Prototypes of the Mind: A Unified Framework for Probing the Visual Brain

Shi Chen[1] **Seoyoung Ahn**[2] **Doris Tsao**[1]

[1]Department of Neuroscience, UC Berkeley
[2]College of AI Convergence, Hankuk University of Foreign Studies

Abstract

Deep learning has driven substantial progress in modeling fMRI responses, yet most existing approaches are optimized for a single objective—such as neural encoding, stimulus decoding, or post hoc interpretation—making it difficult to obtain a coherent account of neural representations. We introduce a unified framework that connects vision foundation models to brain activity while supporting neural analysis from multiple complementary perspectives. Our approach replaces black-box features with a sparse, interpretable representation composed of visual prototypes, and links these prototypes to voxel responses through a linear readout. This design makes the model transparent at both levels: prototype activations expose the visual concepts present in the stimulus, while the learned linear weights characterize each voxel's selectivity. Using this interface, we identify which visual concepts drive neural responses and where they appear in the image, reveal cortical organization through voxel tuning in prototype space, and perform controlled stimulus synthesis to probe cortical selectivity through targeted interventions. Together, these results establish prototype-based representations as a unified and interpretable foundation for studying visual representations in the human brain. Our source code is publicly available.

1 Introduction

Understanding how the human brain represents complex visual scenes remains a fundamental challenge at the intersection of neuroscience and computer vision. Neural measurements are sparse and noisy, while natural scenes are incredibly rich and complex. Modeling the relationship between the two therefore requires representations that are both predictive of neural responses and expressive enough to capture the structure of the visual world.

Advances in deep learning, together with the increasing availability of large-scale neural datasets, have transformed Functional Magnetic Resonance Imaging (fMRI) analysis by enabling increasingly effective computational models of brain responses. One major line of work develops encoding models [1, 2, 3, 4, 5] that seek to predict fMRI responses from visual stimuli with high accuracy. A complementary line of research tackles the inverse problem through decoding models [6, 7, 8, 9, 10, 11], which reconstruct or synthesize stimuli directly from neural activity. Moving beyond black-box features learned by deep networks, several recent studies develop interpretable representations that reveal categorical [12, 13] and spatial [14, 15] selectivity. Collectively, these approaches have produced valuable tools for probing the functional organization of the brain. However, they have largely developed in isolation, each optimized for a single question—prediction, reconstruction, or interpretation—with no shared framework. This fragmentation makes it difficult to develop a comprehensive and coherent picture of neural representations across these complementary dimensions.

In this paper, we investigate a different possibility: can we derive a single unified representation that supports the analysis of neural activity across multiple dimensions? Such a representation can serve

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

as an interpretable bridge between vision foundation models [16, 17, 18] and the brain, grounding neural analysis in the semantic structure learned from large-scale visual data. At the same time, by linking brain responses to a representational space with strong generative properties, it can support more advanced applications that go beyond passive analysis, including stimulus reconstruction and controlled manipulation of visual content.

Realizing this goal requires a representation that is semantically interpretable, structurally organized, and flexible enough to support both analysis and intervention. Guided by the intuition that complex visual scenes can be decomposed into distinct visual concepts [19, 20], we factorize rich features from vision foundation models (*e.g.,* DINOv3 [21]) into spatially grounded prototype units that capture recurring visual patterns in natural images. The resulting activations are sparse and preserve spatial layout, allowing each brain voxel to be associated with a small set of visual concepts using a linear projection. This yields a unified framework for brain analysis that reveals both *what* drives neural activity and *where* those driving concepts appear. Crucially, our compositional representation further enables *targeted stimulus interventions* on interpretable prototypes (*e.g.,* manipulating face and body prototypes), opening a new route to causally probe the relationship between neural activity and visual concepts.

2 Related Work

Recent work on fMRI modeling largely focuses on three complementary directions:

Encoding models. A substantial body of work [22, 23, 2, 3, 4, 1, 5] learns the mappings from visual features to neural responses with encoding models. Early studies [22] demonstrated that voxel activity in early visual cortex can be predicted using Gabor-based features. Subsequent research expanded this framework using richer semantic or deep visual features [23], enabling more accurate prediction in higher visual areas. Recent studies increasingly leverage features from deep networks or multimodal foundation models [1, 2, 3, 4, 5], which capture hierarchical visual representations aligned with cortical organization. While these DNN-based approaches are effective for predicting neural activity and characterizing voxel selectivity, they typically operate as black-box predictors, offering limited interpretability about which specific image properties drive the responses.

Decoding models. Another line of work attempts to reconstruct or synthesize visual stimuli from neural activity using decoding models [6, 24, 25, 26, 7, 8, 9, 10, 11]. Early studies used optimization-based approaches to generate stimuli that maximally activate neurons or brain regions [25, 26]. More recent methods combine fMRI decoding with powerful generative models such as diffusion models [7, 10, 11], enabling reconstruction of naturalistic images from measured neural activity. These approaches provide compelling visualizations of encoded information, but they typically rely on large pretrained generative models and operate through latent spaces that are difficult to interpret or manipulate. As a result, they fall short of probing the causal relationship between neural activity and specific image properties. In contrast, our prototype representation provides a compact and interpretable feature space that supports *direct manipulation of visual concepts*, enabling controlled stimulus interventions such as feature addition, removal, and spatial manipulation.

Spatial selectivity analysis. Several studies investigate the spatial structure of neural responses to natural images [14, 15]. These approaches analyze spatial feature maps extracted from deep networks and relate them to voxel activity patterns to characterize where visual information is processed within an image. For example, recent work has shown that dense feature maps from vision models can help localize image regions driving neural responses across different cortical areas [14]. Such analyses offer valuable insight into spatial attention and receptive field structure, but typically require separate post-hoc analysis pipelines built on top of encoding models. In contrast, our prototype representation preserves spatial structure by design, enabling spatial selectivity to emerge naturally from the model. This allows us to directly visualize how specific visual concepts contribute to neural responses.

3 Methodology

At the core of our method is a two-stage framework that connects visual perception with neural activity through interpretable prototypes. In Stage 1 (§3.1), we learn a dictionary of visual prototypes that

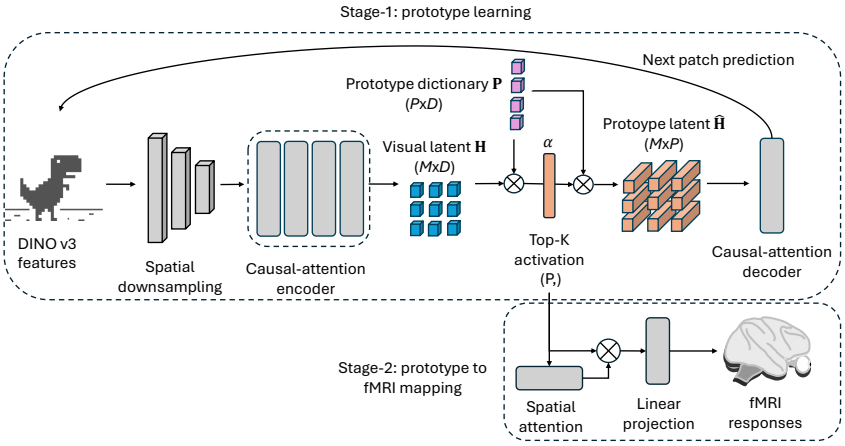

Figure 1: Overview of method: (1) In stage-1, we learn interpretable prototypes with a next patch prediction objective. Given patch-level features (*e.g.,* from DINOv3), we factorize them as a sparse combination of learnable prototypes, which captures distinct semantics and the spatial structure of the visual scene. (2) In stage-2, we train a linear mapping between the activation of prototypes and fMRI responses, enabling brain understanding from complementary perspectives.

decomposes foundation model features into sparse, interpretable components. Using a self-supervised learning paradigm that emphasizes contextual relationships among visual regions, the model supports a rich interpretation of both the concepts present in a scene and their spatial organization. In Stage 2 (§3.2), we learn a lightweight mapping from spatially distributed prototype activations to fMRI voxel responses, producing an interpretable weight matrix that directly characterizes the associations between visual prototypes and brain areas. An overview of our pipeline is shown in Figure 1.

3.1 Stage 1: Reconstructing Visual Scenes with Interpretable Prototypes

Vision foundation models such as DINO [16] have emerged as the *de facto* backbone for predicting fMRI responses [5, 4, 27], delivering state-of-the-art encoding accuracy across visual cortex. Yet, despite their strong predictive performance, the entangled structure of these features makes it challenging to identify which image properties drive the predicted responses in different brain regions. Our objective is to *re-express* these features as sparse combinations of learnable prototypes, each grounded in concrete visual patterns, while retaining the discriminative power of the original features. To achieve this, we introduce an encoder-decoder network with three main components:

Causal encoder with spatial downsampling. Given the patch-level features $\mathbf{X} \in \mathbb{R}^{N \times D}$ extracted from a foundation model (*e.g.,* DINOv3, $N = 16 \times 16$ is the number of spatial patches and $D = 768$ is the feature dimension), we first apply a learnable layer [28] to reduce the feature map resolution from $N = 16 \times 16$ to $M = 8 \times 8$. This step both lowers computational cost and facilitates model optimization:

$$\mathbf{Z} = \texttt{SpatialDown}(\mathbf{X}) \in \mathbb{R}^{M \times D}. \tag{1}$$

The downsampled features are combined with sinusoidal positional embeddings $\mathbf{E}_{pos} \in \mathbb{R}^{M \times D}$, and passed through a stack of four causal self-attention blocks f_{enc} [29, 30] for further refinement:

$$\mathbf{H} = f_{\text{enc}}(\mathbf{Z} + \mathbf{E}_{\text{pos}}) \in \mathbb{R}^{M \times D}, \tag{2}$$

By preserving spatial structure in the foundation model representation, the encoder enables us to identify how image properties at different locations relate to brain responses.

Prototype dictionary and sparse decomposition. A principal idea behind our approach is to decompose foundation model features into interpretable units. For this, we maintain a dictionary of prototypes $\mathbf{P} \in \mathbb{R}^{P \times D}$, where $P = 1000$ denotes the number of prototypes and each row $\mathbf{p}_p \in \mathbb{R}^D$ corresponds to a single prototype. We then factorize the encoded features $\mathbf{H}$ as weighted combinations

of these prototypes. For each encoded patch $\mathbf{h}_m \in \mathbb{R}^D$, the prototype weights $\mathbf{s}_m$ are computed from its cosine similarities with all prototypes:

$$\mathbf{s}_m = \mathbf{h}_m \mathbf{P}^\top \in \mathbb{R}^P. \tag{3}$$

With the overarching goal of developing an interpretable representation, we draw inspirations from sparse autoencoder for analyzing language models [31] and further impose sparsity constraints on prototype selection. Specifically, we retain only the top-k ($k = 128$) prototype weights for each patch and normalize them with a softmax:

$$\tilde{s}_{m,p} = \begin{cases} s_{m,p} & \text{if } s_{m,p} \in \text{top-}k(\mathbf{s}_m) \\ -\infty & \text{otherwise} \end{cases} \qquad \alpha_{m,p} = \frac{\exp(\tilde{s}_{m,p})}{\sum_{j=1}^{P} \exp(\tilde{s}_{m,j})} \tag{4}$$

The resulting prototype activation vector $\boldsymbol{\alpha}_m \in \mathbb{R}^P$ preserves exactly the k most relevant prototypes as nonzero entries while suppressing the remaining $P-k$ prototypes. Each patch is then reconstructed as a weighted combination of prototype vectors:

$$\hat{\mathbf{h}}_m = \sum_{p=1}^{P} \alpha_{m,p} \cdot \mathbf{p}_p \in \mathbb{R}^D. \tag{5}$$

Stacking over all M patches, we obtain the prototype-decomposed representation $\hat{\mathbf{H}} = \mathbf{A}\mathbf{P} \in \mathbb{R}^{M \times D}$, where $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_M]^\top \in \mathbb{R}^{M \times P}$. Each row of $\mathbf{A}$ indicates, for a given spatial token, the strength with which each of the P prototypes is activated, while each column shows how a particular prototype is used across the M spatial tokens. Together, $\mathbf{A}$ provides a sparse, interpretable decomposition of the visual scene in terms of prototypes and their spatial distribution.

Causal decoder and next-patch prediction. The reconstructed representation $\hat{\mathbf{H}}$ is passed through a decoder composed of a single causal self-attention block followed by a linear readout head:

$$\hat{\mathbf{Y}} = g_{\text{read}}\Big(f_{\text{dec}}(\hat{\mathbf{H}})\Big) \in \mathbb{R}^{M \times D'}, \tag{6}$$

Note that the readout projects each decoded patch into a higher-dimensional space (D'=$4D$=3072), to accommodate the 2×2 spatial downsampling within the encoder.

To preserve the discriminative power of the learned prototypes, we adopt an autoregressive training paradigm used in large language [30] and multimodal [32] models. Specifically, we rasterize the visual patches, and task our model to predict features of the next patches based on preceding ones:

$$\mathcal{L}_{proto} = \frac{1}{M-1} \sum_{m=1}^{M-1} \left\| \hat{\mathbf{y}}_m - \mathbf{y}_{m+1} \right\|^2, \tag{7}$$

where $\hat{\mathbf{y}}_m$ is the prediction at position m and $\mathbf{y}_{m+1}$ is the target at position $m{+}1$. This objective encourages the model to capture the spatial structure without the need of additional annotations.

3.2 Stage 2: Bridging Visual Scenes and the Brain with Prototypes

Stage 1 produces, for each image, a sparse activation map $\mathbf{A} \in \mathbb{R}^{M \times P}$ where M=64 spatial locations each activate at most $k = 128$ prototypes. We now leverage this representation to model fMRI responses. Instead of developing a black-box encoding model [3, 4, 1, 5] that poses challenges for interpretation, our mapping from prototypes to neural activity is kept simple: a learned spatial attention aggregates prototype activations across locations, and a single linear layer projects the aggregated prototype vector onto voxel responses. The combination of interpretable prototype features with a linear encoding model yields a transparent interface between visual content and neural activity—each weight directly quantifies how strongly a given prototype drives a given voxel.

Spatial aggregation via attention. As different brain areas may preferentially attend to different spatial locations (*e.g.,* face-selective areas may emphasize central patches), we learn a spatial attention mechanism that aggregates the M=64 patch-level activations into a single image-level prototype

vector. Given the sparse prototype activation map $\mathbf{A} \in \mathbb{R}^{M \times P}$, we reshape it into a 2D feature map $\mathbf{A}_{2D} \in \mathbb{R}^{P \times 8 \times 8}$, and compute the attention map with a convolutional layer:

$$\mathbf{w} = \texttt{softmax}(\texttt{Conv2d}_{3\times3}(\mathbf{A}_{2D})) \in \mathbb{R}^{M \times 1}, \tag{8}$$

The spatially aggregated prototype vector per image is then:

$$\bar{\boldsymbol{\alpha}} = \sum_{m=1}^{M} w_m \, \boldsymbol{\alpha}_m \in \mathbb{R}^P, \tag{9}$$

where w_m is the learned attention weight for spatial position m and $\boldsymbol{\alpha}_m$ is the m-th row of $\mathbf{A}$.

Prototype-to-voxel linear mapping. The aggregated vector $\bar{\boldsymbol{\alpha}}$ is mapped to voxel responses via a linear transformation:

$$\hat{\mathbf{r}} = \mathbf{W}\bar{\boldsymbol{\alpha}} + \mathbf{b} \in \mathbb{R}^V, \tag{10}$$

where $\mathbf{W} \in \mathbb{R}^{V \times P}$ is the learned weight matrix and $\mathbf{b} \in \mathbb{R}^V$ is the bias. Each row $\mathbf{w}_v \in \mathbb{R}^P$ of $\mathbf{W}$ encodes the sensitivity of voxel v to each of the P prototypes, providing a direct and interpretable characterization of voxel selectivity. The model is trained with a composite loss combining mean squared error $\mathcal{L}_{MSE}$ and Pearson's correlation $\mathcal{L}_{corr} = 1 - \rho$:

$$\mathcal{L}_{fMRI} = \mathcal{L}_{MSE} + \mathcal{L}_{corr} \tag{11}$$

both of which are commonly used for fMRI encoding models [27, 4].

4 Experiments

4.1 Datasets

We use MSCOCO dataset [33] for learning our prototypes, which includes images with rich semantics. For the fMRI data, we adopt the processed NSD dataset [34] from Algonauts 2023 [27]. The dataset provides annotations on the functional selectivity of different brain areas (*e.g.,* Occipital Face Area selected for faces, and Extrastriate Body Area selected for bodies), and we choose 12 brain areas with distinct selectivity (*i.e.*, faces, bodies, places, and words) as our testbed to verify the effectiveness of our approach. See Appendix P for details.

4.2 Neural Encoding Performance

We evaluate how well our interpretable prototype representations predict fMRI responses across three visual backbones: JEPA-ViT [17], DINOv3-ViT, and DINOv3-ConvNeXt [21]. These backbones allow us to assess the effectiveness of our method across different training approaches and architectures. To contextualize our results in terms of achievable encoding performance, we also report the noise ceiling and the performance of [4], which uses a transformer prediction network on DINOv2 backbone features. All remaining models use linear prediction, applied either directly to the visual backbone features or to our interpretable prototype representations derived from those features.

Our prototype representations consistently improve prediction across all three backbones and all eight ROIs (Table 1). With DINOv3-ViT, the mean correlation across ROIs increases from 0.338 to 0.394,

	OFA	FFA-1	EBA	FBA-1	OPA	PPA	OWFA	VWFA-1
Noise-ceiling	0.561	0.597	0.600	0.541	0.588	0.563	0.530	0.551
Adeli *et al.*[4]	0.430	0.489	0.481	0.432	0.438	0.485	0.399	0.430
JEPA-VIT	0.305	0.381	0.381	0.348	0.348	0.414	0.244	0.317
Proto JEPA-VIT	0.355	0.403	0.404	0.359	0.399	0.419	0.314	0.350
DINOv3-ConvNeXt	0.273	0.366	0.379	0.319	0.324	0.377	0.209	0.304
Proto DINOv3-ConvNeXt	0.335	0.410	0.42	0.347	0.378	0.408	0.268	0.342
DINOv3-ViT	0.302	0.388	0.396	0.330	0.332	0.386	0.254	0.314
Proto DINOv3-ViT	0.367	0.434	0.433	0.387	0.404	0.443	0.310	0.370

Table 1: Comparison of fMRI encoding across 8 subjects. We report the Pearson's correlation for 8 distinct brain ROIs, averaged across all voxels. See Appendix M for results of other ROIs.

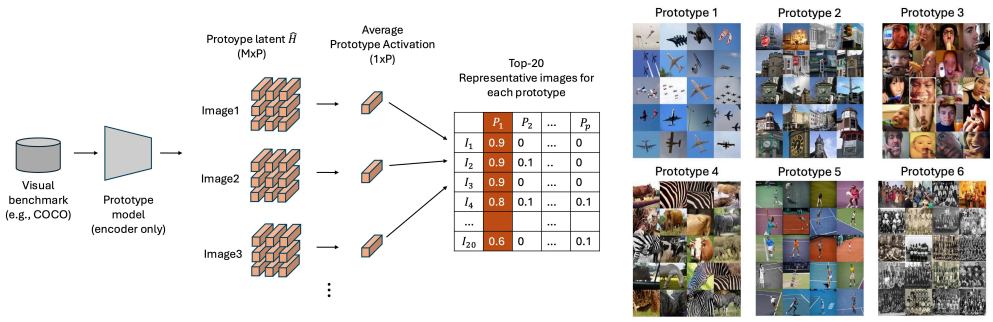

Figure 2: Visualization of learned prototypes. For each prototype, images are ranked by their average prototype activation (Left). We show representative examples among the top-activated images for several prototypes (Right). For visualization, we select 20 representative examples from each set. Additional visualization is provided in Appendix A.

reaching 70% of the noise ceiling and 88% of the correlation achieved by the nonlinear encoding model. These results show that our prototype representation captures substantial neural response information through a simple linear mapping and improve prediction over the original backbone features across different models. We further examine the effects of prototype number and top-k sparsity in Appendix N.

The prototype representation also makes the encoding model directly interpretable: each voxel's predicted response is a weighted combination of prototype activations, allowing us to examine which visual concepts contribute to its response. This structure supports analyses of categorical selectivity, cortical organization, and spatial selectivity, as well as controlled stimulus manipulations. In the following sections, we use the DINOv3-ViT prototype model to investigate these properties.

4.3 Prototypes Capture Semantically Coherent Visual Concepts

We first examine whether the learned prototypes correspond to interpretable visual concepts rather than arbitrary directions in feature space. Specifically, we extract prototype activation maps $\mathbf{A}$ of images from MSCOCO validation set, and identify the representative images with strong activation (averaged across all patches) for each prototype (see Appendix P for details). To enhance consistency among the final representatives, we also include an optional step that computes pairwise cosine similarities between CLIP [35] embeddings of these candidates and applies spectral clustering [36] to select 20 images that are both highly activated and representative of a shared visual concept (see Appendix E for comparison between visualization with and without clustering).

As shown in Figure 2, individual prototypes consistently capture coherent visual concepts. Some prototypes specialize in inanimate objects ("airplanes", "buildings"), while others correspond to animate entities ("faces", "zebra") or activities ("tennis") or abstract scene-level concepts ("crowds"). This structure emerges entirely from the self-supervision, without explicit semantic labels, yet the learned prototypes organize images according to meaningful visual themes. On the other hand, applying the same pipeline to PCA features fails to capture meaningful semantics, indicating that our approach is not tuned for trivial image statistics (see Appendix F).

4.4 Prototypes Reveal Categorical Selectivity of Visual Cortex

Having established that the prototypes correspond to interpretable visual concepts, we next study whether their mapping to neural activity reflects the known categorical selectivity of the visual cortex.

As our model employs a linear mapping between prototypes and fMRI responses, the weight matrix $\mathbf{W} \in \mathbb{R}^{V \times P}$ can provide a direct readout of their relationship: each voxel is associated with a set of weights indicating which prototypes most strongly drive its activity. By aggregating these weights across voxels within predefined regions of interest (ROIs), we obtain a prototype profile that summarizes the relative contribution of each prototype to a given brain area.

We analyze these profiles for several well-established category-selective regions in the ventral visual pathway, including face-selective areas (e.g., Occipital Face Area, OFA), body-selective areas



Figure 3: Prototype profiles (visualized as top-4 prototypes with the highest weights) for category-selective regions of the visual cortex. The profiles closely align with the expected selectivity of diverse brain areas. See Appendix C and D for additional results.

(Extrastriate Body Area, EBA), and scene-selective areas (Parahippocampal Place Area, PPA). If the learned representation captures meaningful visual structure, prototypes corresponding to faces, bodies, or scenes should preferentially contribute to the corresponding ROIs. Figure 3 shows that the resulting prototype profiles closely align with these expected pattern, and these top prototypes also capture a considerable proportion of variance for the ROIs (see Appendix H). These patterns emerge without any explicit supervision for functional localization, suggesting that the learned prototype representation captures visual features that correspond naturally to the selectivity of high-level visual cortex. As a control, we repeat the same ROI–feature analysis using PCA. Results show that it fails to characterize the selectivity of different ROIs, indicating that the ROI selectivity observed above is not a consequence of dimensionality reduction but arises from the learned prototype structure. We include the control results in Appendix F. In addition, our approach also outperforms a standard sparse autoencoder in localizing semantic structures and characterizing selectivity (see Appendix G).

We also examine whether independently fitted encoding models select similar prototypes across subjects (see Appendix B). The substantial overlap in prototypes for the same brain areas, especially face- and body-selective regions, suggests that our prototypes generalize well across subjects.

4.5 Mapping Prototypes to Brain Areas

While the previous analysis examines brain regions and asks which prototypes they prefer, we can also ask the reverse question: for a given prototype, where in the brain is it most strongly represented? This analysis is enabled by the linear projection matrix $\mathbf{W} \in \mathbb{R}^{V \times P}$ that maps prototype activations to voxel responses. Each column $\mathbf{w}_p \in \mathbb{R}^V$ therefore describes how strongly prototype p contributes to the predicted response of each voxel. Figure 4 visualizes these weights on a cortical flatmap. Positive weights (yellow) indicate voxels whose activity is positively driven by the prototype, whereas near-zero or negative weights (blue) indicate weak or suppressed contributions. This analysis is conceptually related to [37], which maps semantic category features onto cortical responses using linear encoding models. Here, the feature space is instead learned directly from self-supervision.

The resulting maps reveal spatially localized and functionally meaningful activation patterns. A face-selective prototype, whose top-activated images depict human faces, produces elevated weights concentrated in face-selective cortex, particularly the occipital face area (OFA) and nearby ventral regions, while place-selective areas such as the occipital place area (OPA) show minimal response. Conversely, a kitchen-selective prototype whose representative images depict kitchen interiors produces strong activation in OPA with little contribution to OFA or EBA.

These results highlight another advantage of our framework: the combination of a simple linear decoder and semantically meaningful prototype units enables direct visualization of how specific concepts map onto cortical responses. This may help generate hypotheses about stimulus–response relationships in specific brain regions and guide the design of targeted neuroscience experiments.

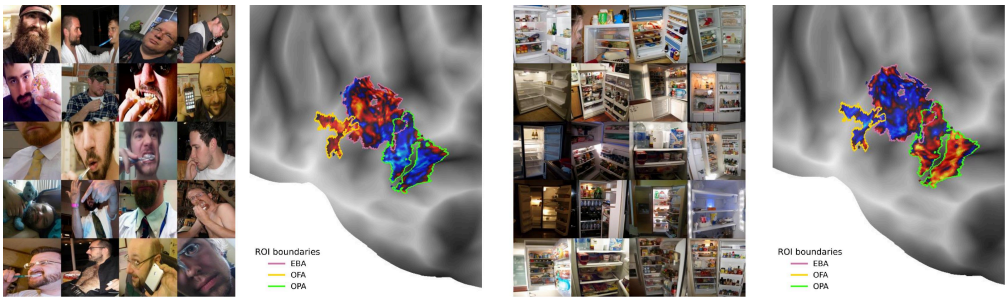

Figure 4: Per-prototype cortical activation maps. Strong and weak activations are denoted with yellow and blue colors, respectively. See Appendix J for additional results.

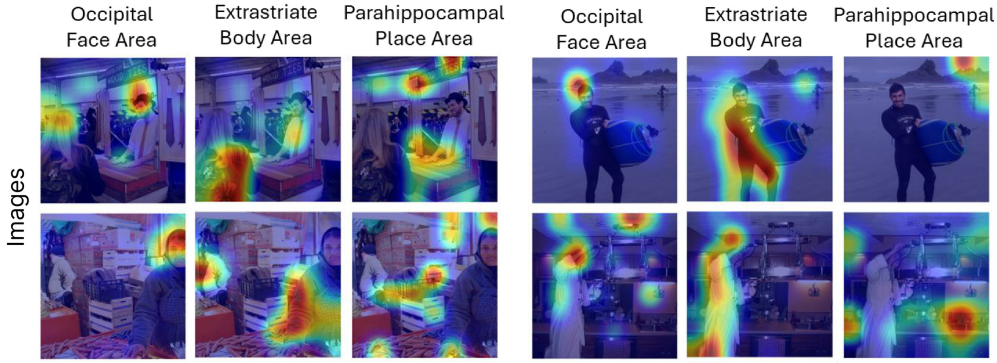

Figure 5: Spatial prototype activation maps for three category-selective regions of the visual cortex. See Appendix I for more examples.

4.6 Prototypes Reveal Spatial Selectivity of Brain Areas

To examine which parts of a visual scene are prioritized by each brain area, we derive spatial attention maps from the prototype activations associated with each ROI. For each ROI, we identify its top 10 prototypes using the linear prototype–voxel projection matrix. We then aggregate the spatial activations of these prototypes from the prototype representation $\mathbf{A} \in \mathbb{R}^{M \times P}$ to obtain a single 8×8 spatial response map. After normalization and upsampling, this produces a continuous attention map highlighting the image regions that most strongly contribute to the predicted activity of the ROI.

The resulting maps reveal distinct spatial selectivity across visual regions (Figure 5). Prototypes associated with the occipital face area (OFA) selectively emphasize facial regions, whereas those linked to the extrastriate body area (EBA) concentrate on body parts and human figures. In contrast, prototypes associated with the parahippocampal place area (PPA) exhibit broader and more distributed patterns that capture scene-level structures such as room layouts. These results demonstrate that the prototype representation preserves spatial structure, enabling the model to reveal not only *what* visual concepts drive neural responses but also *where* those features are located within the scene.

4.7 Prototype Manipulations Enable Controlled Stimulus Interventions

A key advantage of our method is that individual prototypes can be selectively manipulated, enabling targeted manipulation of visual features for testing *causal* hypotheses about selectivity in the brain. We demonstrate this unique capability by training a diffusion model (see Appendix O) that reconstructs the original image based on its prototype activation map (*i.e.,* $\mathbf{A}$ in Equation 5), and selectively permuting the activation map during inference. We train our own model from scratch instead of adopting a generative model pretrained on language data (*e.g.,* [38]), so that our analysis is less prone to potential biases [39]. We experiment with the top-20 prototypes for face-selective areas, and perform three types of permutation for causal intervention (see Appendix K for interventions with body prototypes):

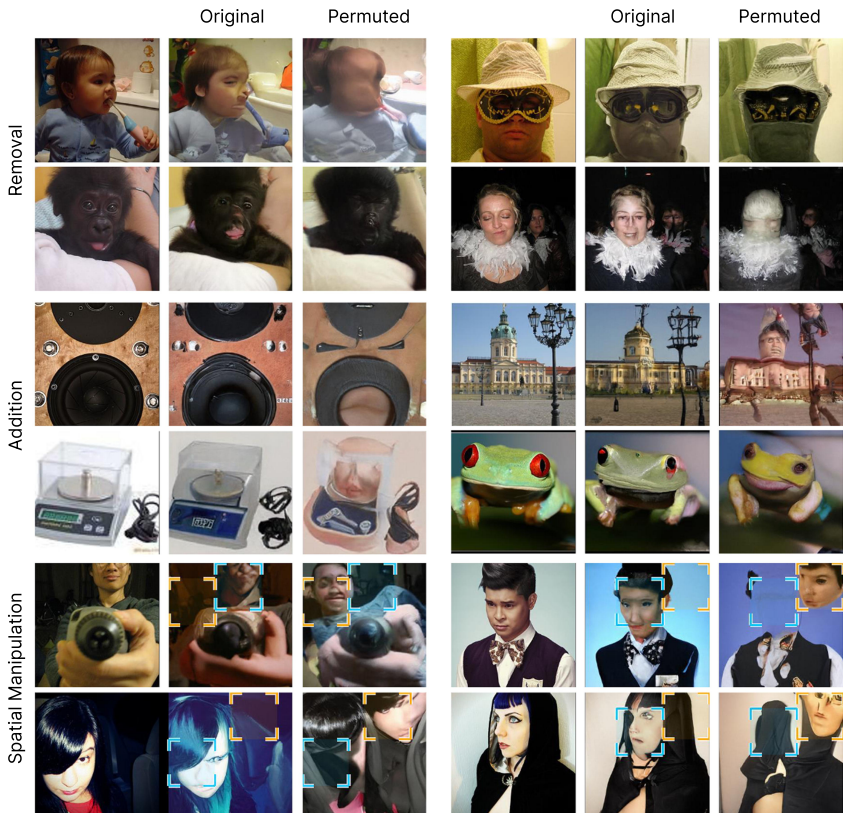

Figure 6: Causal interventions on face prototypes. Each example (from left to right) shows the stimulus, and its reconstruction with original/permuted prototypes. Blue and orange boxes indicate regions with high and low prototype activations that are permuted. See Appendix L for more results.

Removal. We choose images that strongly activate the selected prototypes, and set those prototype activations to zero before generating images from the modified representation. The resulting images (1st-2nd row in Figure 6) show faces selectively removed from the scene while other objects and background elements remain largely intact, indicating that the face prototype captures face-specific information without strongly entangling unrelated visual content.

Addition. Conversely, increasing the activation of face prototypes by their mean values induces the appearance of face-like features in previously non-face stimuli, yielding pareidolia-like objects with visually salient eye- and mouth-like structures (3rd–4th row in Figure 6). In face stimuli, the same manipulation further magnifies these features, for example, by enlarging the mouth.

Spatial Manipulation. We can also spatially manipulate the prototypes, by switching prototype activations of a neighboring region (3×3) with strong face responses (blue box) with one that has lower responses (orange box). Results in the 5th-6th row of Figure 6 demonstrate that we successfully shift the locations of faces, without significantly affecting the content in unaltered regions. These interventions demonstrate that prototypes provide a causally meaningful decomposition of visual content: removing a prototype eliminates its associated features, amplifying it introduces or strengthens them, spatially manipulating them alters the corresponding structure of the scene.

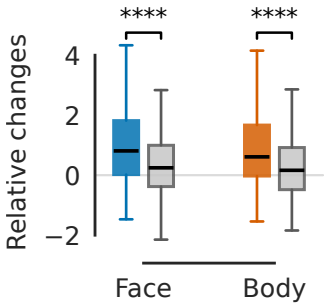

Figure 7: Changes in predicted fMRI responses for targeted permutations. Gray boxes denote controls.

Next, we perform a **close-loop experiment** to study how targeted permutation drives brain activity. Specifically, we pass the images reconstructed with the original and permuted prototypes to our pipeline, and measure the relative

changes in predicted fMRI voxel responses. We consider permutation by increasing prototype activations (*i.e.,* Addition) on face and body ROIs (each with 1000 images), which is relevant to identifying the "optimal stimuli" for brain areas. For control groups, we perform permutation on the same set of images but with randomly selected prototype, instead of those identified based on fMRI encoding models. As visualized in Figure 7, targeted permutations significantly increase predicted responses in corresponding brain ROIs. On the other hand, random permutations (Control) result in much smaller changes (Average relative changes for targeted permutation vs control. Face ROIs: 0.81 vs 0.24, $p < 10^{-4}$; Body ROIs: 0.61 vs 0.16, $p < 10^{-4}$).

Together, the aforementioned results suggest the proposed prototype representation as a practical interface for designing stimulus manipulations to probe the selectivity of visual cortex.

5 Conclusion, Broader Impacts, and Limitations

This paper introduces a prototype-based framework for studying visual representations in the human brain. Our central contribution is to transform dense features from vision foundation models into sparse, spatially grounded prototypes that capture coherent visual concepts and can be linked to fMRI responses through a transparent linear mapping. This design turns an opaque visual representation into an interpretable interface between images and cortex: prototypes describe *what* concepts are present and *where* they occur, while voxel weights reveal *which* concepts drive each brain region. Through this shared interface, we unify analyses that are typically studied separately, including neural encoding, categorical selectivity, cortical organization, spatial attribution, and controlled stimulus manipulation. More broadly, our results show how interpretable representation learning can turn foundation models from black-box predictors into structured tools for probing biological intelligence.

Our work also has room for improvement. First, we evaluate the approach in established brain areas, whereas applying it to less understood regions may enable more novel discoveries. Second, with vision foundation models trained largely with naturalistic images, our method may be less effective for interpreting synthetic or atypical stimuli. Finally, our prototypes summarize recurring visual patterns and provide an interpretable basis for neural prediction, but they do not necessarily correspond to biological prototypes implemented by the brain.

Acknowledgments and Disclosure of Funding

This work was supported by Howard Hughes Medical Institute (to SC and DT), and Hankuk University of Foreign Studies Research Fund (to SA).

References

[1] Talia Konkle and George A. Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, 13(1):491, 2022.

[2] Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. 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, 2023.

[3] Huzheng Yang, James Gee, and Jianbo Shi. Brain decodes deep nets. In *CVPR*, pages 23030–23040, 2024.

[4] Hossein Adeli, Minni Sun, and Nikolaus Kriegeskorte. Transformer brain encoders explain human high-level visual responses. *NeurIPS*, 2025.

[5] Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1):9383, 2024.

[6] Andrew Luo, Margaret Marie Henderson, Leila Wehbe, and Michael J. Tarr. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. In *NeurIPS*, 2023.

[7] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, pages 14453–14463, 2023.

[8] Pinyuan Feng, Hossein Adeli, Wenxuan Guo, Fan L. Cheng, Ethan Hwang, and Nikolaus Kriegeskorte. Towards interpretable visual decoding with attention to brain representations. In *ICLR*, 2026.

[9] Mu Nan, Muquan Yu, Weijian Mai, Jacob S. Prince, Hossein Adeli, Rui Zhang, Jiahang Cao, Benjamin Becker, John A. Pyles, Margaret M. Henderson, Chunfeng Song, Nikolaus Kriegeskorte, Michael J. Tarr, Xiaoqing Hu, and Andrew F. Luo. Meta-learning in-context enables training-free cross subject brain decoding. In *CVPR*, 2026.

[10] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, ethan cohen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, and Tanishq Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In *NeurIPS*, volume 36, pages 24705–24728, 2023.

[11] Roman Beliy, Amit Zalcher, Jonathan Kogman, Navve Wasserman, and Michal Irani. Brain-it: Image reconstruction from fmri via brain-interaction transformer. In *ICLR*, 2026.

[12] Alicia Zeng and Jack Gallant. Disentangling superpositions: Interpretable brain encoding model with sparse concept atoms. *NeurIPS*, 2025.

[13] Navve Wasserman, Matias Cosarinsky, Yuval Golbari, Aude Oliva, Antonio Torralba, Tamar Rott Shaham, and Michal Irani. Brainexplore: Large-scale discovery of interpretable visual representations in the human brain. *arXiv*, 2025.

[14] Gabriel Sarch, Michael Tarr, Katerina Fragkiadaki, and Leila Wehbe. Brain dissection: fmri-trained networks reveal spatial selectivity in the processing of natural images. In *NeurIPS*, volume 36, pages 46255–46283, 2023.

[15] 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. In *ICLR*, 2025.

[16] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9630–9640, 2021.

[17] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pages 15619–15629, 2023.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

[19] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan Su. This looks like that: Deep learning for interpretable image recognition. In *NeurIPS*, 2019.

[20] Pang Wei Koh, Thao Nguyen, Yewen Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, 2020.

[21] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. In *arXiv*, 2025.

[22] 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.

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

[24] Zijin Gu, Keith Wakefield Jamison, Meenakshi Khosla, Emily J. Allen, Yihan Wu, Ghislain St-Yves, Thomas Naselaris, Kendrick Kay, Mert R. Sabuncu, and Amy Kuceyeski. Neurogen: Activation optimized image synthesis for discovery neuroscience. *NeuroImage*, 247:118812, 2022.

[25] Carlos R. Ponce, Will Xiao, Peter F. Schade, Till S. Hartmann, Gabriel Kreiman, and Margaret S. Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009.e10, 2019.

[26] N. Apurva Ratan Murty, Pouya Bashivan, Alex Abate, James J. DiCarlo, and Nancy Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature Communications*, 12(1):5540, 2021.

[27] A. T. Gifford, B. Lahner, S. Saba-Sadiya, M. G. Vilas, A. Lascelles, A. Oliva, K. Kay, G. Roig, and R. M. Cichy. The algonauts project 2023 challenge: How the human brain makes sense of natural scenes. In *arXiv*, 2023.

[28] Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In *ICLR*, 2024.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, page 6000–6010, 2017.

[30] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI*, 2018.

[31] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *ICLR*, 2024.

[32] Sharut Gupta, Shobhita Sundaram, Chenyu Wang, Stefanie Jegelka, and Phillip Isola. Better together: Leveraging unpaired multimodal data for stronger unimodal models. In *ICLR*, 2026.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014.

[34] 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, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763, 2021.

[36] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NeurIPS*, volume 14, pages 849–856, 2002.

[37] 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.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[39] David Mayo, Christopher Wang, Asa Harbin, Abdulrahman Alabdulkareem, Albert Shaw, Boris Katz, and Andrei Barbu. Brainbits: How much of the brain are generative reconstruction methods using? In *NeurIPS*, volume 37, pages 54396–54420, 2024.

[40] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.

[41] Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *arXiv*, 2026.

[42] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[43] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[44] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

A Visualization of More Prototypes

Additional examples of learned prototypes (Figure 8 and Figure 9). For each prototype, we show representative images from the COCO dataset [33] that produce the highest activation scores. Consistent with the examples shown in the main text, the learned prototypes capture a diverse set of semantically coherent visual concepts, including airplanes, buildings, faces, food, sports activities, crowds, ties, electronic appliances, skiing scenes, and fruits.

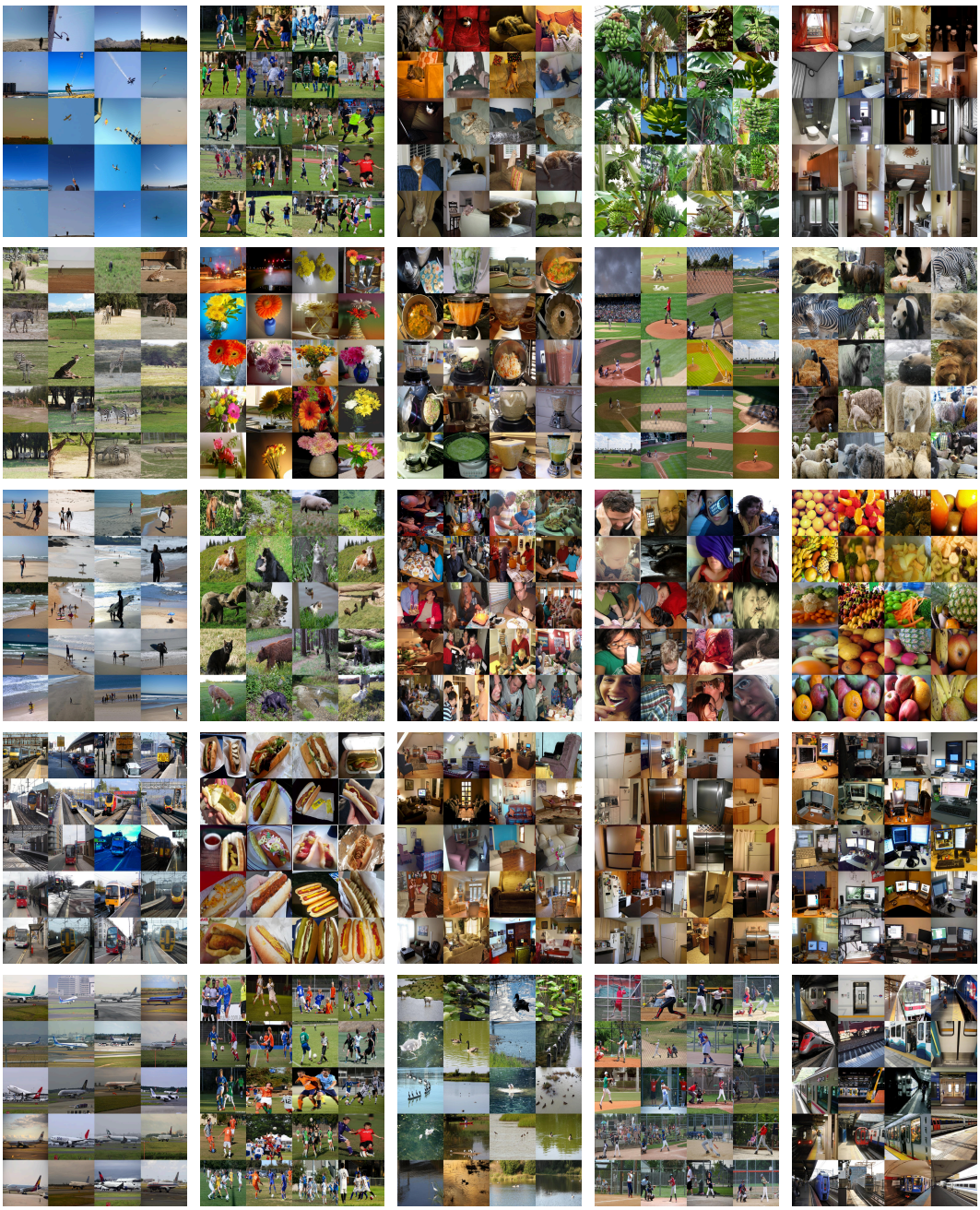

Figure 8: Supplementary results for the visualization of representative images for different prototypes.



Figure 9: Supplementary results for the visualization of representative images for different prototypes (cont'd).

B Shared Prototypes Between Individuals

Cross-subject consistency of selected prototypes (Figure 10). For each subject and ROI, the top-10 prototypes are identified from independently trained encoding models. The histogram shows the number of prototypes shared across different numbers of subjects. Many prototypes are consistently selected across individuals, particularly for face- and body-selective regions, suggesting that these prototypes capture visual features that generalize across brains.

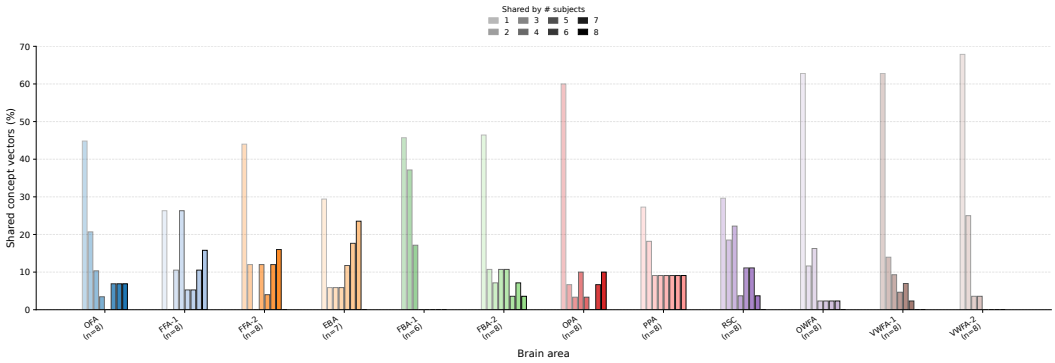

Figure 10: Histogram for the number of prototypes shared across multiple subjects.

C More ROI-Prototype Visualizations

More visualizations of representative prototypes for all twelve visual ROIs (Figure 11). The twelve visual ROIs include face-selective regions (occipital face area, OFA; fusiform face areas, FFA-1 and FFA-2), body-selective regions (extrastriate body area, EBA; fusiform body areas, FBA-1 and FBA-2), scene-selective regions (occipital place area, OPA; parahippocampal place area, PPA; retrosplenial cortex, RSC), and word-selective regions (occipital word-form area, OWFA; visual word-form areas, VWFA-1 and VWFA-2).

These visualizations reveal additional patterns consistent with the known categorical organization of the ventral visual stream. For example, the occipital word-form area (OWFA) is driven by prototypes containing objects with prominent lettering (e.g., clocks, stop signs, and bus fronts with visible text). The visual word-form area (VWFA) unexpectedly shows a prevalence of food-related prototypes. One possible explanation is that these images share visual properties, such as dense textures, repeated circular elements, or fine-grained patterns that resemble the spatial statistics of letter strings, which are known to drive responses in this region.



Figure 11: Top prototypes that characterize all 12 selected brain ROIs.

D Visualization of Prototypes with Diverse Weights to ROIs

Our results in the main paper and Appendix C visualize the most preferred prototypes for individual ROIs, demonstrating the ability of our method to capture the dominant selectivity of brain ROIs. In this section, we present complementary results showing the full spectrum of prototype preferences—ranging from the most to the least preferred prototypes for predicting responses in a given ROI.

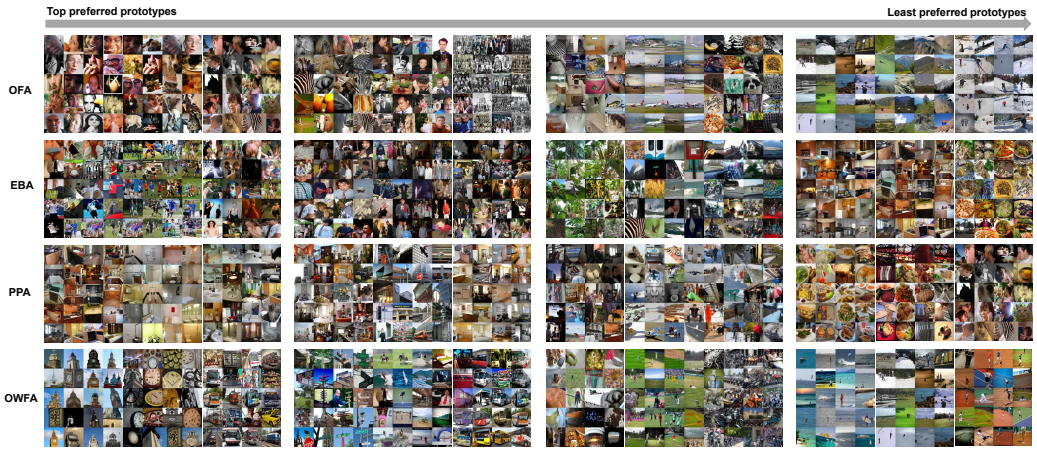

Figure 12: Prototypes with different preference levels for four brain ROIs with distinct selectivity.

As shown in Figure 12, our method further characterizes the continuous tuning properties of brain ROIs. For the occipital face area (OFA), moving from the most to the least preferred prototypes reveals a gradual transition from frontal faces, to human figures in varied poses, and finally to natural scenes. A similar trend holds for the extrastriate body area (EBA), where prototypes shift from sports scenes with prominent body movements, to group photos, and eventually to indoor scenes without people. Together, these results show that our method characterizes brain ROIs beyond their dominant selectivity, revealing a graded preference structure that reflects their tuning.

E Effects of Spectral Clustering on Prototype Visualization

Following [13] that uses CLIP embedding for image selection and pattern summarization, we employ spectral clustering with CLIP embedding to select representative images for each prototype. This step is optional and for visualization only. To evaluate its influence, we compare the selected images for the same prototypes before and after clustering in Figure 13. Results show that, while clustering affects selection of specific images, it does not change the semantic meanings of the corresponding prototypes, *e.g.,* frontal faces for OFA, sports scenes for EBA, and texts for OWFA.

F PCA Fails to Characterize Brain ROIs

As a control baseline for prototypes, we apply Principal Component Analysis (PCA) on the raw DINOv3 [21] features, using the same training data as our method. We then train a fMRI encoding model based on the PCA features, and identify the top principal components (PCs) [40] for different brain areas. In Figure 14, we visualize top-20 images with the highest values for each of these selected PCs (top-4 PCs shown for each brain ROI). The results show that, unlike our visual prototypes, PCA features do not exhibit clear category-selective patterns across visual regions. In addition, the top images tuned for each PC also do not have coherent meanings.

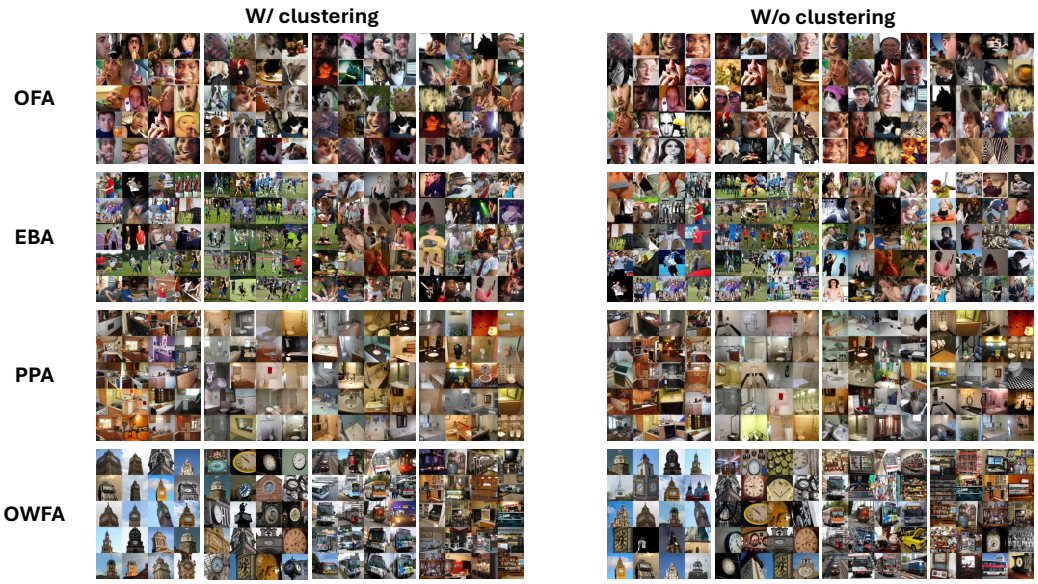

Figure 13: Representative images for the same prototypes (*i.e.,* Top-4 prototypes for different brain ROIs) with and without clustering.

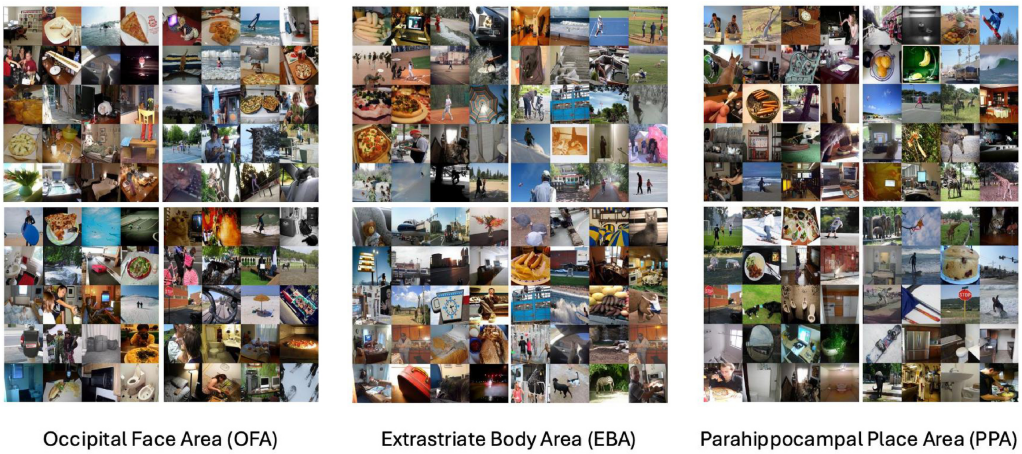

Figure 14: Representative images of top principal components (for PCA features) that characterize different brain ROIs.

G Comparison between Prototypes and Sparse Autoencoder for Brain ROI Characterization

In the main paper, we show that our proposed prototype method achieves better fMRI encoding performance compared to a sparse autoencoder (SAE) learned on top of DINOv3 features. In this section, we further demonstrate the advantages of our method in characterizing the selectivity of brain ROIS. In particular, we visualize the representative images for the top-5 prototypes (our method) or SAE features (SAE) with the highest weights for predicting fMRI responses in different brain ROIs (Occipital Face Area, Extrastriate Body Area, and Parahippocampal Place Area).

As shown the in Figure 15, SAE roughly captures the selectivity of certain brain ROIs (*e.g.,* top SAE features for Parahippocampal Place Area mostly correspond to outdoor scenes). However, unlike our method that characterize a brain ROI with prototypes with coherent meanings (*e.g.,* close-up faces for Occipital Face Area and sports scenes for Extrastriate Body Area), top SAE features for a brain

ROI typically have mixed semantics (*e.g.,* top SAE features for Occipital Face Area are associated to both text and faces, top SAE features for Extrastriate Body Area show close-up faces or animals with limited movement), resulting in difficulties in accurately elaborating the brain selectivity.

These results demonstrate the effectiveness of our approach. More importantly, they highlight the advantages of explicitly factorizing visual features into compositional prototypes compared to learning implicit sparse features.

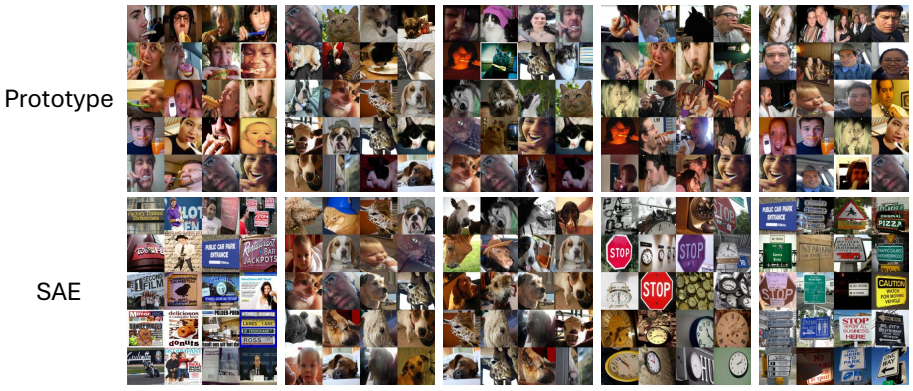

Figure 15: Comparison between the proposed prototype method (Prototype) and sparse autoencoder (SAE) for characterizing selectivity of different brain ROIs. Each row visualizes the representative images for top-5 prototypes/SAE features.

H Neural Encoding Performance of Top Preferred Prototypes

Our analysis in the main paper show that prototypes with the largest positive weights successfully reveal an ROI's dominant preferences (*e.g.,* face, body, scene), and the full collection of prototypes show improved fMRI encoding performance compared to the corresponding baselines. In this section, we ask a complimentary question: how much predictive performance is retained when the encoding model is restricted to its top-(k) preferred prototypes? We address this question by ranking prototype features by their learned encoding weights and fitting a separate linear regression model for each voxel using only the top-(k) features.

	OFA	FFA-1	FFA-2	EBA	FBA-1	FBA-2	OPA	PPA	RSC	OWFA	VWFA-1	VWFA-2	Average
Top-4	0.184	0.283	0.358	0.295	0.183	0.327	0.118	0.126	0.262	0.100	0.161	0.169	0.214
Top-8	0.192	0.318	0.375	0.313	0.242	0.329	0.157	0.197	0.333	0.153	0.154	0.213	0.248
Top-16	0.225	0.332	0.393	0.333	0.256	0.359	0.155	0.206	0.334	0.187	0.258	0.228	0.272
Top-32	0.274	0.342	0.411	0.347	0.263	0.372	0.180	0.286	0.342	0.225	0.261	0.261	0.297
Top-64	0.304	0.375	0.427	0.365	0.292	0.394	0.251	0.327	0.365	0.281	0.281	0.279	0.329
Full model	0.371	0.466	0.516	0.506	0.403	0.504	0.414	0.424	0.414	0.299	0.390	0.293	0.416

Table 2: fMRI encoding with top-k preferred prototypes, results are reported on subject 1.

The results in Table 2 show that the top-4 prototypes capture a considerable component of voxel responses, achieving a mean correlation of (r=0.214), or approximately 52% of the full-model correlation. Performance increases progressively as more prototypes are included: the top 16, 32, and 64 retain approximately 65%, 71%, and 79% of the full-model correlation (with all 1000 prototypes), respectively. Together, these results indicate that a relatively small subset of the highest-weighted prototypes accounts for most of the ROI-level prediction, although the full set is needed to achieve the best predictive performance.

I Additional Examples for Spatial Selectivity of Brain ROIs

In Figure 16, we provide additional examples for characterizing the spatial selectivity of three distinct brain ROIs (Occipital Face Area, Extrastriate Body Area, and Parahippocampal Place Area). The results further demonstrate that our approach accurately reveals the spatial locations that tightly aligned with the functional selectivity of brain ROIs.

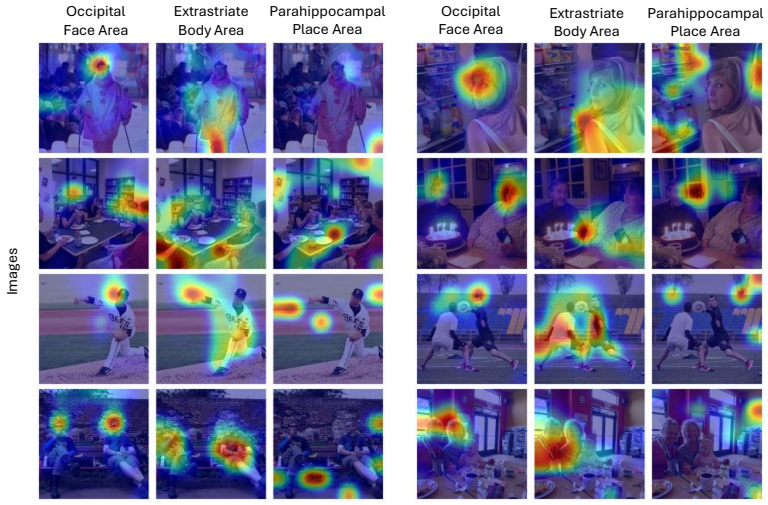

Figure 16: Additional examples for spatial selectivity of brain ROIs.

J More per-prototype cortical activation maps

Per-prototype cortical activation maps and representative image examples (Figure 17). For each prototype, we visualize the top activating images (left) and the corresponding cortical activation pattern across ROIs (right). Warm colors indicate stronger predicted responses, while cool colors indicate weaker responses. ROI boundaries for EBA, OFA, and OPA are outlined.

Prototypes in the first row selectively activate the extrastriate body area (EBA), corresponding to images containing body parts such as legs, feet, and hands interacting with objects (e.g., shoes, phones). The second row shows prototypes associated with the occipital place area (OPA), capturing indoor environments such as kitchens and transportation scenes. The third row highlights prototypes that jointly activate OFA and OPA, often corresponding to scenes containing a single prominent object, e.g., airplanes or trains embedded within a scene. The final row illustrates prototypes that produce distributed activation across multiple ROIs, corresponding to complex scene categories such as outdoor environments or farm scenes.

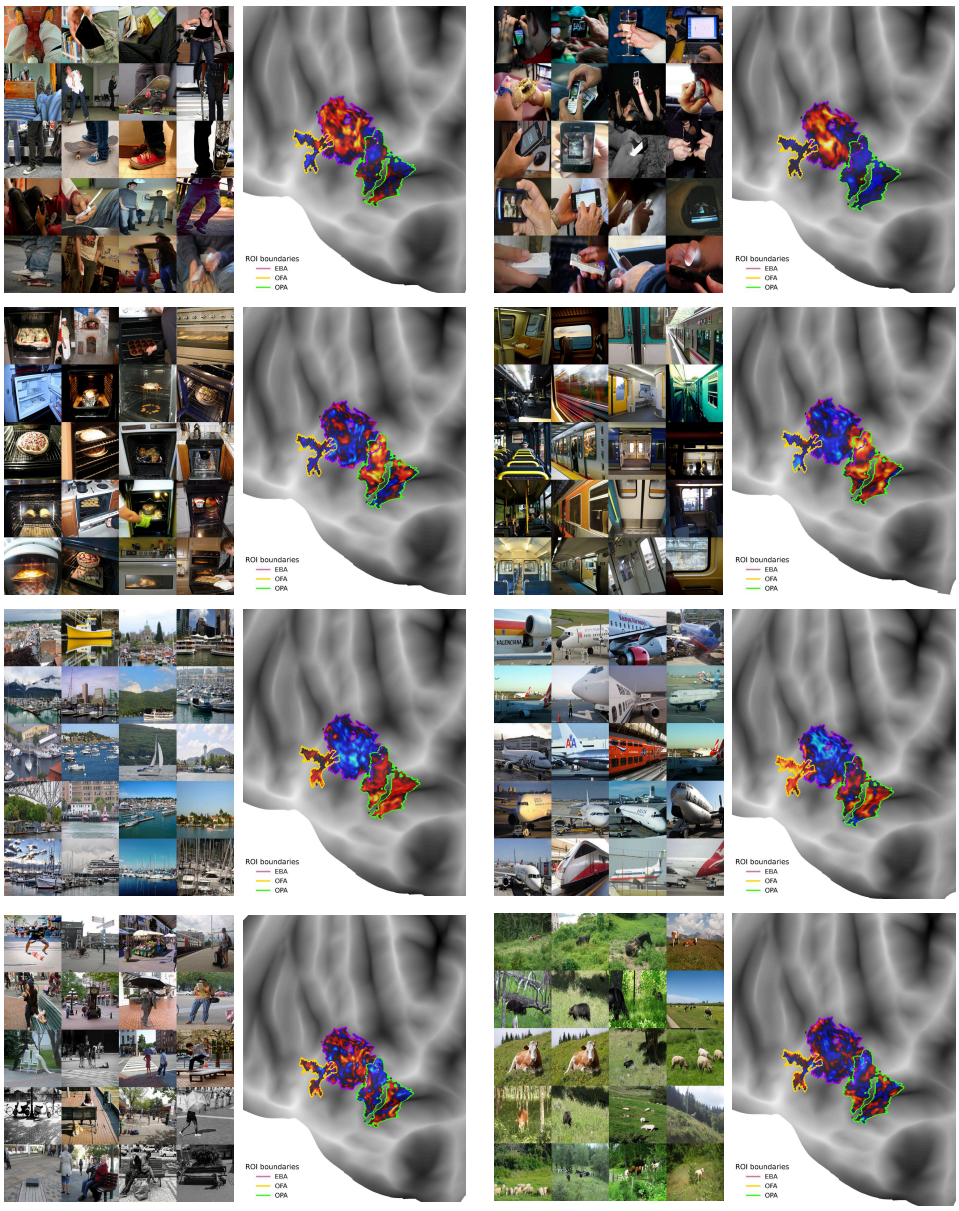

Figure 17: Additional results on prototype weights for different brain ROIs.

K Controlled Stimulus Interventions with Body Prototypes

The main paper uses prototypes contributing to face-selective brain ROIs for causal intervention with our approach. In this section, we additionally experiment with top-20 prototypes that have the highest weights for body-selective ROIs, and perform the same three types of permutations. As visualized in Figure 18, manipulating these body prototypes robustly alters the targeted visual concepts (*e.g.,* hands and arms) while keeping the other image properties relatively unchanged. Specifically: (1) Setting activations of body prototypes (*i.e.,* Removal) to zeros removes hands and arms within the images, or distorts the shape of fingers; (2) On the other hand, increasing the activations (*i.e.,* Addition) leads to emergence of hands and fingers, or alters the colors of objects to make them more skin-like; (3) Spatially permuting (*i.e.,* Spatial manipulation) the body prototypes (blue/orange boxes denote regions with high/low body prototype activations that are switched) allows us to shift the locations of corresponding concepts. Taken together, these results complement our findings in the main paper, and further validate the usefulness of our approach for probing the causal relationships between neural activity and visual concepts.

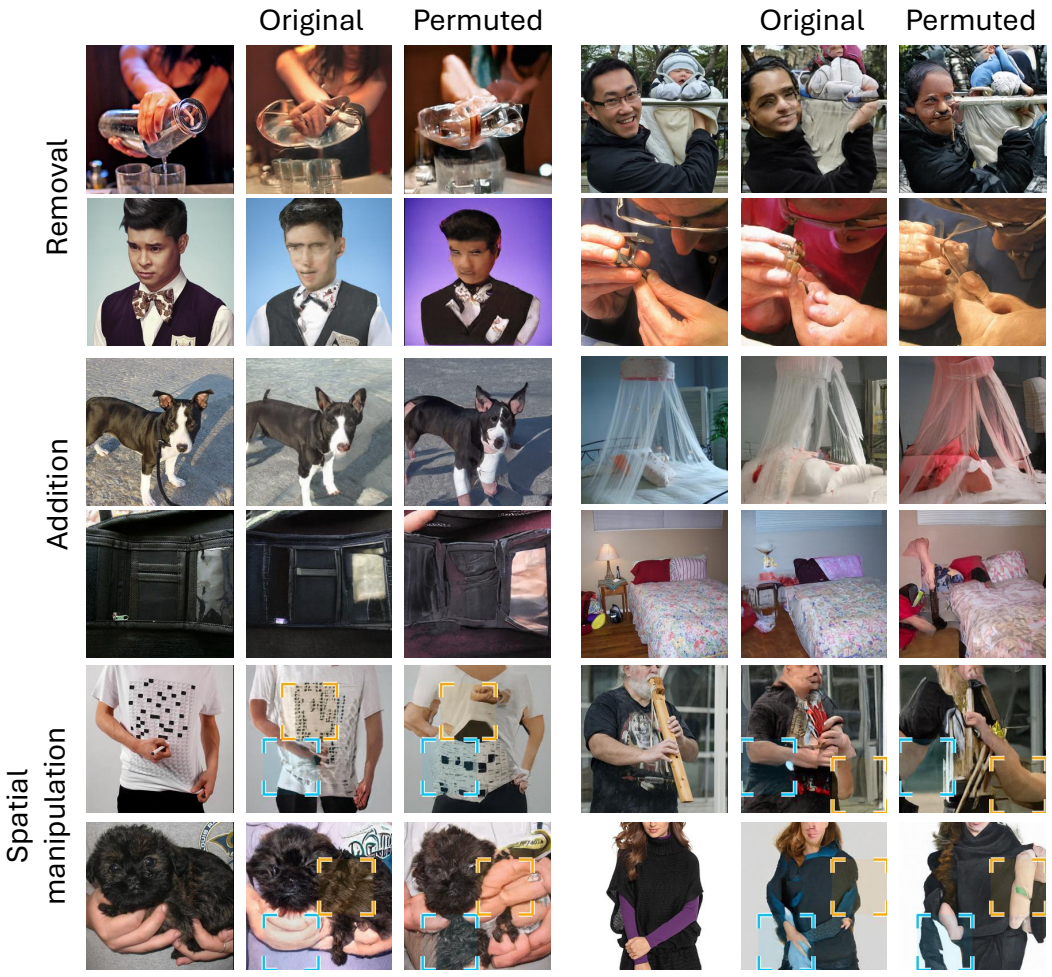

Figure 18: Causal stimuli generated results with targeted permutations on body areas.

L Additional Examples of Face Prototype Interventions

In this section, we provide additional examples (Figure 19) for causal intervention with the three types of permutations for face prototypes. The results agree with our observations in the main paper.

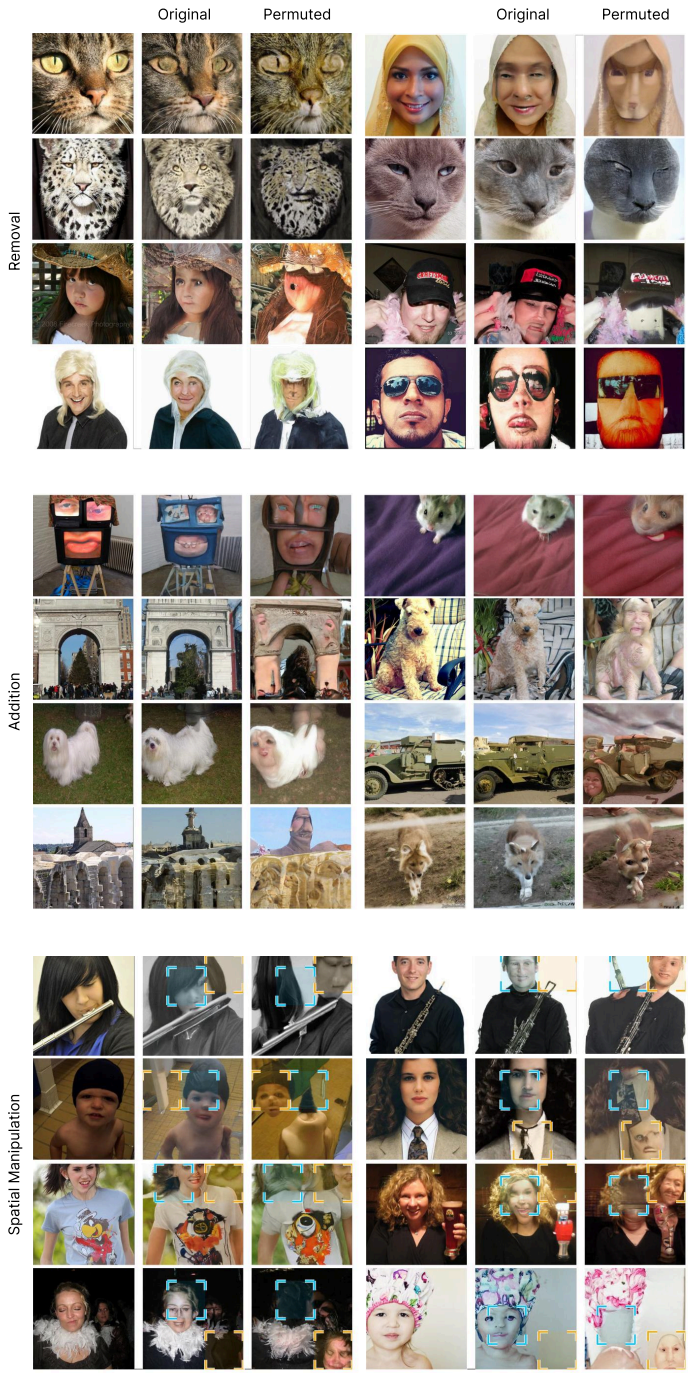

Figure 19: Additional results on causal stimuli generation with permutations of body areas.

M Supplementary Neural Encoding Results

Table 3 reports the fMRI encoding performance of different models, across all 12 brain ROIs used in our experiments.

	OFA	FFA-1	FFA-2	EBA	FBA-1	FBA-2	OPA	PPA	RSC	OWFA	VWFA-1	VWFA-2
Noise-ceiling	0.561	0.597	0.541	0.600	0.541	0.588	0.563	0.585	0.542	0.530	0.551	0.500
Adeli *et al.* [4]	0.430	0.489	0.428	0.481	0.432	0.476	0.438	0.485	0.444	0.399	0.430	0.363
JEPA-VIT	0.305	0.381	0.324	0.381	0.348	0.381	0.348	0.414	0.368	0.244	0.317	0.259
Proto JEPA-VIT	0.355	0.403	0.346	0.404	0.359	0.400	0.399	0.419	0.353	0.314	0.350	0.270
DINOv3-ConvNeXt	0.273	0.366	0.308	0.379	0.319	0.369	0.324	0.377	0.330	0.209	0.304	0.256
Proto DINOv3-ConvNeXt	0.335	0.410	0.340	0.421	0.347	0.404	0.378	0.408	0.310	0.268	0.342	0.267
DINOv3-VIT	0.302	0.388	0.326	0.396	0.330	0.379	0.332	0.386	0.325	0.254	0.314	0.255
Proto DINOv3-VIT	0.367	0.434	0.376	0.433	0.387	0.432	0.404	0.443	0.387	0.310	0.370	0.296

Table 3: Comparison of fMRI encoding across 8 subjects. We report the voxel-wise Pearson's correlation for all 12 brain ROIs.

N Ablation Study on the Hyperparameters

Our method has two hyperparameters, *i.e.,* the number of prototypes P and k for sparsity constraint (only top-k prototypes are activated). Here we show a comprehensive comparison of the fMRI encoding performance with different parameter choices, reported as the voxel-wise correlation on Subject 1. For experiments with different numbers of prototypes, we fix the sparsity constraint as $k = 128$. For experiments with sparsity constraint, we fix the number of prototypes as $P = 1000$.

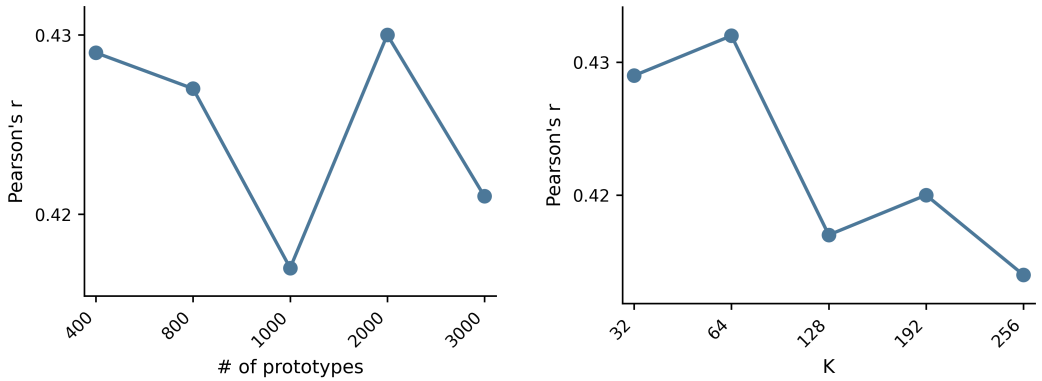

Figure 20: fMRI encoding performance averaged across 12 different brain ROIs. For experiments with different numbers of prototypes (left), we fix the sparsity constraint as $k = 128$. For experiments with sparsity constraint (right), we fix the number of prototypes as $P = 1000$

Results in Figure 20 suggest that our approach is relatively robust to the selection of these two hyperparameters: increasing the number of prototypes or loosening the sparsity constraints leads to a minor drop of performance (*i.e.,* around 0.01 drop in Pearson's correlation). Based on these observation, we choose sparsity constraint $k = 128$ and $P = 1000$ for a good balance between interpretability (*i.e.,* reserving sufficient number of prototypes with reasonable granularity) and performance.

O Diffusion Model for Causal Intervention Experiment

Different from studies [6, 7, 8, 10, 11] that rely on generative models pretrained with language data (*e.g.,* Stable Diffusion [38] and CLIP [35]), we opt to train our own generative models from scratch with only visual data. This enables us to analyze the patterns of the visual cortex without assuming prior structures, *e.g.,* without imposing language structure on representations in the visual cortex.

Specifically, as shown in Figure 21, we choose the recently introduced JiT [41] diffusion model as our testbed, which follows a minimalist design while achieving state-of-the-art performance. While

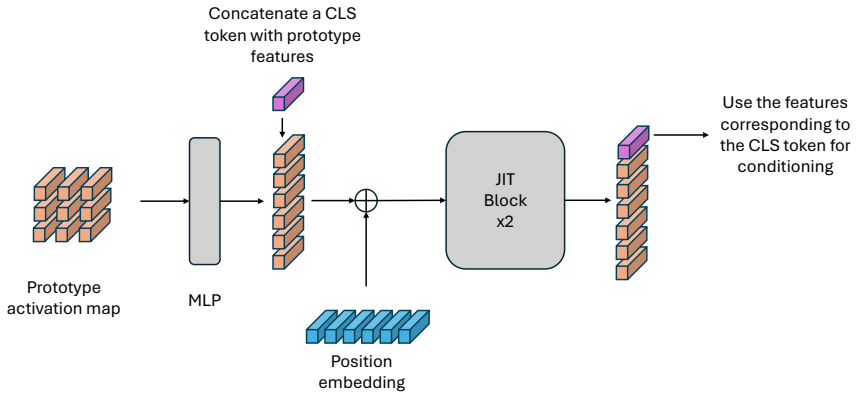

Figure 21: Schematic for the design of prototype encoder used in the diffusion models.

the original goal of JiT is to *synthesize* images given class labels, here our objective is to *reconstruct* images based on their prototype activation maps. Therefore, we substitute the class encoder with a prototype encoder (see the figure below), and also disable the classifier-free guidance [42]. We train the model on ImageNet [43] with the standard parameters used for the original model, except that we reduce the noise level by half to facilitate learning of reconstruction.

P Experiment Details

Vision foundation model. We use ViT-B16 and ConvNeXt-Base from DINOv3 [21], and ViT-H from JEPA [17] as our backbone vision foundation models. For feature decomposition, we take the normalized patch features from the last layer.

Brain ROIs. For this study, we focus on 12 established brain ROIs with distinct selectivity, including Occipital Face Area (OFA) and Fusiform Face Area (FFA-1 and FFA-2) selective for faces; Extrastriate Body Area (EBA) and Fusiform Body Area (FBA-1 and FBA-2) selective for bodies; Occipital Place Area (OPA), Parahippocampal Place Area (PPA), and Retrosplenial Cortex (RSC) selective for places; Occipital Word Form Area (OWFA) and Visual Word Form Area (VWFA-1 and VWFA-2) selective for words.

Visualization of prototypes. To visualize the representative images for each prototype, we first identify the top-50 images with the highest activation averaged across all patches. To ensure consistency among the final representatives, we then compute pairwise cosine similarities between CLIP [35] embeddings of these candidates and apply spectral clustering [36] to select 20 images that are both highly activated and representative of a shared visual concept associated with the prototype.

Optimization for prototypes. We use Adam [44] as our optimizer for training the prototypes (stage-1). The model is trained for 15 epochs with the official training and validation split from MSCOCO [33]. The learning rate is set to 10^{-3}, with exponential decay. Batch size is set to 128. The best model is selected based on the validation performance.

Optimization for fMRI encoding model. Upon learning the prototypes, we extracted the prototype activation maps for images used in the NSD dataset [34]. The same Adam optimizer is used to train the fMRI encoding model, which takes the fixed prototype activations and predicts the voxel responses. We create our own training, validation, and test set for each subject by splitting the official training set into $80\%/10\%/10\%$. We train our encoding model on a subject basis, and select the voxel from a single or multiple brain ROIs, depending on the analysis (*e.g.,* for categorical selectivity

in §4.4, we train an encoding model for each brain ROI, while for prototype mapping in §4.5, we train the model to predict responses for multiple ROIs.) The learning rate is set to 10^{-2}, with batch size 32, and number of epochs 20.

Computational resource. Our main model is relatively lightweight, and trained on a customized Linux server with two RTX 4090 GPUs (24GB VRAM per GPU). For the diffusion model used in our causal intervention analysis, we use another Linux server with one H100 and two RTX 6000 (96GB VRAM per GPU).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We carry out a collection of experiments to analyze fMRI data from six distinct perspectives, which validates the effectiveness of our approach on learning a unified representation for brain understanding.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations and potential extensions of our method.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: Our paper does not have theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the detailed design (*e.g.,* number of layers and their hyperparameters) in the main paper and appendix, alongside with training configuration. Our source code will be publicly available upon acceptance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will open source our code upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the experiment details, and also perform ablation study on the hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the standard deviation of our fMRI encoding performance across different subjects.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the instrument used for model training in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All authors follow the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impact of our works is discussed

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper does not introduce data or model that has a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All corresponding creators are cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide details of our model and dataset being used. Comprehensive documentation will be available, with our source code upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our study does not involve crowd-sourcing, and instead depends on public datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our study does not involve crowd-sourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Our method does not involve components that rely on large language models.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.