BrainACTIV: Identifying visuo-semantic properties driving cortical selectivity using diffusion-based image manipulation

Diego García Cerdas[1,*], **Christina Sartzetaki**[1], **Magnus Petersen**[2], **Gemma Roig**[2], **Pascal Mettes**[1] **and Iris Groen**[1]

[1]*Informatics Institute, University of Amsterdam, The Netherlands*
[2]*Department of Computer Science, Goethe Universität Frankfurt, Germany*

Corresponding author: *E-mail address:* diego.gcerdas@gmail.com.

Figure 1. BrainACTIV manipulates a reference image to maximize or minimize the activity of a region in the human visual cortex. By analyzing these image variations, we can quantify visuo-semantic representations that underlie selective responses in the human brain. In the examples above, manipulations enhance hypothesized preferred categories in specific brain regions known to exhibit selectivity for faces, bodies, words, and places, respectively. Reference images belong to the MS-COCO dataset (Lin et al., 2014).

The human brain is thought to efficiently represent visual inputs through specialized neural populations that selectively respond to specific image categories. Advancements in generative modeling have enabled data-driven discovery of neural selectivity using brain-optimized image synthesis. However, current methods independently generate one sample at a time, making it hard to discern which image features drive neural response selectivity. To address this issue, we introduce Brain Activation Control Through Image Variation (BrainACTIV), a method for manipulating a reference image to enhance or suppress activity in a target cortical region using pretrained diffusion models. Starting from a reference image allows for fine-grained and reliable identification of optimal visuo-semantic properties. We show that our manipulations effectively modulate predicted fMRI responses and agree with hypothesized preferred categories in established regions of interest, while remaining structurally close to the reference image. In addition, we describe how two hyperparameters allow a trade-off between semantic variation and low-level structural control, and how both variations affect predicted brain activity. Finally, we demonstrate how our method accentuates differences between brain regions that are selective to the same category. We propose that BrainACTIV holds the potential to formulate robust hypotheses about brain representation as well as produce controllable naturalistic stimuli for neuroscientific experiments.

1. Introduction

The discovery that human visual cortex contains neural populations that selectively respond to specific image categories has intrigued neuroscientists for decades (Grill-Spector and Weiner, 2014). While traditional approaches of measuring neural selectivity rely on hand-selected images, recent studies guide generative models with human brain encoder gradients for data-driven exploration of cortical selectivity (Ratan Murty et al., 2021; Gu et al., 2022; Luo et al., 2023; Pierzchlewicz et al., 2023). These studies optimize random noise vectors to synthesize maximum-activating images in target category-selective regions of interest (ROIs), allowing the formulation of new hypotheses about the representations in each ROI. However, because the initial noise vectors are independently sampled for every image, this process naturally leads to a varied set of images of which some characteristics are preferred by ROIs and others are randomly produced by the generative model. Disentangling these factors is essential for understanding the neural representations underlying category-selectivity, and for determining the relative contribution of visual versus semantic features to neural representation, which is a much-debated question across the domains of scene (Groen et al., 2017), object (Bracci et al., 2017), face (Vinken et al., 2023) and word (Janini et al., 2022) perception.

Here, we introduce *Brain Activation Control Through Image Variation* (BrainACTIV), a method for manipulating a reference image to enhance or suppress predicted brain activity in a target cortical region, see Figure 1. BrainACTIV uses IP-Adapter (Ye et al., 2023) to prompt a pretrained diffusion model with brain-optimal embeddings obtained through spherical interpolation in CLIP space. Initial diffusion latents are computed with SDEdit (Meng et al., 2022) to retain structural fidelity to the reference image. The manipulation of a reference image (Goetschalckx et al., 2019; Papale et al., 2024) ensures a reliable comparison point for the synthesized stimuli, isolating the effect of brain optimality on the latter. Besides allowing straightforward visual interpretation, our method opens up the possibility of quantifying differences in visuo-semantic and mid-level image features using computer vision techniques, highlighting those preferred by a brain region of interest.

We validate BrainACTIV by targeting fMRI responses in well-established category-selective ROIs, confirming that their predicted activation is successfully modulated by our image variations and the visuo-semantic properties highlighted agree with previous neuroscientific findings. Additionally, we describe two hyperparameters that researchers can use to balance semantic variation and low-level structural control when using BrainACTIV for experimental stimulus design. Finally, we demonstrate how our method can accentuate differences between similar regions of interest, providing insights into their specific roles in visual processing. Our contributions are:

- The use of image manipulation to maximize or minimize responses in higher visual cortex: this guarantees that the changes made to the original image come from the objective of increasing or decreasing the brain activation, rather than stochasticity in the image generation process, with the original image serving for activity baseline comparison.
- The introduction of tunable hyperparameters that control the degree of structural similarity and semantic variation with regard to the reference image, allowing for manipulation of the degree of low-level visual changes when generating image variations.
- The use of automated methods to quantify semantic category presence and mid-level image features to characterize each ROI in finer detail, circumventing the need for human behavioral assessments.
- The identification of differences in stimulus representation between similar brain regions beyond category selectivity, by accentuating these differences in a reference image.

2. Related Work

2.1. Category Selectivity in the Higher Visual Cortex

Different regions in high-level areas of the human visual cortex exhibit selectivity for specific semantic categories like faces, bodies, places, and words (Kanwisher et al., 1997; McCarthy et al., 1997; Downing et al., 2001; Peelen and Downing, 2005; Epstein and Kanwisher, 1998; McCandliss et al., 2003). Reliable characterization of each region requires measuring neural responses to large sets of images and finding those that elicit maximal activity (Ratan Murty et al., 2021). However, experimental constraints and the high dimensionality of image space make it implausible to test all potential stimuli. Neuroscientists have traditionally narrowed this search by focusing on hand-selected stimuli, but this risks overlooking relevant features that could not be conceived a priori. Deep neural networks (DNNs) trained on large-scale datasets of brain recordings have been adopted as "brain encoders" to make rapid and highly accurate predictions of neural responses to large volumes of images (Khosla et al., 2021). Moreover, deep generative models such as diffusion models Ho et al. (2020); Song et al. (2020) can synthesize novel stimuli by sampling from rich image priors constrained to the domain of natural images. Our work combines brain encoders and diffusion models to highlight semantic properties that drive functional selectivity in the visual cortex, enabling the formulation of new hypotheses more robustly and objectively than current data-driven approaches.

2.2. Image Variation with Diffusion Models and CLIP

Diffusion models treat the data generation process as iterative noise removal, progressively refining random noise $\boldsymbol{x}_T \sim \mathcal{N}(0,1)$ into structured data $\boldsymbol{x}_{T-1}, ..., \boldsymbol{x}_{t+1}, \boldsymbol{x}_t, \boldsymbol{x}_{t-1}, ..., \boldsymbol{x}_0$ through a trained denoising network. This process can be guided to synthesize samples from a conditional distribution, as done by text-to-image (T2I) models (Nichol et al., 2022; Saharia et al., 2022). Stable Diffusion (Rombach et al., 2022) enables efficient T2I synthesis by representing data in a lower-dimensional latent space. Image prompting allows for generating variations of a reference image $\mathcal{I}$, preserving its style and content. Early diffusion-based approaches supported image prompting by inverting a CLIP image encoder (Radford et al., 2021; Ramesh et al., 2022) or fine-tuning T2I models on CLIP image embeddings (Rombach et al., 2023). Alternatively, IP-Adapter Ye et al. (2023) introduces additional cross-attention layers in the denoising network of pretrained T2I models, incorporating information extracted by a CLIP image encoder. To preserve low-level structural fidelity to the reference image, SDEdit (Meng et al., 2022) initializes the denoising process at an intermediate step by injecting noise to $\mathcal{I}$ up to timestep $t_0 = \gamma \cdot T$ with $\gamma \in [0, 1]$ and using $\boldsymbol{x}_{t_0-1}$ as starting point. Our work employs IP-Adapter and SDEdit on Stable Diffusion to generate image variations conditioned on brain-derived CLIP embeddings.

2.3. Optimal Visual Stimulus Generation

Previous studies have successfully used gradients from DNN-based brain encoders to produce stimuli that maximally activate parts of the macaque and mouse visual cortex (Bashivan et al., 2019; Walker et al., 2019; Ponce et al., 2019). Later approaches steered random noise vectors within generative models using encoder gradients to synthesize optimal stimuli for category-selective visual regions in the human brain: Gu et al. (2022) and Ratan Murty et al. (2021) used GANs (Goodfellow et al., 2014), while BrainDiVE (Luo et al., 2023) improved stimulus quality and semantic specificity by using diffusion models and a CLIP-based brain encoder. Because the noise vectors are randomly sampled for each synthesized image, this process leads to a varied stimulus set that shares some characteristics (i.e., those preferred by the region) and differs in others (i.e., those randomly produced by the generative model). This introduces the need for human behavioral studies to interpret large image sets to disentangle these features. In contrast, our method of brain-targeted image variation ensures a point of comparison for each synthesized stimulus, directly disentangling the effect of brain activity optimality and opening up the possibility of quantifying semantic and mid-level image features relevant to the targeted cortical region using computer vision techniques. Concurrent work by Prince et al. (2024) explores the accentuation of features in an image to modulate brain activations; however, they edit images in pixel space through brain encoder gradients while we leverage generative models and interpolation in CLIP's latent space to modify the entire image's semantics without the need for gradient-based optimization.

3. Methods

Given a reference image $\mathcal{I}$, we aim to produce variations highlighting semantic selectivity properties of a target cortical region. First, we explain how to condition diffusion models on a brain-derived signal to synthesize variations that drive or suppress predicted activations (Figure 2). Then, we describe how to quantify differences in semantic and mid-level image features to identify properties preferred by each region.

3.1. Brain-Targeted Image Variation

CLIP's semantically rich image embeddings have displayed high representational similarity to the higher visual cortex (Conwell et al., 2023; Wang et al., 2023), making them a suitable choice for representing and manipulating semantic content in the original image $\mathcal{I}$. Specifically, we move the image embedding $\boldsymbol{z}_{\mathcal{I}} = \text{CLIP}_{\text{img}}(\mathcal{I})$ towards optimal endpoints that we derive from paired images and fMRI recordings. We refer to these endpoints as *modulation embeddings*.

First, similarly to BrainDiVE (Luo et al., 2023), we fit a brain encoder $f : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}$ that transforms images $\mathcal{I}$ into brain activations y, where the latter are single values representing the average of voxel-wise beta values belonging to the region of interest (ROI)[1]. The brain encoder consists of two parts. The first is a frozen CLIP image encoder that outputs D-dimensional vectors. The second is a linear

[1]While we use ROI-wise averaged brain responses, this method could be straightforwardly adapted to smaller cortical regions or even single voxels.

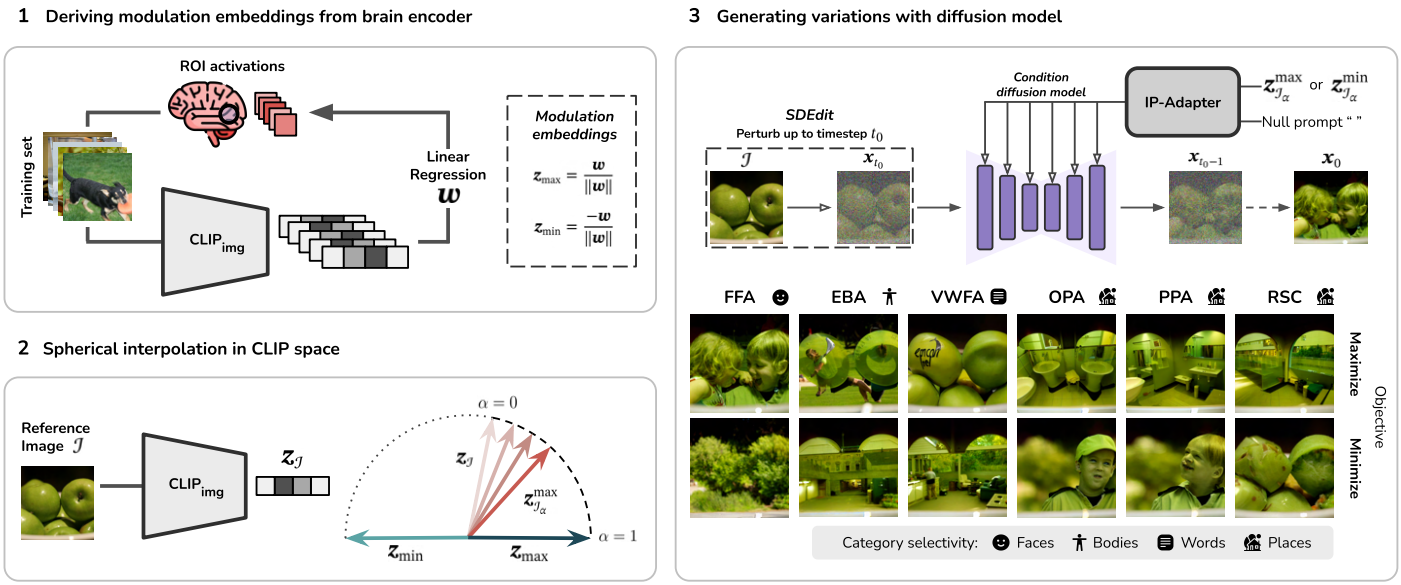

Figure 2. Brain-targeted image variation pipeline. **(1)** Modulation embeddings are derived from a CLIP-based brain encoder. **(2)** For a reference image, we produce intermediate embeddings using spherical interpolation in CLIP space. **(3)** An IP-Adapter conditions a pretrained diffusion model on the intermediate embeddings to generate images that maximize or minimize activity in category-selective ROIs; SDEdit is used to obtain the initial diffusion latents, helping to retain low-level structural similarity to the reference image. Reference image $\mathcal{I}$ belongs to the MS-COCO dataset (Lin et al., 2014).

regression model on normalized CLIP embeddings:

$$\left[\frac{\text{CLIP}_{\text{img}}(\mathcal{I})}{\|\text{CLIP}_{\text{img}}(\mathcal{I})\|} \cdot \boldsymbol{w} + b\right] \Rightarrow y. \tag{1}$$

Due to the linear relationship between normalized embeddings and activations, $\boldsymbol{w} \in \mathbb{R}^D$ can be thought of as a vector in CLIP space that points in the direction of maximal activity for an ROI. Likewise, the negated weights $-\boldsymbol{w}$ point in the direction that minimizes it. Therefore, we define two ROI-specific modulation embeddings, $\boldsymbol{z}_{\text{max}}$ and $\boldsymbol{z}_{\text{min}}$, through the unit-norm weight vector:

$$\boldsymbol{z}_{\text{max}} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \quad \boldsymbol{z}_{\text{min}} = \frac{-\boldsymbol{w}}{\|\boldsymbol{w}\|}. \tag{2}$$

Luo et al. (2024) explain how to close the modality gap between CLIP embeddings of natural images and $\boldsymbol{z}_{\text{max}}$. First, for each image M_i in a set of K natural images $\mathbb{M} = \{M_1, M_2, \cdots, M_K\}$, a softmax score with temperature τ is computed through

$$\text{score}_i = \frac{\exp(S_{\cos}(\boldsymbol{z}_{\text{max}}, e_i)/\tau)}{\sum_{k=1}^{K} \exp(S_{\cos}(\boldsymbol{z}_{\text{max}}, e_k)/\tau)}, \tag{3}$$

where $e_i = \text{CLIP}_{\text{img}}(M_i)$ and $S_{\cos}(\cdot,\cdot)$ is the cosine similarity function. Then, $\boldsymbol{z}_{\text{max}}$ is projected to the space of CLIP embeddings for natural images through a decoupled weighted sum of the image embeddings:

$$\boldsymbol{z}^{\text{proj}}_{\text{max}} = \left(\sum_{k=1}^{K} \text{score}_k \cdot \|e_k\|\right) \cdot \left(\sum_{k=1}^{K} \text{score}_k \cdot \frac{e_k}{\|e_k\|}\right). \tag{4}$$

A similar procedure can be followed for $\boldsymbol{z}_{\text{min}}$. In the following, we assume both $\boldsymbol{z}_{\text{max}}$ and $\boldsymbol{z}_{\text{min}}$ are projected unless otherwise stated.

Next, we produce intermediate embeddings $\boldsymbol{z}^{\text{max}}_{\mathcal{I}_\alpha}$ using spherical interpolation:

$$\boldsymbol{z}^{\text{max}}_{\mathcal{I}_\alpha} = \frac{\sin((1-\alpha)\cdot\theta)}{\sin(\theta)}\boldsymbol{z}_{\mathcal{I}} + \frac{\sin(\alpha\cdot\theta)}{\sin(\theta)}\boldsymbol{z}'_{\text{max}}, \tag{5}$$

where $\theta = \cos^{-1}(\frac{\boldsymbol{z}_{\mathcal{I}} \cdot \boldsymbol{z}'_{\text{max}}}{\|\boldsymbol{z}_{\mathcal{I}}\| \cdot \|\boldsymbol{z}'_{\text{max}}\|})$ is the angle between the vectors, $\alpha \in [0,1]$ indicates the extent of rotation, and $\boldsymbol{z}'_{\text{max}} = \|\boldsymbol{z}_{\mathcal{I}}\| \cdot \boldsymbol{z}_{\text{max}}$. Larger values of α are thus expected to increase activations in the target ROI. An analogous operation with $\boldsymbol{z}_{\text{min}}$ yields $\boldsymbol{z}^{\text{min}}_{\mathcal{I}_\alpha}$.

Finally, we perform guided image synthesis with Stable Diffusion (Rombach et al., 2022) to generate the image variations $\mathcal{I}_\alpha$. To incorporate the semantic information from the modulation embeddings, we use an IP-Adapter (Ye et al., 2023) to prompt the diffusion model with $\boldsymbol{z}^{\text{max}}_{\mathcal{I}_\alpha}$ or $\boldsymbol{z}^{\text{min}}_{\mathcal{I}_\alpha}$ (skipping the adapter's prepended image encoder). To retain fidelity to $\mathcal{I}$ to use it as a point of comparison, we obtain the initial diffusion latents $\boldsymbol{x}_{t_0}$ through SDEdit with $t_0 = \gamma \cdot T$, where T is the total number of denoising steps and $\gamma \in [0,1]$. The hyperparameters α and γ create a trade-off between semantic variation and structural control (subsection 4.4).

3.2. Quantifying information in activity-maximizing and minimizing images

We identify the effect of brain optimization in the image variations $\mathcal{I}_\alpha$ by quantifying differences in category presence and mid-level image features with respect to the reference $\mathcal{I}$. We focus on 16 categories based on previous research on cortical representation and behavioral judgments (Huth et al., 2012; King et al., 2019; Hebart et al., 2020): *faces*, *hands*, *feet*, *people*, *animals*, *plants*, *food*, *furniture*, *tools*, *clothing*, *electronics*, *vehicles*, *landscapes*, *buildings*, *rooms*, and *text*. For each category, we build a representation embedding with CLIP. A challenge in doing so is that single-word descriptions are typically insufficient to capture all possible category instances. Therefore, we build the embeddings using an overcomplete set of concrete nouns from WordNet (Miller, 1995) classified by a large language model (details in appendix subsection A.1). Hence, we measure category presence through cosine similarity between an image embedding and the category's embedding.

Further, we focus on four mid-level image-computable features: *entropy*, the minimum number of bits needed to encode the gray level distribution in a local neighborhood, as a loose quantification of texture; *metric depth*, estimated with the ZoeDepth network (Bhat et al., 2023); *Gaussian curvature* $K = k_1 \cdot k_2$, where principal curvatures k_1, k_2 are estimated with the XTC network (Zamir et al., 2020); and *surface normals*, computed with the XTC network.

4. Results

4.1. Setup

We use the Natural Scenes Dataset (NSD) (Allen et al., 2022), a large dataset of whole-brain high-resolution fMRI responses from eight

human subjects. Each subject viewed ~10,000 natural images from the MS COCO dataset (Lin et al., 2014) repeated three times across multiple scanning sessions. The fMRI beta values are z-scored within their original session and averaged across repetitions. Following standard practice in fMRI encoding (van Gerven, 2017; Naselaris et al., 2011), we split the data into a shared test set consisting of the 1,000 images seen by all subjects and a subject-specific training set with the remaining images. The training sets are used to analytically derive weights $\boldsymbol{w}$ during modulation embedding derivation, as well as for embedding projection (using $\tau = 0.01$). We use the functional localizer masks included in NSD (thresholding at $t > 5$) to define cortical regions of interest.

We employ a pretrained Stable Diffusion model (`stable-diffusion-v1-5`) (Rombach et al., 2022) for guided image synthesis and a pretrained IP-Adapter (`ip-adapter_sd15`) (Ye et al., 2023) for image embedding conditioning. For consistency with these models, we use OpenCLIP's `ViT-H/14` CLIP architecture with `LAION2B-S32B-B79K` pretrained weights (Radford et al., 2021; Ilharco et al., 2021; Schuhmann et al., 2022).

We use a separate brain encoder to predict activations in our experiments. Its architecture consists of DINOv2 (Oquab et al., 2023) as a feature extractor, followed by an ensemble of single-layer vision transformers (ViTs) (Dosovitskiy et al., 2021a) and multilayer perceptrons, inspired by Adeli et al. (2023). To ensure the robustness of our method, we employ an additional feature-weighted receptive field encoder (St-Yves and Naselaris, 2018; Allen et al., 2022) (available through the Neural Encoding Dataset (Gifford and Cichy, 2024)) in our validation procedure. We refer to these as *DINO-ViT encoder* and *fwRF encoder*, respectively. Details on architecture and prediction performance can be found in the appendix subsection A.2. Importantly, neither encoder is CLIP-based; hence, they do not share the same latent space as the encoders used to derive the modulation embeddings.

4.2. Modulating activity in brain regions of interest

We validate BrainACTIV by targeting six previously identified regions of interest in the higher visual cortex: fusiform face area (FFA), extrastriate body area (EBA), visual word form area (VWFA), occipital place area (OPA), parahippocampal place area (PPA), and retrosplenial cortex (RSC) (precise location can be found in Allen et al. (2022)). First, we identify six mutually exclusive subsets of images in NSD that share broadly similar semantic contexts: *wild animals*, *birds*, *vehicles*, *people in sports*, *food*, and *furniture*. We define each subset by filtering the pixel-wise category annotations made available with COCO (appendix subsection A.3). These subsets are employed to enforce the diversity of image selection in our experiment.

For each subject and each ROI, we select the 20 test images from each subset with measured responses closest to baseline activation (i.e., the average ROI activation across all test images). Because initial experiments showed that modulation embeddings are highly similar across subjects, we opt to average the embeddings over all subjects, excluding the subject on which predictions are made. Hence, we modulate brain activity in each subject through a signal derived exclusively from the rest of the subjects' data. We manipulate each of the 120 selected test images with interpolation values $\alpha \in \{0.1, 0.2, ..., 0.9, 1\}$, producing 20 variations in total for each image. For SDEdit, we use an exponential warm-up schedule for γ up to a value of $\gamma = 0.6$ for $\alpha = 1$. Next, we predict activations for each of them using the appropriate subject- and ROI-specific DINO-ViT encoder and fwRF encoder. To compute the predicted difference in activation for each variation, we subtract the prediction of the reference image.

Figure 3 (A) displays example variations for each ROI, along with the predicted differences in activation. Note that the effect of α on the magnitude of these differences varies for each image, as an effect of its features and the ROI's sensitivity to these. To study a region's selectivity, we look for features that consistently appear or disappear

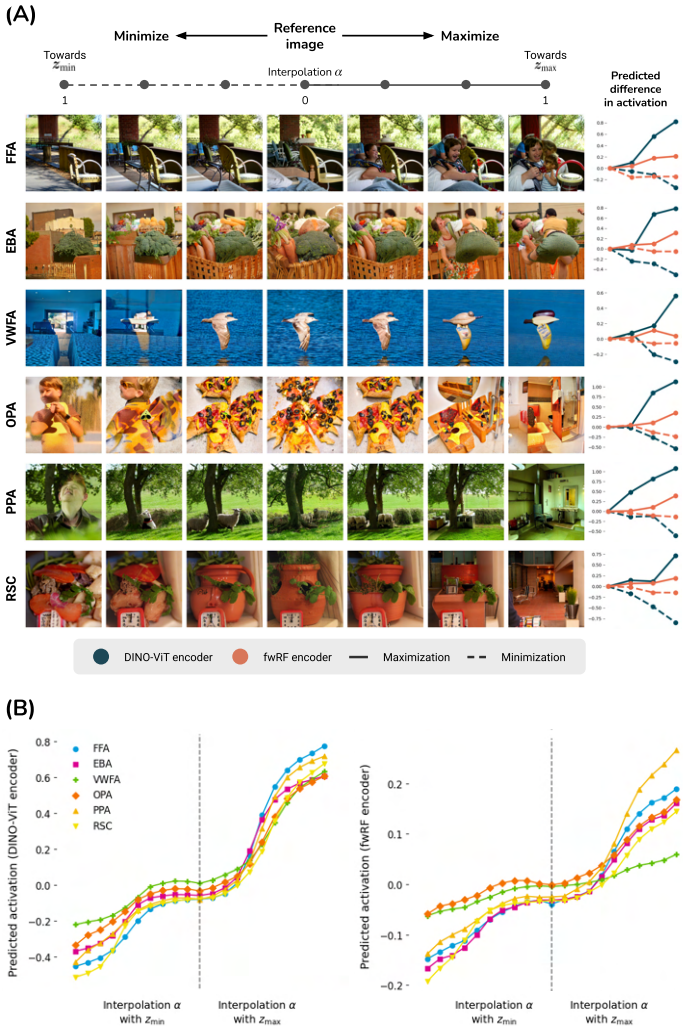

Figure 3. Example variations and modulation results. **(A)** Example image variations show the effect of activation maximization and minimization in each ROI; line plots show predicted differences in activation with respect to the reference image. Reference images belong to the MS-COCO dataset (Lin et al., 2014). **(B)** ROI activations predicted by DINO-ViT encoder (top) and fwRF encoder (bottom) as a function of interpolation α with each modulation embedding, averaged across test images and subjects.

over a wide range of contexts. Thus, our analyses focus on the general effect of BrainACTIV over the whole selection of test images. Additional examples can be found in the appendix subsection A.4.

First, we verify that the reference images serve as a reliable point of comparison by measuring fidelity to low-level structure in our variations. Following Meng et al. (2022), we compute image L_2 distance and LPIPS (Zhang et al., 2018) between reference and variations, averaged over images and subjects. As a baseline, we compute these metrics between 1000 random pairs in the test set. **??** shows that structural similarity is preserved on maximization and minimization ($\alpha = 1$) for all ROIs.

Next, we look at the effect of our variations on the DINO-ViT and fwRF encoder outputs to verify that BrainACTIV successfully increases and decreases predicted ROI responses. Figure 3 (B) shows these predictions as a function of α for all ROIs (averaged over images and subjects). We observe a stable increase and decrease across ROIs for both encoders, confirming that our method modulates predicted activations. The plots show an expected lag in activity increase/decrease up to $\alpha \approx 0.4$ due to our γ schedule since we intended the initial variations to be close to the reference image.

4.3. Quantifying visuo-semantic changes in image variations

In this section, we verify that the category selectivity suggested by BrainACTIV for each ROI agrees with established neuroscientific

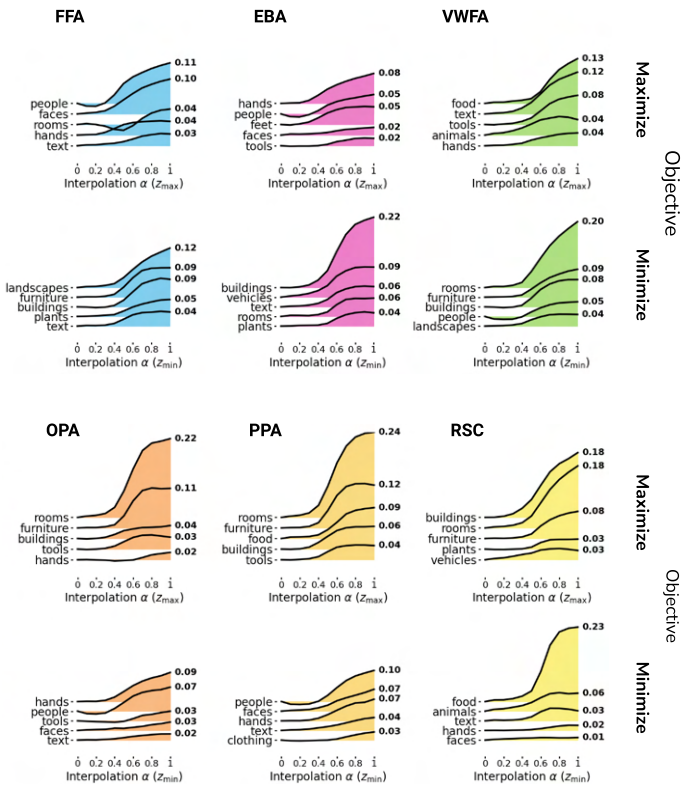

Figure 4. Top categories for each region. Each plot displays the difference in category presence (see subsection 3.2) with respect to the reference image as a function of α. Top-5 categories per ROI for maximization (top) and minimization objective (bottom) are ranked by the highest measured difference. Results agree with hypothesized preferred categories.

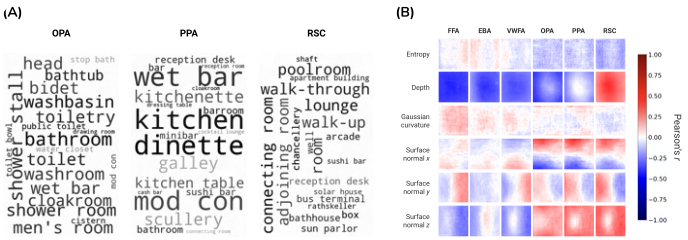

Figure 5. Scene-selective nouns and mid-level features. **(A)** Top nouns emphasize differences between scene-selective ROIs. **(B)** Correlation between predicted ROI activation differences and mid-level feature differences at each location in an image.

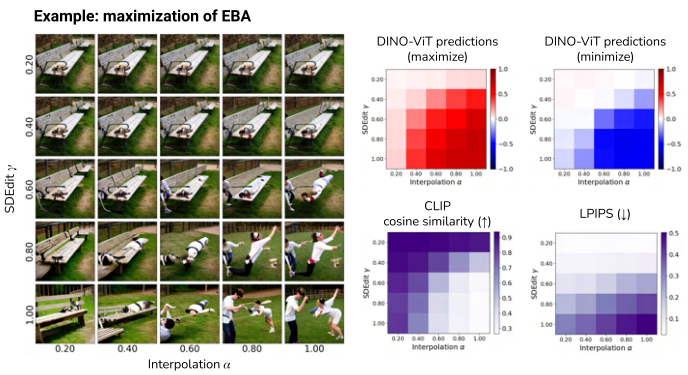

Figure 6. Effect of α and γ. (Left) Example variations for different values of α and γ display a trade-off between structural fidelity and semantic variation. (Right) Predicted activations are more strongly modulated when semantic content is closer to $\boldsymbol{z}_{\text{max}}$ or $\boldsymbol{z}_{\text{min}}$; lower α and γ result in higher semantic similarity and structural fidelity to the reference image. Reference image belongs to the MS-COCO dataset (Lin et al., 2014).

findings. To this end, we look at the categories whose presence is increased when activations are maximized (Figure 4, top row). FFA increases the presence of faces, agreeing with Kanwisher et al. (1997) and McCarthy et al. (1997); EBA increases body parts (Downing et al., 2001); VWFA increases text (McCandliss et al., 2003) and food, potentially suggesting a preference for text on small objects [2]; OPA, PPA, and RSC increase manmade structures/scenes (Kamps et al., 2016; Epstein and Kanwisher, 1998; Mitchell et al., 2018). Our results also emphasize differences between scene-selective regions, evidenced by the increased presence of individual WordNet nouns (Figure 5 (A)): OPA prefers local scene elements, while PPA and RSC prefer more global views (Kamps et al., 2016; Henderson et al., 2008); additionally, RSC prefers corridor-like scenes, potentially related to its suggested role in navigation (Mitchell et al., 2018). Moreover, category presence increased during activation minimization (Figure 4, bottom row) suggests that FFA, EBA, and VWFA respond minimally for scenes/structures (particularly for FFA, landscapes), while OPA and PPA respond minimally to people, and RSC to food.

To investigate finer-grained differences between ROIs, we look at the correlation between predicted activation differences and mid-level feature differences for different locations in an image (Figure 5 (B)). Results suggest an important role of surface orientation in differentiating scene- and object-selective regions, with FFA, EBA, and VWFA preferring surfaces pointing outwards and OPA, PPA, and RSC preferring surfaces pointing inward. Moreover, the enhanced correlations between predicted ROI activitation differences and surface normals in scene-selective regions is consistent with reported sensitivity of scene-selective regions to 3D configurations (Lescroart and Gallant, 2019). In addition, depth correlations further emphasize differences between scene-selective regions, with RSC showing higher correlation with deeper depth values. Together with section 4.2, these results demonstrate the validity of BrainACTIV as a data-driven method for fine-grained and reliable hypothesis formulation in visual cortical regions.

4.4. Balancing semantic variation and structural control

Two hyperparameters–interpolation α and SDEdit γ—jointly control a trade-off between manipulating semantic content in the reference image and preserving its low-level structure. Because one of these properties might be preferred when designing hypothesis-driven experiments, we briefly illustrate their role and the effect on the predicted ROI activations. Figure 6 (left) shows example variations for different combinations of α and γ. Cosine similarity values between CLIP image embeddings of the reference image and those of variations show that lower α and lower γ lead to higher semantic similarity to the reference image (Figure 6 (right), averaged over images, ROIs, and subjects). Similarly, lower α and γ lead to higher structural fidelity to the reference image (here measured with LPIPS). At the same time, we observe that lower semantic similarity and lower structural fidelity result in higher changes in predicted activations. This is to be expected from the design of our modulation embeddings. However, two distinct alternatives exist: choosing lower α and higher γ results in variations that mostly retain the semantic content of the reference image while the spatial arrangement differs (depending on how well CLIP can capture it). Conversely, higher α and lower γ favor the low-level structure of the reference image while differing in semantic content.

4.5. Accentuating differences between similar ROIs

As illustrated in Figure 5, regions that are selective for the same broad semantic category can differ in their sensitivity to visual stimulus features. Identifying what distinguishes one region from another

[2] It is important to note that the projection of modulation embeddings to the space of CLIP image embeddings necessarily biases the representativity of particular features towards objects in the projection set that most frequently hold these features (e.g., small size → food).

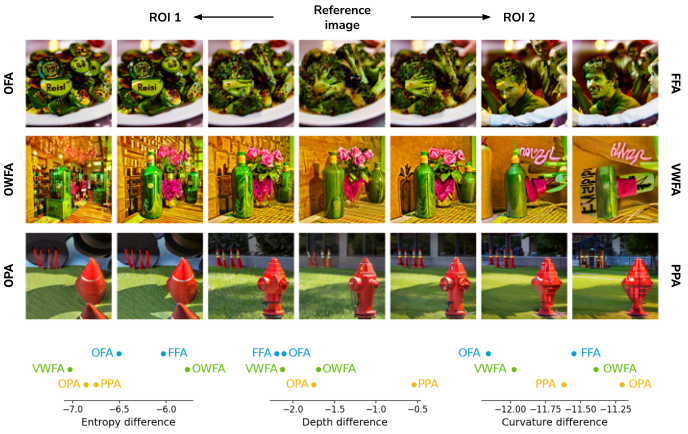

Figure 7. Accentuation of pairwise differences. (Top) Example variations accentuating differences for three ROI pairs: OFA-FFA, OWFA-VWFA, OPA-PPA. (Bottom) Differences in mid-level feature values (with respect to reference image) of each ROI, averaged and z-scored over test images. Reference images belong to the MS-COCO dataset (Lin et al., 2014).

beyond category selectivity is an important step toward understanding broader functional organization principles in the visual cortex. This section demonstrates how BrainACTIV can accentuate these differences by manipulating a reference image. We target three pairs of ROIs that are selective to the same category: (1) occipital face area (OFA) and fusiform face area (FFA), (2) occipital word form area (OWFA) and visual word form area (VWFA), and (3) occipital place area (OPA) and parahippocampal place area (PPA).

For each pair, we create *accentuation embeddings* by subtracting the modulation embeddings of each ROI from one another. We randomly sample 50 images from the test set and manipulate each of them with interpolation values $\alpha \in \{0.1, 0.2, ..., 0.9, 1\}$ toward the accentuation embeddings. For SDEdit, we use an exponential warm-up schedule up to $\gamma = 0.6$ for $\alpha = 1$. Figure 7 displays example variations for each ROI pair and measured differences in entropy, depth, and curvature. The features accentuated on each side represent preferred visual properties that distinguish the regions.

We observe a higher preference for text in OFA and a higher preference for faces in FFA, despite both of them being face-selective. For OWFA, we identify a higher preference for cluttered coarse-grained elements, evidenced by higher entropy values; VWFA shows a higher preference for text on small items. Finally, OPA and PPA differ in their sensitivity to depth as analyzed in subsection 4.2. While these differences are suggestive, they will need to be validated in future experiments: for example, images such as these could be used as experimental stimuli to test theories about hierarchical relations between regions as for example proposed for faces by Pitcher et al. (2011).

5. Discussion

We introduced BrainACTIV, a method for modulating predicted brain responses through image manipulation. To our knowledge, we are the first work to use generative models—particularly, diffusion models—to manipulate reference images with the goal of maximizing or minimizing activations in the human visual cortex. Our results show the potential of our approach for fine-grained and reliable identification of visuo-semantic properties preferred by specialized neural populations. This information can be used to formulate new hypotheses about visual representations in the brain.

For example, our generative framework can be employed by neuroscientists to design precisely controlled and innovative experimental paradigms, for example to disambiguate the role of low-, mid- and high-level features, whose inherent correlations in natural images complicates the ability to isolate their effect on brain responses (Malcolm et al., 2016; Lescroart et al., 2015). While we here demonstrate our approach on brain regions with known category-selectivity, we envision that BrainACTIV may help interpret so-called 'no-mans land' regions of visual cortex (Bao et al., 2020) whose functional specialization and representations are less well understood. Finally, we believe that BrainACTIV holds potential to enhance current GAN-based approaches for exploration of selectivity using closed-loop paradigms in which activations are continuously updated along an optimization trajectory (e.g., Ponce et al. (2019) with more photorealistic stimuli.

Our method has some limitations. First, because we employ pretrained models for image synthesis and representation, our results are subjected to biases in their training data. These biases might over-represent certain categories through correlations with specific lower-level image features, producing misleading results. We encourage future work to use fine-tuned models and domain-specific representation spaces to explore finer-grained selectivity within smaller specialized cortical regions. Second, our work relies on brain encoders to validate the effective modulation of brain activity. While we have taken measures to ensure the robustness of our results, future work should test the effect of BrainACTIV inside an fMRI scanner to validate the predicted activations against novel brain responses. In summary, BrainACTIV unlocks the possibility to test existing and generate novel hypotheses about neural representations in visual cortex using brain-guided image diffusion with structural control.

References

Adeli, H., Minni, S., and Kriegeskorte, N. (2023). Predicting brain activity using transformers. *bioRxiv preprint*.

Allen, E. J. et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25:116–126.

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.

Bashivan, P., Kar, K., and DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364.

Bhat, S. F. et al. (2023). Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint*.

Bracci, S., Ritchie, J. B., and de Beeck, H. O. (2017). On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia*, 105:153–164.

Conwell, C. et al. (2023). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv preprint*.

Dosovitskiy, A. et al. (2021a). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.

Dosovitskiy, A. et al. (2021b). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

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

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

Gifford, A. T. and Cichy, R. M. (2024). The neural encoding dataset. *In preparation*.

Goetschalckx, L., Andonian, A., Oliva, A., and Isola, P. (2019). Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 5744–5753.

Goodfellow, I. et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:139–144.

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.

Groen, I. I., Silson, E. H., and Baker, C. I. (2017). Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714):20160102.

Gu, Z. et al. (2022). NeuroGen: Activation optimized image synthesis for discovery neuroscience. *NeuroImage*, 247:118812.

Hebart, M. H., Zheng, C. Y., Pereira, F., and Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4:1173–1185.

Henderson, J. M., Larson, C. L., and Zhu, D. C. (2008). Full scenes produce more activation than close-up scenes and scene-diagnostic objects in parahippocampal and retrosplenial cortex: an fmri study. *Brain and Cognition*, 66:40–49.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

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

Ilharco, G. et al. (2021). OpenCLIP. *Zenodo (software package)*.

Janini, D., Hamblin, C., Deza, A., and Konkle, T. (2022). General object-based features account for letter perception. *PLoS computational biology*, 18(9):e1010522.

Kamps, F. S. et al. (2016). The occipital place area represents the local elements of scenes. *NeuroImage*, 132:417–424.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: A module inhuman extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17:4302–4311.

Khosla, M. et al. (2021). Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Science Advances*, 7.

King, M. L. et al. (2019). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, 197:368–382.

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

Lescroart, M. D., Stansbury, D. E., and Gallant, J. L. (2015). Fourier power, subjective distance, and object categories all provide plausible models of bold responses in scene-selective visual areas. *Frontiers in computational neuroscience*, 9:135.

Lewis, M. et al. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Lin, T.-Y. et al. (2014). Microsoft COCO: Common objects in context. *European Conference on Computer Vision*, pages 740–755.

Luo, A. F., Henderson, M. M., Tarr, M. J., and Wehbe, L. (2024). BrainSCUBA: Fine-grained natural language captions of visual cortex selectivity. *International Conference on Learning Representations (ICLR)*.

Luo, A. F., Henderson, M. M., Wehbe, L., and Tarr, M. J. (2023). Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *37th Conference on Neural Information Processing Systems (NeurIPS)*.

Malcolm, G. L., Groen, I. I., and Baker, C. I. (2016). Making sense of real-world scenes. *Trends in cognitive sciences*, 20(11):843–856.

McCandliss, B. D., Cohen, L., and Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, 7:293–299.

McCarthy, G., Puce, A., Gore, J. C., and Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience*, 9:605–610.

Meng, C. et al. (2022). SDEdit: Guided image synthesis and editing with stochastic differential equations. *International Conference on Learning Representations (ICLR)*.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Mitchell, A. S. et al. (2018). Retrosplenial cortex and its role in spatial cognition. *Brain and Neuroscience Advances*, 2:1–13.

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

Nichol, A. et al. (2022). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *International Conference on Machine Learning*.

Oquab, M. et al. (2023). DINOv2: Learning robust visual features without supervision. *arXiv preprint, arXiv:2304.07193*.

Papale, P., De Luca, D., and Roelfsema, P. R. (2024). Deep generative networks reveal the tuning of neurons in it and predict their influence on visual perception. *bioRxiv*, pages 2024–10.

Peelen, M. V. and Downing, P. E. (2005). Selectivity for the human body in the fusiform gyrus. *Journal of Neurophysiology*, 93:603–608.

Pierzchlewicz, P. A. et al. (2023). Energy guided diffusion for generating neurally exciting images. *bioRxiv preprint*.

Pitcher, D., Walsh, V., and Duchaine, B. (2011). The role of the occipital face area in the cortical face perception network. *Experimental brain research*, 209:481–493.

Ponce, C. R. et al. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177:999–1009.

Prince, J. S. et al. (2024). Dissecting visual population codes with brain-guided feature accentuation. *Cognitive Computational Neuroscience Conference*.

Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*.

Ramesh, A. et al. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint, arXiv:2204.06125*.

Ratan Murty, N. A. et al. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature Communications*, 12:55409.

Rombach, R. et al. (2022). High-resolution image synthesis with latent diffusion models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Rombach, R. et al. (2023). Stable unCLIP. *https://huggingface.co/stabilityai/stable-diffusion-2-1-unclip*.

Saharia, C. et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *36th Conference on Neural Information Processing Systems*.

Schuhmann, C. et al. (2022). LAION-5b: An open large-scale dataset for training next generation image-text models. *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Song, Y. et al. (2020). Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*.

St-Yves, G. and Naselaris, T. (2018). The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage*, 180:188–202.

van Gerven, M. A. (2017). A primer on encoding models in sensory neuroscience. *Journal of Mathematical Psychology*, 76:172–183.

Vinken, K., Prince, J. S., Konkle, T., and Livingstone, M. S. (2023). The neural code for "face cells" is not face-specific. *Science Advances*, 9(35):eadg1736.

Walker, E. Y. et al. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, 22:2060–2065.

Wang, A. Y. et al. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5:1415–1426.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Ye, H. et al. (2023). IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint, arXiv:2308.06721*.

Yin, W., Hay, J., and Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923.

Zamir, A. R. et al. (2020). Robust learning through cross-task consistency. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 11197–11206.

Zhang, R. et al. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Computer Vision and Pattern Recognition Conference (CVPR)*.

A. Appendix

A.1. Category Representations in CLIP

We build category-specific CLIP embeddings to quantify the presence of each category (*faces*, *hands*, *feet*, *people*, *animals*, *plants*, *food*, *furniture*, *tools*, *clothing*, *electronics*, *vehicles*, *landscapes*, *buildings*, *rooms*, and *text*) in an image through cosine similarity with the image embedding. First, we gather a large list of $N_{\text{nouns}} = 17,086$ concrete nouns from WordNet (Miller, 1995) by collecting all hyponyms of the following synsets:

- `amphibian.n.03`
- `article.n.02`
- `bird.n.01`
- `body_of_water.n.01`
- `building.n.01`
- `commodity.n.01`
- `correspondence.n.01`
- `external_body_part.n.01`
- `facility.n.01`
- `fish.n.01`
- `food.n.02`
- `geological_formation.n.01`
- `instrumentality.n.03`
- `land.n.04`
- `person.n.01`
- `placental.n.01`
- `plant.n.02`
- `plaything.n.01`
- `publication.n.01`
- `reptile.n.01`
- `room.n.01`
- `sign.n.02`
- `vehicle.n.01`
- `way.n.06`
- `written_record.n.01`

Then, we perform zero-shot classification of each noun into one of the categories using `facebook/bart-large-mnli`, a version of the language model BART (Lewis et al., 2020) trained on the MultiNLI dataset (Williams et al., 2018; Yin et al., 2019). Instead of using the category names as labels, we build custom labels:

- faces: "related to faces, eyes, nose, mouth"
- hands: "related to hands, arms, fingers"
- feet: "related to feet, legs, toes"
- people: "related to people, humans, persons"
- animals: "related to animals, creatures, fauna"
- plants: "related to plants, greenery, flora"
- food: "related to food, meals, eating"
- furniture: "related to furniture, household items"
- tools: "related to tools, equipment, instruments"
- clothing: "related to clothing, textiles, garments"
- electronics: "related to electronics, gadgets, devices"
- vehicles: "related to vehicles, transportation, travel"
- landscapes: "related to natural areas, landscapes, outdoors"
- buildings: "related to urban areas, buildings, structures"
- rooms: "related to indoors, rooms, interiors"
- text: "related to written text, signs"

The resulting class probabilities are gathered in a matrix $\boldsymbol{Y}_{\text{prob}} \in [0,1]^{N_{\text{nouns}} \times 16}$ where each row sums up to 1. We weigh the probabilities by the salience of each category with respect to the rest for each noun to obtain $\boldsymbol{Y}_{\text{sal}}$:

$$[\boldsymbol{Y}_{\text{sal}}]_{i,j} = [\boldsymbol{Y}_{\text{prob}}]_{i,j} \cdot \frac{[\boldsymbol{Y}_{\text{prob}}]_{i,j}}{\sum_k [\boldsymbol{Y}_{\text{prob}}]_{i,k}}.$$

Next, we compute embeddings for each of the nouns using CLIP's text encoder. To make these more robust, we average the embeddings obtained through 18 prompt templates (e.g., *"a photo of a {}."* or *"a good photo of the {}."*) used originally by CLIP for image classification (Radford et al., 2021). We normalize these embeddings and gather them in a matrix $\boldsymbol{Z}_{\text{nouns}} \in \mathbb{R}^{N_{\text{nouns}} \times 1024}$.

Finally, we use a regularized linear regression model on $\boldsymbol{Z}_{\text{nouns}}$ to predict $\boldsymbol{Y}_{\text{sal}}$ and analytically derive the weights $\boldsymbol{W}_{\text{nouns}} \in \mathbb{R}^{D \times 16}$. Each column in the weight matrix then functions as our representation for each category. We notice that the *text* category is difficult to represent through this method; therefore, we instead compute its embedding by encoding the phrase "text on an object" using each of the 18 prompt templates and averaging them. Figure 8 displays each category's representative examples from the Natural Scenes Dataset (NSD) Allen et al. (2022), as well as salient WordNet nouns.

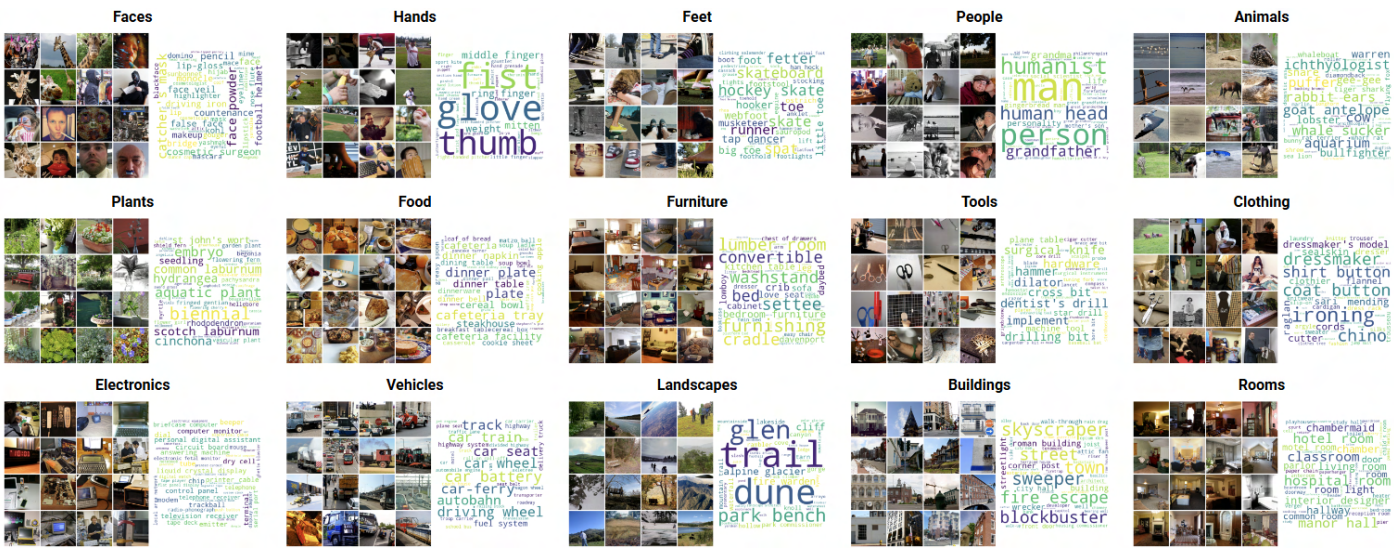

Figure 8. Representative examples (high cosine similarity) from NSD for each category, together with top WordNet nouns as classified by `facebook/bart-large-mnli`. All images belong to the MS-COCO dataset (Lin et al., 2014).

A.2. Brain Encoder Details and Performance

DINO-ViT Encoder

Adeli et al. (2023) explored the use of a pretrained 12-layer DINOv2 model (Oquab et al., 2023) as a feature extractor for a single-layer transformer that learns ROI-specific queries, which are later linearly mapped to voxel activations. They train 22 models per subject, differing on the choice of layer used to extract features from DINOv2 and targeted ROIs. Finally, they employ an ensemble approach to produce their voxel-wise predictions. We simplify this process by fixing a single architecture and training it once for each category-selective ROI (9 models in total per subject). Specifically, we extract the output of each of the 12 layers in DINOv2 and pass each of them through a separate single-layer vision transformer (ViT) (Dosovitskiy et al., 2021b). The output CLS token of each ViT is used by a multilayer perception (MLP) to predict voxel-wise activations for the ROI. The outputs of the 12 MLPs are aggregated through a learnable linear layer to produce our final predictions. All models are trained for 15 epochs with early stopping, using a learning rate of 1e-4 and a batch size of 64 samples.

fwRF Encoder

The Neural Encoding Dataset (NED) (Gifford and Cichy, 2024) provides pretrained brain encoders for the NSD dataset (Allen et al., 2022). These encoders are feature-weighted receptive field encoding models (St-Yves and Naselaris, 2018), neural networks trained end-to-end to predict neural responses.

Encoding performance

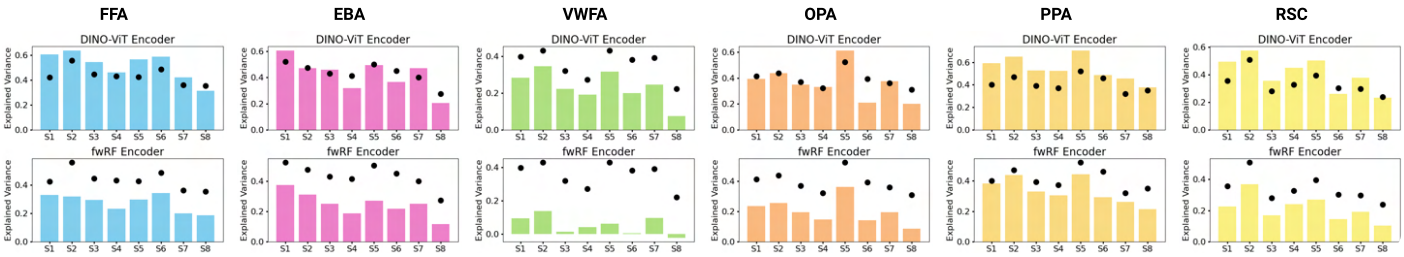

Figure 9. Encoding performance (explained variance) of both encoders for each subject and each region of interest. Black dots indicate estimated noise ceiling from Allen et al. (2022) (averaged over ROI voxels).

A.3. NSD Image Subsets

We identify six mutually exclusive subsets of images in NSD (Figure 10) to enforce diversity in our validation experiment. We filter the pixel-wise category annotations from COCO (Lin et al., 2014) as specified in Figure 11; each row also indicates the size of each subset for training and test sets.

Figure 10. Example images per subset. All images belong to the MS-COCO dataset (Lin et al., 2014).

Subset	Size (training set)	Size (test set)	COCO Categories
Wild animals	773±38	150	**include:** horse, sheep, cow, elephant, bear, zebra, giraffe **exclude:** person, bicycle, car, motorcycle, airplane, bus, train, truck, boat, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake
Birds	125±14	24	**include:** bird **exclude:** person, bicycle, car, motorcycle, airplane, bus, train, truck, boat, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, horse, sheep, cow, elephant, bear, zebra, giraffe, cat, dog
Vehicles	1000±37	123	**include:** bicycle, car, motorcycle, airplane, bus, train, truck, boat **exclude:** person, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, horse, sheep, cow, elephant, bear, zebra, giraffe, bird, cat, dog
People in sports	839±31	101	**include:** person AND frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket **exclude:** horse, sheep, cow, elephant, bear, zebra, giraffe, bird, cat, dog, bicycle, car, motorcycle, airplane, bus, train, truck, boat, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake
Food	509±33	52	**include:** banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake **exclude:** person, horse, sheep, cow, elephant, bear, zebra, giraffe, bird, cat, dog, bicycle, car, motorcycle, airplane, bus, train, truck, boat
Furniture	883±40	108	**include:** chair, couch, potted, bed, toilet **exclude:** person, horse, sheep, cow, elephant, bear, zebra, giraffe, bird, cat, dog, bicycle, car, motorcycle, airplane, bus, train, truck, boat, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, dining table, umbrella

Figure 11. Overview of the size and COCO categories used to define each subset. Upper rows (in the rightmost column) indicate categories present in all images. Bottom rows indicate categories that were explicitly excluded.

A.4. Additional Image Variations

Fusiform Face Area (FFA)

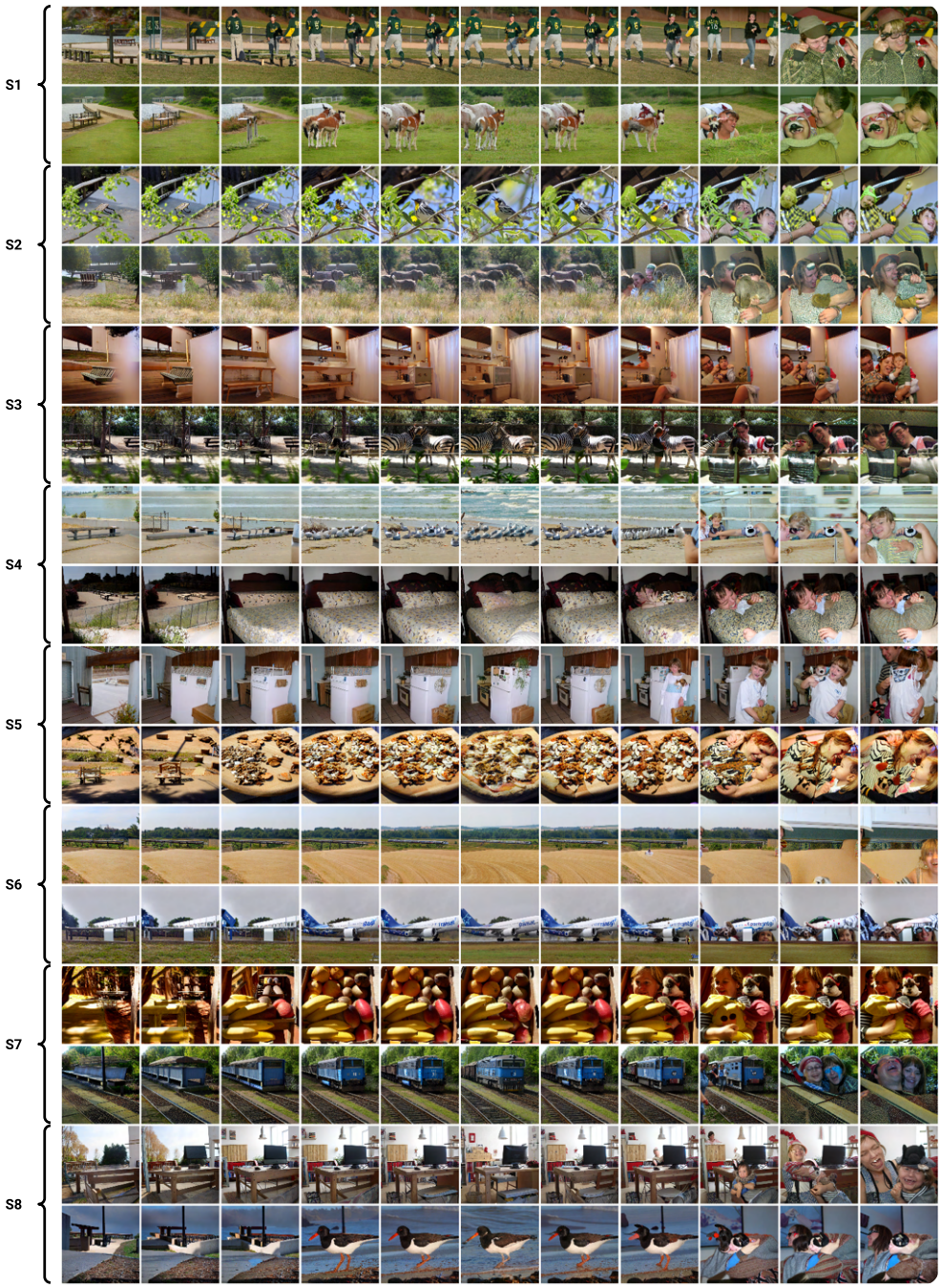

Figure 12. Example image variations (two per subject). Middle column: reference images. Left: minimization of FFA. Right: maximization of FFA. Reference images belong to the MS-COCO dataset (Lin et al., 2014).

Extrastriate Body Area (EBA)

Figure 13. Example image variations (two per subject). Middle column: reference images. Left: minimization of EBA. Right: maximization of EBA. Reference images belong to the MS-COCO dataset (Lin et al., 2014).

Visual Word Form Area (VWFA)

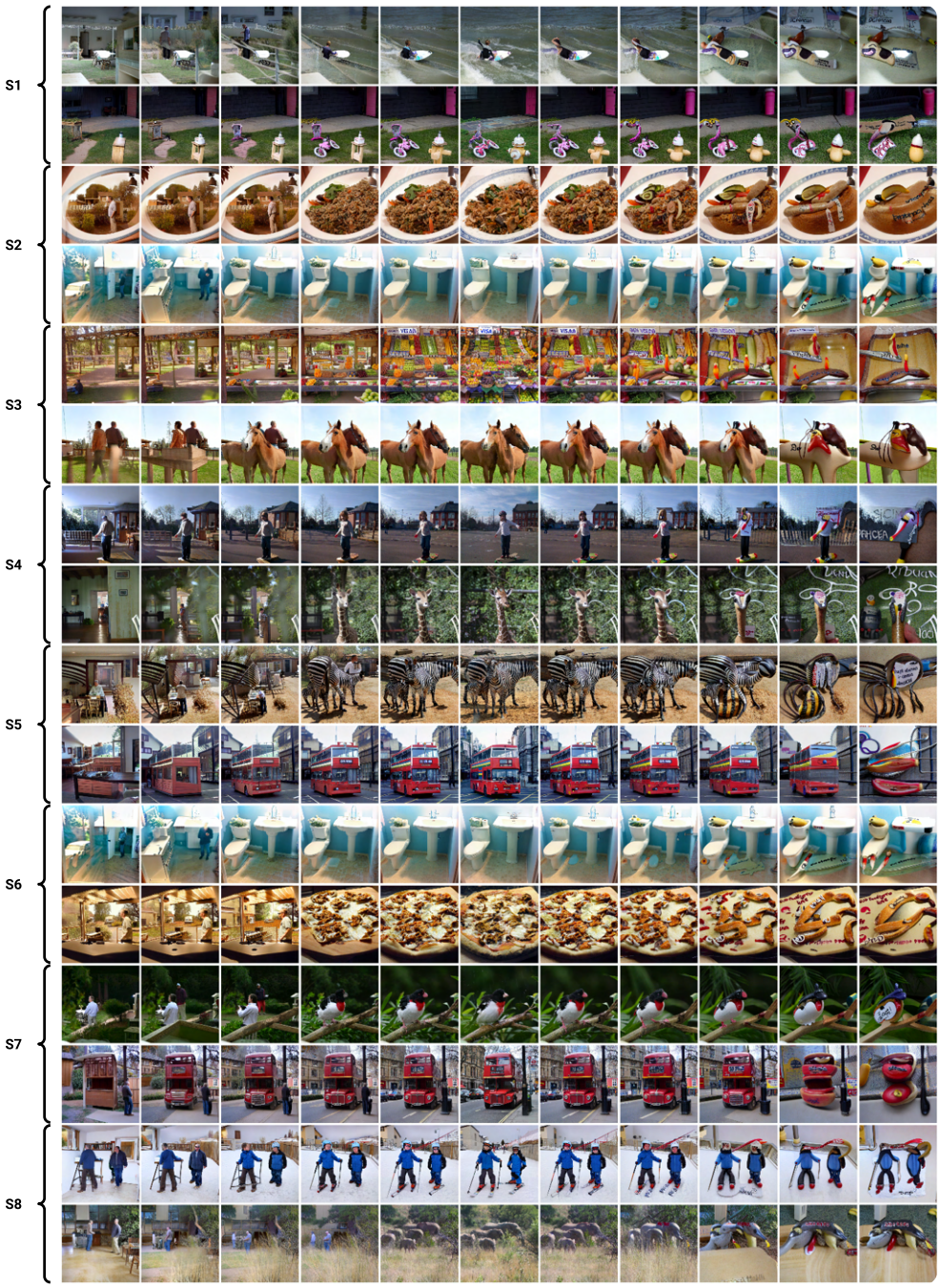

Figure 14. Example image variations (two per subject). Middle column: reference images. Left: minimization of VWFA. Right: maximization of VWFA. Reference images belong to the MS-COCO dataset (Lin et al., 2014).

Occipital Place Area (OPA)

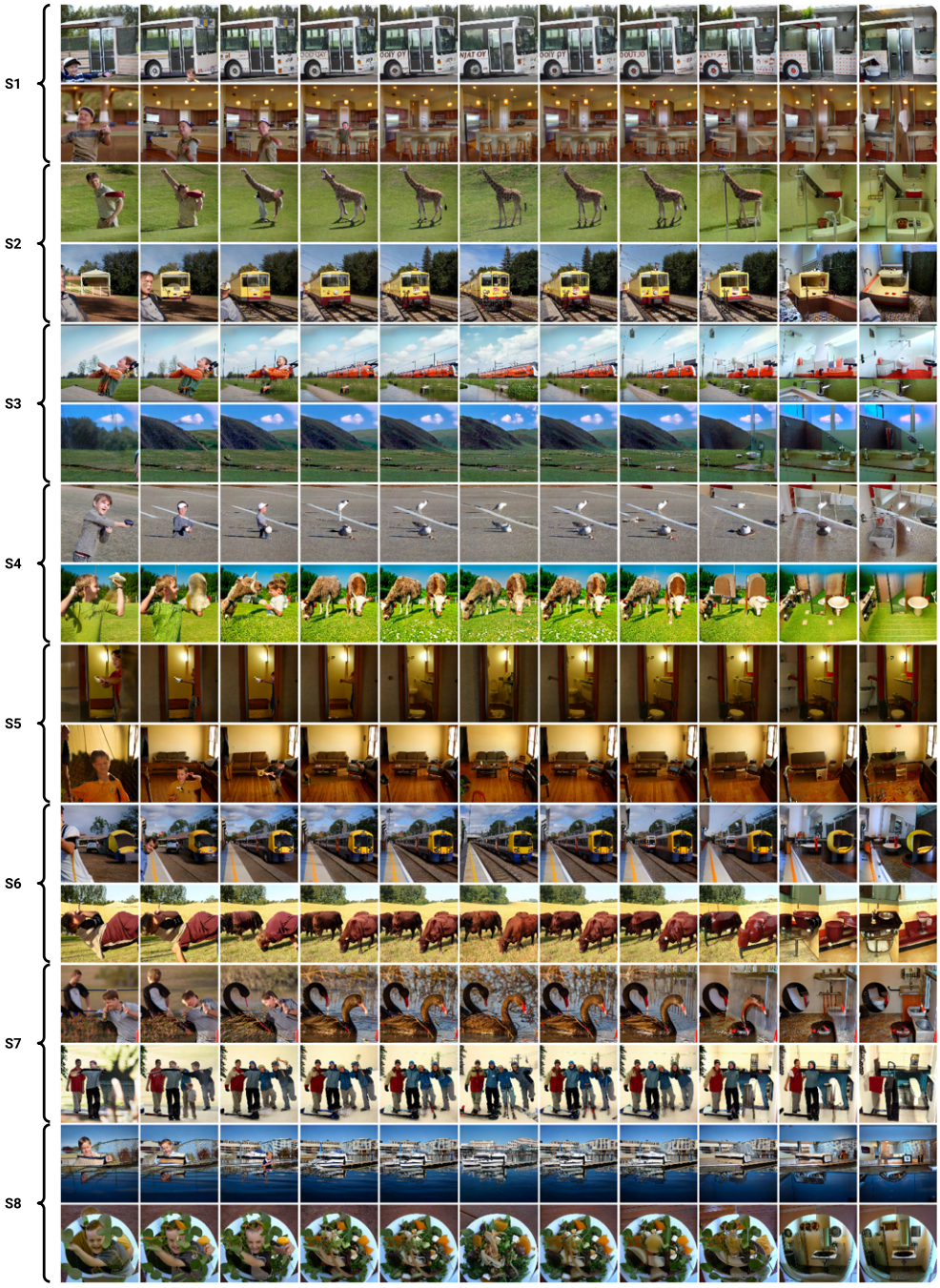

Figure 15. Example image variations (two per subject). Middle column: reference images. Left: minimization of OPA. Right: maximization of OPA. Reference images belong to the MS-COCO dataset (Lin et al., 2014).

Parahippocampal Place Area (PPA)

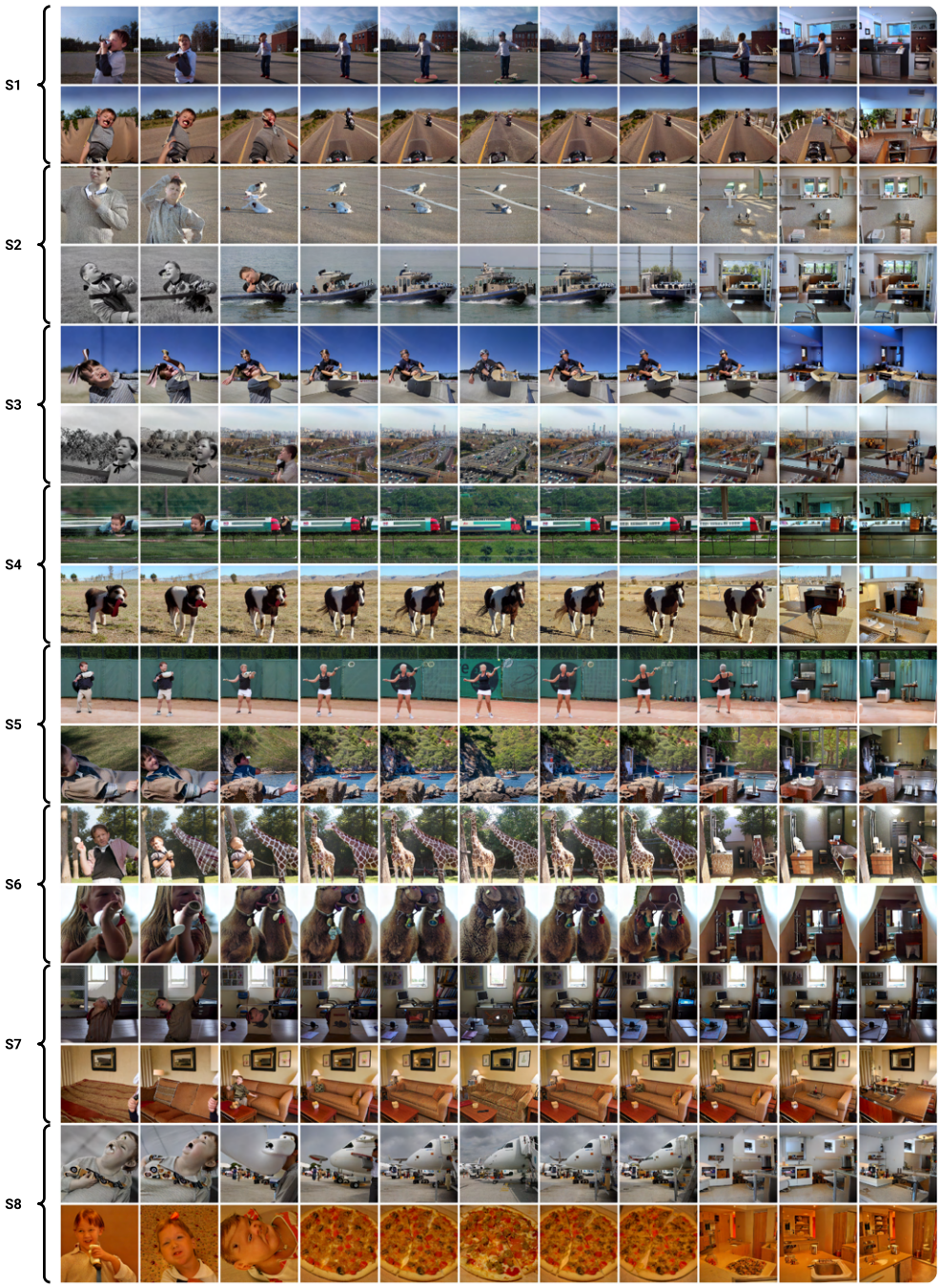

Figure 16. Example image variations (two per subject). Middle column: reference images. Left: minimization of PPA. Right: maximization of PPA. Reference images belong to the MS-COCO dataset (Lin et al., 2014).

Retrosplenial Cortex (RSC)

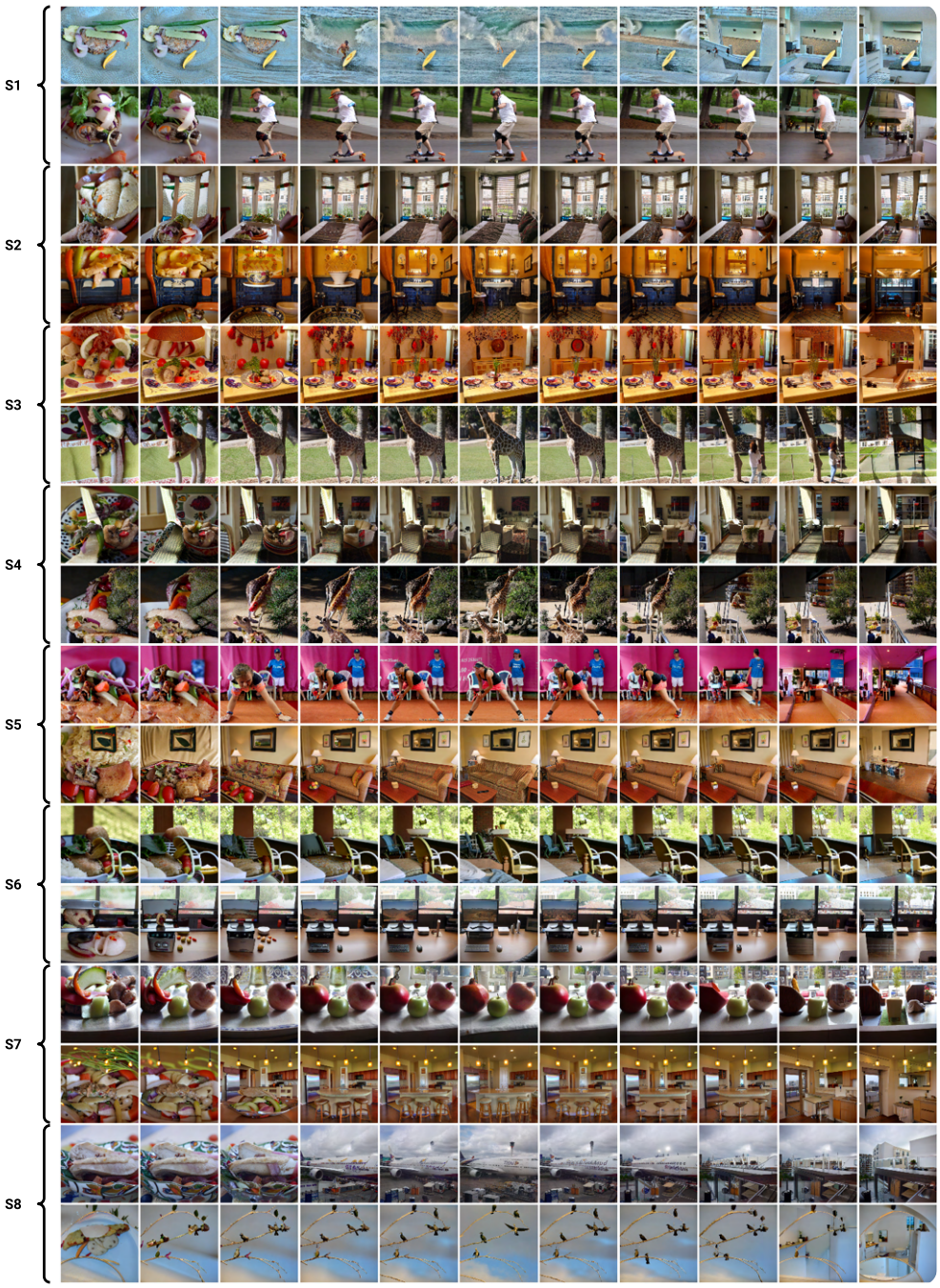

Figure 17. Example image variations (two per subject). Middle column: reference images. Left: minimization of RSC. Right: maximization of RSC. Reference images belong to the MS-COCO dataset (Lin et al., 2014).

A.5. Difference in low-level image features

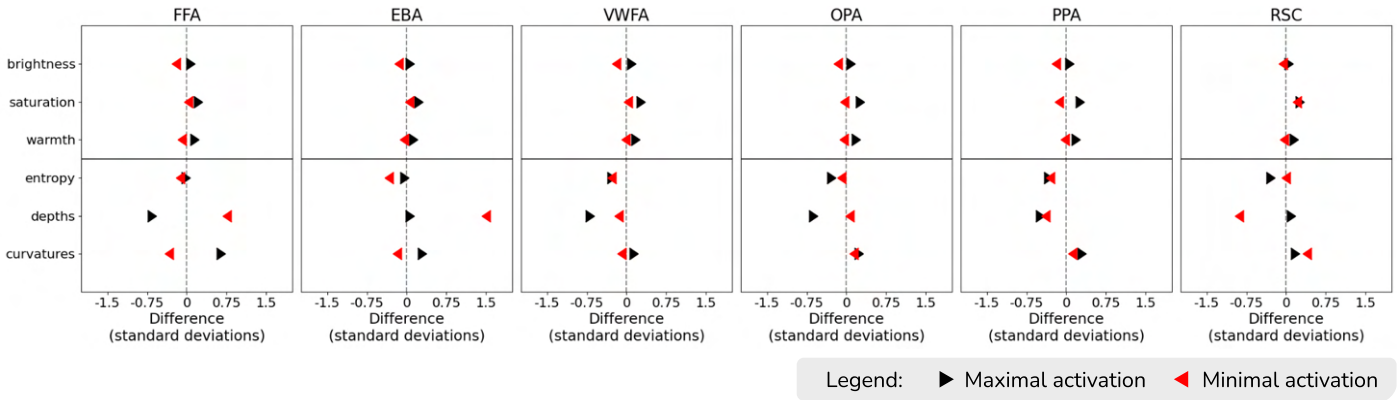

Figure 18. Quantification of low-level properties (brightness, saturation, color warmth) shows that these remain unaffected by variations at maximal (black) and minimal (red) activation, unlike mid-level features.