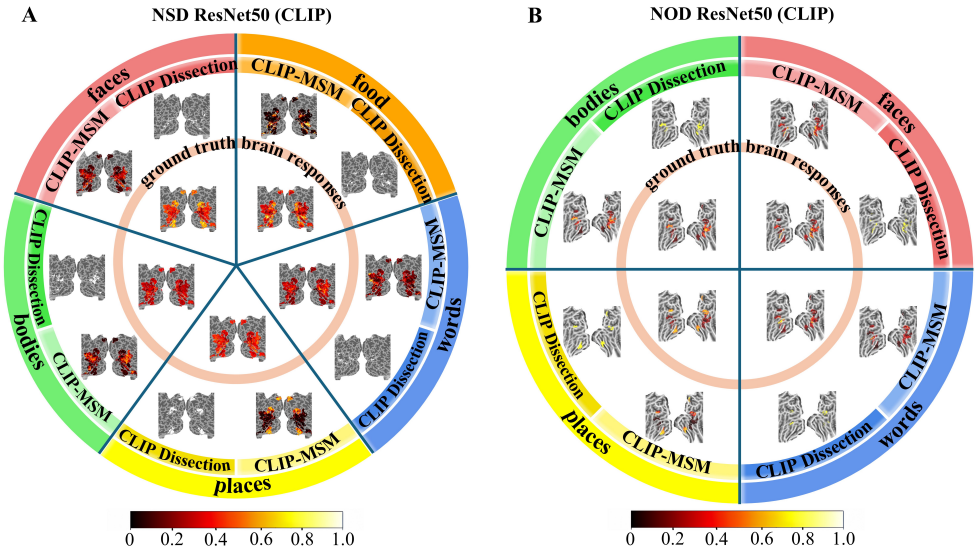


Figure 5: **CLIP-MSM for hypothesis-free analysis of neural selectivity in human high-level visual cortex.** We validate our CLIP-MSM for hypothesis-free analysis in Selective ROI for the subject S5 in the NSD (A) and the subject S4 in the NOD (B) on the ResNet50$_{\text{CLIP}}$ model. We also performed the CLIP Dissection for hypothesis-free analysis for comparative analysis. We visualized the semantic mapping scores of each voxel for each label on flatmaps for both the CLIP-MSM and CLIP Dissection in the outer circle. All semantic mapping scores were normalized to scale $[0, 1]$ for comparison. In the inner circle, the ground truth brain responses of the faces, places, bodies, and word concepts were extracted from the beta value of the floc experiment and normalized to scale $[0, 1]$ for both the NSD and NOD. In NOD, there is no food concept analysis in the floc experiment. In NSD, the ground truth of brain responses to the food concept was extracted from the beta value of the shared images which are labeled as food images by (Jain et al., 2023).

hypothesis-free analysis using Pycortex and Connectome Workbench.

**Performance Analysis** We compared the performance of CLIP-MSM with the state-of-the-art method, Network Dissection, using the same pipeline as (Bau et al. 2017; Xue et al. 2024). This comparison evaluates their effectiveness in capturing semantic mapping.

## Experiments

In this section, we trained encoding models ResNet50$_{\text{CLIP}}$, ViT-32$_{\text{CLIP}}$, ResNet50$_{\text{ImageNet}}$ and AlexNet$_{\text{ImageNet}}$ on NSD and NOD to predict brain responses. We found CLIP models significantly outperformed ImageNet-pretrained models. Building on ResNet50$_{\text{CLIP}}$ and ViT-32$_{\text{CLIP}}$, we proposed CLIP-MSM, applying CLIP Dissection to automatically inspect voxel functionality. Then, we investigate the accuracy between CLIP-MSM and Network Dissection by measuring the similarity between the labels of semantic mapping generated by these explainable frameworks and the voxel-wise weights obtained during training. Finally, we validated CLIP-MSM's accuracy by comparing reconstructed brain responses to ground truth activations for categories like faces, bodies, places, words, and food.

### Multimodal Embeddings Enhanced the Prediction of High-level Visual Cortex

We constructed voxel-wise encoding models to predict human high-level visual cortical responses to natural images. Voxel-wise prediction performance ($R^2$) was significantly higher for ResNet50$_{\text{CLIP}}$ and ViT-32$_{\text{CLIP}}$ compared to ResNet50$_{\text{ImageNet}}$ and AlexNet$_{\text{ImageNet}}$ on NSD subject S5, in functionally defined regions (Fig. 2A). Only voxels with significantly better predictions than chance ($P < 0.05$, one-sided test, FDR-corrected) are shown. We also assessed model performance as a function of the training data volume. Results showed that performance stabilized when the data reached 34 $\times$ 250 images (Fig. 2B). A statistical analysis based on the model trained with the maximum data volume revealed that CLIP models outperformed ImageNet-pretrained models (Fig. 2C, $P < 0.001$, paired $t$-test). To further assess model performance, we compared $R^2$ scores with the noise ceiling, which indicates the percentage of voxel response variance due to the stimulus rather than measurement noise. The results show that CLIP models predicted voxels closer to their noise ceiling than ImageNet models, with many CLIP-predicted voxels exceeding the 85% noise ceiling (Fig. 2D). Full results for S1-S8 in NSD and S1-S9 in NOD are provided in the Appendix.

### CLIP-MSM Captures More Accurate Semantic Mapping Than Network Dissection

We further applied CLIP Dissection to the interpretability of the ResNet50$_{\text{CLIP}}$ and ViT-32$_{\text{CLIP}}$ models to construct our CLIP-MSM. Then, we conducted a comparative analysis between CLIP-MSM and Network Dissection. To control labels in the two dissection procedures, we use the same image

dataset and label set from **Broden** (Bau et al. 2017). After model dissection, we first visualize labels and corresponding images for an identical voxel in each selective region generated by two dissection methods (Fig. 3A). Our results show that CLIP-MSM enables richer semantic mapping, allowing for broader and fine-grained image category. Then, we compared the performance of CLIP-MSM with Network Dissection for capturing semantic mapping in selective regions. Specifically, we calculate the cosine similarity between the text embeddings corresponding to labels generated by the two explainable methods for all voxels in each selective region and the voxel-wise weights obtained in CLIP models. We find that the cosine similarity between text embedding and voxel-wise weights was higher in CLIP-MSM in most selective regions for the NSD subject S5 (Fig. 3B). The results corresponding to S1-S8 in NSD and S1-S9 in NOD are in the Appendix.

## CLIP-MSM Enables Precise Mapping of High-level Visual Cortex across Different Categories

We then compared the category labels that have the maximum number of responsive voxels in each region with those from previous studies that identified distinct regions causally engaged in the perception of faces, words, bodies, places, and food. The results indicate that in both the ResNet50$_{\text{CLIP}}$ and ViT-32$_{\text{CLIP}}$ models, the categories eliciting the strongest responses in various brain regions were highly consistent with those identified in previous studies with hypothesis-driven analysis (Fig. 4A). These include the FFA showing the strongest response to face concepts, the EBA to body concepts, the RSC to scene concepts, the VWFA to word concepts, and the food-selective region to food concepts. We further visualized the top 3 images with the highest score within the concept that corresponds to the highest number of voxels in each selective region to demonstrate the accuracy of CLIP-MSM (Fig. 4B). We observed that the label of each of these top 3 images exhibits a clear and more descriptive concept for each selective regions. The results corresponding to S1-S8 in NSD and S1-S9 in NOD are in the Appendix.

## CLIP-MSM for Fine-grained Hypothesis-free Analysis of Human High-level Visual Cortex

Through normalizing the semantic mapping score to soft maximize the establishment of brain responses to concept labels, our CLIP-MSM achieves voxel-wise multi-semantic mapping. We then verify this method by developing a hypothesis-free analysis of brain activation in response to categories of faces, bodies, places, words, and food. Our results show that CLIP-MSM for hypothesis-free analysis in Selective ROI for the subject S5 of the NSD (Fig. 5A) and the subject S4 of the NOD (Fig. 5B) on the ResNet50$_{\text{CLIP}}$ model exhibit a higher similarity with the ground truth of the brain activation than the CLIP Dissection of maximizing the semantic mapping score. We further calculate the Pearson correlation coefficient between the multi-semantic mapping across concepts obtained by CLIP-MSM and the ground truth brain responses. Then, we performed a statistical analysis of the correlation coefficients calculated using CLIP-MSM and CLIP Dissection with ground truth. Our results show that CLIP-MSM provides significant accuracy in the prediction of visual responses than CLIP Dissection in places, food, faces, bodies and words concepts in NSD (Fig. 6A). In NOD, CLIP-MSM provides significant accuracy in the prediction of visual responses than CLIP Dissection in faces, bodies and words concepts (Fig. 6B). The results of ViT-32$_{\text{CLIP}}$ model and other subjects can found in Appendix.

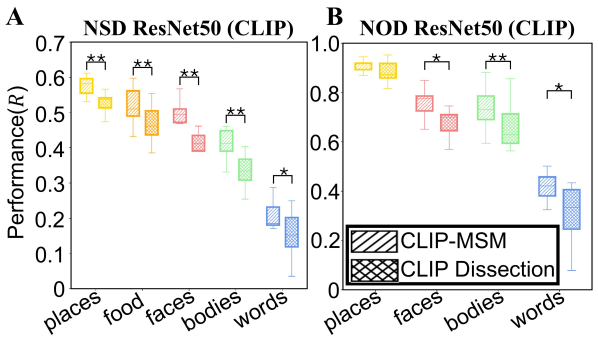


Figure 6: **Semantic mapping performance.** Pearson correlation coefficient between the semantic mapping scores for both CLIP-MSM and CLIP Dissection and the ground truth of brain responses across selective regions. A. In NSD, CLIP-MSM provides significant accuracy in the prediction of visual responses than CLIP Dissection in places, food, faces, bodies, and words concepts. B. In NOD, CLIP-MSM provides significant accuracy in the prediction of visual responses than CLIP Dissection in faces, bodies, and words concepts ($^{*}P < 0.05$, $^{**}P < 0.001$, paired $t$-test).

## Conclusion

We propose the CLIP-MSM, a multi-semantic mapping framework, for hypothesis-free analysis in the human high-level visual cortex. By combining CLIP models with CLIP Dissection, our CLIP-MSM enables granular examination of concept preferences for the human visual cortex with a broad range of natural scene categories. We observed higher performance of interpretability accuracy in CLIP-MSM than in Network Dissection. Furthermore, through calculating the accuracy between CLIP-MSM's reconstructed brain activation patterns in response to categories of faces, bodies, places, words, and food, and ground truth of brain activation, our study revealed that CLIP-MSM's reconstructed brain activation exhibited more precise alignment with the ground truth compared to CLIP Dissection. To further demonstrate the reliability of our CLIP-MSM, we validated our results on two fMRI datasets NSD and NOD, which showed a high degree of reproducibility. Our study focused on high-level visual cortical, future work can extend CLIP-MSM to analyze brain responses to more nuanced concepts.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (grants number 82302175, 62071049 and 62336002); the National Science and Technology Innovation 2030 Program (grant number 2021ZD0200500); the Beijing Municipal Science and Technology Commission (grants number Z171100000117012 and Z181100001518003); the Beijing Municipal Natural Science Foundation Project (grant number 4222018); the China Postdoctoral Science Foundation (grant number 2021M700015).

## 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.

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.

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*.

Conwell, C.; Prince, J. S.; Kay, K. N.; Alvarez, G. A.; and Konkle, T. 2024. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1): 9383.

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.

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*.

Fischl, B.; Sereno, M. I.; and Dale, A. M. 1999. Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. *Neuroimage*, 9(2): 195–207.

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. 2002. The fusiform face area: a module in human extrastriate cortex specialized for face perception.

Kanwisher, N.; and Yovel, G. 2006. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1476): 2109–2128.

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.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Luo, A.; Henderson, M.; Wehbe, L.; and Tarr, M. 2024. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *Advances in Neural Information Processing Systems*, 36.

Luo, A. F.; Henderson, M. M.; Tarr, M. J.; and Wehbe, L. 2023. BrainSCUBA: Fine-Grained Natural Language Captions of Visual Cortex Selectivity. *arXiv preprint arXiv:2310.04420*.

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

Marcus, D. S.; Harms, M. P.; Snyder, A. Z.; Jenkinson, M.; Wilson, J. A.; Glasser, M. F.; Barch, D. M.; Archie, K. A.; Burgess, G. C.; Ramaratnam, M.; et al. 2013. Human Connectome Project informatics: quality control, database services, and data visualization. *Neuroimage*, 80: 202–219.

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.

Oikarinen, T.; and Weng, T.-W. 2022. Clip-dissect: Automatic description of neuron representations in deep vision networks. *arXiv preprint arXiv:2204.10965*.

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.

Quiroga, R. Q.; Reddy, L.; Kreiman, G.; Koch, C.; and Fried, I. 2005. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045): 1102–1107.

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.

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.

Thorpe, S.; Fize, D.; and Marlot, C. 1996. Speed of processing in the human visual system. *nature*, 381(6582): 520–522.

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.

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.