Computational Exploration of Higher Visual Selectivity in the Human Brain

Andrew F. Luo

Sept 2024

Machine Learning Department & Neuroscience Institute

School of Computer Science & Dietrich College of Humanities and Social Sciences

Carnegie Mellon University

Pittsburgh, PA 15213

Thesis Committee:

Michael J. Tarr (Co-Chair)

Leila Wehbe (Co-Chair)

Deva Ramanan

Maggie Henderson

Eero P. Simoncelli

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy.

A deep gratitude to all that helped me along the way.

Abstract

A fundamental goal of cognitive neuroscience has been understanding how the human visual cortex supports perceiving and interpreting visual information in the world around us. Traditional approaches to mapping the visual cortex have relied on manually assembled stimulus sets, often employing isolated objects in artificial contexts with simplified backgrounds. These approaches do not fully capture the complexity and richness of real-world visual experience, potentially biasing results and limiting our understanding of visual processing. My thesis introduces a suite of computational approaches leveraging naturalistic image stimuli to identify and characterize the high-level organization of visual information in the human brain. Specifically, I present:

1. Brain Diffusion for Visual Exploration (**BrainDiVE**): A method utilizing gradient guidance from a differentiable image-to-fMRI encoder and a pre-trained image diffusion model to generate naturalistic "most-exciting-inputs" that maximally activate specific brain regions.
2. Semantic Captioning Using Brain Alignments (**BrainSCUBA**): A technique unifying the embedding spaces of CLIP image and text embeddings with fMRI encoder weights to drive a vision-language model. This enables the generation of natural language descriptions of voxel-wise selectivity within the visual cortex.
3. Semantic Attribution and Image Localization (**BrainSAIL**): An approach employing vision foundation models and dense semantic features to localize activating objects within complex naturalistic images across higher-level visual areas.

These computational methods are complemented by human validation experiments using synthetically generated stimuli. Overall, my thesis work demonstrates the power

of combining naturalistic stimuli with advanced computational techniques to reveal the fine-grained organization of the human visual cortex. In addition to providing a detailed overview of the computational models I have developed, I outline future computational and fMRI experiments designed to further validate and extend these findings. My research paves the way towards a more comprehensive and ecologically valid understanding of visual processing, with implications for building more accurate models of the brain and contributing to novel applications in artificial intelligence.

Acknowledgments

I'd like to thank all my collaborators and lab mates.

Contents

Chapter 1

Introduction

Understanding how the human visual cortex extracts and organizes semantic information from the complex visual stimulus is a long-standing question in cognitive neuroscience, with implications for understanding human perception, cognition, and behavior. The brain's ability to efficiently process and interpret the complex stream of visual information is fundamental to our ability to recognize objects, navigate the world around us, and perform social interactions. Traditional approaches to investigating this process have relied on controlled experiments employing simplified stimuli, such as isolated images to identify preferential responses of specific brain regions to broad semantic categories. While these studies have provided valuable insights into the functional organization of the visual cortex, the hand-crafted nature of these stimuli limits the ecological validity of such findings. Unlike simplified experimental paradigms, natural scenes are inherently complex, characterized by multiple co-occurring objects, textures, and contextual associations. This discrepancy raises concerns about the generalizability of findings based on hand-crafted stimuli to real-world visual processing. The reliance on pre-defined stimulus categories introduces researcher bias, potentially obscuring novel functional organizations or fine-grained selectivities not encompassed by existing hypotheses. Relying solely on pre-defined categories may overlook subtle but meaningful distinctions within categories and may fail to capture the impact of contextual information on object recognition. To overcome these limitations and gain a more com-

prehensive understanding of how the visual cortex processes semantic information in real-world contexts, this thesis leverages recent advancements in computer vision, particularly models trained on massive image datasets. These models provide a powerful tool for extracting semantically rich, human-aligned representations from natural images, capturing the complexity and nuances present in everyday visual experiences. By integrating these models with fMRI data collected during naturalistic image viewing, I introduce three novel methodologies that advance our understanding of visual cortex organization: **First**, I introduce Brain Diffusion for Visual Exploration ("BrainDiVE"), a method for synthesizing novel, naturalistic images specifically designed to maximally activate a given brain region using diffusion models. This data-driven approach circumvents the need for pre-defined category-specific stimuli, enabling the exploration of functional organization and fine-grained selectivities without relying on a priori assumptions. **Second**, I introduce Semantic Captioning using Brain Alignments ("BrainSCUBA"). This technique enables the generation of natural language captions that describe the optimal stimuli for individual voxels. By leveraging contrastive vision-language models and large-language models, this method provides concrete, interpretable descriptions of voxel-wise semantic selectivity, offering a nuanced understanding of feature preferences across visual sub-regions. **Third**, I introduce Semantic Attribution and Image Localization ("BrainSAIL"), which can spatially attribute higher visual cortex selectivity within natural images. This technique extracts dense, per-pixel semantic embeddings and integrates them with whole-image representations to identify the specific image regions driving activation in different cortical areas. This approach elucidates the neural mechanisms underlying semantic visual processing in ecologically valid contexts.

This thesis leverages novel methodologies to uncover the fine-grained semantic organization of the visual cortex, grounded in naturalistic stimuli and interpretable outputs. By demonstrating the power of these techniques to reveal novel functional organizations, identify voxel-level semantic preferences, and link image features to brain activity, this work contributes to a more nuanced and ecologically valid understanding of visual cognition. Ultimately, these insights aim to bridge the gap between artificial and biological vision systems, potentially informing the development of

advanced artificial intelligence and deepening our understanding of the fundamental principles of human visual perception.

Chapter 2

Brain Diffusion for Visual Exploration: Cortical Discovery using Large Scale Generative Models

2.1 Introduction

The human visual cortex plays a fundamental role in our ability to process, interpret, and act on visual information. While previous studies have provided important evidence that regions in the higher visual cortex preferentially process complex semantic categories such as faces, places, bodies, words, and food [Epstein and Kanwisher, 1998, Grill-Spector and Malach, 2004, Jain et al., 2023, Kanwisher et al., 1997, Khosla et al., 2022a, Pennock et al., 2023a, Sergent et al., 1992a], these important discoveries have been primarily achieved through the use of researcher-crafted stimuli. However, hand-selected, synthetic stimuli may bias the results or may not accurately capture the complexity and variability of natural scenes, sometimes leading to debates about the interpretation and validity of identified functional regions [Ishai et al., 1999]. Furthermore, mapping selectivity based on responses to a fixed set of stimuli is necessarily limited, in that it can only identify selectivity for the stimulus properties that are sampled. For these

Figure 2.1: **Images generated using BrainDiVE .** Images are generated using a diffusion model with maximization of voxels identified from functional localizer experiments as conditioning. We find that brain signals recorded via fMRI can guide the synthesis of images with high semantic specificity, strengthening the evidence for previously identified category selective regions. Select images are shown, please see below for uncurated images.

reasons, data-driven methods for interpreting high-dimensional neural tuning are complementary to traditional approaches. We introduce Brain Diffusion for Visual Exploration ("BrainDiVE"), a *generative* approach for synthesizing images that are predicted to activate a given region in the human visual cortex. Several recent studies have yielded intriguing results by combining deep generative models with brain guidance [Gu et al., 2022, Ponce et al., 2019, Ratan Murty et al., 2021]. BrainDiVE, enabled by the recent availability of large-scale fMRI datasets based on natural scene images [Allen et al., 2022, Chang et al., 2019], allows us to further leverage state-of-the-art diffusion models in identifying fine-grained functional specialization in an objective and data-driven manner. BrainDiVE is based on image diffusion models which are typically driven by text prompts in order to generate synthetic stimuli [Nichol et al., 2021]. We replace these prompts with maximization of voxels in given brain areas. The result being that the resultant synthesized images are tailored to targeted regions in higher-order visual areas. Analysis of these images enables data-driven exploration of the underlying feature preferences for different visual cortical sub-regions. Importantly, because the synthesized images are optimized to maximize the response of a given sub-region, these images emphasize and isolate critical feature preferences beyond what was present in the original stimulus images used in collecting the brain data. To validate our findings, we further performed several human behavioral studies that confirmed the semantic identities of our synthesized images.

More broadly, we establish that BrainDiVE can synthesize novel images (Figure 2.1) for

category-selective brain regions with high semantic specificity. Importantly, we further show that BrainDiVE can identify ROI-wise differences in selectivity that map to ecologically relevant properties. Building on this result, we are able to identify novel functional distinctions within sub-regions of existing ROIs. Such results demonstrate that BrainDiVE can be used in a data-driven manner to enable new insights into the fine-grained functional organization of the human visual cortex.

2.2 Related work

Mapping High-Level Selectivity in the Visual Cortex. Certain regions within the higher visual cortex are believed to specialize in distinct aspects of visual processing, such as the perception of faces, places, bodies, food, and words [Cohen et al., 2000, Desimone et al., 1984, Downing et al., 2001, Epstein and Kanwisher, 1998, Grill-Spector and Malach, 2004, Jain et al., 2023, Kanwisher et al., 1997, Khosla et al., 2022b, McCandliss et al., 2003, Pennock et al., 2023b]. Many of these discoveries rely on carefully handcrafted stimuli specifically designed to activate targeted regions. However, activity under natural viewing conditions is known to be different [Gallant et al., 1998]. Recent efforts using artificial neural networks as image-computable encoders/predictors of the visual pathway [Conwell et al., 2022a, Eickenberg et al., 2017, Khaligh-Razavi and Kriegeskorte, 2014, Kubilius et al., 2019, la Tour et al., 2022, Naselaris et al., 2011, Wang et al., 2022, Wen et al., 2018, Yamins et al., 2014] have facilitated the use of more naturalistic stimulus sets. Our proposed method incorporates an image-computable encoding model in line with this past work.

Deep Generative Models. The recent rise of learned generative models has enabled sampling from complex high dimensional distributions. Notable approaches include variational autoencoders [Kingma and Welling, 2013, Van Den Oord et al., 2017], generative adversarial networks [Goodfellow et al., 2020], flows [Dinh et al., 2014, Rezende and Mohamed, 2015], and score/energy/diffusion models [Ho et al., 2022, Hyvärinen and Dayan, 2005, Sohl-Dickstein et al., 2015, Song et al., 2020b]. It is possible to condition the model on category [Brock et al., 2018,

Mirza and Osindero, 2014], text [Ramesh et al., 2022, Reed et al., 2016], or images [Rombach et al., 2022]. Recent diffusion models have been conditioned with brain activations to reconstruct observed images [Chen et al., 2022, Kneeland et al., 2023, Lu et al., 2023, Ozcelik and VanRullen, 2023, Takagi and Nishimoto, 2022]. Unlike BrainDiVE, these approaches tackle reconstruction but not synthesis of novel images that are predicted to activate regions of the brain.

Brain-Conditioned Image Generation. The differentiable nature of deep encoding models inspired work to create images from brain gradients in mice, macaques, and humans [Bashivan et al., 2019, Khosla and Wehbe, 2022, Walker et al., 2019]. Without constraints, the images recovered are not naturalistic. Other approaches have combined deep generative models with optimization to recover natural images in macaque and humans [Gu et al., 2022, Ponce et al., 2019, Ratan Murty et al., 2021]. Both [Gu et al., 2022, Ratan Murty et al., 2021] utilize fMRI brain gradients combined with ImageNet trained BigGAN. In particular [Ratan Murty et al., 2021] performs end-to-end differentiable optimization by assuming a soft relaxation over the $1,000$ ImageNet classes; while [Gu et al., 2022] trains an encoder on the NSD dataset [Allen et al., 2022] and first searches for top-classes, then performs gradient optimization within the identified classes. Both approaches are restricted to ImageNet images, which are primarily images of single objects. Our work presents major improvements by enabling the use of diffusion models [Rombach et al., 2022] trained on internet-scale datasets [Schuhmann et al., 2022a] over three magnitudes larger than ImageNet. Concurrent work by [Pierzchlewicz et al., 2023] explore the use of gradients from macaque V4 with diffusion models, however their approach focuses on early visual cortex with grayscale image outputs, while our work focuses on higher-order visual areas and synthesize complex compositional scenes. By avoiding the search-based optimization procedures used in [Gu et al., 2022], our work is not restricted to images within a fixed class in ImageNet. Further, to the authors' knowledge we are the first work to use image synthesis methods in the identification of functional specialization in sub-parts of ROIs.

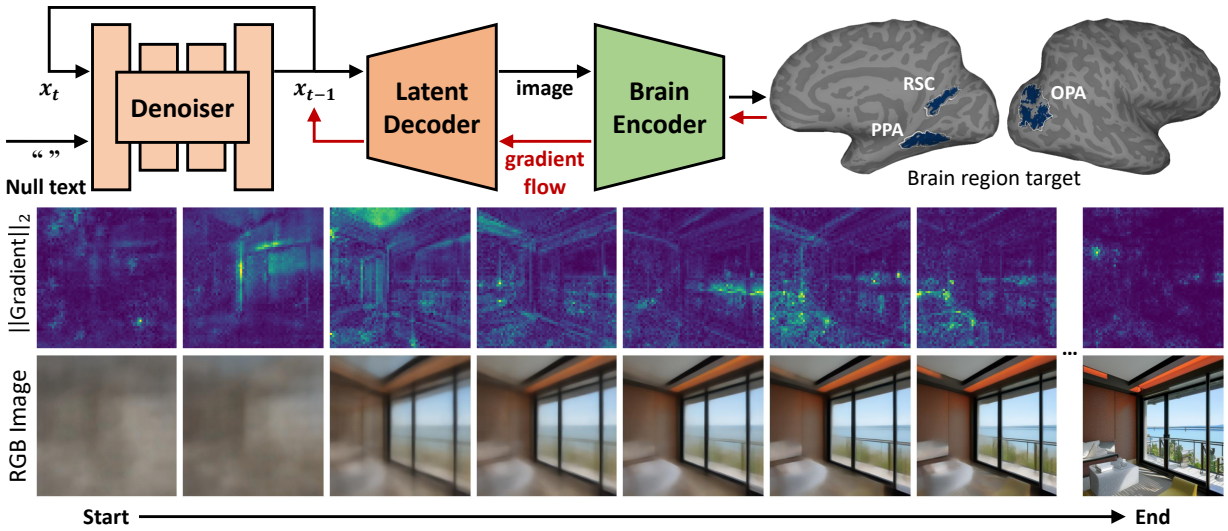


Figure 2.2: **Architecture of brain guided diffusion (BrainDiVE). Top:** Our framework consists of two core components: **(1)** A diffusion model trained to synthesize natural images by iterative denoising; we utilize pretrained LDMs. **(2)** An encoder trained to map from images to cortical activity. Our framework can synthesize images that are predicted to activate any subset of voxels. Shown here are scene-selective regions (RSC/PPA/OPA) on the right hemisphere. **Bottom:** We visualize every 4 steps the magnitude of the gradient of the brain w.r.t. the latent and the corresponding "predicted x_0" [Song et al., 2020a] when targeting scene selective voxels in both hemispheres. We find clear structure emerges.

2.3 Methods

We aim to generate stimuli that maximally activate a given region in visual cortex using paired natural image stimuli and fMRI recordings. We first review relevant background information on diffusion models. We then describe how we can parameterize encoding models that map from images to brain data. Finally, we describe how our framework (Figure 2.2) can leverage brain signals as guidance to diffusion models to synthesize images that activate a target brain region.

2.3.1 Background on Diffusion Models

Diffusion models enable sampling from a data distribution $p(x)$ by iterative denoising. The sampling process starts with $x_T \sim \mathcal{N}(0, \mathbb{I})$, and produces progressively denoised samples $x_{T-1}, x_{T-2}, x_{T-3} \ldots$ until a sample x_0 from the target distribution is reached. The noise level varies by timestep t, where the sample at each timestep is a weighted combination of x_0 and

$\epsilon \sim \mathcal{N}(0, \mathbb{I})$, with $x_t = \sqrt{\alpha_t} x_0 + \epsilon\sqrt{1-\alpha_t}$. The value of α interpolates between $\mathcal{N}(0, \mathbb{I})$ and $p(x)$.

In the noise prediction setting, an autoencoder network $\epsilon_\theta(x_t, t)$ is trained using a mean-squared error $\mathbb{E}_{(x,\epsilon,t)}\left[\|\epsilon_\theta(x_t, t) - \epsilon\|_2^2\right]$. In practice, we utilize a pretrained latent diffusion model (LDM) [Rombach et al., 2022], with learned image encoder E_Φ and decoder D_Ω, which together act as an autoencoder $\mathcal{I} \approx D_\Omega(E_\Phi(\mathcal{I}))$. The diffusion model is trained to sample x_0 from the latent space of E_Φ.

2.3.2 Brain-Encoding Model Construction

A learned voxel-wise brain encoding model is a function M_θ that maps an image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$ to the corresponding brain activation fMRI beta values represented as an N element vector $B \in \mathbb{R}^N$: $M_\theta(\mathcal{I}) \Rightarrow B$. Past work has identified later layers in neural networks as the best predictors of higher visual cortex [Wang et al., 2021, 2022], with CLIP trained networks among the highest performing brain encoders [Conwell et al., 2022a, Sun et al., 2023]. As our target is the higher visual cortex, we utilize a two component design for our encoder. The first component consists of a CLIP trained image encoder which outputs a K dimensional vector as the latent embedding. The second component is a linear adaptation layer $W \in \mathcal{R}^{N\times K}, b \in \mathcal{R}^N$, which maps euclidean normalized image embeddings to brain activation.

$$B \approx M_\theta(\mathcal{I}) = W \times \frac{\texttt{CLIP}_{\texttt{img}}(\mathcal{I})}{\|\texttt{CLIP}_{\texttt{img}}(\mathcal{I})\|_2} + b$$

Optimal W^*, b^* are found by optimizing the mean squared error loss over images. We observe that use of a normalized CLIP embedding improves stability of gradient magnitudes w.r.t. the image.

2.3.3 Brain-Guided Diffusion Model

BrainDiVE seeks to generate images conditioned on maximizing brain activation in a given region. In conventional text-conditioned diffusion models, the conditioning is done in one of two ways. The first approach modifies the function ϵ_θ to further accept a conditioning vector c, resulting in $\epsilon_\theta(x_t, t, c)$. The second approach uses a contrastive trained image-to-concept encoder, and seeks to maximize a similarity measure with a text-to-concept encoder.

Conditioning on activation of a brain region using the first approach presents difficulties. We do not know *a priori* the distribution of other non-targeted regions in the brain when a target region is maximized. Overcoming this problem requires us to either have a prior $p(B)$ that captures the joint distribution for all voxels in the brain, to ignore the joint distribution that can result in catastrophic effects, or to use a handcrafted prior that may be incorrect [Ozcelik and VanRullen, 2023]. Instead, we propose to condition the diffusion model via our image-to-brain encoder. During inference we perturb the denoising process using the gradient of the brain encoder *maximization* objective, where γ is a scale, and $S \subseteq N$ are the set of voxels used for guidance. We seek to maximize the average activation of S predicted by M_θ:

$$\epsilon'_{theta} = \epsilon_{theta} - \sqrt{1-\alpha_t}\nabla_{x_t}(\frac{\gamma}{|S|}\sum_{i\in S} M_\theta(D_\Omega(x'_t))_i)$$

Like [Dhariwal and Nichol, 2021, Li et al., 2022b, Nichol et al., 2021], we observe that convergence using the current denoised x_t is poor without changes to the guidance. This is because the current image (latent) is high noise and may lie outside of the natural image distribution. We instead use a weighted reformulation with an euler approximation [Li et al., 2022b, Song et al., 2020a] of the final image:

$$\hat{x}_0 = \frac{1}{\sqrt{\alpha}}(x_t - \sqrt{1-\alpha}\epsilon_t)$$
$$x'_t = (\sqrt{1-\alpha})\hat{x}_0 + (1-\sqrt{1-\alpha})x_t$$

By combining an image diffusion model with a differentiable encoding model of the brain, we are able to generate images that seek to maximize activation for any given brain region.

2.4 Results

In this section, we use BrainDiVE to highlight the semantic selectivity of pre-identified category-selective voxels. We then show that our model can capture subtle differences in response properties between ROIs belonging to the same broad category-selective network. Finally, we utilize BrainDiVE to target finer-grained sub-regions within existing ROIs, and show consistent divisions based on semantic and visual properties. We quantify these differences in selectivity across regions using human perceptual studies, which confirm that BrainDiVE images can highlight differences in tuning properties. These results demonstrate how BrainDiVE can elucidate the functional properties of human cortical populations, making it a promising tool for exploratory neuroscience.

2.4.1 Setup

We utilize the Natural Scenes Dataset (NSD; Allen et al. [2022]), which consists of whole-brain 7T fMRI data from 8 human subjects, 4 of whom viewed $10,000$ natural scene images repeated $3\times$. These subjects, S1, S2, S5, and S7, are used for analyses in the main paper (see Supplemental for results for additional subjects). All images are from the MS COCO dataset. We use beta-weights (activations) computed using GLMSingle [Prince et al., 2022] and further normalize each voxel to $\mu = 0, \sigma = 1$ on a per-session basis. We average the fMRI activation across repeats of the same image within a subject. The $\sim 9,000$ unique images for each subject [Allen et al., 2022] are used to train the brain encoder for each subject, with the remaining $\sim 1,000$ shared images used to evaluate R^2. Image generation is on a per-subject basis and done on an `Nvidia V100` using $1,500$ compute hours. As the original category ROIs in NSD are very generous, we utilize a stricter $t > 2$ threshold to reduce overlap unless otherwise noted. The final category and ROI masks used in our experiments are derived from the logical `AND` of the official NSD masks with

the masks derived from the official *t*-statistics.

We utilize `stable-diffusion-2-1-base`, which produces images of 512×512 resolution using ϵ-prediction. Following best practices, we use multi-step 2nd order DPM-Solver++ [Lu et al., 2022] with 50 steps and apply 0.75 SAG [Hong et al., 2022]. We set step size hyperparameter $\gamma = 130.0$. Images are resized to 224×224 for the brain encoder. "" (null prompt) is used as the input prompt, thus the diffusion performs unconditional generation without brain guidance. For the brain encoder we use `ViT-B/16`, for CLIP probes we use `CoCa ViT-L/14`. These are the highest performing `LAION-2B` models of a given size provided by `OpenCLIP` [Ilharco et al., 2021, Radford et al., 2021, Schuhmann et al., 2022b, Yu et al., 2022]. We train our brain encoders on each human subject separately to predict the activation of all higher visual cortex voxels. See Supplemental for visualization of test time brain encoder R^2. To compare images from different ROIs and sub-regions (OFA/FFA in 2.4.3, two clusters in 2.4.4), we asked human evaluators select which of two image groups scored higher on various attributes. We used 100 images from each group randomly split into 10 non-overlapping subgroups. Each human evaluator performed 80 comparisons, across 10 splits, 4 NSD subjects, and for both fMRI and generated images. See Supplemental for standard error of responses. Human evaluators provided written informed consent and were compensated at $\$12.00$/hour. The study protocol was approved by the institutional review board at the authors' institution.

2.4.2 Broad Category-Selective Networks

In this experiment, we target large groups of category-selective voxels which can encompass more than one ROI (Figure 2.3). These regions have been previously identified as selective for broad semantic categories, and this experiment validates our method using these identified regions. The face-, place-, body-, and word- selective ROIs are identified with standard localizer stimuli [Stigliani et al., 2015]. The food-selective voxels were obtained from [Jain et al., 2023]. The same voxels were used to select the top activating NSD images (referred to as "NSD") and to guide the generation of BrainDiVE images.

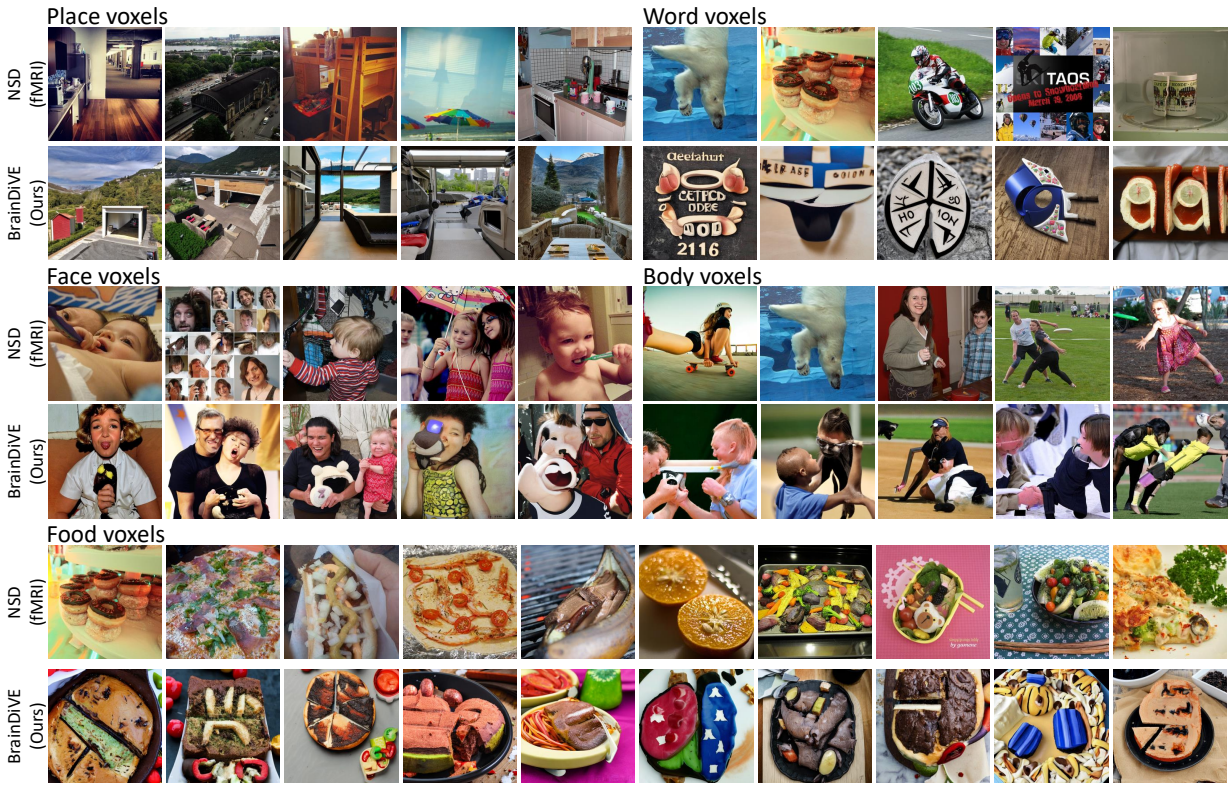


Figure 2.4: **Results for category selective voxels (S1).** We identify the top-5 images from the stimulus set or generated by our method with highest average activation in each set of category selective voxels for the face/place/word/body categories, and the top-10 images for the food selective voxels.

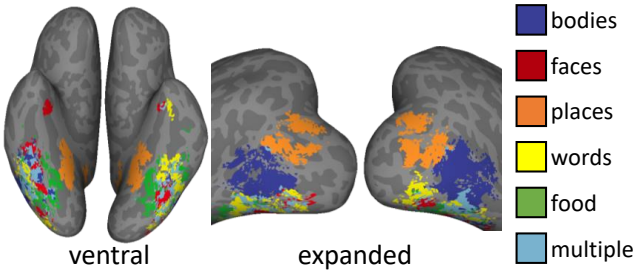


Figure 2.3: **Visualizing category-selective voxels in S1.** See text for details on how category selectivity was defined.

In Figures 2.4 we visualize, for place-, face-, word-, and body- selective voxels, the top-5 out of $10,000$ images from the fMRI stimulus set (NSD), and the top-5 images out of $1,000$ total images as evaluated by the encoding component of BrainDiVE. For food selective voxels, the top-10 are visualized. A visual inspection indicates that our method is able to generate diverse images that semantically represent the target category. We further use CLIP to perform semantic probing of the images, and force the images to be classified into one of five categories. We measure the percentage of images that match the preferred category for a given set of voxels (Table 2.1). We

find that our top-10% and 20% of images exceed the top-1% and 2% of natural images in accuracy, indicating our method has high semantic specificity.

	Faces		Places		Bodies		Words		Food		Mean	
	S1↑	S2↑	S1↑	S2↑	S1↑	S2↑	S1↑	S2↑	S1↑	S2↑	S1↑	S2↑
NSD all stim	17.4	17.2	29.9	29.5	31.6	31.8	10.3	10.6	10.8	10.9	20.0	20.0
NSD top-200	42.5	41.5	66.5	80.0	56.0	65.0	31.5	34.5	68.0	85.5	52.9	61.3
NSD top-100	40.0	45.0	68.0	79.0	49.0	60.0	30.0	49.0	78.0	85.0	53.0	63.6
BrainDiVE-200	**69.5**	**70.0**	**97.5**	**100**	**75.5**	68.5	**60.0**	57.5	89.0	94.0	**78.3**	75.8
BrainDiVE-100	61.0	68.0	**97.0**	**100**	75.0	**69.0**	**60.0**	**62.0**	**92.0**	**95.0**	77.0	**78.8**

Table 2.1: **Evaluating semantic specificity with zero-shot CLIP classification.** We use CLIP to classify images from each ROI into five semantic categories: face/place/body/word/food. Shown is the percentage where the classified category of the image matches the preferred category of the brain region. We show this for each subject's entire NSD stimulus set ($10,000$ images for S1&S2); the top-200 and top-100 images (top-2% and top-1%) evaluated by mean true fMRI beta, and the top-200 and top-100 (20% and 10%) of BrainDiVE images as self-evaluated by the encoding component of BrainDiVE. BrainDiVE generates images with higher semantic specificity than the top 1% of natural images for each brain region.

2.4.3 Individual ROIs

In this section, we apply our method to individual ROIs that are selective for the same broad semantic category. We focus on the occipital face area (OFA) and fusiform face area (FFA), as initial tests suggested little differentiation between ROIs within the place-, word-, and body-selective networks. In this experiment, we also compare our results against the top images for FFA and OFA from NeuroGen [Gu et al., 2022], using the top 100 out of 500 images provided by the authors. Following NeuroGen, we also generate 500 total images, targeting FFA and OFA separately (Figure 2.5). We observe that both diffusion-generated and NSD images have very high face content in FFA, whereas NeuroGen has higher animal face content. In OFA, we observe both NSD and BrainDiVE images have a strong face component, although we also observe text selectivity in S2 and animal face selectivity in S5. Again NeuroGen predicts a higher animal component than face for S5. By avoiding the use of fixed categories, BrainDiVE images are more diverse than those of NeuroGen. This trend of face and animals appears at $t > 2$ and the much stricter $t > 5$ threshold for identifying face-selective voxels ($t > 5$ used for

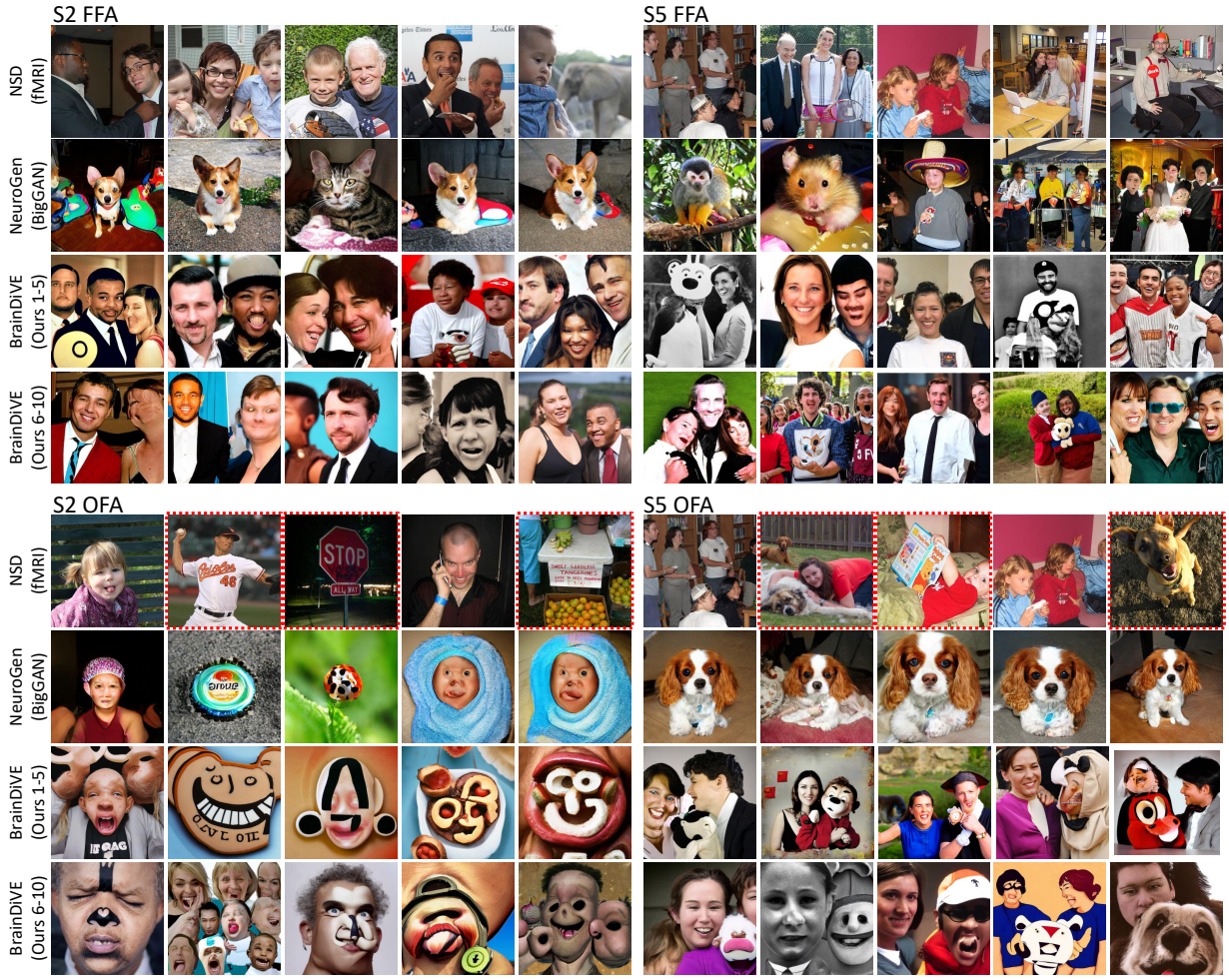


Figure 2.5: **Results for face-selective ROIs.** For each ROI (OFA, FFA) we visualize the top-5 images from NSD and NeuroGen, and the top-10 from BrainDiVE. NSD images are selected using the fMRI betas averaged within each ROI. NeuroGen images are ranked according to their official predicted ROI activity means. BrainDiVE images are ranked using our predicted ROI activities from 500 images. Red outlines in the NSD images indicate examples of responsiveness to non-face content.

visualization/evaluation). The differences in images synthesized by BrainDiVE for FFA and OFA are consistent with past work suggesting that FFA represents faces at a higher level of abstraction than OFA, while OFA shows greater selectivity to low-level face features and sub-components, which could explain its activation by off-target categories [Liu et al., 2010, Pitcher et al., 2011, Tsantani et al., 2021].

To quantify these results, we perform a human study where subjects are asked to compare the

Which ROI has more...	photorealistic faces				animals				abstract shapes/lines			
	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7
FFA-NSD	**45**	**43**	**34**	**41**	34	34	17	15	21	6	14	22
OFA-NSD	25	22	21	18	**47**	**36**	**65**	**65**	**24**	**44**	**28**	**25**
FFA-BrainDiVE	**79**	**89**	**60**	**52**	17	13	21	19	6	11	18	20
OFA-BrainDiVE	11	4	15	22	**71**	**61**	**52**	**50**	**80**	**79**	**40**	**39**

Table 2.2: **Human evaluation of the difference between face-selective ROIs**. Evaluators compare groups of images corresponding to OFA and FFA; comparisons are done within GT and generated images respectively. Questions are posed as: "Which group of images has more X?"; options are FFA/OFA/Same. Results are in %. Note that the "Same" responses are not shown; responses across all three options sum to 100.

top-100 images between FFA & OFA, for both NSD and generated images. Results are shown in Table 2.2. We find that OFA consistently has higher animal and abstract content than FFA. Most notably, this difference is on average more pronounced in the images from BrainDiVE, indicating that our approach is able to highlight subtle differences in semantic selectivity across regions.

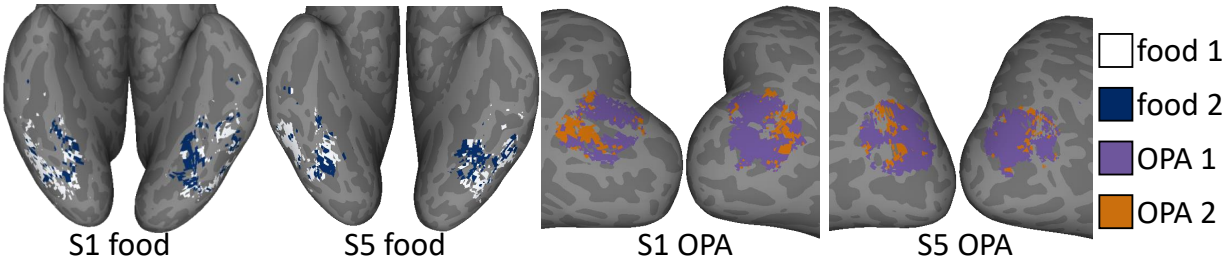


Figure 2.6: **Clustering within the food ROI and within OPA**. Clustering of encoder model weights for each region is shown for two example subjects on an inflated cortical surface.

2.4.4 Semantic Divisions within ROIs

In this experiment, we investigate if our model can identify novel sub-divisions within existing ROIs. We first perform clustering on normalized per-voxel encoder weights using vmf-clustering [Banerjee et al., 2005]. We find consistent cosine difference between the cluster centers in the food-selective ROI as well as in the occipital place area (OPA), clusters shown in Figure 2.6. In all four subjects, we observe a relatively consistent anterior-posterior split of OPA. While the clusters within the food ROI vary more anatomically, each subject appears to have a more

Which cluster is more ...	vegetables/fruits				healthy				colorful				far away			
	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7
Food-1 NSD	17	21	27	36	28	22	29	40	19	18	13	27	32	24	23	28
Food-2 NSD	**65**	**56**	**56**	**49**	**50**	**47**	**54**	**45**	**42**	**52**	**53**	**42**	**34**	**39**	**36**	**42**
Food-1 BrainDiVE	11	10	8	11	15	16	20	17	6	9	11	16	24	18	27	18
Food-2 BrainDiVE	**80**	**75**	**67**	**64**	**68**	**68**	**46**	**51**	**79**	**82**	**65**	**61**	**39**	**51**	**39**	**40**

Table 2.3: **Human evaluation of the difference between food clusters**. Evaluators compare groups of images corresponding to food cluster 1 (Food-1) and food cluster 2 (Food-2), with questions posed as "Which group of images has/is more X?". Comparisons are done within NSD and generated images respectively. Note that the "Same" responses are not shown; responses across all three options sum to 100. Results are in %.

medial and a more lateral cluster. We visualize the images for the two food clusters in Figure 2.7, and for the two OPA clusters in Figure 2.8. We observe that for both the food ROI and OPA, the BrainDiVE-generated images from each cluster have noticeable differences in their visual and semantic properties. In particular, the BrainDiVE images from food cluster-2 have much higher color saturation than those from cluster-1, and also have more objects that resemble fruits and vegetables. In contrast, food cluster-1 generally lacks vegetables and mostly consist of bread-like foods. In OPA, cluster-1 is dominated by indoor scenes (rooms, hallways), while 2 is overwhelmingly outdoor scenes, with a mixture of natural and man-made structures viewed from a far perspective. Some of these differences are also present in the NSD images, but the differences appear to be highlighted in the generated images.

To confirm these effects, we perform a human study (Table 2.3, Table 2.4) comparing the images from different clusters in each ROI, for both NSD and generated images. As expected from visual inspection of the images, we find that food cluster-2 is evaluated to have higher vegetable/fruit content, judged to be healthier, more colorful, and slightly more distant than food cluster-1. We find that OPA cluster-1 is evaluated to be more angular/geometric, include more indoor scenes, to be less natural and consisting of less distant scenes. Again, while these trends are present in the NSD images, they are more pronounced with the BrainDiVE images. This not only suggests that our method has uncovered differences in semantic selectivity within pre-existing ROIs, but also reinforces the ability of BrainDiVE to identify and highlight core

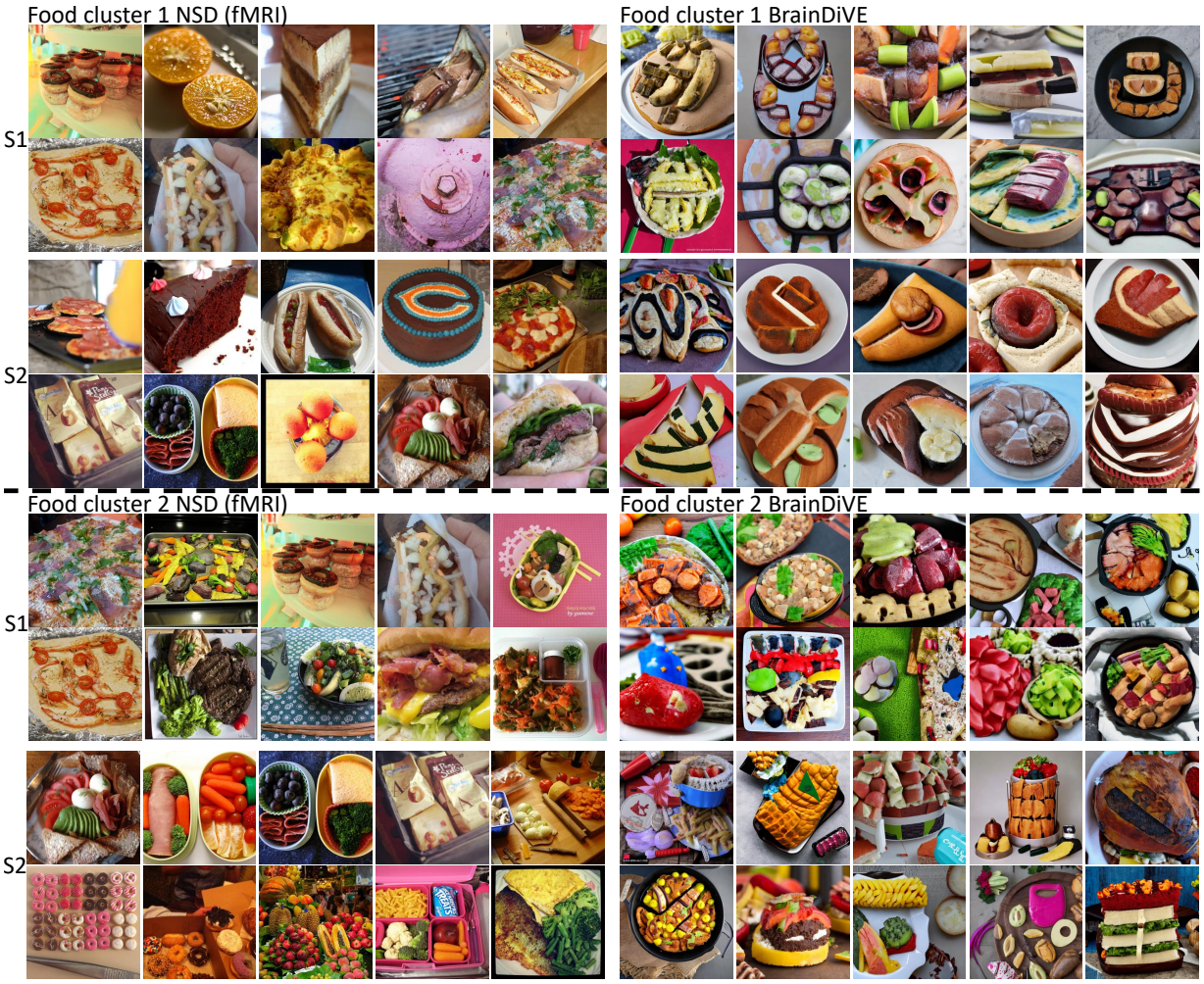


Figure 2.7: **Comparing results across the food clusters.** We visualize top-10 NSD fMRI (out of 10,000) and diffusion images (out of 500) for *each cluster*. While the first cluster largely consists of processed foods, the second cluster has more visible high color saturation foods, and more vegetables/fruit like objects. BrainDiVE helps highlight the differences between clusters.

functional differences across visual cortex regions.

2.4.5 fMRI Scanning with Human Subjects

To investigate neural responses to different visual categories, we conducted an fMRI experiment using a 3T Prisma scanner. Stimuli were presented using a mini-block design, with each block consisting of 12 images presented over 6 seconds. Each image was shown for 400 milliseconds, followed by a 100 millisecond grey background. A red fixation dot was continuously displayed at

Which cluster is more...	angular/geometric				indoor				natural				far away			
	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7
OPA-1 NSD	**45**	**58**	**49**	**51**	**71**	**88**	**80**	**79**	14	3	9	10	10	1	6	8
OPA-2 NSD	13	12	14	16	7	8	11	14	**73**	**89**	**71**	**81**	**69**	**93**	**81**	**85**
OPA-1 BrainDiVE	**76**	**87**	**88**	**76**	**89**	**90**	**90**	**85**	6	6	9	6	1	3	3	8
OPA-2 BrainDiVE	12	3	4	10	7	7	5	8	**91**	**91**	**83**	**90**	**97**	**92**	**91**	**88**

Table 2.4: **Human evaluation of the difference between OPA clusters**. Evaluators compare groups of images corresponding to OPA cluster 1 (OPA-1) and OPA cluster 2 (OPA-2), with questions posed as "Which group of images is more X?". Comparisons are done within NSD and generated images respectively. Note that the "Same" responses are not shown; responses across all three options sum to 100. Results are in %.

the center of the screen.

Functional localizer images were taken from Jain et al. [2023] and Stigliani et al. [2015]. These images were greyscale, and consisted of a single object each on scrambled backgrounds. The classes consisted of faces (adult), bodies, places (houses), words, and food.

Natural images were selected from the Natural Scenes Dataset (NSD) for each category selective region (faces, places, words, bodies, food). For each subject, images were ranked based on their average beta values within the corresponding category selective voxels (i.e., voxels previously identified as preferentially responding to that category). We then calculated the average rank across four subjects (S1, S2, S5, S7) and selected the top 100 images for each category.

For synthetic images, we generated 1000 images per category using BrainDiVE. We then used the encoder to obtain voxel-wise predictions for each image. As with the NSD images, we ranked the synthetic images based on the average encoder predictions across the four subjects and selected the top 100. Subjects were instructed to press a button if the same image was presented consecutively (1-back task). We perform motion correction using SPM12, followed by surface reconstruction using freesurfer. Functional data was aligned to the anatomical data using bbregister. We visualize preliminary results in Figure 2.9 and Figure 2.10. These results show that BrainDiVE images can generally trigger higher activations than the NSD images in the functional areas identified by the localizer, and yield more concentrated and less diffuse patterns on the brain, more specific to the region that we are targeting.

2.5 Discussion

Limitations and Future Work Here, we show that BrainDiVE generates diverse and realistic images that can probe the human visual pathway. This approach relies on existing large datasets of natural images paired with brain recordings. In that the evaluation of synthesized images is necessarily qualitative, it will be important to validate whether our generated images and candidate features derived from these images indeed maximize responses in their respective brain areas. As such, future work will focus on additional collection of human fMRI recordings using both our synthesized images and more focused stimuli designed to test our qualitative observations. Future work may also explore the images generated when BrainDiVE is applied to additional sub-region, new ROIs, or mixtures of ROIs.

Conclusion We introduce a novel method for guiding diffusion models using brain activations – BrainDiVE – enabling us to leverage generative models trained on internet-scale image datasets for data driven explorations of the brain. This allows us to better characterize fine-grained preferences across the visual system. We demonstrate that BrainDiVE can accurately capture the semantic selectivity of existing characterized regions. We further show that BrainDiVE can capture subtle differences between ROIs within the face selective network. Finally, we identify and highlight fine-grained subdivisions within existing food and place ROIs, differing in their selectivity for mid-level image features and semantic scene content. We validate our conclusions with extensive human evaluation of the images.

Figure 2.8: **Comparing results across the OPA clusters.** We visualize top-10 NSD fMRI (out of 10,000) and diffusion images (out of 500) for *each cluster*. While both consist of scene images, the first cluster have more indoor scenes, while the second has more outdoor scenes. The BrainDiVE images help highlight the differences in semantic properties.

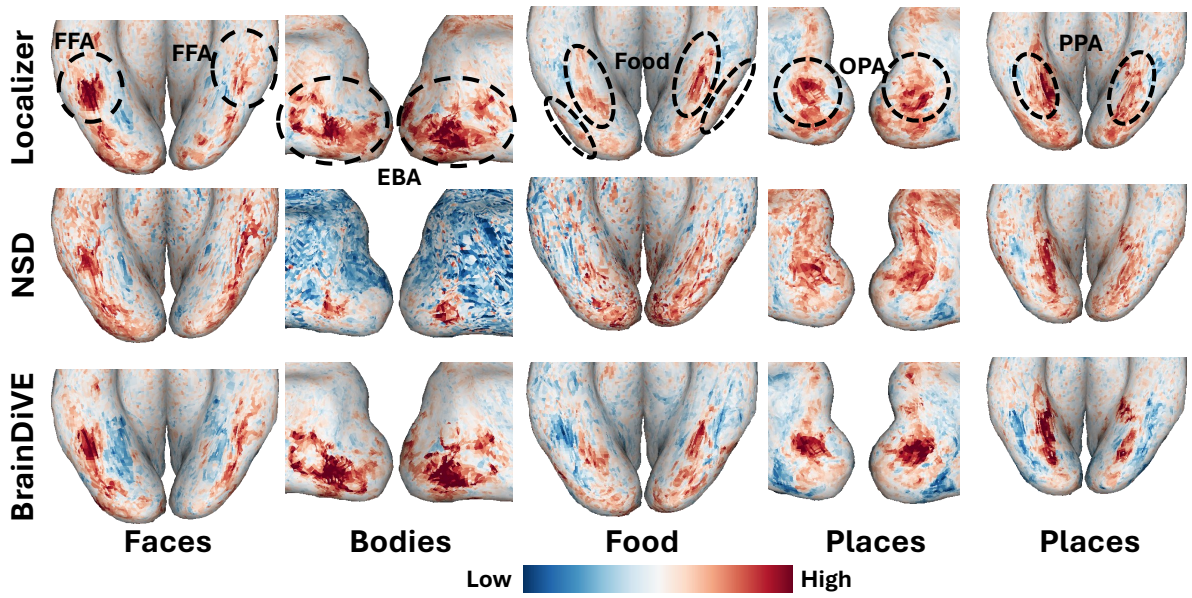


Figure 2.9: **Preliminary fMRI beta values for Subj1.**. **Top:** Results using greyscale functional localizer images. **Middle:** Results using top NSD images (natural). **Bottom:** Results using top BrainDiVE images.

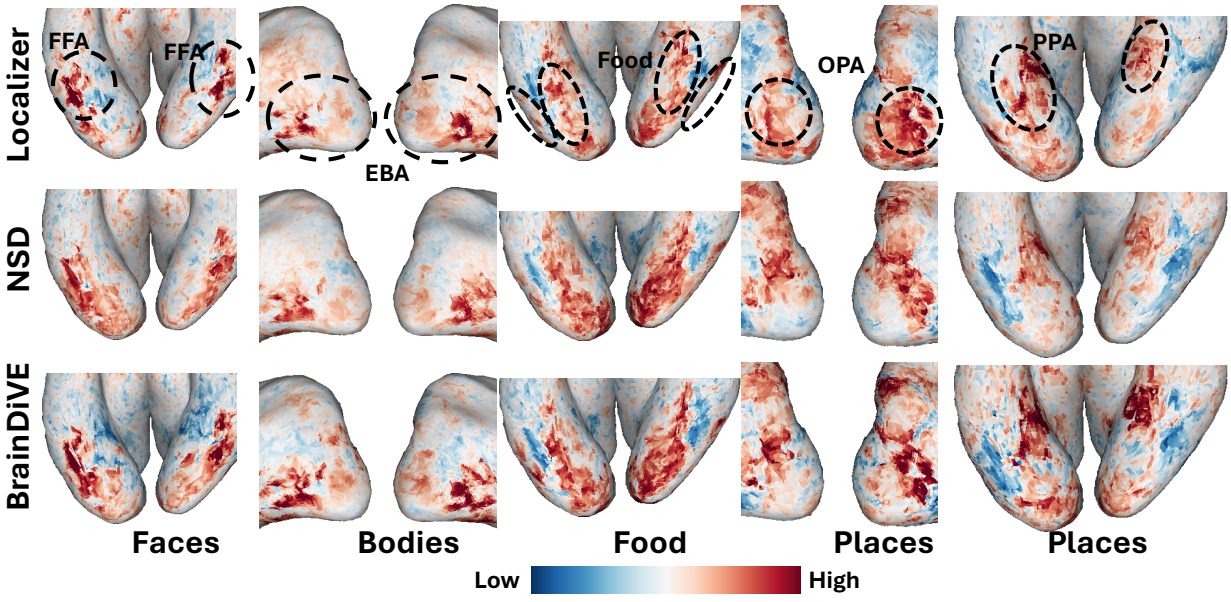


Figure 2.10: **Preliminary fMRI beta values for Subj2.**. **Top:** Results using greyscale functional localizer images. **Middle:** Results using top NSD images (natural). **Bottom:** Results using top BrainDiVE images.

2.6 Additional Results for BrainDiVE

2.6.1 Visualization of each subject's category selective voxel images

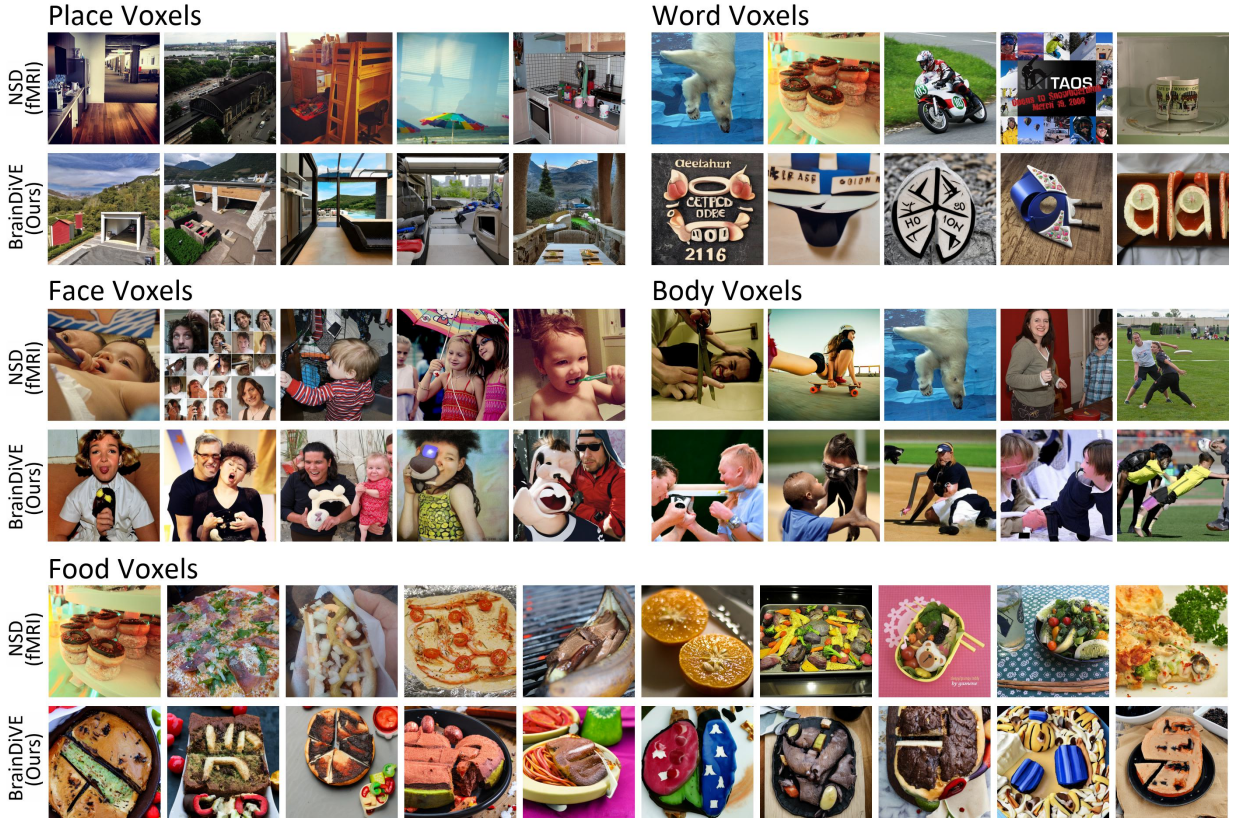


Figure 2.11: **Results for category selective voxels (S1).** We identify the top-5 images from the stimulus set or generated by our method with highest average activation in each set of category selective voxels for the face/place/word/body categories, and the top-10 images for the food selective voxels. Note the top NSD body voxel image for S1 was omitted from the main paper due to content.

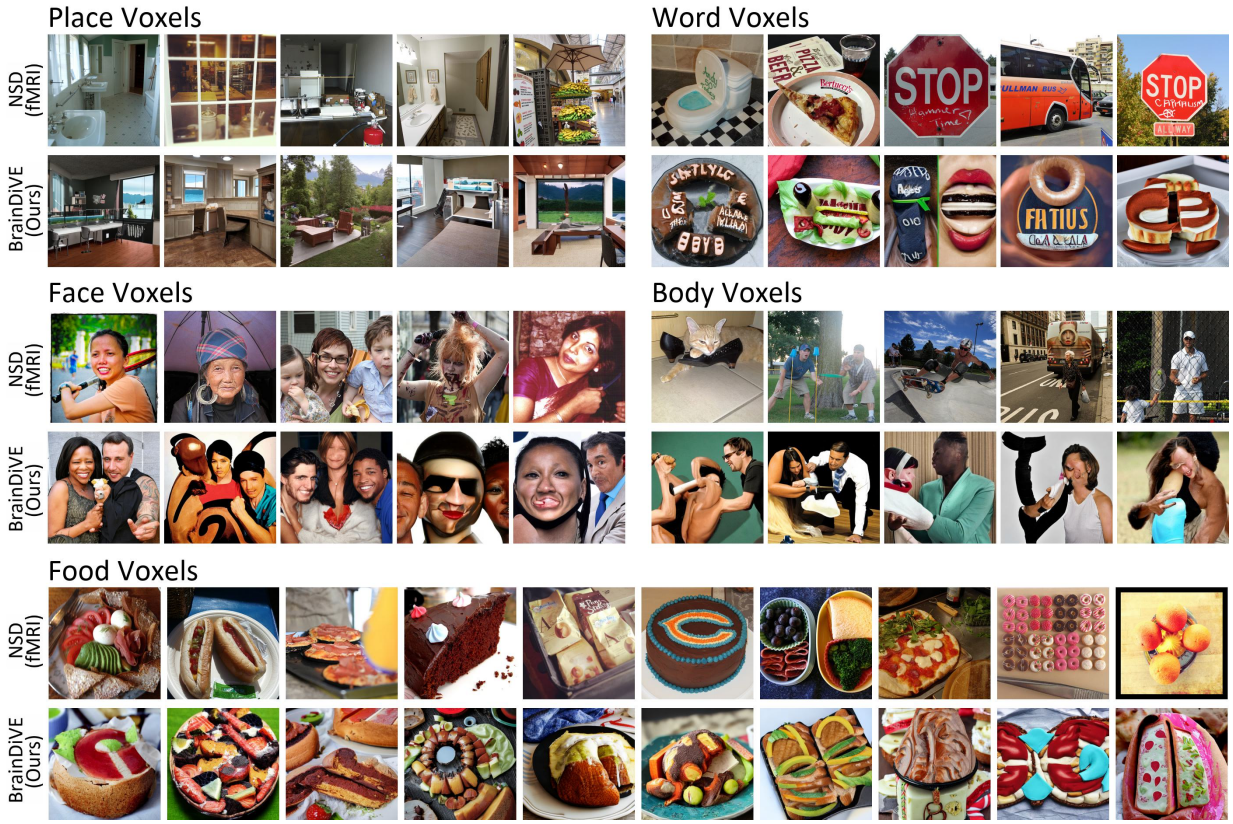


Figure 2.12: **Results for category selective voxels (S2).** We identify the top-5 images from the stimulus set or generated by our method with highest average activation in each set of category selective voxels for the face/place/word/body categories, and the top-10 images for the food selective voxels.

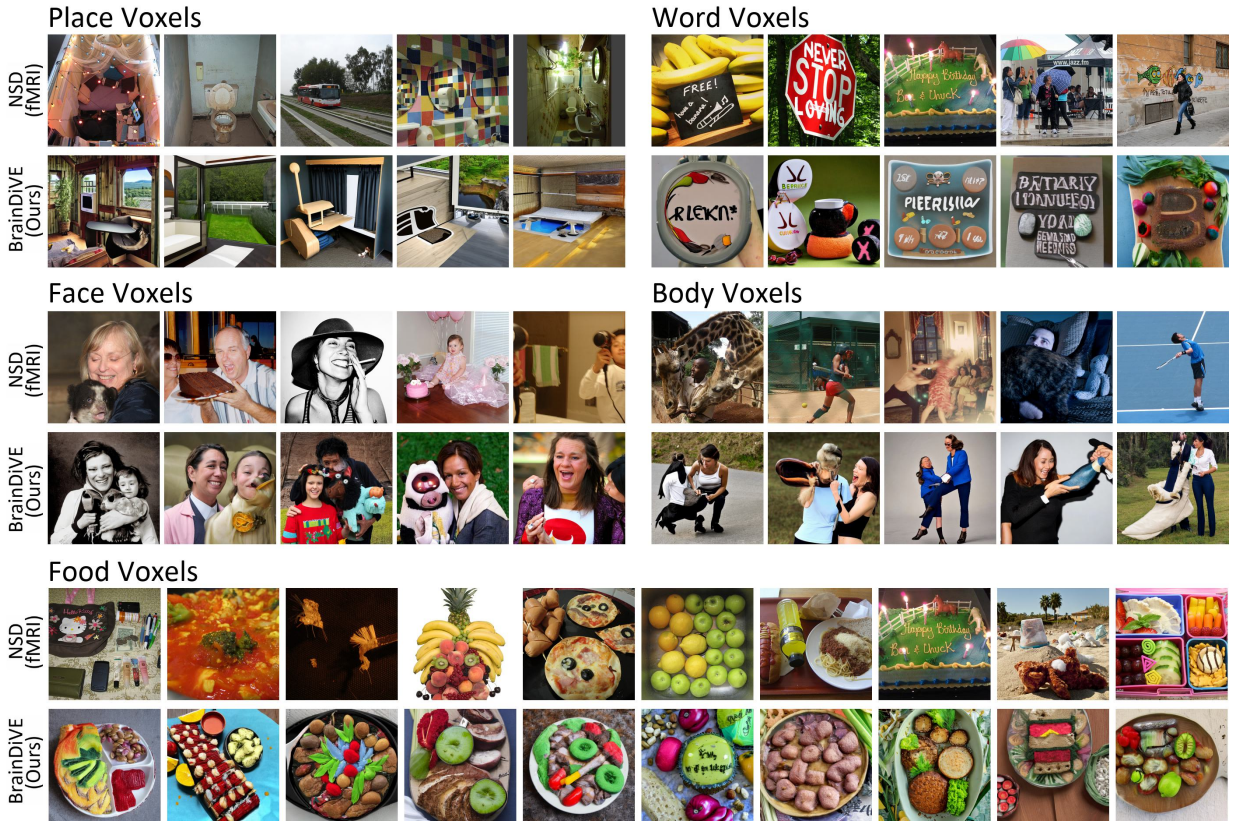


Figure 2.13: **Results for category selective voxels (S3).** We identify the top-5 images from the stimulus set or generated by our method with highest average activation in each set of category selective voxels for the face/place/word/body categories, and the top-10 images for the food selective voxels.

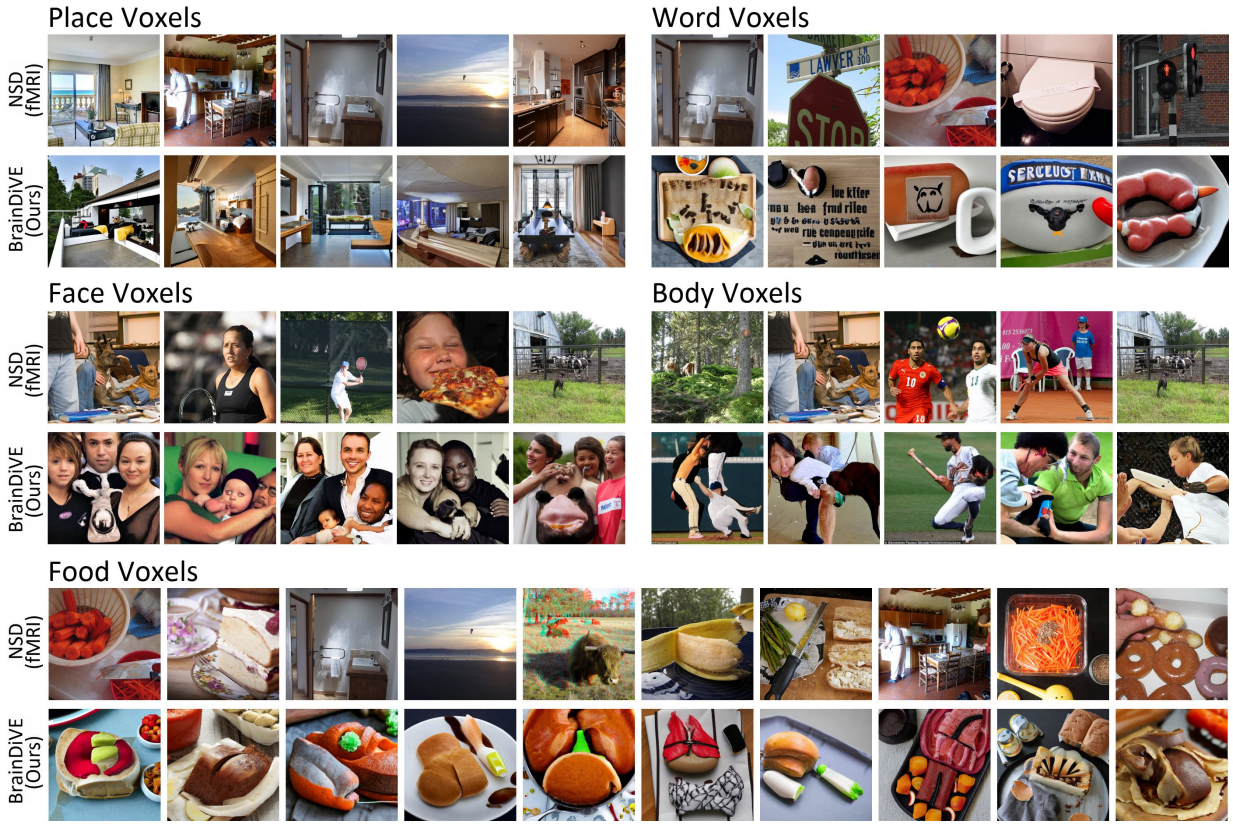


Figure 2.14: **Results for category selective voxels (S4).** We identify the top-5 images from the stimulus set or generated by our method with highest average activation in each set of category selective voxels for the face/place/word/body categories, and the top-10 images for the food selective voxels.

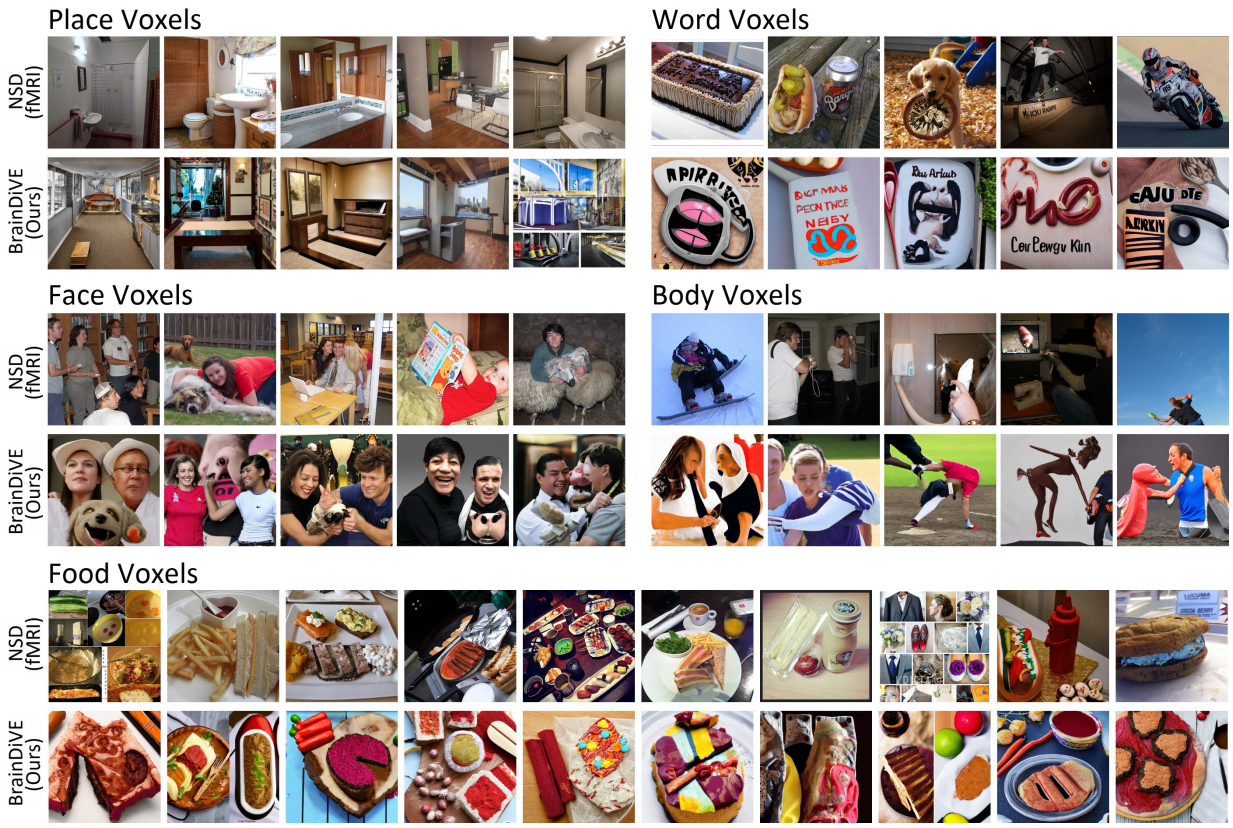


Figure 2.15: **Results for category selective voxels (S5).** We identify the top-5 images from the stimulus set or generated by our method with highest average activation in each set of category selective voxels for the face/place/word/body categories, and the top-10 images for the food selective voxels.

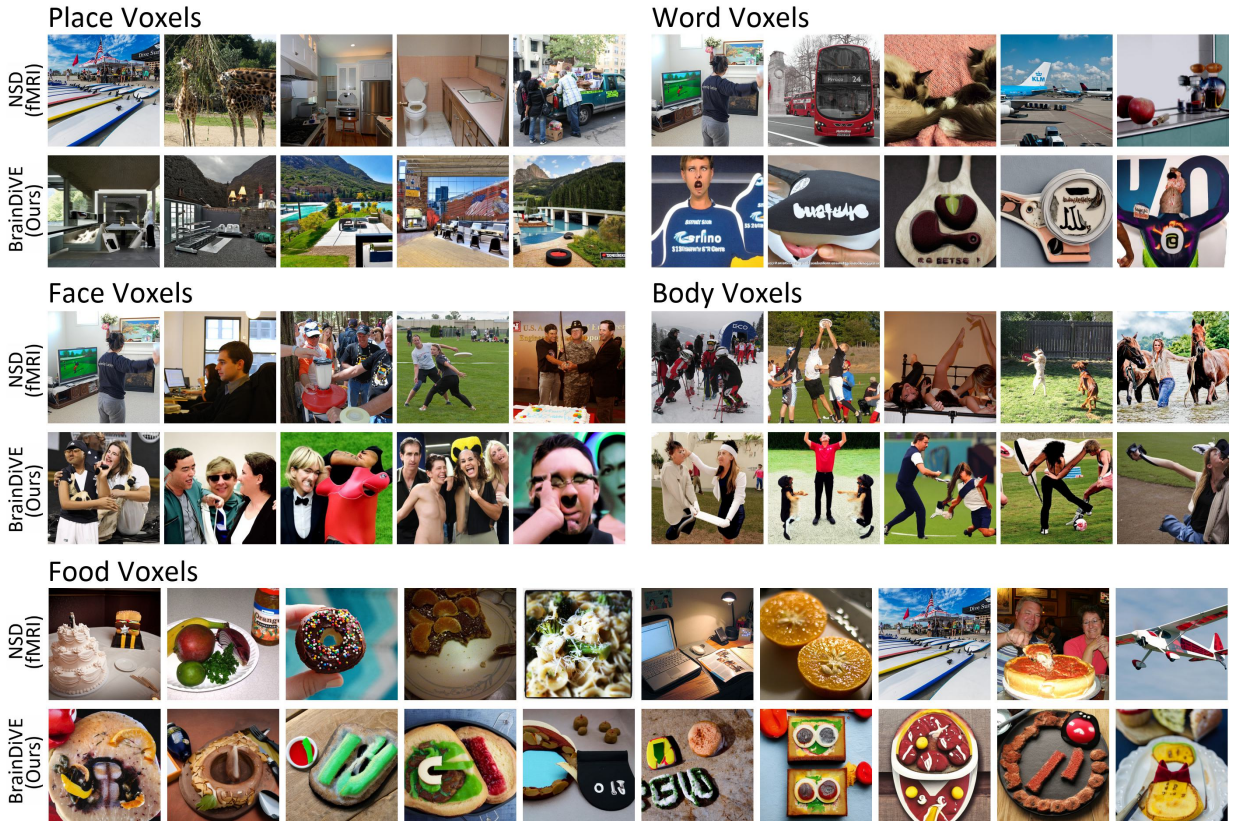


Figure 2.16: **Results for category selective voxels (S6).** We identify the top-5 images from the stimulus set or generated by our method with highest average activation in each set of category selective voxels for the face/place/word/body categories, and the top-10 images for the food selective voxels.

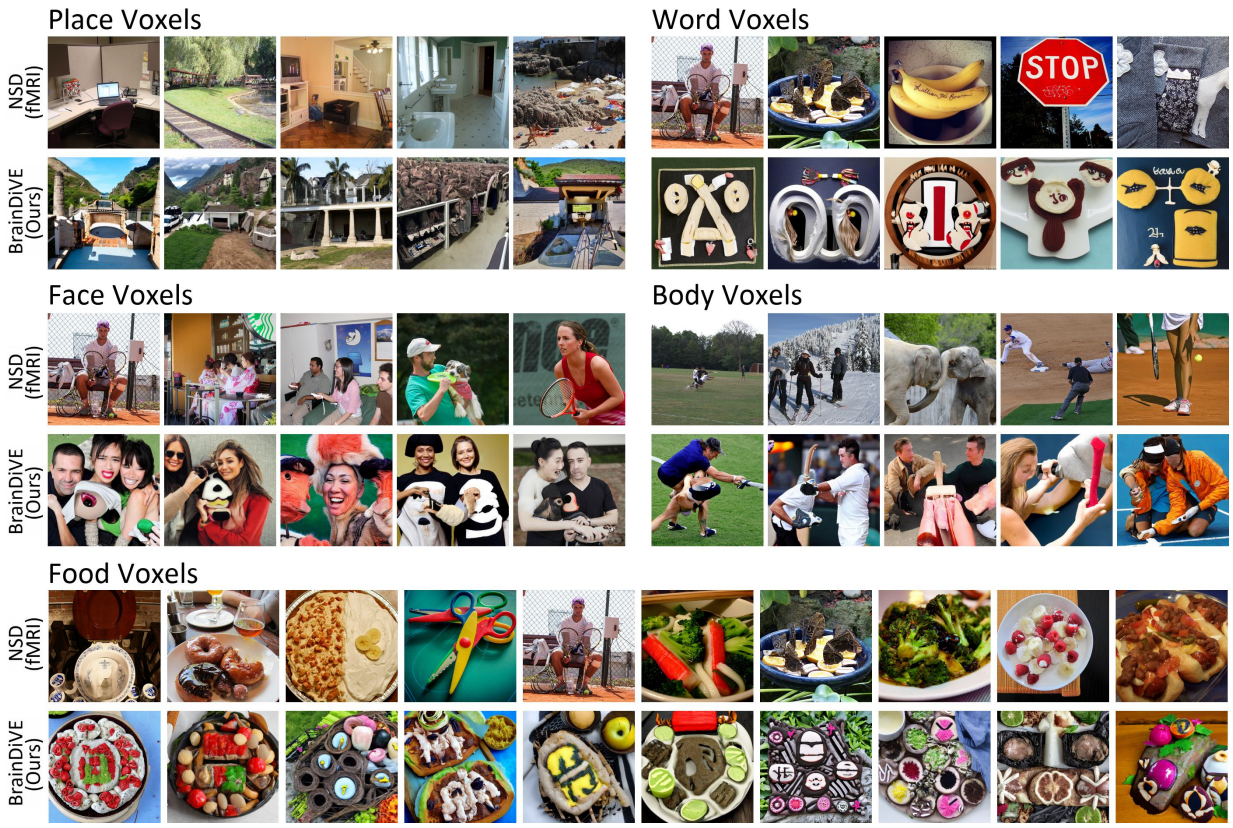


Figure 2.17: **Results for category selective voxels (S7).** We identify the top-5 images from the stimulus set or generated by our method with highest average activation in each set of category selective voxels for the face/place/word/body categories, and the top-10 images for the food selective voxels.

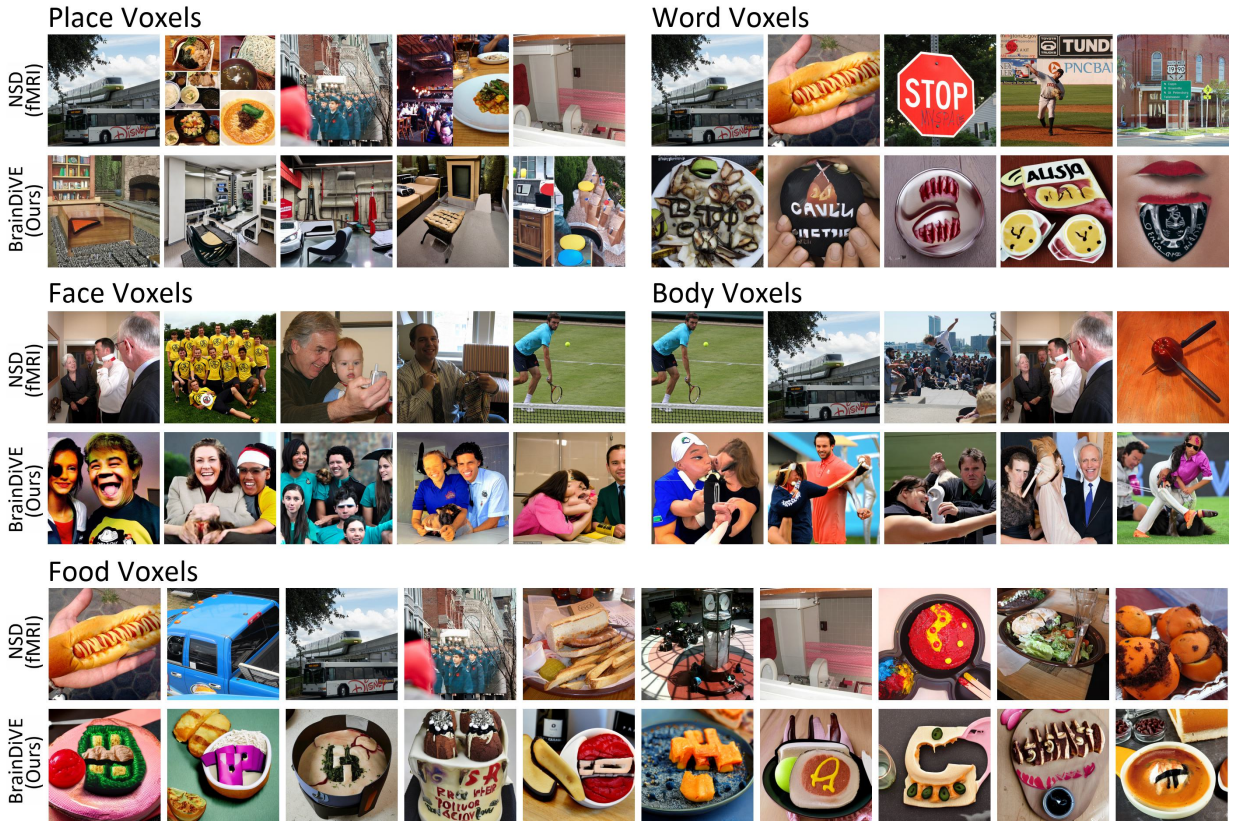


Figure 2.18: **Results for category selective voxels (S8).** We identify the top-5 images from the stimulus set or generated by our method with highest average activation in each set of category selective voxels for the face/place/word/body categories, and the top-10 images for the food selective voxels.

2.6.2 CLIP zero-shot classification

In this section we show the CLIP classification results for S1 – S8, where Table 2.5 in this Supplementary material matches that of Table 1 in the main paper. We use CLIP Radford et al. [2021] to classify images from each ROI into five semantic categories: face/place/body/word/food. Shown is the percentage where the classified category of the image matches the preferred category of the brain region. We show this for the top-200 and top-100 images (top-2% and top-1%) evaluated by mean true fMRI beta, and the top-200 and top-100 (20% and 10%) of BrainDiVE images as self-evaluated by the encoding component of BrainDiVE. Please see Supplementary Section 2.6.13 for the prompts we use for CLIP classification.

	Faces		Places		Bodies		Words		Food		Mean	
	S1↑	S2↑	S1↑	S2↑	S1↑	S2↑	S1↑	S2↑	S1↑	S2↑	S1↑	S2↑
NSD top-200	42.5	41.5	66.5	80.0	56.0	65.0	31.5	34.5	68.0	85.5	52.9	61.3
NSD top-100	40.0	45.0	68.0	79.0	49.0	60.0	30.0	49.0	78.0	85.0	53.0	63.6
BrainDiVE-200	**69.5**	**70.0**	**97.5**	**100**	**75.5**	68.5	**60.0**	57.5	89.0	94.0	**78.3**	75.8
BrainDiVE-100	61.0	68.0	**97.0**	**100**	75.0	**69.0**	**60.0**	**62.0**	**92.0**	**95.0**	77.0	**78.8**

Table 2.5: **Evaluating semantic specificity with zero-shot CLIP classification for S1 and S2**

	Faces		Places		Bodies		Words		Food		Mean	
	S3↑	S4↑	S3↑	S4↑	S3↑	S4↑	S3↑	S4↑	S3↑	S4↑	S3↑	S4↑
NSD top-200	33.0	39.0	74.5	71.5	57.9	47.5	27.0	20.5	49.5	53.5	48.4	46.4
NSD top-100	38.0	41.0	81.0	72.0	60.0	49.0	30.0	25.0	46.0	57.9	51.0	49.0
BrainDiVE-200	**67.5**	**73.5**	99.0	**100**	**59.0**	66.5	**61.0**	31.0	85.0	89.0	74.3	72.0
BrainDiVE-100	67.0	71.0	**100**	**100**	**59.0**	**72.0**	**61.0**	**34.0**	**89.0**	**93.0**	**75.2**	**74.0**

Table 2.6: **Evaluating semantic specificity with zero-shot CLIP classification for S3 and S4**

	Faces		Places		Bodies		Words		Food		Mean	
	S5↑	S6↑	S5↑	S6↑	S5↑	S6↑	S5↑	S6↑	S5↑	S6↑	S5↑	S6↑
NSD top-200	41.0	38.5	89.5	56.9	57.9	56.5	33.5	34.0	77.0	55.5	59.8	48.3
NSD top-100	45.0	46.0	93.0	55.0	54.0	61.0	33.0	32.0	85.0	56.9	62.0	50.2
BrainDiVE-200	**67.0**	**63.0**	99.5	96.0	74.0	66.0	75.0	68.0	83.5	79.0	79.8	74.4
BrainDiVE-100	64.0	57.9	**100**	**99.0**	**77.0**	**72.0**	**80.0**	**75.0**	**87.0**	**83.0**	**81.6**	**77.4**

Table 2.7: **Evaluating semantic specificity with zero-shot CLIP classification for S5 and S6**

	Faces		Places		Bodies		Words		Food		Mean	
	S7↑	S8↑	S7↑	S8↑	S7↑	S8↑	S7↑	S8↑	S7↑	S8↑	S7↑	S8↑
NSD top-200	38.5	34.0	71.0	57.5	61.0	56.5	20.5	24.5	52.0	36.5	48.6	41.8
NSD top-100	35.0	36.0	76.0	48.0	63.0	61.0	26.0	21.0	56.0	37.0	51.2	40.6
BrainDiVE-200	**73.0**	**77.5**	93.5	**94.5**	**65.0**	64.5	**31.0**	**56.5**	85.5	55.5	**69.6**	69.7
BrainDiVE-100	69.0	72.0	**94.0**	94.0	**65.0**	**67.0**	25.0	56.0	**92.0**	**74.0**	69.0	**72.6**

Table 2.8: **Evaluating semantic specificity with zero-shot CLIP classification for S7 and S8.**

2.6.3 Image gradients and synthesis process

In this section, we show examples of the image at each step of the synthesis process. We perform this visualization for face-, place-, body-, word-, and food- selective voxels. Two visualizations are shown for each set of voxels, we use S1 for all visualizations in this section. The diffusion model is guided only by the objective of maximizing a given set of voxels. We observe that coarse image structure emerges very early on from brain guidance. Furthermore, the gradient and diffusion model sometimes work against each other. For example in Figure 2.24 for body voxels, the brain gradient induces the addition of an extra arm, while the diffusion has already generated three natural bodies. Or in Figure 2.25 for word voxels, where the brain gradient attempts to add horizontal words, but they are warped by the diffusion model. Future work could explore early guidance only, as described in "SDEdit" and "MagicMix" Liew et al. [2022], Meng et al. [2021].

2.6.4 Face voxels

We show examples where the end result contains multiple faces (Figure 2.19), or a single face (Figure 2.20).

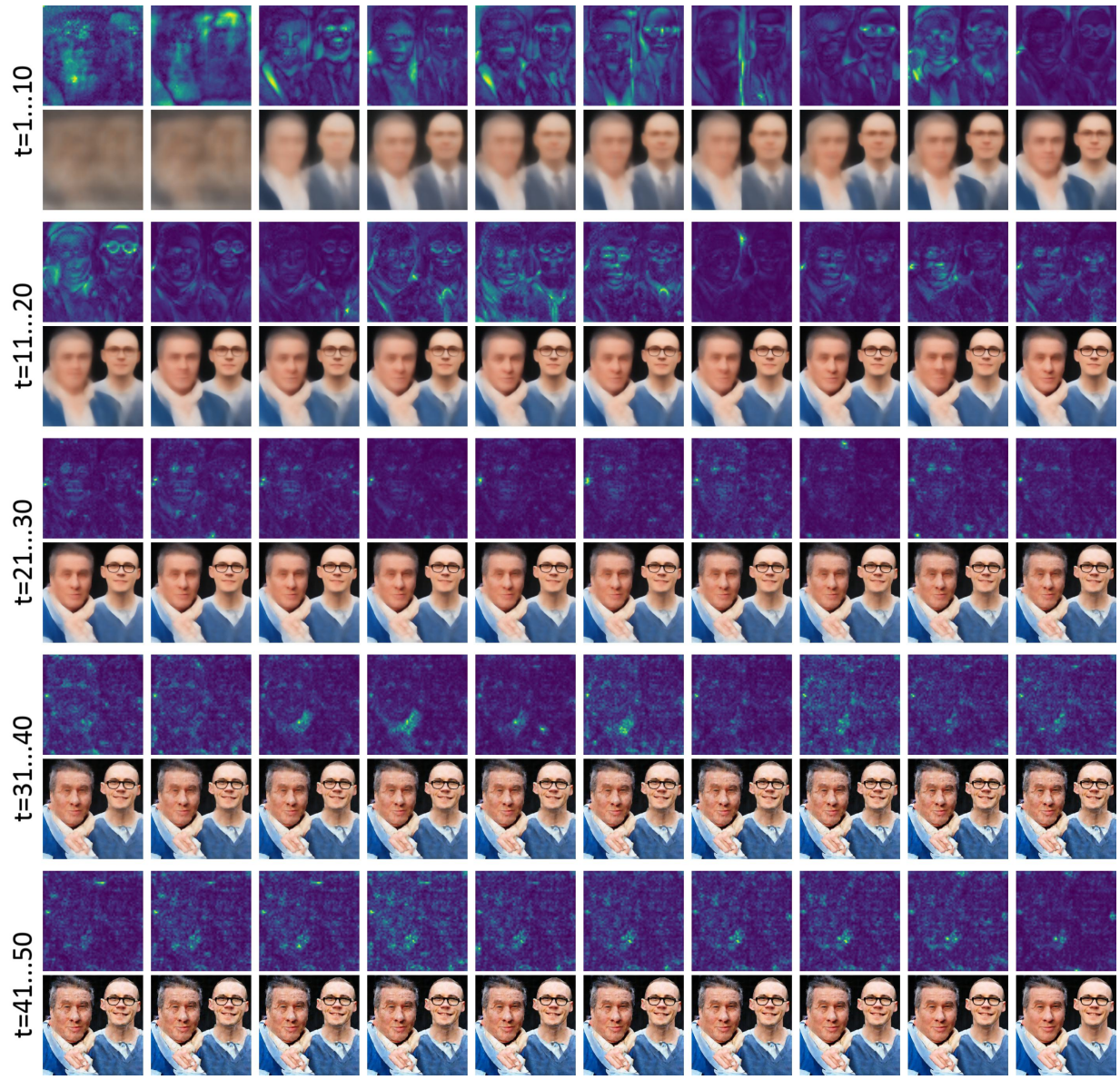


Figure 2.19: **Example 1 of face voxel guided image synthesis for S1.** We utilize 50 steps of Multistep DPM-Solver++. We visualize the gradient magnitude w.r.t. the latent (top, normalized at each step for visualization) and the weighted euler RGB image that the brain encoder accepts (bottom).

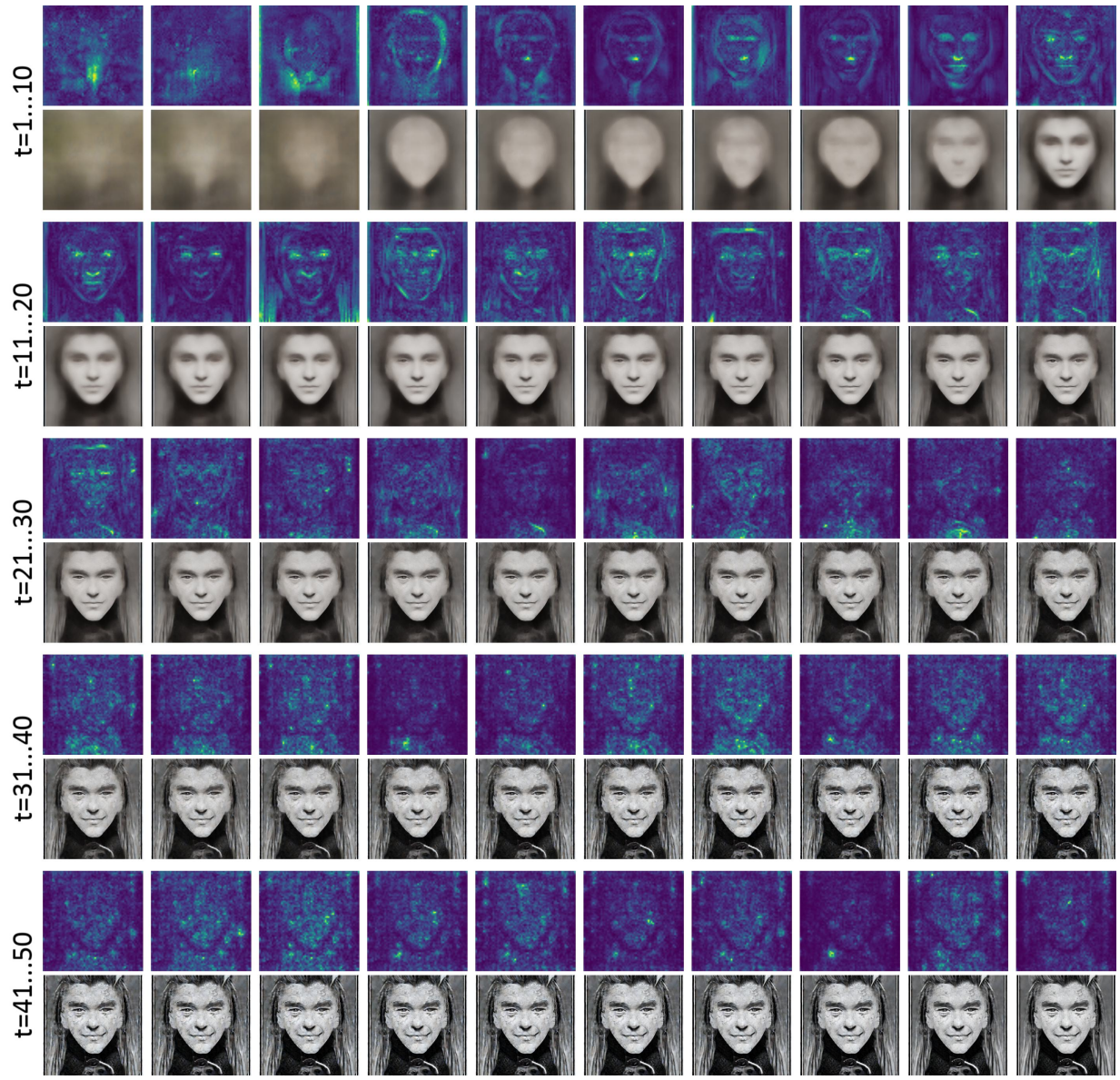


Figure 2.20: **Example 2 of face voxel guided image synthesis for S1.** We utilize 50 steps of Multistep DPM-Solver++. We visualize the gradient magnitude w.r.t. the latent (top, normalized at each step for visualization) and the weighted euler RGB image that the brain encoder accepts (bottom)

2.6.5 Place voxels

We show examples where the end result contains an indoor scene (Figure 2.21), or an outdoor scene (Figure 2.22).

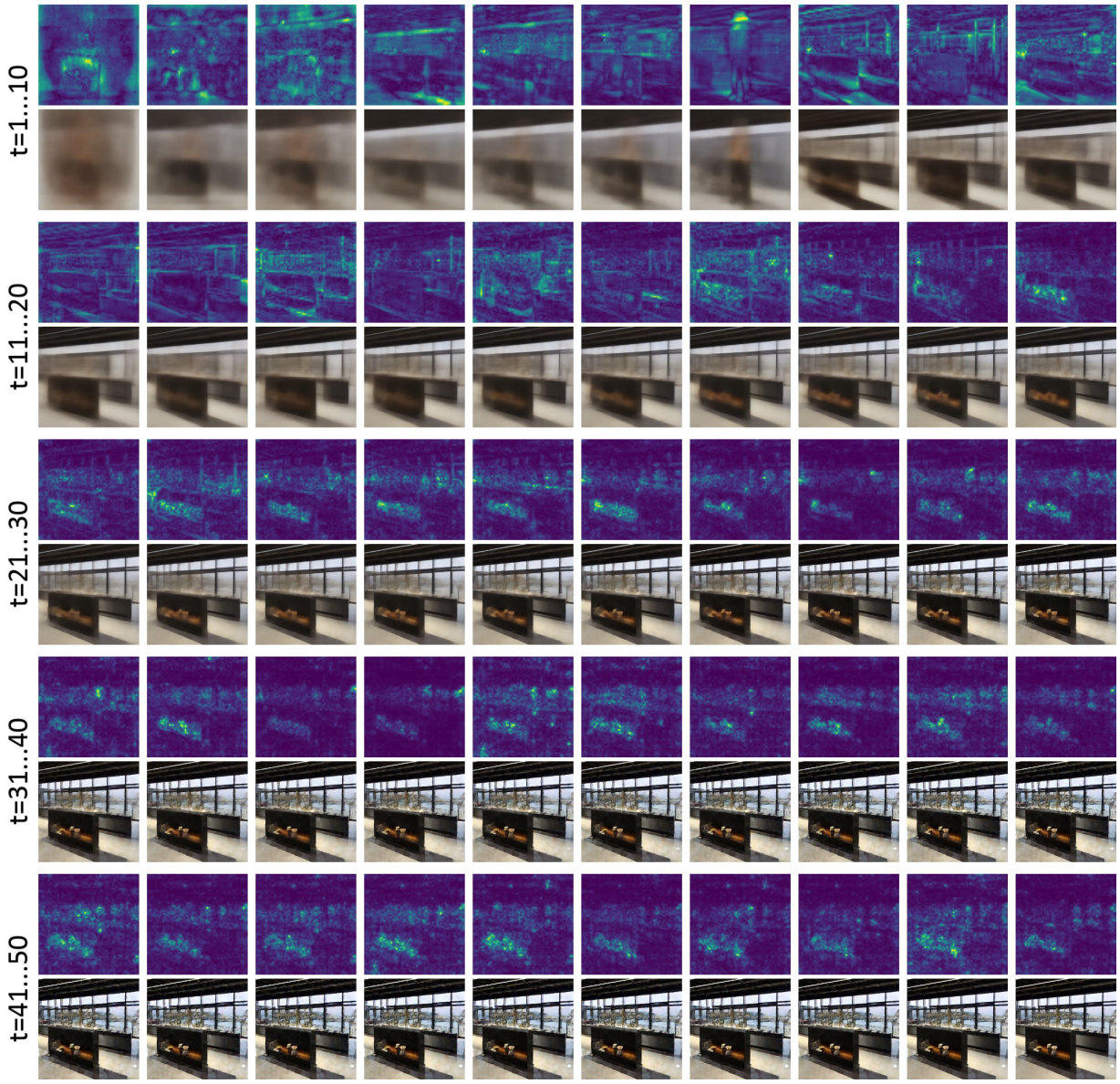


Figure 2.21: **Example 1 of place voxel guided image synthesis for S1.** We utilize 50 steps of Multistep DPM-Solver++. We visualize the gradient magnitude w.r.t. the latent (top, normalized at each step for visualization) and the weighted euler RGB image that the brain encoder accepts (bottom).

Figure 2.22: **Example 2 of place voxel guided image synthesis for S1.** We utilize 50 steps of Multistep DPM-Solver++. We visualize the gradient magnitude w.r.t. the latent (top, normalized at each step for visualization) and the weighted euler RGB image that the brain encoder accepts (bottom).

2.6.6 Body voxels

We show examples where the end result contains an single person's body (Figure 2.23), or an multiple people (Figure 2.24).

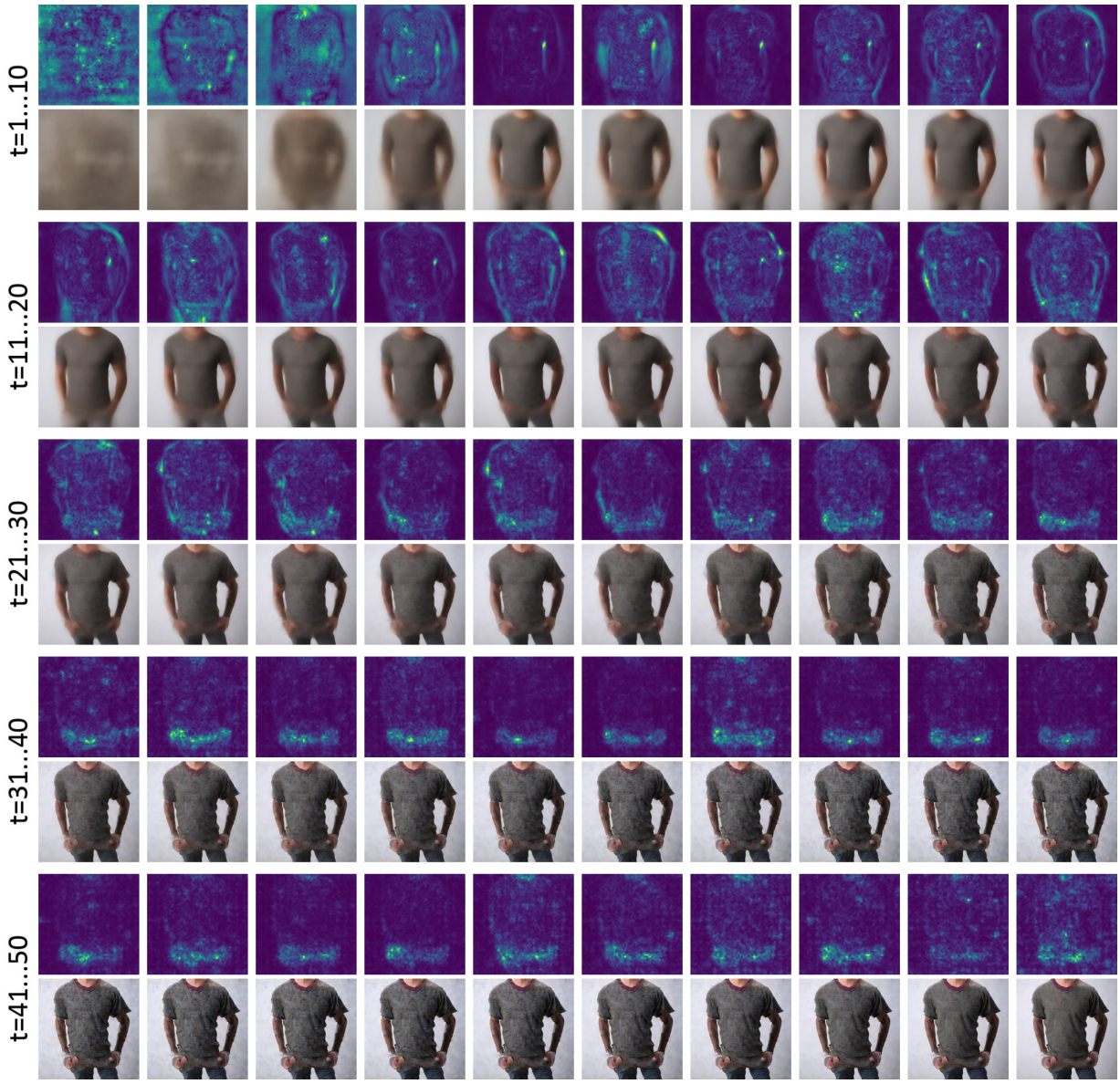


Figure 2.23: **Example 1 of body voxel guided image synthesis for S1.** We utilize 50 steps of Multistep DPM-Solver++. We visualize the gradient magnitude w.r.t. the latent (top, normalized at each step for visualization) and the weighted euler RGB image that the brain encoder accepts (bottom).

Figure 2.24: **Example 2 of body voxel guided image synthesis for S1.** We utilize 50 steps of Multistep DPM-Solver++. We visualize the gradient magnitude w.r.t. the latent (top, normalized at each step for visualization) and the weighted euler RGB image that the brain encoder accepts (bottom).

2.6.7 Word voxels

We show examples where the end result contains recognizable words (Figure 2.25), or glyph like objects (Figure 2.26).

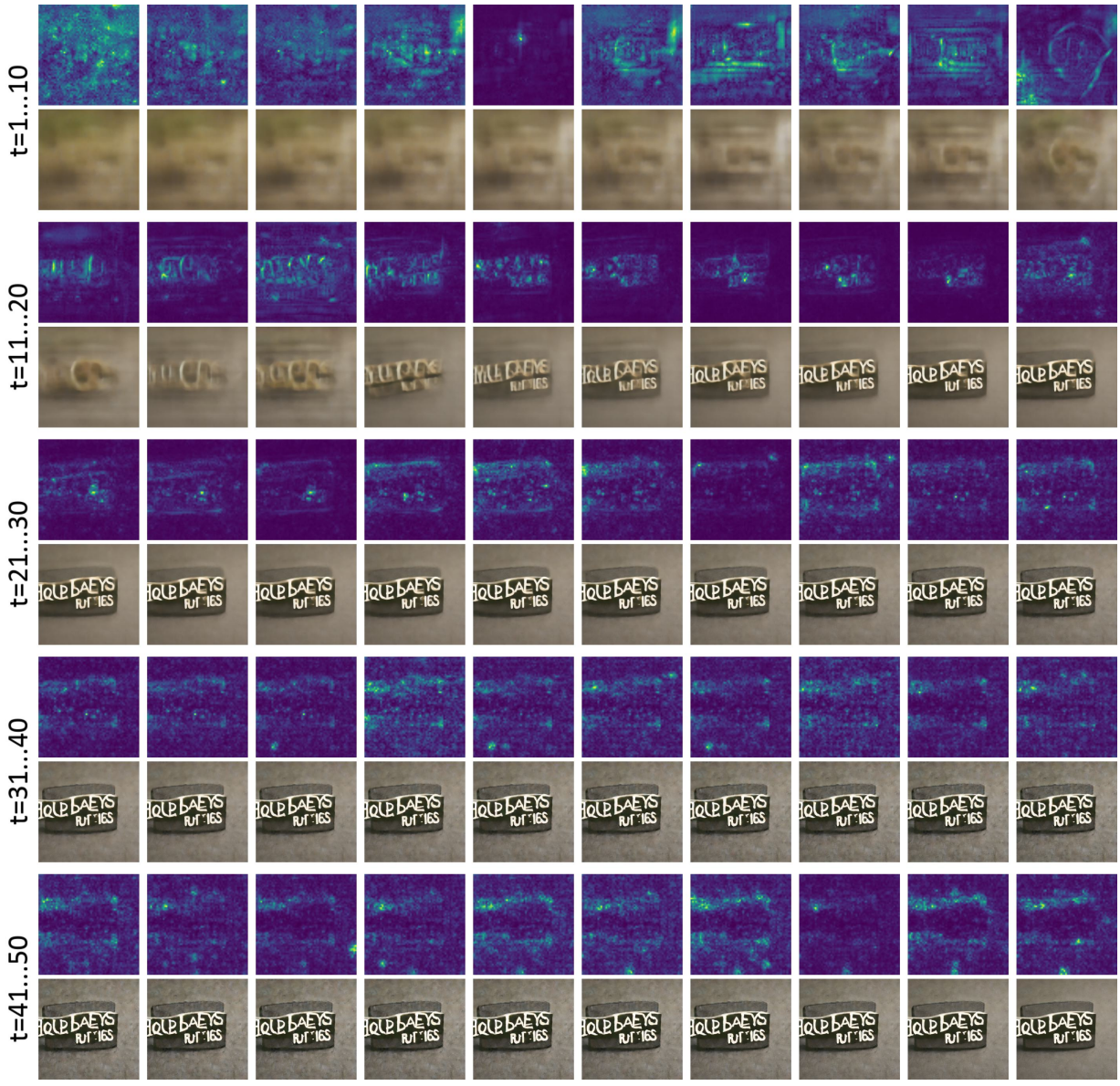


Figure 2.25: **Example 1 of word voxel guided image synthesis for S1.** We utilize 50 steps of Multistep DPM-Solver++. We visualize the gradient magnitude w.r.t. the latent (top, normalized at each step for visualization) and the weighted euler RGB image that the brain encoder accepts (bottom).

Figure 2.26: **Example 2 of word voxel guided image synthesis for S1.** We utilize 50 steps of Multistep DPM-Solver++. We visualize the gradient magnitude w.r.t. the latent (top, normalized at each step for visualization) and the weighted euler RGB image that the brain encoder accepts (bottom).

2.6.8 Food voxels

We show examples where the end result contains highly processed foods (Figure 2.27, showing what appears to be a cake), or cooked food containing vegetables (Figure 2.28).

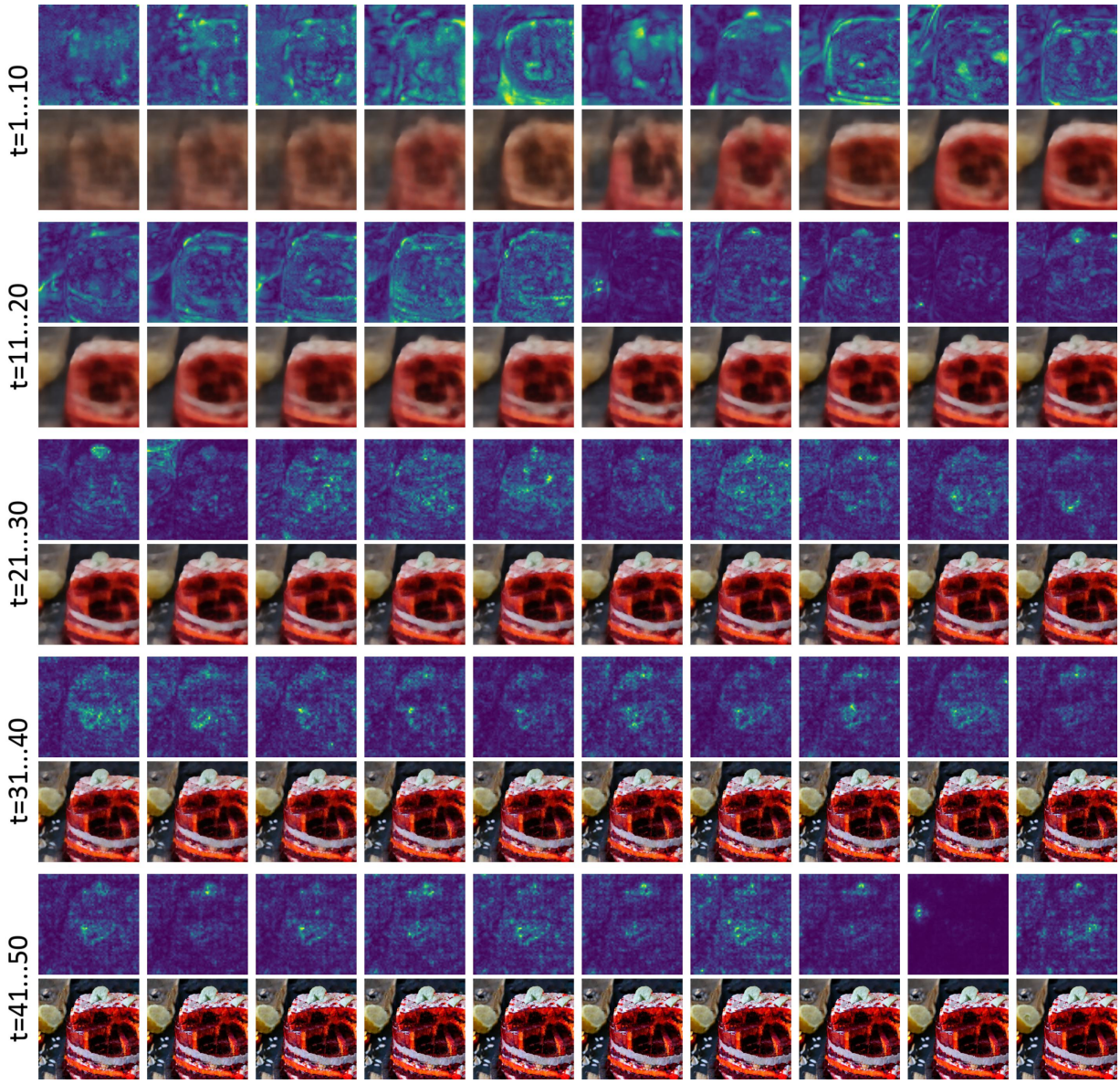


Figure 2.27: **Example 1 of food voxel guided image synthesis for S1.** We utilize 50 steps of Multistep DPM-Solver++. We visualize the gradient magnitude w.r.t. the latent (top, normalized at each step for visualization) and the weighted euler RGB image that the brain encoder accepts (bottom).

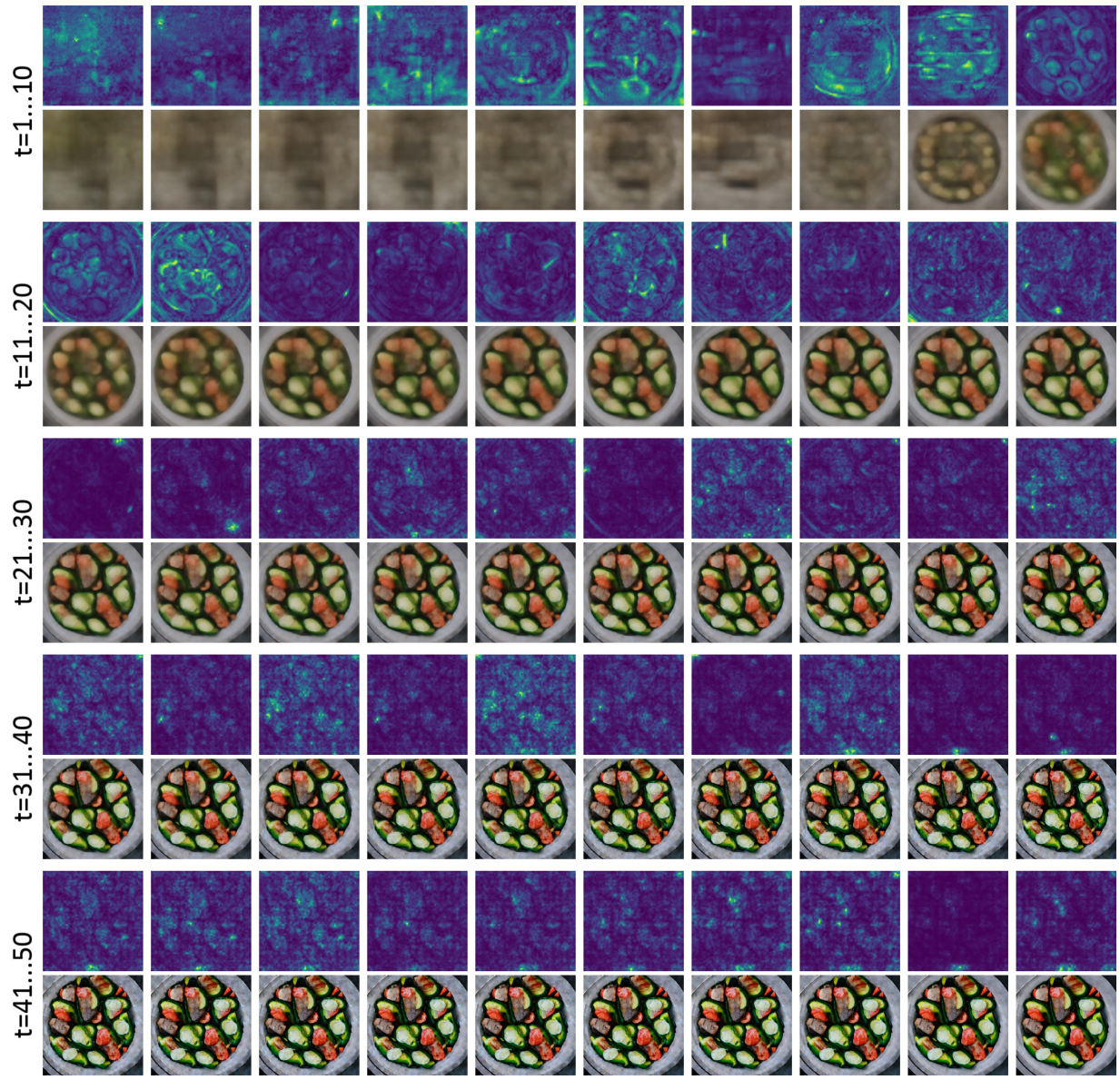


Figure 2.28: **Example 2 of food voxel guided image synthesis for S1.** We utilize 50 steps of Multistep DPM-Solver++. We visualize the gradient magnitude w.r.t. the latent (top, normalized at each step for visualization) and the weighted euler RGB image that the brain encoder accepts (bottom).

2.6.9 Human behavioral study standard error

In this section, we show the human behavioral study results along with the standard error of the responses. Each question was answered by exactly 10 subjects from `prolific.co`. In each table, the results are show in the following format: **Mean(SEM)**. Where **Mean** is the average response, while **SEM** is the standard error of the mean ratio across 10 subjects: (SEM $= \frac{\sigma}{\sqrt{10}}$).

Which ROI has more...	photorealistic faces				animals				abstract shapes/lines			
	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7
FFA-NSD	**45**(7.2)	**43**(8.3)	**34**(6.2)	**41**(6.5)	34(4.5)	34(3.5)	17(4.0)	15(3.8)	21(6.8)	6(4.0)	14(2.9)	22(6.6)
OFA-NSD	25(5.1)	22(6.4)	21(5.6)	18(5.3)	**47**(3.2)	**36**(2.5)	**65**(5.7)	**65**(6.4)	**24**(8.5)	**44**(9.2)	**28**(8.1)	**25**(6.4)
FFA-BrainDiVE	**79**(7.8)	**89**(4.8)	**60**(5.3)	**52**(5.3)	17(5.6)	13(3.5)	21(3.9)	19(2.2)	6(3.2)	11(6.4)	18(4.9)	20(6.6)
OFA-BrainDiVE	11(5.7)	4(2.5)	15(2.9)	22(5.1)	**71**(8.4)	**61**(8.2)	**52**(5.1)	**50**(3.5)	**80**(5.8)	**79**(7.4)	**40**(5.8)	**39**(7.1)

Table 2.9: **Human evaluation of the difference between face-selective ROIs**. Evaluators compare groups of images corresponding to OFA and FFA; comparisons are done within GT and generated images respectively. Questions are posed as: "Which group of images has more X?"; options are FFA/OFA/Same. Results are in %. Note that the "Same" responses are not shown; responses across all three options sum to 100.

Which cluster is more...	vegetables/fruits				healthy			
	S1	S2	S5	S7	S1	S2	S5	S7
Food-1 NSD	17(4.3)	21(4.8)	27(5.1)	36(3.5)	28(5.8)	22(3.7)	29(6.2)	40(4.0)
Food-2 NSD	**65**(7.2)	**56**(6.4)	**56**(5.7)	**49**(3.9)	**50**(7.1)	**47**(4.9)	**54**(6.0)	**45**(4.3)
Food-1 BrainDiVE	11(7.0)	10(6.0)	8(6.6)	11(6.5)	15(6.2)	16(6.0)	20(7.2)	17(7.1)
Food-2 BrainDiVE	**80**(7.3)	**75**(8.0)	**67**(9.8)	**64**(7.4)	**68**(7.7)	**68**(7.3)	**46**(9.3)	**51**(7.8)

Which cluster is more...	colorful				far away			
	S1	S2	S5	S7	S1	S2	S5	S7
Food-1 NSD	19(5.5)	18(6.1)	13(2.8)	27(3.5)	32(6.6)	24(4.7)	23(6.5)	28(4.2)
Food-2 NSD	**42**(6.4)	**52**(5.6)	**53**(6.5)	**42**(6.4)	**34**(7.0)	**39**(8.1)	**36**(7.9)	**42**(7.3)
Food-1 BrainDiVE	6(3.8)	9(5.7)	11(5.7)	16(4.9)	24(6.8)	18(6.4)	27(8.9)	18(6.0)
Food-2 BrainDiVE	**79**(7.9)	**82**(6.9)	**65**(7.6)	**61**(8.9)	**39**(10.1)	**51**(9.0)	**39**(8.8)	**40**(8.8)

Table 2.10: **Human evaluation of the difference between food clusters**. Evaluators compare groups of images corresponding to food cluster 1 (Food-1) and food cluster 2 (Food-2), with questions posed as "Which group of images has/is more X?". Comparisons are done within NSD and generated images respectively. Note that the "Same" responses are not shown; responses across all three options sum to 100. Results are in %.

Which cluster is more...	angular/geometric				indoor			
	S1	S2	S5	S7	S1	S2	S5	S7
OPA-1 NSD	**45**(7.2)	**58**(9.4)	**49**(7.7)	**51**(9.0)	**71**(5.6)	**88**(4.9)	**80**(5.1)	**79**(5.6)
OPA-2 NSD	13(4.0)	12(2.4)	14(2.9)	16(4.5)	7(3.2)	8(3.7)	11(3.0)	14(4.3)
OPA-1 BrainDiVE	**76**(7.8)	**87(8.6)**	**88**(6.6)	**76**(7.8)	**89**(5.6)	**90**(5.7)	**90**(4.7)	**85**(5.3)
OPA-2 BrainDiVE	12(4.9)	3(2.0)	4(1.5)	10(4.2)	7(3.2)	7(3.2)	5(2.1)	8(2.4)

Which cluster is more...	natural				far away			
	S1	S2	S5	S7	S1	S2	S5	S7
OPA-1 NSD	14(3.8)	3(2.0)	9(4.1)	10(2.8)	10(2.4)	1(0.9)	6(2.9)	8(2.4)
OPA-2 NSD	**73**(3.4)	**89**(7.4)	**71**(6.4)	**81**(6.1)	**69**(4.6)	**93**(3.8)	**81**(6.5)	**85**(5.5)
OPA-1 BrainDiVE	6(3.2)	6(1.5)	9(3.6)	6(2,9)	1(0.9)	3(2.8)	3(2.8)	8(5.6)
OPA-2 BrainDiVE	**91**(5.7)	**91**(3.6)	**83**(6.9)	**90**(5.5)	**97**(2.8)	**92**(6.6)	**91**(5.6)	**88**(7.4)

Table 2.11: **Human evaluation of the difference between OPA clusters**. Evaluators compare groups of images corresponding to OPA cluster 1 (OPA-1) and OPA cluster 2 (OPA-2), with questions posed as "Which group of images is more X?". Comparisons are done within NSD and generated images respectively. Note that the "Same" responses are not shown; responses across all three options sum to 100. Results are in %.

2.6.10 Brain encoder R^2

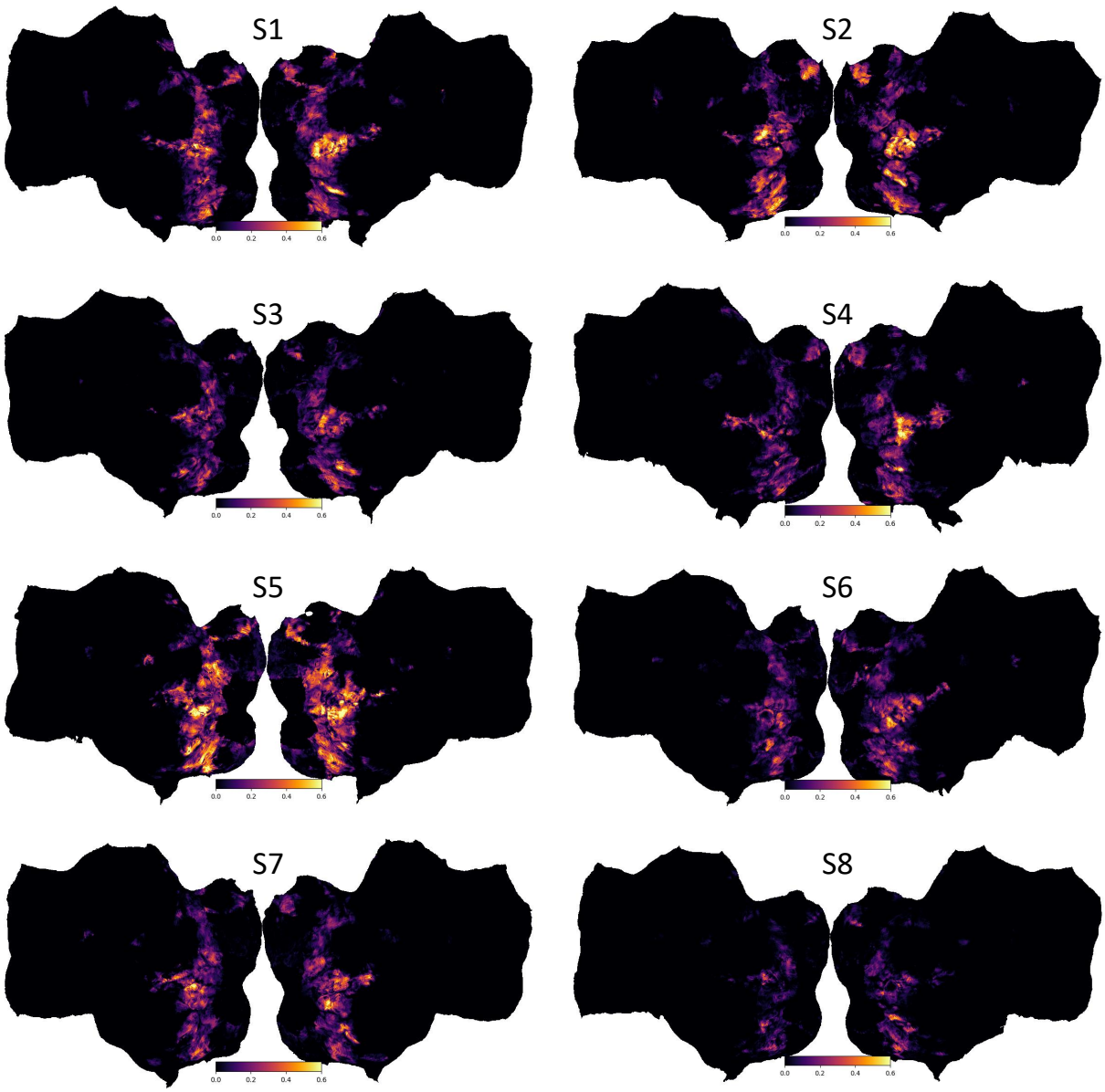


Figure 2.29: **Visualization of R^2 on test set images.** We evaluate R^2 on the $\sim$ 1000 images shared by all subjects. Note that voxels in early visual or outside of higher visual are not modeled.

In Figure 2.6.10 we show the R^2 of the brain encoder as evaluated on the test images. Our brain encoder consists of a CLIP backbone and a linear adaptation layer. We do not model voxels in the early visual cortex, nor do we model voxels outside of higher visual. Our model can

generally achieve high R^2 in regions in known regions of visual semantic selectivity.

2.6.11 OFA and FFA visualizations

In this section, we visualize the top-10 NSD and BrainDiVE images for OFA and FFA. NSD images are selected using the fMRI betas averaged within each ROI. BrainDiVE images are ranked using our predicted ROI activities from 500 images.

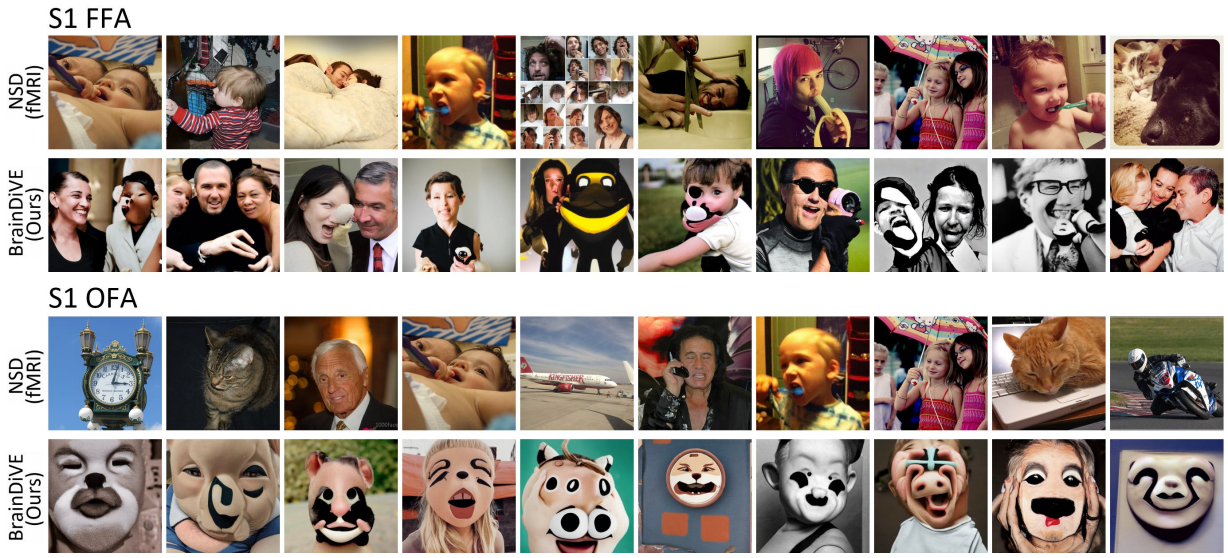


Figure 2.30: **Results for face-selective ROIs in S1.**

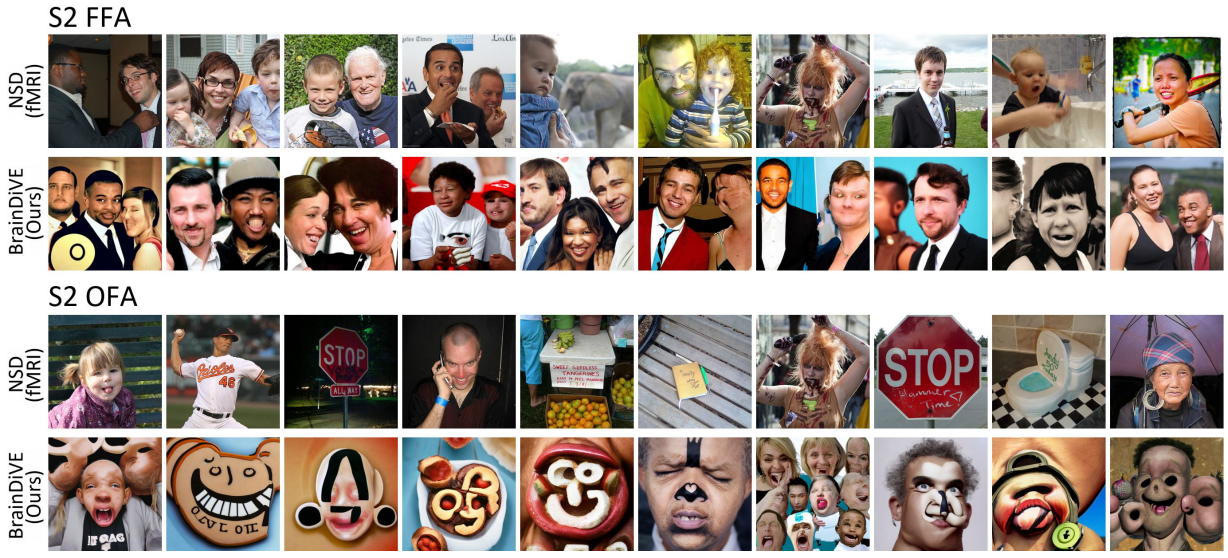


Figure 2.31: **Results for face-selective ROIs in S2.**

Figure 2.32: **Results for face-selective ROIs in S5.**

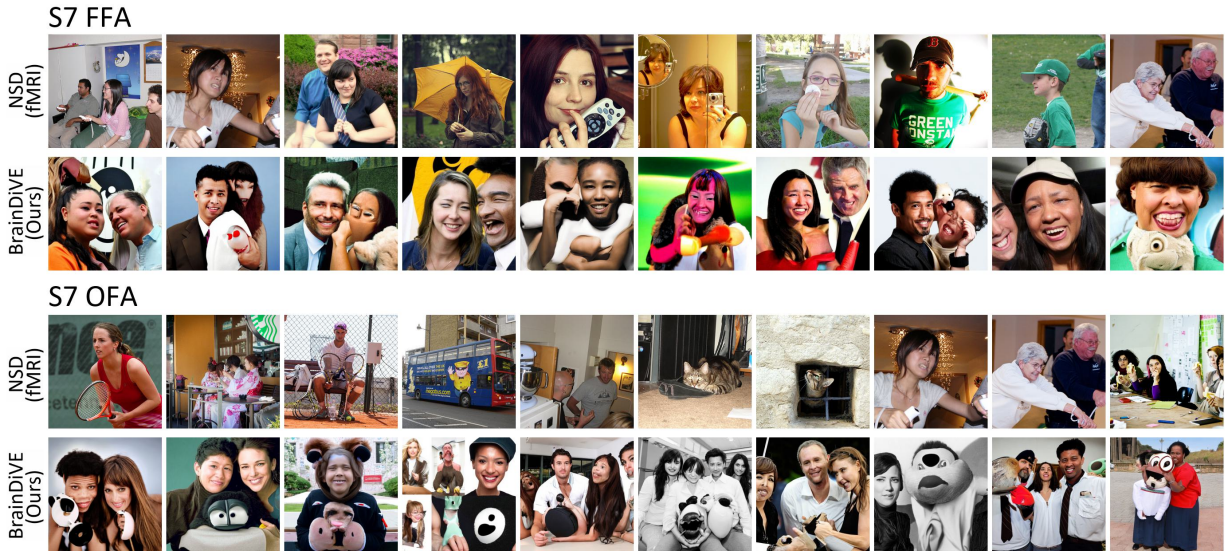


Figure 2.33: **Results for face-selective ROIs in S7.**

2.6.12 OPA and food visualizations

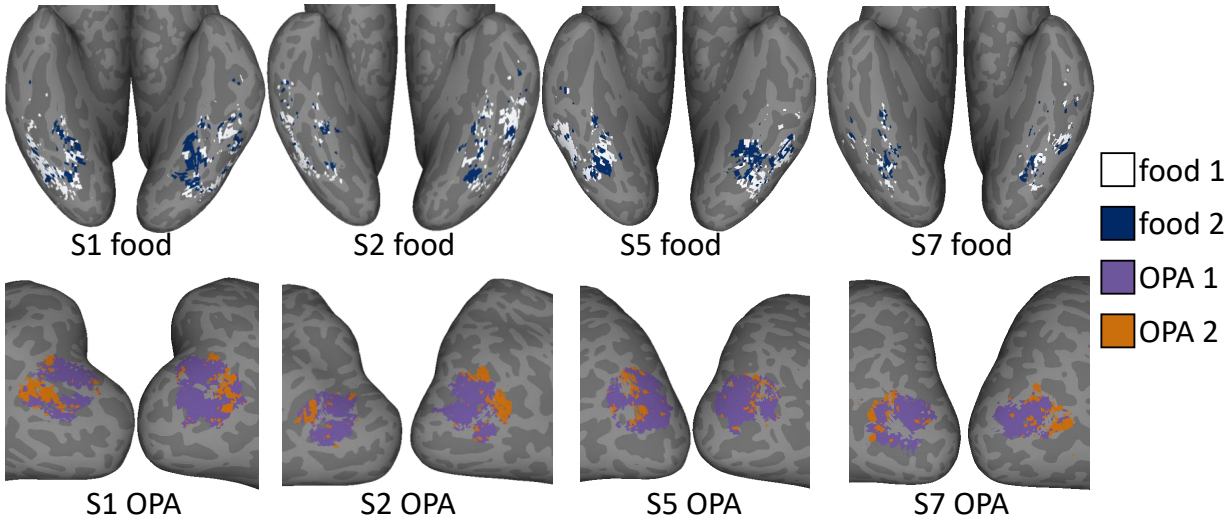


Figure 2.34: **Clustering within the food ROI and within OPA.** Clustering of encoder model weights for each region is shown for four subjects on an inflated cortical surface.

Consistent with Jain et al. [2023], we observe that the food voxels themselves are anatomically variable across subjects, while the two food clusters form alternating patches within the food patches. OPA generally yields anatomically consistent clusters in the four subjects we investigated, with all four subjects showing an anterior-posterior split for OPA.

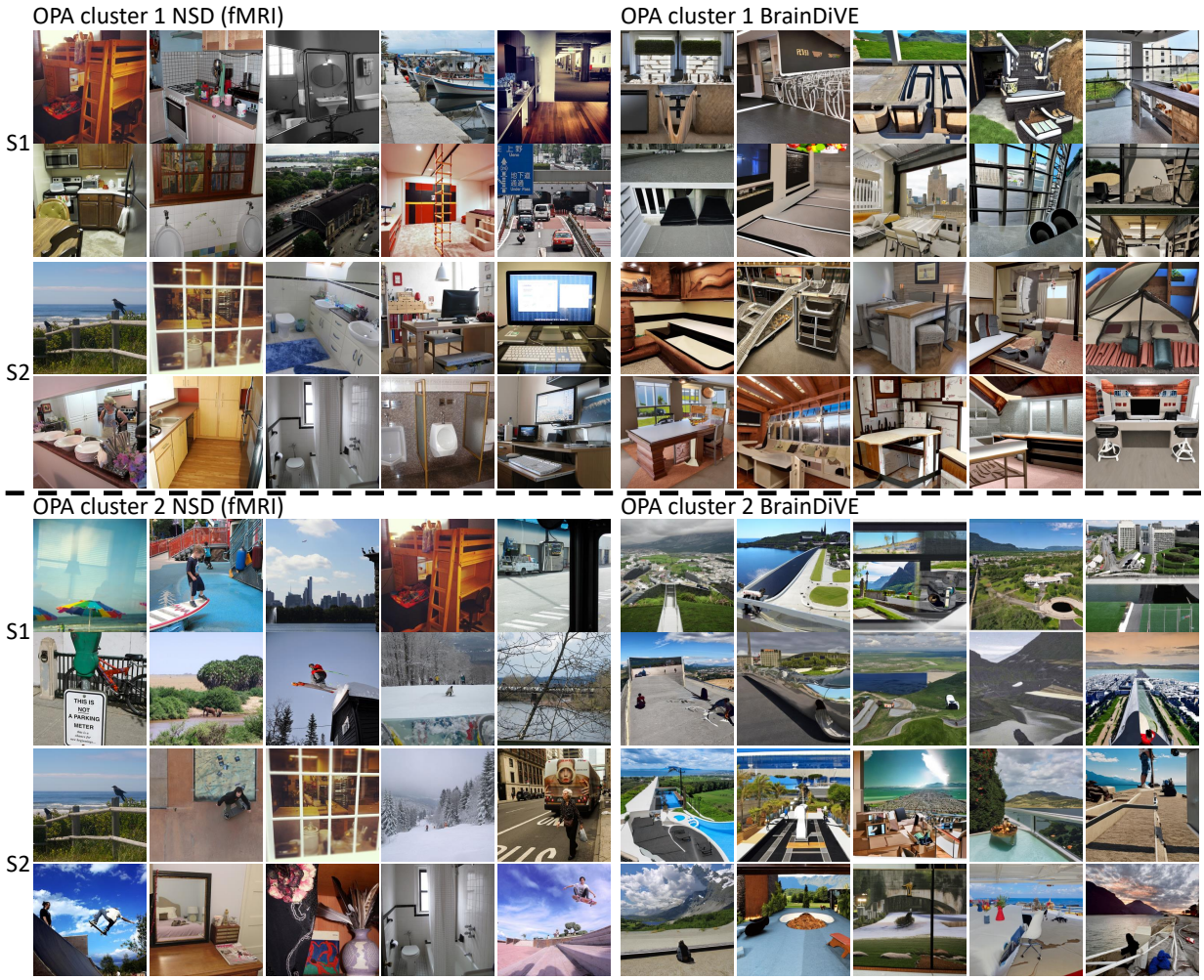


Figure 2.35: **Comparing results across the OPA clusters for S1 and S2.**

Figure 2.36: **Comparing results across the OPA clusters for S5 and S7.**

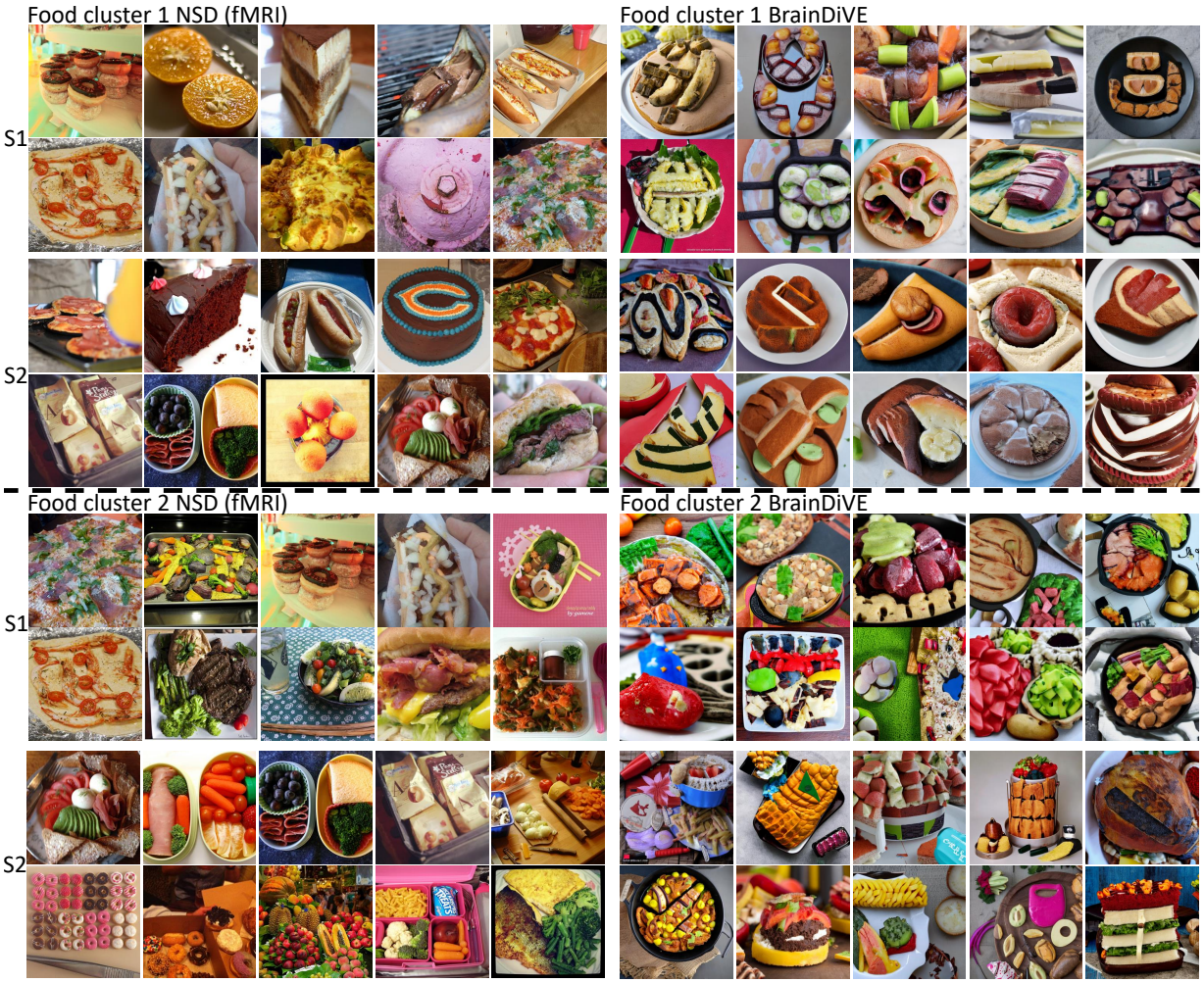


Figure 2.37: **Comparing results across the food clusters for S1 and S2.**

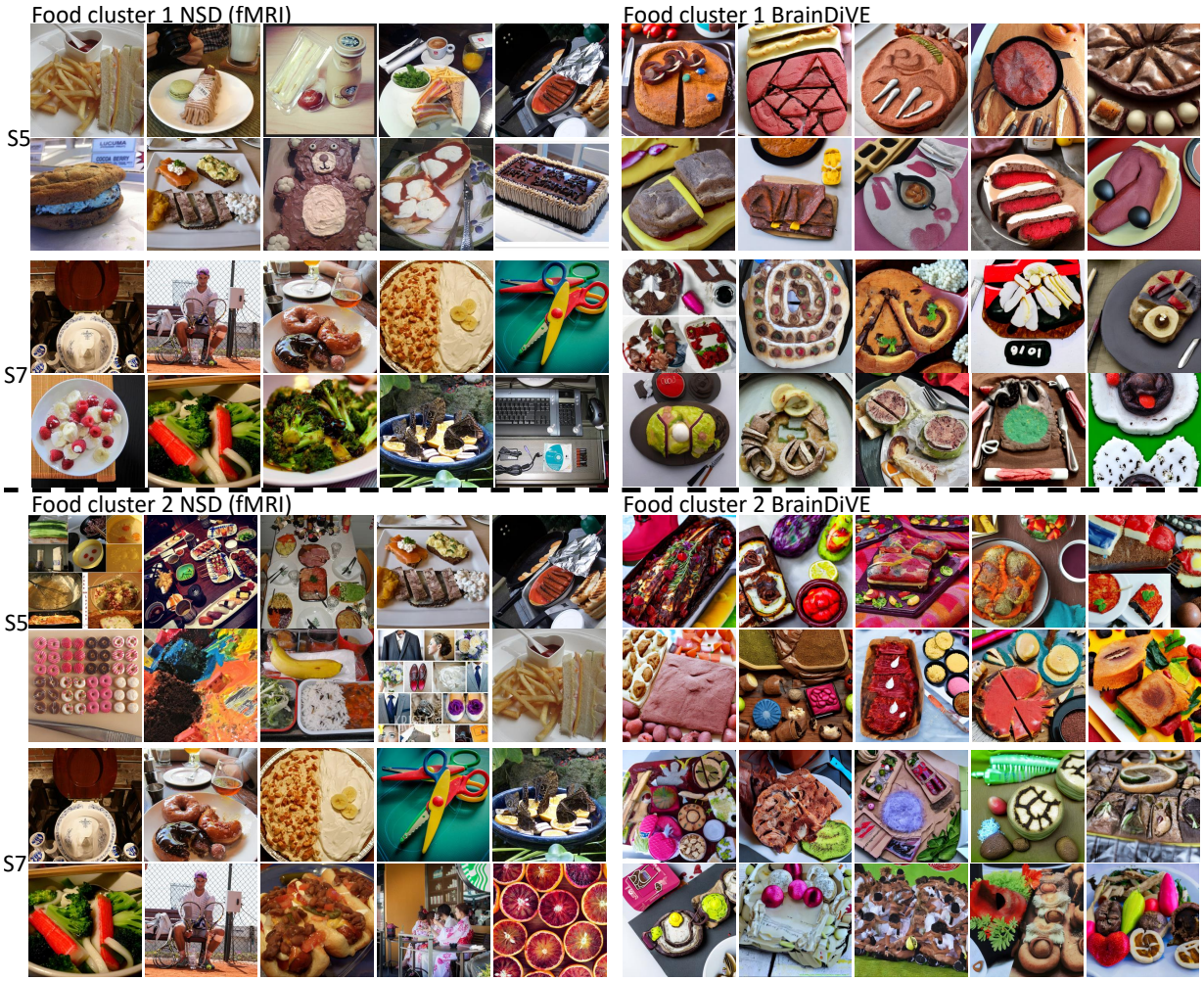


Figure 2.38: **Comparing results across the food clusters for S5 and S7.**

2.6.13 Training, inference, and compute details

Encoder training. Our encoder backbone uses `ViT-B/16` with CLIP pretrained weights `laion2b_s34b_b88k` provided by OpenCLIP [Ilharco et al., 2021, Schuhmann et al., 2022b]. The ViT [Dosovitskiy et al., 2020] weights for the brain encoder are frozen. We train a linear layer consisting of weight and bias to map from the 512 dimensional vector to higher visual voxels B. The CLIP image branch outputs are normalized to the unit sphere.

$$M_\theta(\mathcal{I}) = W \times \frac{\texttt{CLIP}_{\texttt{img}}(\mathcal{I})}{\|\texttt{CLIP}_{\texttt{img}}(\mathcal{I})\|_2} + b$$

Training is done using the `Adam` optimizer [Kingma and Ba, 2014] with learning rate $\texttt{lr}_{\texttt{init}} = 3e-4$ and $\texttt{lr}_{\texttt{end}} = 1.5e-4$, with learning rate adjusting exponentially each epoch. We train for 100 epochs. Decoupled weight decay [Loshchilov and Hutter, 2017] of magnitude $\texttt{decay} = 2e-2$ is applied. Each subject is trained independently using the ~ 9000 images unique to each subject's stimulus set, with R^2 evaluated on the ~ 1000 images shared by all subjects.

During training of the encoder weights, the image is resized to 224×224 to match the input size of `ViT-B/16`. We augment the images by first randomly scaling the pixels by a value between $[0.95, 1.05]$, then normalize the image using OpenCLIP ViT image mean and variance. Prior to input to the network, we further randomly offset the image spatially by up to 4 pixels along the height and width dimensions. The empty pixels are filled in using edge value padding. A small amount of gaussian noise $\mathcal{N}(0, 0.05^2)$ is added to each pixel prior to input to the encoder backbone.

Objective. For all experiments, the objective used is the maximization of a selected set of voxels. Here we will further draw a link between the optimization objective we use and the traditional CLIP text prompt guidance objective [Li et al., 2022b, Nichol et al., 2021]. Recall that M_θ is our brain activation encoder that maps from the image to per-voxel activations. It accepts as input an image, passes it through a ViT backbone, normalizes that vector to the unit sphere,

then applies a linear mapping to go to per-voxel activations. $S \in N$ are the set of voxels we are currently trying to maximize (where N is the set of all voxels in the brain), γ is a step size parameter, and D_Ω is the decoder from the latent diffusion model that outputs an RGB image (we ignore the euler approximation for clarity). Also recall that we use a diffusion model that performs ϵ-prediction.

In the general case, we perturb the denoising process by trying to maximize a set of voxels S:

$$\epsilon'_{theta} = \epsilon_{theta} - \sqrt{1-\alpha_t}\nabla_{x_t}(\frac{\gamma}{|S|}\sum_{i\in S} M_\theta(D_\Omega(x'_t))_i)$$

For the purpose of this section, we will focus on a single voxel first, then discuss the multi-voxel objective.

In our case, the single voxel perturbation is (assuming W is a vector, and that $\langle\cdot,\cdot\rangle$ is the inner product):

$$\begin{aligned}
\epsilon'_{theta} &= \epsilon_{theta} - \sqrt{1-\alpha_t}\nabla_{x_t}(\gamma M_\theta(D_\Omega(x'_t))) \\
&= \epsilon_{theta} - \sqrt{1-\alpha_t}\nabla_{x_t}(\gamma M_\theta(\mathcal{I}_{\text{gen}})) \\
&= \epsilon_{theta} - \gamma\sqrt{1-\alpha_t}\nabla_{x_t}(\langle W, \frac{\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})}{\|\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})\|_2}\rangle + b)
\end{aligned}$$

We can ignore b, as it does not affect optimal $\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})$

$$\equiv \epsilon_{theta} - \gamma\sqrt{1-\alpha_t}\nabla_{x_t}(\langle W, \frac{\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})}{\|\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})\|_2}\rangle)$$

Now let us consider the typical CLIP guidance objective for diffusion models, where $\mathcal{P}_{\text{text}}$ is the guidance prompt, and $\texttt{CLIP}_{\texttt{text}}$ is the text encoder component of CLIP:

$$\epsilon'_{theta} = \epsilon_{theta} - \gamma\sqrt{1-\alpha_t}\nabla_{x_t}(\langle\frac{\texttt{CLIP}_{\texttt{text}}(\mathcal{P}_{\text{text}})}{\|\texttt{CLIP}_{\texttt{text}}(\mathcal{P}_{\text{text}})\|_2}, \frac{\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})}{\|\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})\|_2}\rangle)$$

As such, the W that we find by linearly fitting CLIP image embeddings to brain activation plays the role of a text prompt. In reality, $\|W\|_2 \neq 1$ (but norm is a constant for each voxel), and there is likely no computationally efficient way to "invert" W directly into a human interpretable text prompt. By performing brain guidance, we are essentially using the diffusion model to synthesize an image $\mathcal{I}_{\text{gen}}$ where in addition to satisfying the natural image constraint, the image also attempts to satisfy:

$$\frac{\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})}{\|\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})\|_2} = \frac{W}{\|W\|_2}$$

Or put another way, it generates images where the CLIP latent is aligned with the direction of W. Let us now consider the multi-voxel perturbation, where W_i, b_i is the per-voxel weight vector and bias:

$$\epsilon'_{theta} = \epsilon_{theta} - \sqrt{1-\alpha_t}\nabla_{x_t}(\frac{\gamma}{|S|}\sum_{i\in S} M_\theta(D_\Omega(x'_t))_i)$$

We move $\frac{\gamma}{|S|}$outside of the gradient operation

$$= \epsilon_{theta} - \frac{\gamma}{|S|}\sqrt{1-\alpha_t}\nabla_{x_t}(\sum_{i\in S} M_\theta(D_\Omega(x'_t))_i)$$

$$= \epsilon_{theta} - \frac{\gamma}{|S|}\sqrt{1-\alpha_t}\nabla_{x_t}(\sum_{i\in S}\left[\langle W_i, \frac{\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})}{\|\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})\|_2}\rangle + b_i\right])$$

We again ignore b_i as it does not affect gradient

$$\equiv \epsilon_{theta} - \frac{\gamma}{|S|}\sqrt{1-\alpha_t}\nabla_{x_t}(\sum_{i\in S}\langle W_i, \frac{\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})}{\|\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})\|_2}\rangle)$$

We can move $\sum$ outside due to the distributive nature of gradients

$$= \epsilon_{theta} - \frac{\gamma}{|S|}\sqrt{1-\alpha_t}\sum_{i\in S}\left[\nabla_{x_t}(\langle W_i, \frac{\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})}{\|\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})\|_2}\rangle)\right]$$

Thus from a gradient perspective, the total gradient is the average of gradients from all voxels. Recall that the inner product is a bilinear function, and that the CLIP image latent is on the unit sphere. Then we are generating an image that

$$\frac{\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})}{\|\texttt{CLIP}_{\texttt{img}}(\mathcal{I}_{\text{gen}})\|_2} = \frac{\sum_{i \in S} W_i}{\|\sum_{i \in S} W_i\|_2}$$

Where the optimal image has a CLIP latent that is aligned with the direction of $\sum_{i \in S} W_i$.

Compute. We perform our experiments on a cluster of `Nvidia V100` GPUs in either 16GB or 32GB VRAM configuration, and all experiments consumed approximately $1,500$ compute hours. Each image takes between 20 and 30 seconds to synthesize. All experiments were performed using PyTorch, with cortex visualizations done using PyCortex [Gao et al., 2015].

CLIP prompts. Here we list the text prompts that are used to classify the images for Table 1. in the main paper.

```
face_class = ["A face facing the camera", "A photo of a face", "A
photo of a human face", "A photo of faces", "A photo of a person's
face", "A person looking at the camera", "People looking at the camera","A
portrait of a person", "A portrait photo"]

body_class = ["A photo of a torso", "A photo of torsos", "A photo
of limbs", "A photo of bodies", "A photo of a person", "A photo of
people"]

scene_class = ["A photo of a bedroom", "A photo of an office","A
photo of a hallway", "A photo of a doorway", "A photo of interior
design", "A photo of a building", "A photo of a house", "A photo
```

```
of nature", "A photo of landscape", "A landscape photo", "A photo
of trees", "A photo of grass"]

food_class = ["A photo of food"]

text_class = ["A photo of words", "A photo of glyphs", "A photo of
a glyph", "A photo of text", "A photo of numbers", "A photo of a
letter", "A photo of letters", "A photo of writing", "A photo of
text on an object"]
```

We classify an image as belonging to a category if the image's CLIP latent has highest cosine similarity with the CLIP latent of a prompt belonging to a given category. The same prompts are used to classify the NSD and generated images.

Chapter 3

BrainSCUBA: Fine-Grained Natural Language Captions of Visual Cortex Selectivity

3.1 Introduction

The recognition of complex objects and semantic visual concepts is supported by a network of regions within higher visual cortex. Past research has identified the specialization of certain regions in processing semantic categories such as faces, places, bodies, words, and food [Downing et al., 2001, Epstein and Kanwisher, 1998, Grill-Spector, 2003, Jain et al., 2023, Kanwisher et al., 1997, Khosla et al., 2022a, Maguire, 2001, McCarthy et al., 1997, Pennock et al., 2023b, Puce et al., 1996]. Notably, the discovery of these regions has largely relied on a hypothesis-driven approach, whereby the researcher hand-selects stimuli to study a specific hypothesis. This approach risk biasing the results as it may fail to capture the complexity and variability inherent in real-world images, which can lead to disagreements regarding a region's functional selectivity [Gauthier et al., 1999].

To better address these issues, we introduce ***BrainSCUBA*** (Semantic Captioning Using Brain

Alignments), an approach for synthesizing *per-voxel* natural language captions that describe voxel-wise *preferred stimuli*. Our method builds upon the availability of large-scale fMRI datasets [Allen et al., 2022] with a natural image viewing task, and allows us to leverage contrastive vision-language models and large-language models in identifying fine-grained voxel-wise functional specialization in a data-driven manner. BrainSCUBA is conditioned on weights from an image-computable fMRI encoder that maps from image to voxel-wise brain activations. The design of our encoder allows us to extract the optimal encoder embedding for each voxel, and we use a training-free method to close the modality gap between the encoder-weight space and natural images. The output of BrainSCUBA describes (in words) the visual stimulus that maximally activates a given voxel. Interpretation and visualization of these captions facilitates data-driven investigation into the underlying feature preferences across various visual sub-regions in the brain.

In contrast to earlier studies that decode text from the brain activity related to an image, we demonstrate *voxel-wise captions* of semantic selectivity. Concretely, we show that our method captures the categorical selectivity of multiple regions in visual cortex. Critically, the content of the captions replicates the field's pre-existing knowledge of each region's preferred category. We further show that BrainSCUBA combined with a text-to-image model can generate images semantically aligned with targeted brain regions and yield high predicted activations when evaluated with a different encoder backbone. Finally, we use BrainSCUBA to perform data-driven exploration for the coding of the category "person", finding evidence for person-selective regions outside of the commonly recognized face/body domains and discovering new finer-grained selectivity within known body-selective areas.

3.2 Related Work

Several recent studies have yielded intriguing results by using large-scale vision-language models to reconstruct images and text-descriptions from brain patterns when viewing images [Chen et al., 2022, Doerig et al., 2022, Ferrante et al., 2023, Liu et al., 2023, Ozcelik and VanRullen,

2023, Takagi and Nishimoto, 2022], or to generate novel images that are predicted to activate a given region [Gu et al., 2022, Luo et al., 2023, Ratan Murty et al., 2021]. Broadly speaking, these approaches require conditioning on broad regions of the visual cortex, and have not demonstrated the ability to scale down and enable voxel-level understanding of neural selectivity. Additionally, these methods produce images rather than interpretable captions. Work on artificial neurons [Borowski et al., 2020, Zimmermann et al., 2021] have shown that feature visualization may not be more informative than top images in artificial neural networks. In contrast, our work tackles biological networks which have more noisy top-images that are less conducive to direct analysis, and the synthesis of novel images/captions can act as a source of stimuli for future hypothesis-driven neuroscience studies.

Semantic Selectivity in Higher Visual Cortex. Higher visual cortex in the human brain contains regions which respond selectively to specific categories of visual stimuli, such as faces, places, bodies, words, and food [Cohen et al., 2000, Desimone et al., 1984, Downing et al., 2001, Epstein and Kanwisher, 1998, Grill-Spector, 2003, Jain et al., 2023, Kanwisher et al., 1997, Khosla et al., 2022a, Maguire, 2001, McCarthy et al., 1997, Pennock et al., 2023b, Puce et al., 1996]. These discoveries have predominantly relied on the use of hand-selected stimuli designed to trigger responses of distinct regions. However the handcrafted nature of these stimuli may misrepresent the complexity and diversity of visual information encountered in natural settings [Felsen and Dan, 2005, Gallant et al., 1998]. In contrast, the recent progress in fMRI encoders that map from stimulus to brain response have enabled data-driven computational tests of brain selectivity in vision [Conwell et al., 2023, Eickenberg et al., 2017, Huth et al., 2012, Kubilius et al., 2019, Naselaris et al., 2011, Wang et al., 2022, Wen et al., 2018, Yamins et al., 2014], language [Deniz et al., 2019, Huth et al., 2016], and at the interface of vision and language [Popham et al., 2021]. Here, based on Conwell et al. [2023]'s evaluation of the brain alignment of various pre-trained image models, we employ CLIP as our encoder backbone.

Image-Captioning with CLIP and Language Models. Vision-language models trained with a contrastive loss demonstrate remarkable capability across many discriminative tasks [Cherti et al., 2023, Radford et al., 2021, Sun et al., 2023]. However, due to the lack of a text-decoder, these models are typically paired with an adapted language model in order to produce captions. When captioning, some models utilize the full spatial CLIP embedding [Li et al., 2023a, Shen et al., 2021], whilst others use only the vector embedding [Li et al., 2023b, Mokady et al., 2021, Tewel et al., 2022]. By leveraging the multi-modal latent space learned by CLIP, we are able to generate voxel-wise captions without human-annotated voxel-caption data.

Brain-Conditioned Image and Caption Generation. There are two broad directions when it comes to brain conditioned generative models for vision. The first seeks to decode (reconstruct) visual inputs from the corresponding brain activations, including works that leverage retrieval, variational autoencoders (VAEs), generative adversarial networks (GANs), and score/energy/diffusion models [Chen et al., 2023, Han et al., 2019, Kamitani and Tong, 2005, Lu et al., 2023, Ozcelik and VanRullen, 2023, Ren et al., 2021, Seeliger et al., 2018, Shen et al., 2019, Takagi and Nishimoto, 2022]. Some approaches further utilize or generate captions that describe the observed visual stimuli [Doerig et al., 2022, Ferrante et al., 2023, Liu et al., 2023, Mai and Zhang, 2023, Scotti et al., 2024].

The second approach seeks to generate stimuli that *activates* a given region rather than exactly reconstructing the input [Bashivan et al., 2019, Walker et al., 2019]. Some of these approaches utilize GANs or Diffusion models to constrain the synthesized output [Gu et al., 2022, Luo et al., 2023, Ponce et al., 2019, Ratan Murty et al., 2021]. BrainSCUBA falls under the broad umbrella of this second approach. But unlike prior methods which were restricted to modeling broad swathes of the brain, our method can be applied at voxel-level, and can output concrete interpretable captions.

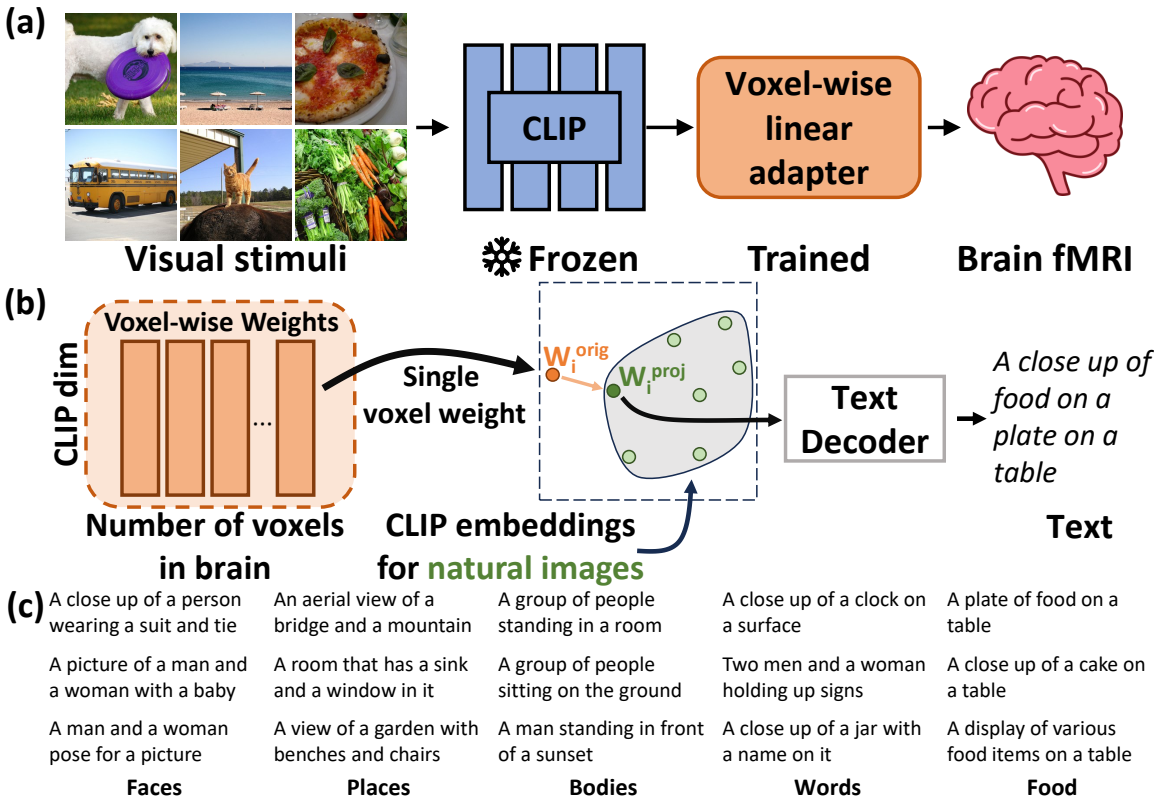


Figure 3.1: **Architecture of BrainSCUBA**. **(a)** Our framework relies on an fMRI encoder trained to map from images to voxel-wise brain activations. The encoder consists of a frozen CLIP image network with a unit norm output and a linear probe. **(b)** We decode the voxel-wise weights by projecting the weights into the space of CLIP embeddings for natural images followed by sentence generation. **(c)** Select sentences from each region, please see experiments for a full analysis.

3.3 Methods

We aim to generate fine-grained (voxel-level) natural language captions that describe a visual scene which maximally activate a given voxel. We first describe the parameterization and training of our voxel-wise fMRI encoder which goes from images to brain activations. We then describe how we can analytically derive the optimal CLIP embedding given the encoder weights. Finally, we describe how we close the gap between optimal CLIP embeddings and the natural image embedding space to enable voxel-conditioned caption generation. We illustrate our framework in Figure 3.1.

3.3.1 Image-to-Brain Encoder Construction

An image-computable brain encoder is a learned function F_θ that transforms an image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$ to voxel-wise brain activation beta values represented as a $1D$ vector of N brain voxels $B \in \mathbb{R}^{1 \times N}$, where $F_\theta(\mathcal{I}) \Rightarrow B$. Recent work identified models trained with a contrastive vision-language objective as the highest performing feature extractor for visual cortex, with later CLIP layers being more accurate for higher visual areas [Conwell et al., 2023, Wang et al., 2022]. As we seek to solely model higher-order visual areas, we utilize a two part design for our encoder. First is a frozen CLIP [Radford et al., 2021] backbone which outputs a $R^{1 \times M}$ dimensional embedding vector for each image. The second is a linear probe $W \in \mathcal{R}^{M \times N}$ with bias $b \in \mathcal{R}^{1 \times N}$, which transform a unit-norm image embedding to brain activations.

$$\left[\frac{\text{CLIP}_{\text{img}}(\mathcal{I})}{\|\text{CLIP}_{\text{img}}(\mathcal{I})\|_2} \times W + b \right] \Rightarrow B \tag{3.1}$$

After training with MSE loss, we evaluate the encoder on the test set in Figure 3.2(a) and find that our encoder can achieve high R^2.

3.3.2 Deriving the Optimal Embedding and Closing the Gap

The fMRI encoder we construct utilizes a linear probe applied to a unit-norm CLIP embedding. It follows from the design that the maximizing embedding e_i^* for a voxel i can be derived efficiently from the weight, and the predicted activation is upper bounded by $\|W_i\|_2 + b$ when

$$e_i^* = \frac{W_i}{\|W_i\|_2} \tag{3.2}$$

In practice, a natural image $\mathcal{I}^*$ that achieves $\frac{\text{CLIP}_{\text{img}}(\mathcal{I}^*)}{\|\text{CLIP}_{\text{img}}(\mathcal{I}^*)\|_2} = e_i^*$ does not typically exist. There is a modality gap between the CLIP embeddings of natural images and the optimal embedding derived from the linear weight matrix. We visualize this gap in Figure 3.2(b) in a joint UMAP [McInnes et al., 2018] fitted on CLIP ViT-B/32 embeddings and fMRI encoder weights, both normalized to

unit-norm. To close this modality gap, we utilize a softmax weighted sum to project the voxel weights onto the space of natural images. Let the original voxel weight be $W_i^{\text{orig}} \in R^{1\times M}$, which we will assume to be unit-norm for convenience. We have a set with K natural images $M = \{M_1, M_2, M_3, \cdots, M_K\}$. For each image, we compute the CLIP embedding $e_j = \text{CLIP}_{\text{img}}(M_j)$. Given W_i^{orig}, we use cosine similarity followed by softmax with temperature τ to compute a score that sums to 1 across all images. For each weight W_i^{orig} and example image M_j:

$$\text{Score}_{i,j} = \frac{\exp(W_i^{\text{orig}} e_j^T/\tau)}{\exp(\sum_{k=1}^{K} W_i^{\text{orig}} e_k^T/\tau)} \tag{3.3}$$

We parameterize W_i^{proj} using a weighted sum derived from the scores, applied to the norms and directions of the image embeddings:

$$W_i^{\text{proj}} = \left(\sum_{k=1}^{K} \text{Score}_{i,k} * \|e_k\|_2\right) * \left(\sum_{k=1}^{K} \text{Score}_{i,k} * \frac{e_k}{\|e_k\|_2}\right) \tag{3.4}$$

In Figure 3.2(c) we show the cosine similarity between W_i^{orig} and W_i^{proj} as we increase the size of M. This projection operator can be treated as a special case of dot-product attention [Vaswani et al., 2017], with query $= W_i^{\text{orig}}$, key $= \{e_1, e_2, \cdots, e_K\}$, and value equal to norm or direction of $\{e_1, e_2, \cdots, e_K\}$. A similar approach is leveraged by Li et al. [2023b], which shows a similar operator outperforms nearest neighbor search for text-only caption inference. As W_i^{proj} lies in the space of CLIP embeddings for natural images, this allows us to leverage any existing captioning system that is solely conditioned on the final CLIP embedding of an image. We utilize a frozen CLIPCap network, consisting of a projection layer and finetuned GPT-2 [Mokady et al., 2021].

3.4 Results

In this section, we utilize BrainSCUBA to generate voxel-wise captions and demonstrate that it can capture the selectivity in different semantic regions in the brain. We first show that the generated

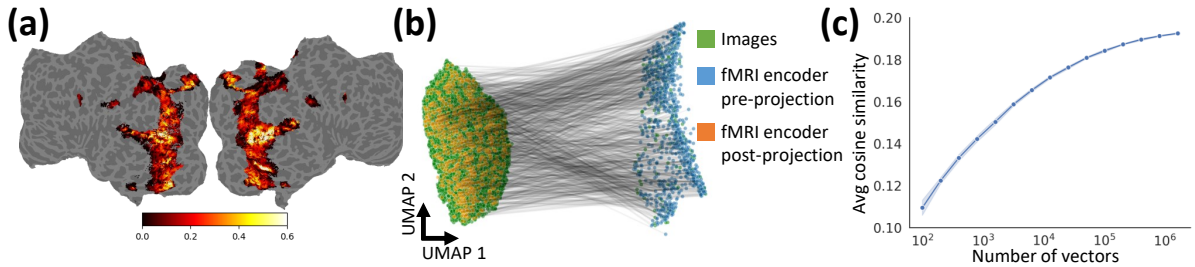


Figure 3.2: **Projection of fMRI encoder weights**. **(a)** We validate the encoder R^2 on a test set, and find it can achieve high accuracy in the higher visual cortex. **(b)** The joint-UMAP of image CLIP embeddings, and pre-/post-projection of the encoder. All embeddings are normalized before UMAP. **(c)** We measure the average cosine similarity between pre-/post-projection weights, and find it increases as the images used are increased. Standard deviation of 5 projections shown in light blue.

nouns are interpretable across the entire brain and exhibit a high degree of specificity within pre-identified category-selective regions. Subsequently, we use the captions as input to text-to-image diffusion models to generate novel images, and confirm the images are semantically consistent within their respective regions. Finally, we utilize BrainSCUBA to analyze the distribution of person representations across the brain to offer novel neuroscientific insight. These results illustrate BrainSCUBA's ability to characterize human visual cortical populations, rendering it a promising framework for exploratory neuroscience.

3.4.1 Setup

We utilize the Natural Scenes Dataset (NSD; Allen et al. [2022]), the largest whole-brain 7T human visual stimuli dataset. Of the 8 subjects, 4 subjects viewed the full $10,000$ image set repeated $3\times$. We use these subjects, S1, S2, S5, S7, for experiments in the main paper, and present additional results in the appendix. The fMRI activations (betas) are computed using GLMSingle [Prince et al., 2022], and further normalized so each voxel's response is $\mu = 0, \sigma^2 = 1$ on a session basis. The response across repeated viewings of the same image is averaged. The brain encoder is trained on the ~ 9000 unique images for each subject, while the remaining ~ 1000 images viewed by all are used to validate R^2.

The unpaired image projection set is a 2 million combination of LAION-A v2 ($6+$ subset) and Open Images [Kuznetsova et al., 2020, Schuhmann et al., 2022a]. We utilize OpenAI's ViT-B/32

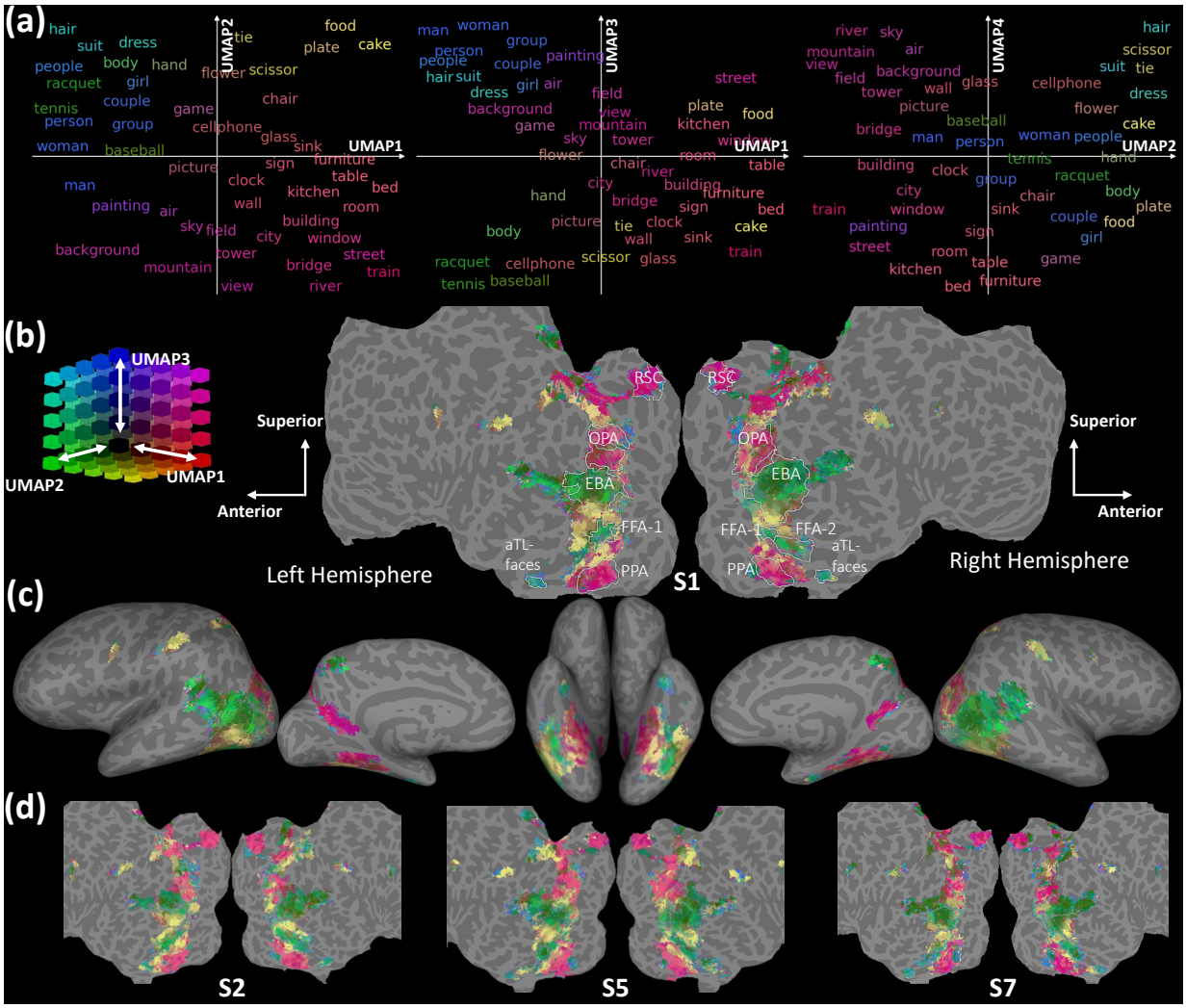


Figure 3.3: **Interpreting the nouns generated by BrainSCUBA** . We take the projected encoder weights and fit a UMAP transform that goes to 4-dims. **(a)** The 50 most common noun embeddings across the brain are projected & transformed using the fMRI UMAP. **(b)** Flatmap of **S1** with ROIs labeled. **(c)** Inflated view of **S1**. **(d)** Flatmaps of **S2, S5, S7**. We find that BrainSCUBA nouns are aligned to previously identified functional regions. Shown here are body regions (EBA), face regions (FFA-1/FFA-2/aTL-faces), place regions (RSC/OPA/PPA). Note that the yellow near FFA match the food regions identified by Jain et al. [2023]. The visualization style is inspired by Huth et al. [2016].

for the encoder backbone and embedding computation as this is the standard for CLIP conditioned caption generation. For image generation, we use the same model as used by Luo et al. [2023] in BrainDiVE, `stable-diffusion-2-1-base` with 50 steps of second order DPM-Solver++. In order to ensure direct comparability with BrainDiVE results, OpenCLIP's CoCa ViT-L/14 is used for image retrieval and zero-shot classification. We define face/place/body/word regions

using independent category localizer data provided with the NSD by Allen et al. [2022] (threshold of $t > 2$), and use the masks provided by Jain et al. [2023] to define the food regions. For details on the human study, please see the appendix.

3.4.2 Voxel-Wise Text Generations

In this section, we first investigate how BrainSCUBA outputs conceptually tile the higher visual cortex. We perform part-of-speech (POS) tagging and lemmatization of the BrainSCUBA output for four subjects, and extract the top-50 nouns. To extract noun specific CLIP embeddings, we reconstitute them into sentences of the form "A photo of a/an [NOUN]" as suggested by CLIP. Both the noun embeddings and the brain encoder voxel-wise weights are projected to the space of CLIP image embeddings and normalized to the unit-sphere for UMAP. We utilize UMAP fit on the encoder weights for S1. Results are shown in Figure 3.3. We observe that the nouns generated by BrainSCUBA are conceptually aligned to pre-identified functional regions. Namely, voxels in extrastriate body area (EBA) are selective to nouns that indicate bodies and activities (green), fusiform face area (FFA-1/FFA-2) exhibits person/body noun selectivity (blue-green), place regions – retrosplenial cortex (RSC), occipital place area (OPA), and parahippocampal place area (PPA) – show selectivity for scene elements (magenta), and the food regions (yellow; Jain et al. [2023]) surrounding FFA exhibit selectivity for food-related nouns. These results show that our framework can characterize the broad semantic selectivity of visual cortex in a zero-shot fashion.

We further quantify the top-10 nouns within each broad category selective region (Figure 3.4). We observe that BrainSCUBA generates nouns that are conceptually matched to the expected preferred category of each region. Note the multimodal selectivity for words/people/food within the word region has also been observed by Khosla and Wehbe [2022], Mei et al. [2010].

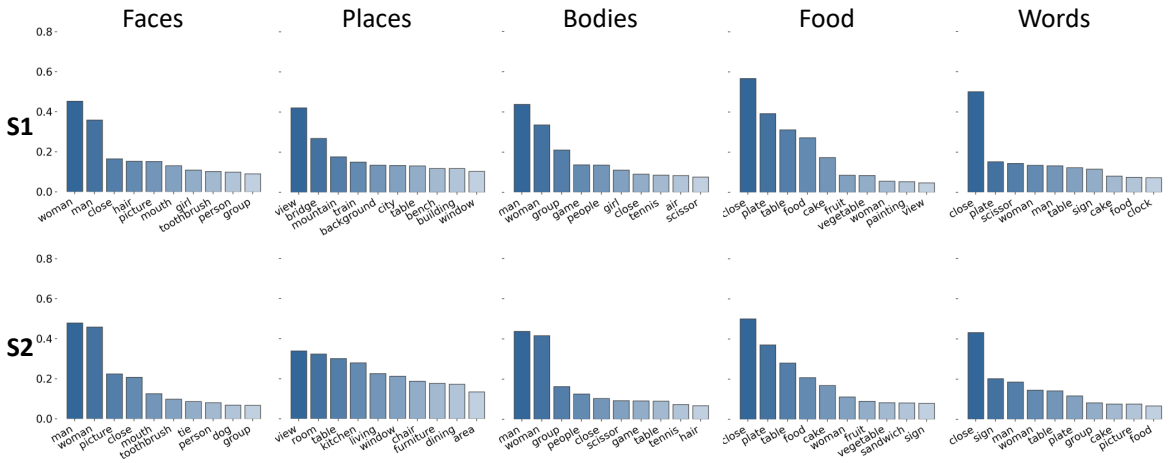


Figure 3.4: **Top BrainSCUBA nouns via voxel-wise captioning in broad category selective regions.** We perform part-of-speech tagging and lemmatization to extract the nouns, $y-$axis normalized by voxel count. We find that the generated captions are semantically related to the functional selectivity of broad category selective regions. Note that the word "close" tended to appear in the noun phrase "close-up", which explains its high frequency in the captions from food- and word-selective voxels.

	Faces		Places		Bodies		Words		Food		Mean	
	S2	S5	S2	S5	S2	S5	S2	S5	S2	S5	S2	S5
NSD all stim	17.1	17.5	29.4	30.7	31.5	30.3	11.0	10.1	10.9	11.4	20.0	20.0
NSD top-100	45.0	43.0	78.0	93.0	59.0	55.0	48.0	33.0	86.0	83.0	63.2	61.4
BrainDiVE-100	68.0	64.0	100	100	69.0	77.0	61.0	80.0	94.0	87.0	78.4	81.6
BrainSCUBA-100	**67.0**	**62.0**	**100**	**99.0**	**54.0**	**73.0**	**55.0**	**34.0**	**97.0**	**92.0**	**74.6**	**72.0**

Table 3.1: **Semantic evaluation of images with zero-shot CLIP.** We use CLIP to perform zero-shot 5-way classification. Show here is the percentage where category of the image matches the preferred category for a brain region. This is shown for each subject's NSD stimulus set ($10,000$ images for S2&S5); the top-100 images (top-1%) evaluated by average region true fMRI, the top-100 (10%) of BrainDiVE and BrainSCUBA (**bolded**) as evaluated by their respective encoders. BrainSCUBA has selectivity that is closer to the true NSD top 1%.

3.4.3 Text-Guided Brain Image Synthesis

Visualization of the captions can be helpful in highlighting subtle co-occurrence statistics, with novel images critical for future hypothesis driven investigations of the visual cortex [Gu et al., 2023, Jain et al., 2023, Ratan Murty et al., 2021]. We utilize a text-to-image diffusion model, and condition the synthesis process on the voxel-wise captions within an ROI (Figure 3.5). We perform 1000 generations per-ROI, subsampling without replacement when the number of voxels/captions in a ROI exceed 1000, and randomly sample the gap when there are fewer than 1000. For face-, place-, word-, body-selective regions, we visualize the top-5 out of $10,000$ images ranked by

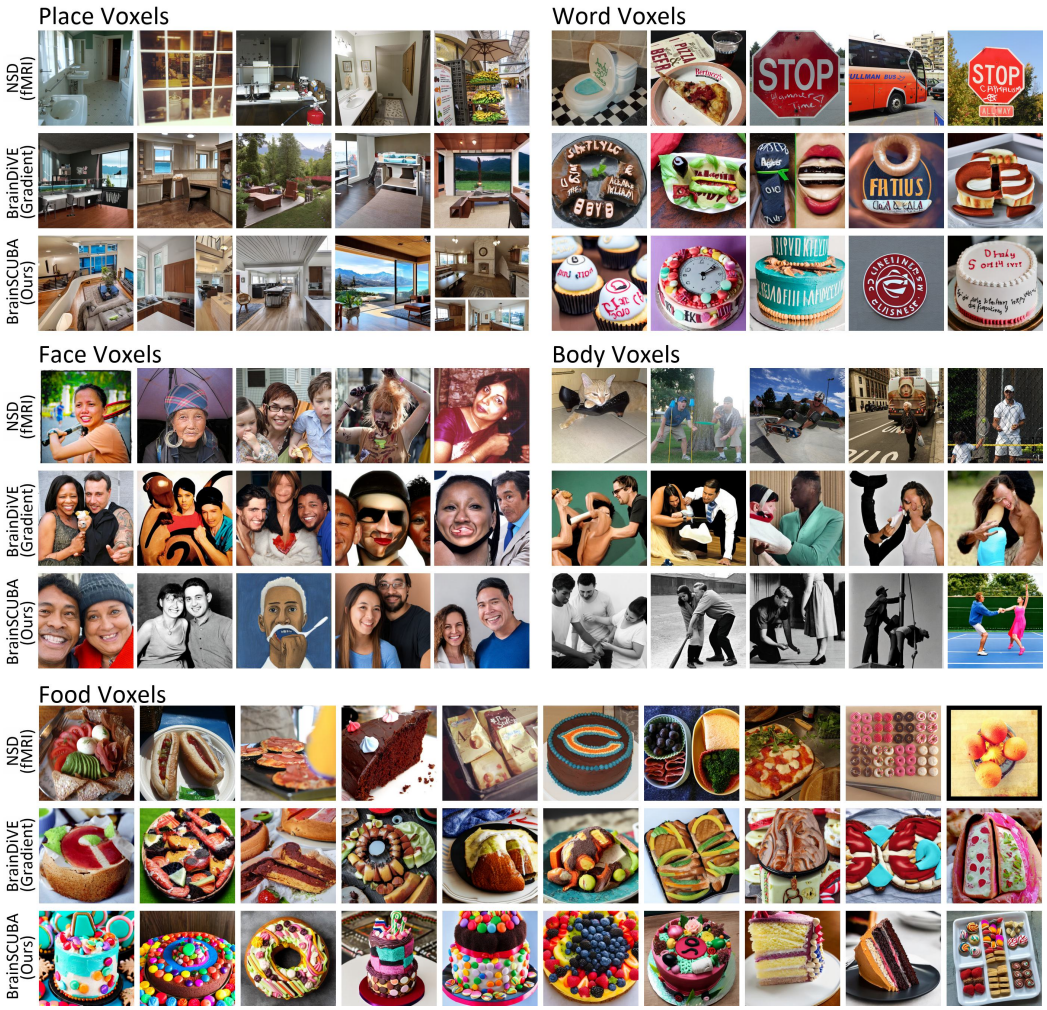


Figure 3.5: **Novel images for category selective voxels in S2**. We visualize the top-5 images from the fMRI stimuli and generated images for the place/word/face/body regions, and the top-10 images for the food region. We observe that images generated with BrainSCUBA appear more coherent.

real average ROI response from the fMRI stimuli (NSD), and the top-5 out of $1,000$ generations ranked by predicted response using the respective BrainDiVE [Luo et al., 2023] and BrainSCUBA encoders. BrainDiVE is used for comparison as it is the state of the art method for synthesizing activating images in the higher visual cortex, and we follow their evaluation procedure. For the food region, we visualize the top-10. Predicted activation is shown in Figure 3.6, with semantic classification shown in Table 3.1. Visual inspection suggests our method can generate diverse images semantically aligned with the target category. Our images are generally more visually coherent than those generated by BrainDiVE, and contain more clear text in word voxels, and fewer degraded faces and bodies in the respective regions. This is likely because our images are

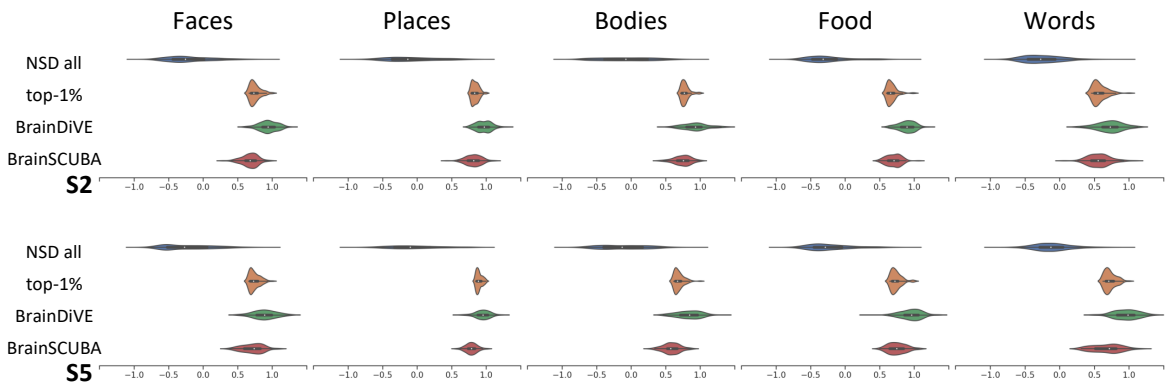


Figure 3.6: **Evaluating the distribution of BrainSCUBA captions with a different encoder**. We train an encoder with a different backbone (EVA02-CLIP-B-16) from both BrainDiVE and BrainSCUBA. For each region, we evaluate the response to all images a subject saw in NSD, the response of the top-1% of images in NSD stimuli ranked using `EVA02`, the top-10% of images generated by BrainDiVE and BrainSCUBA and ranked by their respective encoders. Each region is normalized to $[-1, 1]$ using the min/max of the predicted responses to NSD stimuli. BrainSCUBA can achieve high predicted responses despite *not* performing explicit gradient based maximization like BrainDiVE, and yields concretely interpretable captions. BrainSCUBA is also $\sim 10\times$ faster per image.

conditioned on text, while BrainDiVE utilizes the gradient signal alone.

3.4.4 Investigating the Brain's Social Network

The intrinsic social nature of humans significantly influences visual perception. This interplay is evident in the heightened visual sensitivity towards social entities such as faces and bodies [Downing et al., 2001, Kanwisher et al., 1997, Pitcher and Ungerleider, 2021]. In this section, we explore if BrainSCUBA can provide insights on the finer-grained coding of people in the brain. We use a rule based filter and count the number of captions that contain one of 140 nouns that describe people (person, man, woman, child, boy, girl, family, occupations, and plurals). We visualize the voxels whose captions contain people in Figure 3.7, and provide a quantitative evaluation in Table 3.2. We observe that our captions can correctly identify non-person-selective scene, food, and word regions as having lower person content than person-selective ROIs like the FFA or the EBA. Going beyond traditional functional ROIs, we find that the precuneus visual area (PCV) and the temporoparietal junction (TPJ) have a very high density of captions with people.

The precuneus has been implicated in third-person mental representations of self [Cavanna and Trimble, 2006, Petrini et al., 2014], while the TPJ has been suggested to be involved in theory of mind and social cognition [Saxe and Kanwisher, 2013]. Our results lend support to these hypotheses.

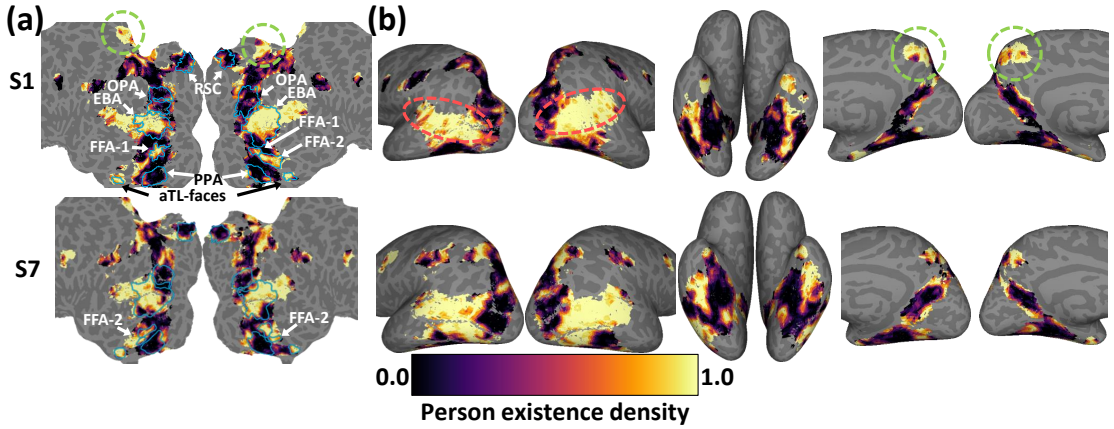


Figure 3.7: **Presence of people in captions**. We perform rule-based filtering and identify voxels where a caption contains at least one person. Data is surface smoothed for visualization. Dotted orange oval shows approximate location of TPJ, which is linked to theory of mind; green circle shows location of PCV, associated with third-person perspective of social interactions. Note that TPJ is HCP defined following Igelström and Graziano [2017], while PCV is HCP atlas region 27 [Glasser et al., 2016].

	Non-Person					Person		Other	
	RSC	OPA	PPA	Food	Word	EBA	FFA	PCV	TPJ
S1	12.9	17.3	10.6	11.5	32.0	87.2	88.5	89.7	92.1
S2	5.58	8.15	2.70	20.0	34.8	81.4	87.2	70.8	89.1
S5	9.31	6.43	1.95	17.8	38.4	79.5	89.4	78.5	79.9
S7	7.14	9.87	5.99	10.7	36.9	84.3	89.5	84.2	90.3
Mean	8.72	10.4	5.30	15.0	35.5	83.1	88.6	80.8	87.8

Table 3.2: **Percentage of captions in each region that contain people.** We observe a sharp difference between non-person regions (Scene RSC/OPA/PPA, Food, Word), and regions that are believed to be person selective (body EBA, face FFA). We also observe extremely high person density in PCV — a region involved in third-person social interactions, and TPJ — a region involved in social self-other distinction.

A close visual examination of Figure 3.7 suggests a divide within EBA. We perform spherical k-means clustering on joint encoder weights for $t > 2$ EBA from S1/S2/S5/S7, and identify two stable clusters. These clusters are visualized in Figure 3.8. Utilizing the rule parser, we labels

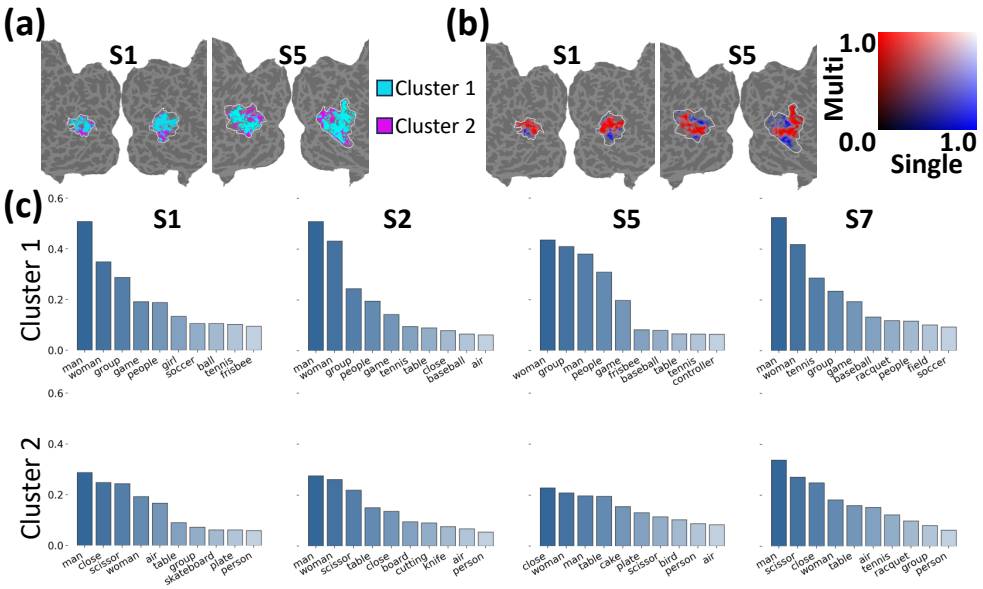


Figure 3.8: **Clusters within EBA. (a)** The EBA clusters for two subjects are shown on a flatmap. **(b)** Number of people mentioned in each caption. **(c)** Top nouns within each cluster, y-axis is normalized to the number of voxels within a cluster. Compared to Cluster-1, Cluster-2 has less emphasis on multiple people and more emphasis on objects that can be held.

Which cluster is more...	people per-img				inanimate objs				far away				sports			
	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7
EBA-1 (Cluster 1)	**88**	**84**	**91**	**78**	15	11	12	13	**62**	**72**	**78**	**63**	**75**	**79**	**85**	**76**
EBA-2 (Cluster 2)	5	10	4	13	**72**	**80**	**81**	**65**	21	21	14	25	9	12	6	11

Table 3.3: **Human evaluation comparing two EBA clusters.**. Evaluators compare the top 100 images for each cluster, with questions like "Which group of images is more X?", answers include EBA-1/EBA-2/Same. We do not show "Same"; responses sum to 100 across all three options. Results in %.

the voxels into those that contain a single individual or multiple people, and further visualize the top-nouns within each of these two clusters. While both clusters include general person words like “man” and “woman”, cluster 1 has more nouns that suggest groups of people interacting together (group, game, people), and cluster 2 has words that suggest close-ups of individuals with objects that may be hand-held. To validate our findings, we perform a study where subjects are asked to evaluate the top-100 images from each of the clusters. Results are shown in Table 3.3. Aligned with the top-nouns, the study suggests that cluster-1 has more groups of people, fewer inanimate objects, and consists of larger scenes. This intriguing novel finding about the fine-grained distinctions in EBA can lead to new hypotheses about its function. This finding also

demonstrates the ability of BrainSCUBA to uncover broad functional differences across the visual cortex.

3.5 Discussion

Limitations and Future Work. Although our methods can generate semantically faithful descriptions for the broad category selective regions, our approach ultimately relies on a pre-trained captioning model. Due to this, our method reflects the biases of the captioning model. It is further not clear if the most selective object in each region can be perfectly captured by language. Future work could explore the use of more unconstrained captioning models [Tewel et al., 2022] or more powerful language models [Touvron et al., 2023].

Conclusion. To summarize, in this paper we propose BrainSCUBA, a method which can generate voxel-wise captions to describe each voxel's semantic selectivity. We explore how the output tiles the higher visual cortex, perform text-conditioned image synthesis with the captions, and apply it to uncover finer-grained patterns of selectivity in the brain within the person class. Our results suggest that BrainSCUBA may be used to facilitate data-driven exploration of the visual cortex.

3.6 Additional Results for BrainSCUBA

3.6.1 Visualization of each subject's top-nouns for category selective voxels

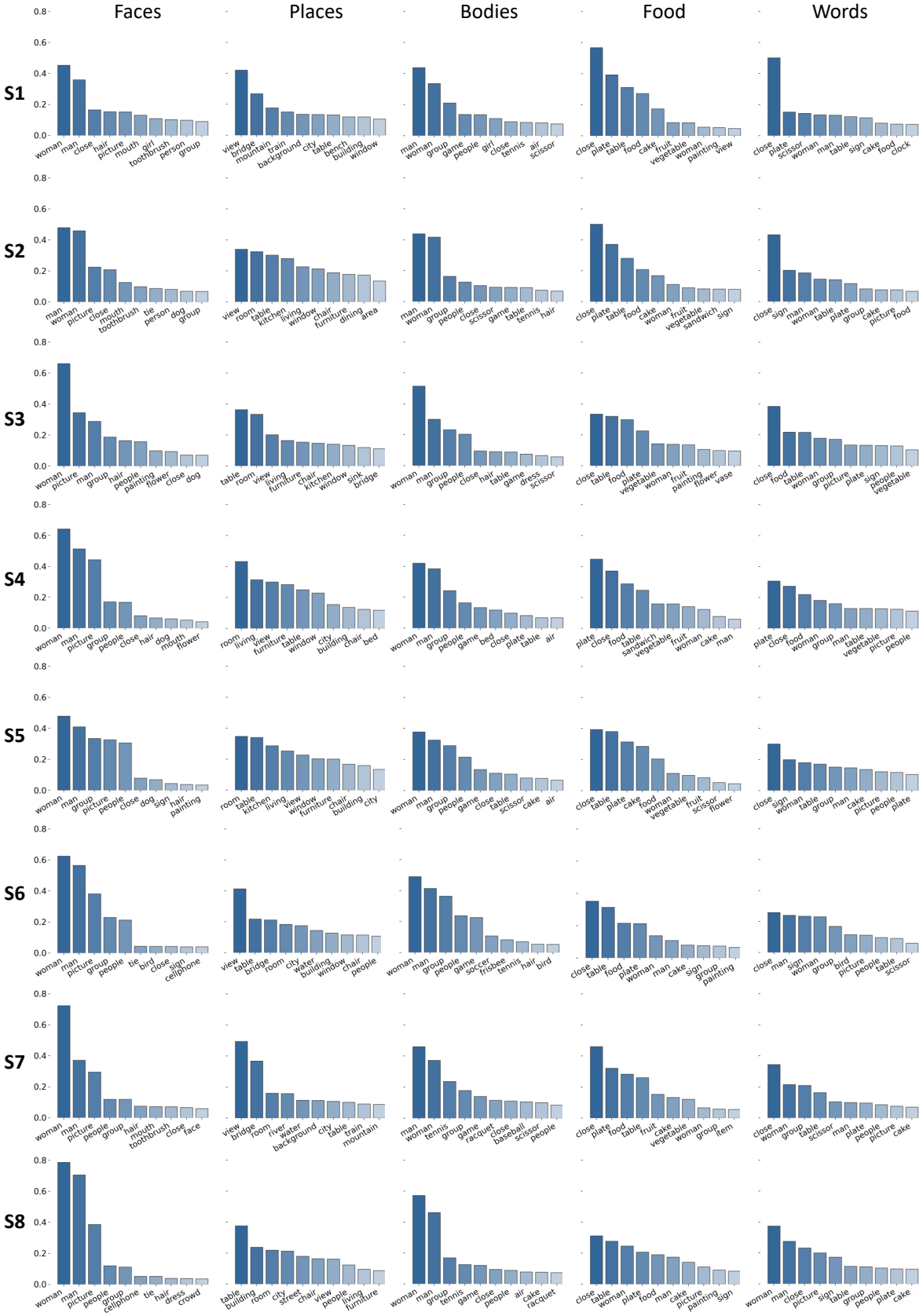


Figure 3.9: **Top BrainSCUBA nouns via voxel-wise captioning in broad category selective regions for all subjects.** We see broad semantic alignment between the top-nouns and the semantic selectivity of a region. Note that the category selective voxels were derived from the intersection of official NSD functional localizer values $t > 2$ and their provided region masks.

3.6.2 Visualization of UMAPs for all subjects

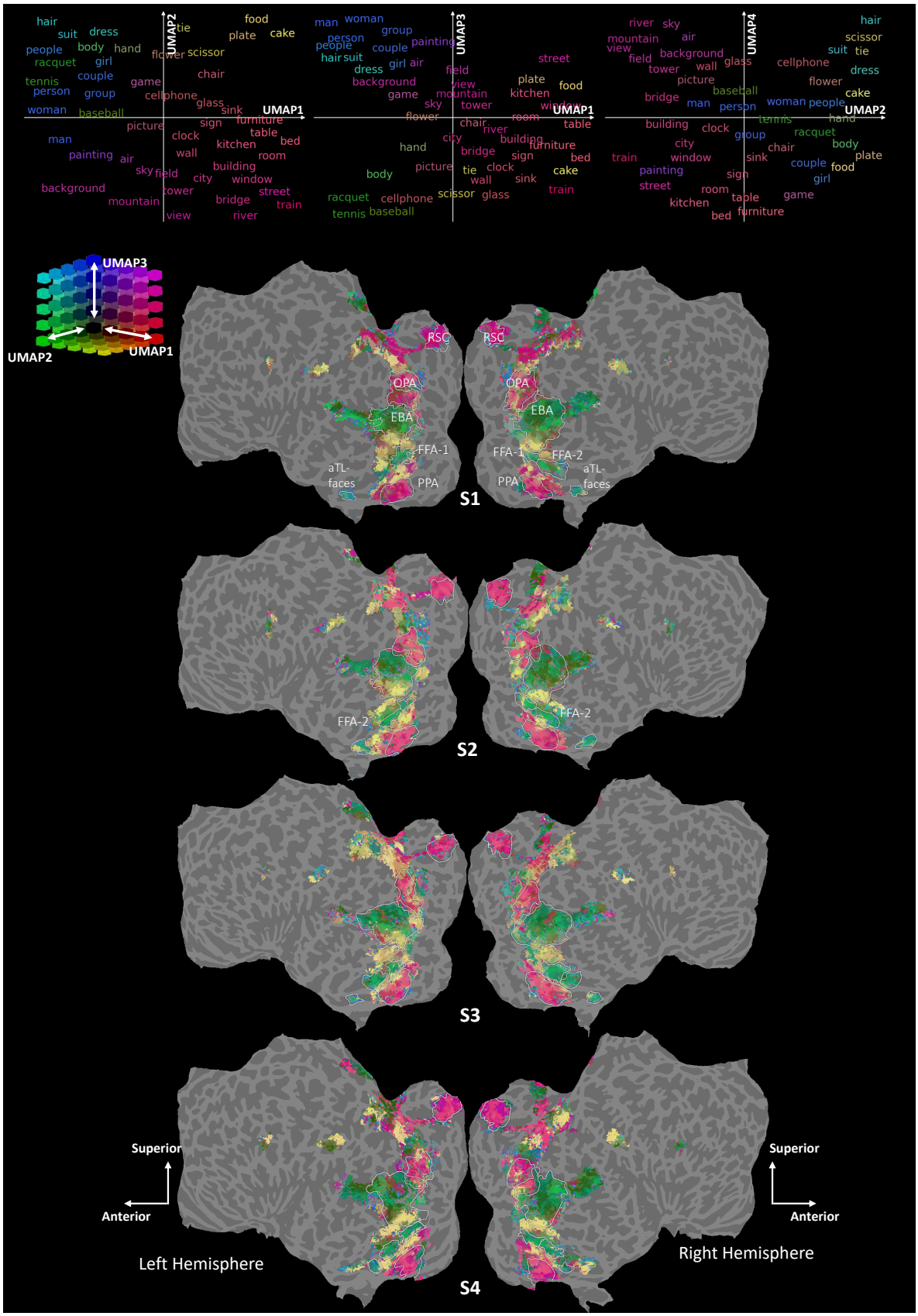


Figure 3.10: **UMAP transform results for S1-S4.** All vectors are normalized to unit norm prior to UMAP. UMAP is fit on S1. Both word and voxel vectors are projected onto the space of natural images prior to transform using softmax weighted sum. Nouns are the most common across all subjects.

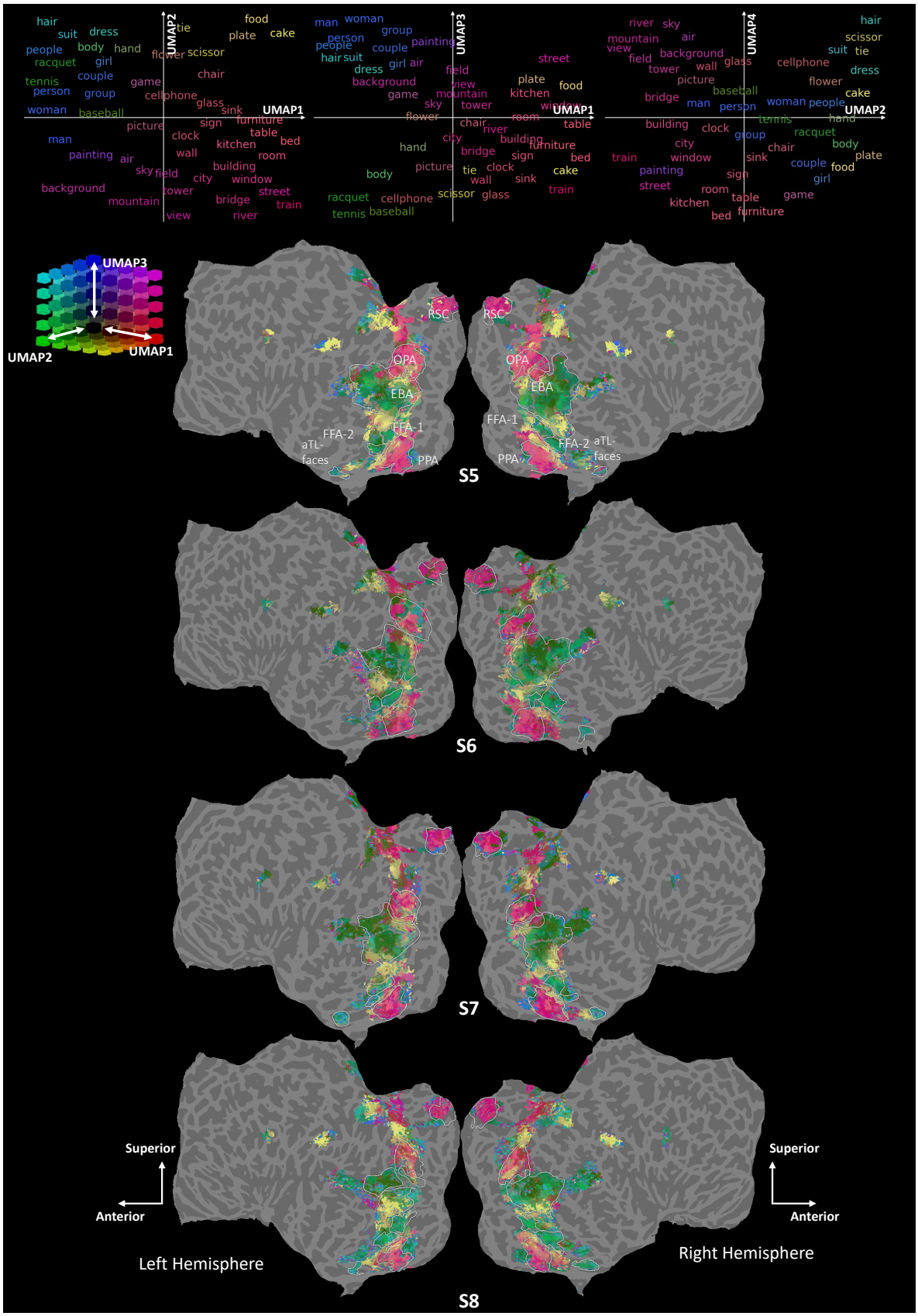


Figure 3.11: **UMAP transform results for S5-S8.** All vectors are normalized to unit norm prior to UMAP. UMAP is fit on S1. Both word and voxel vectors are projected onto the space of natural images prior to transform using softmax weighted sum. Nouns are the most common across all subjects.

3.6.3 Novel image generation for all subjects

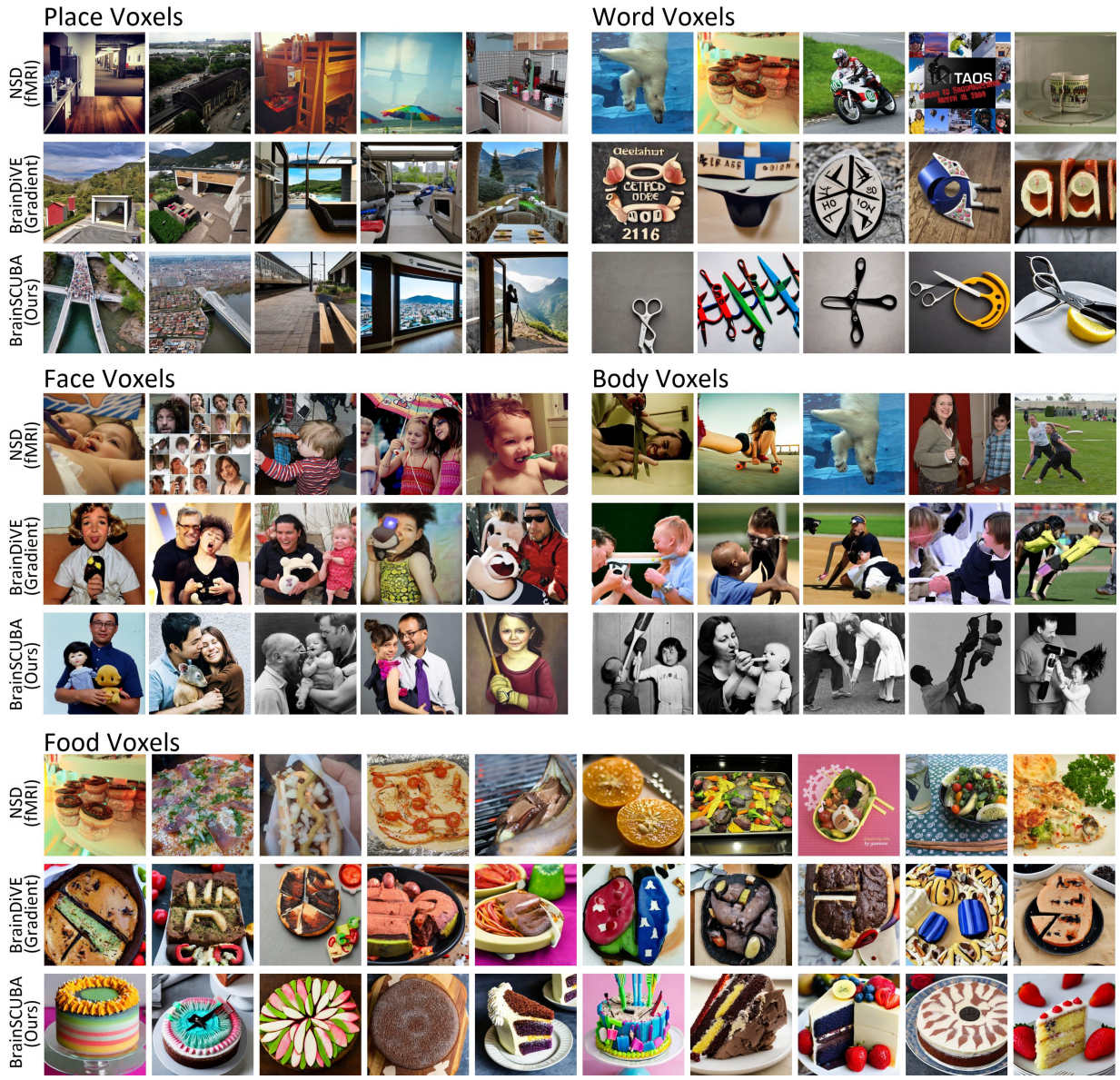


Figure 3.12: **Image generation for S1.** We visualize the top-5 for face/place/body/word categories, and the top-10 for food. NSD images are ranked by ground truth response. BrainDiVE and BrainSCUBA are ranked by their respective encoders. BrainSCUBA images have more recognizable objects and fewer artifacts, likely due to the use of captions rather than gradients as in BrainDiVE.

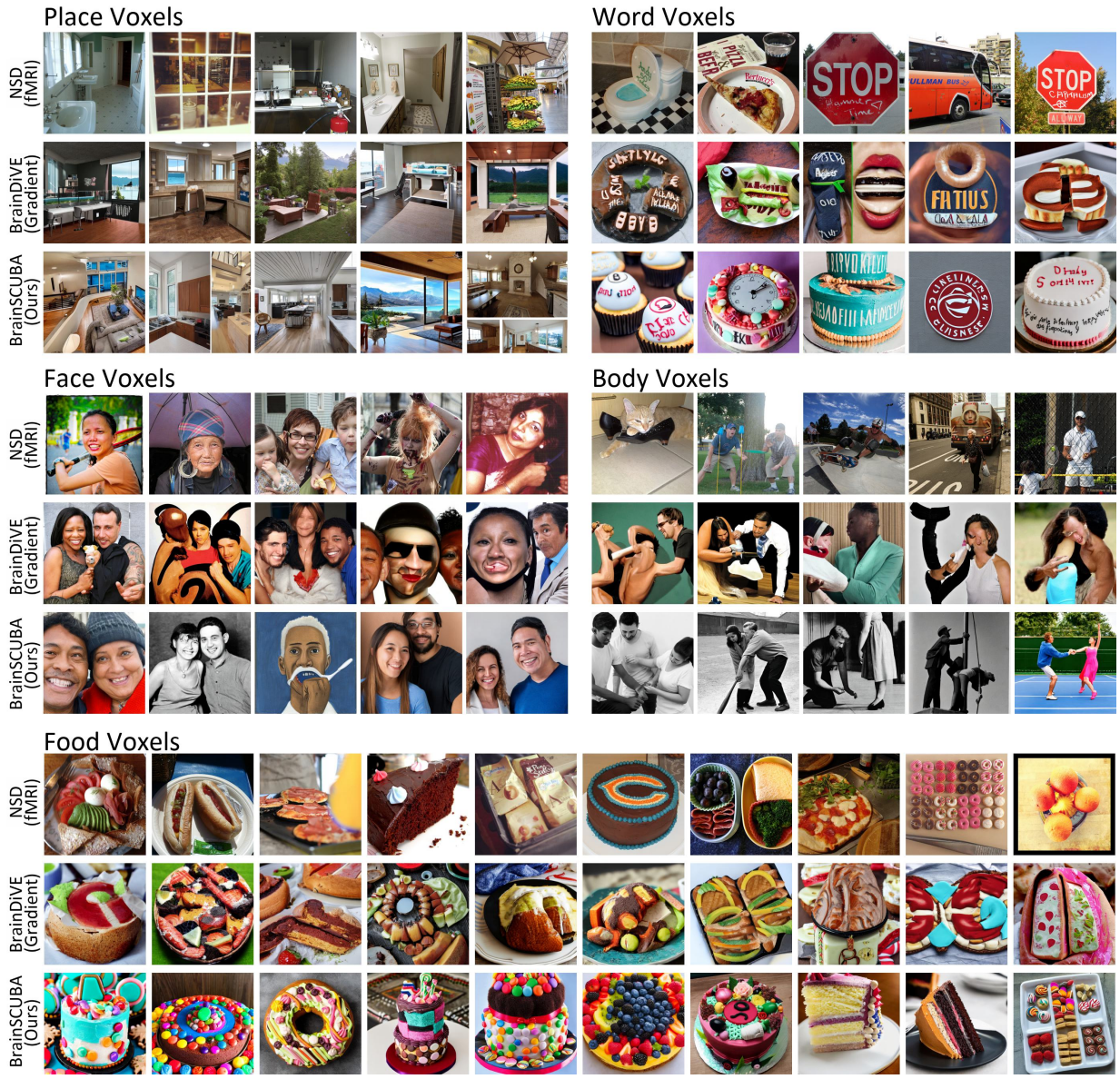


Figure 3.13: **Image generation for S2.** We visualize the top-5 for face/place/body/word categories, and the top-10 for food. NSD images are ranked by ground truth response. BrainDiVE and BrainSCUBA are ranked by their respective encoders. BrainSCUBA images have more recognizable objects and fewer artifacts, likely due to the use of captions rather than gradients as in BrainDiVE.

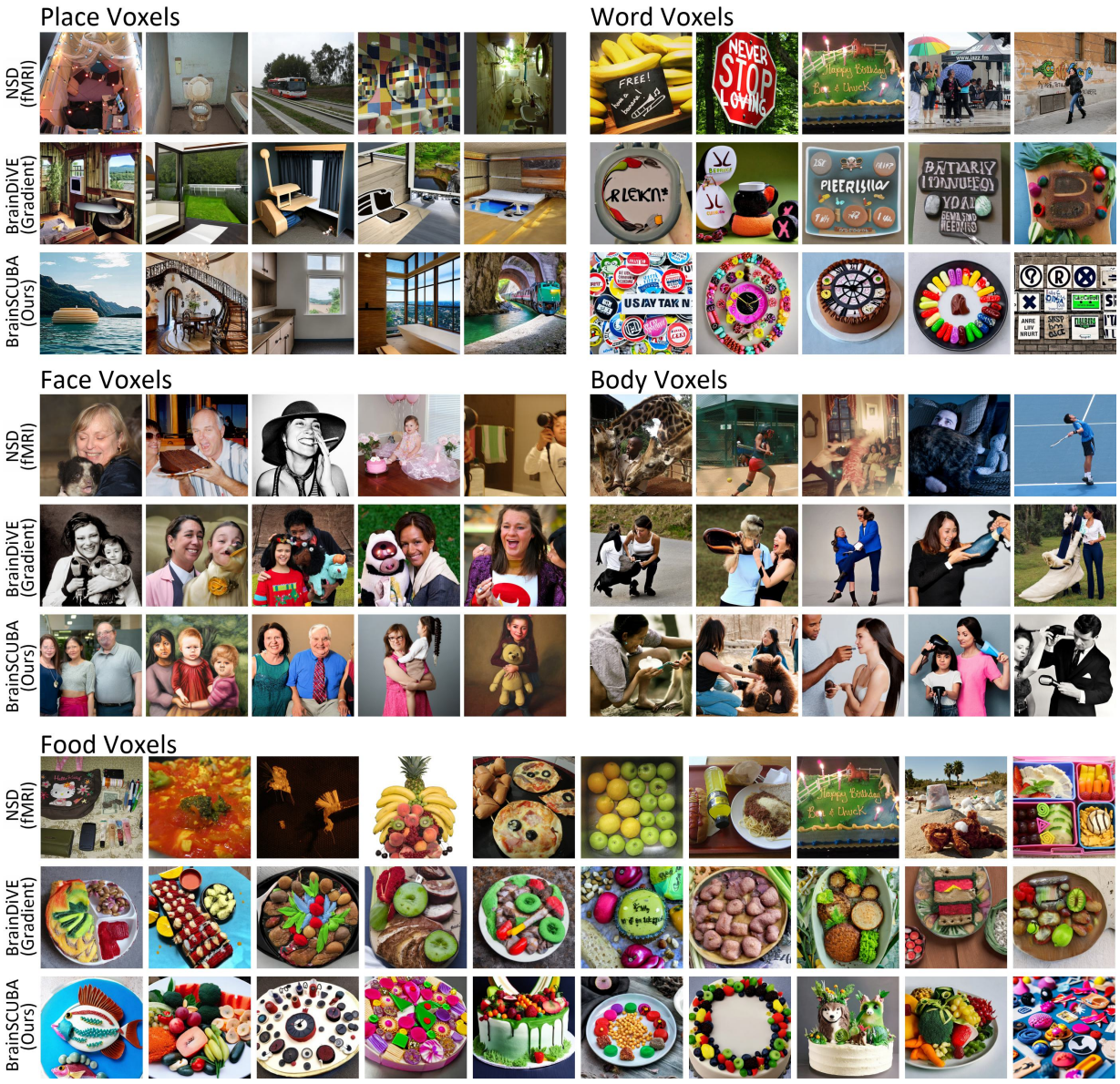


Figure 3.14: **Image generation for S3.** We visualize the top-5 for face/place/body/word categories, and the top-10 for food. NSD images are ranked by ground truth response. BrainDiVE and BrainSCUBA are ranked by their respective encoders. BrainSCUBA images have more recognizable objects and fewer artifacts, likely due to the use of captions rather than gradients as in BrainDiVE.

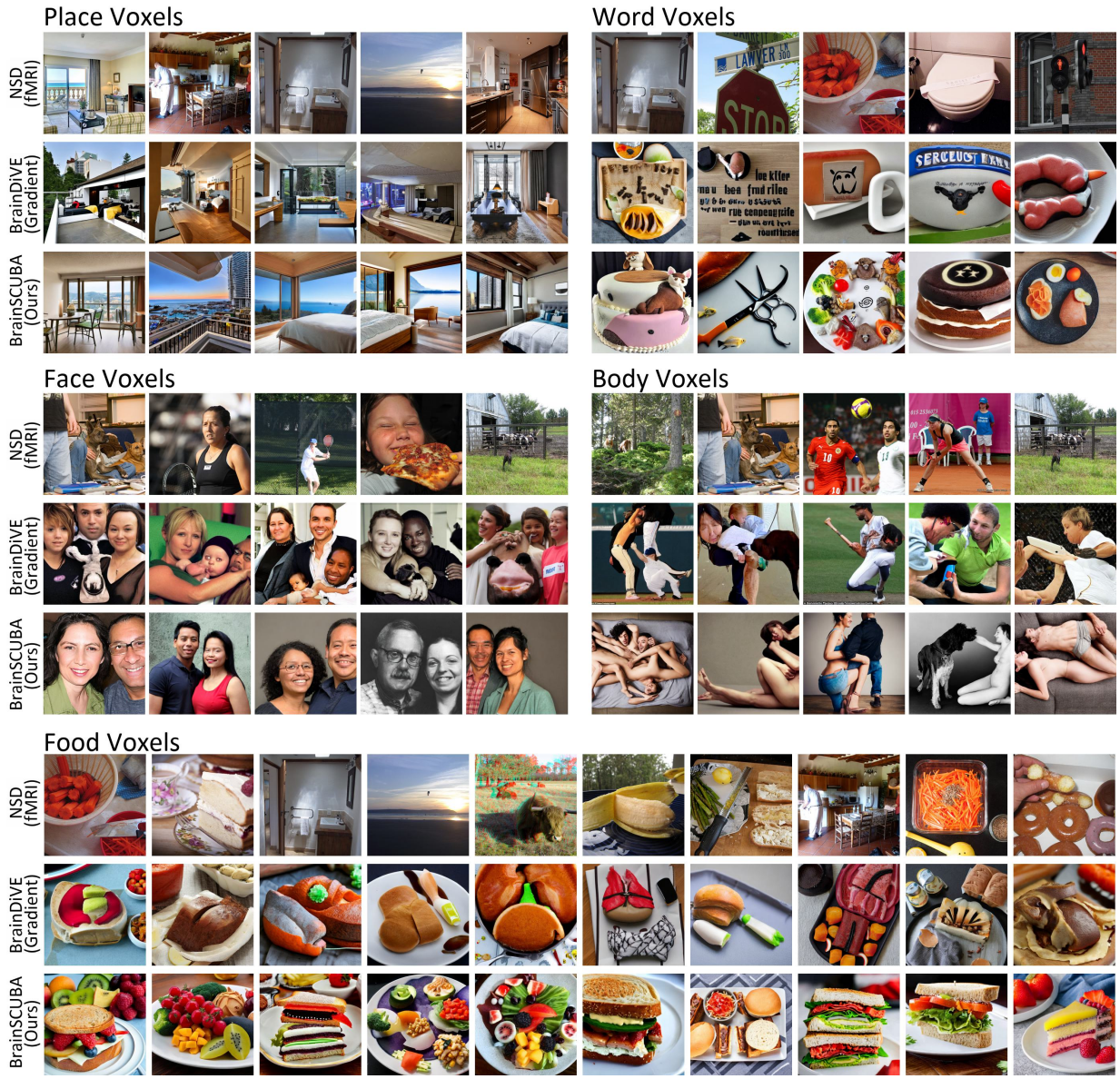


Figure 3.15: **Image generation for S4.** We visualize the top-5 for face/place/body/word categories, and the top-10 for food. NSD images are ranked by ground truth response. BrainDiVE and BrainSCUBA are ranked by their respective encoders. BrainSCUBA images have more recognizable objects and fewer artifacts, likely due to the use of captions rather than gradients as in BrainDiVE.

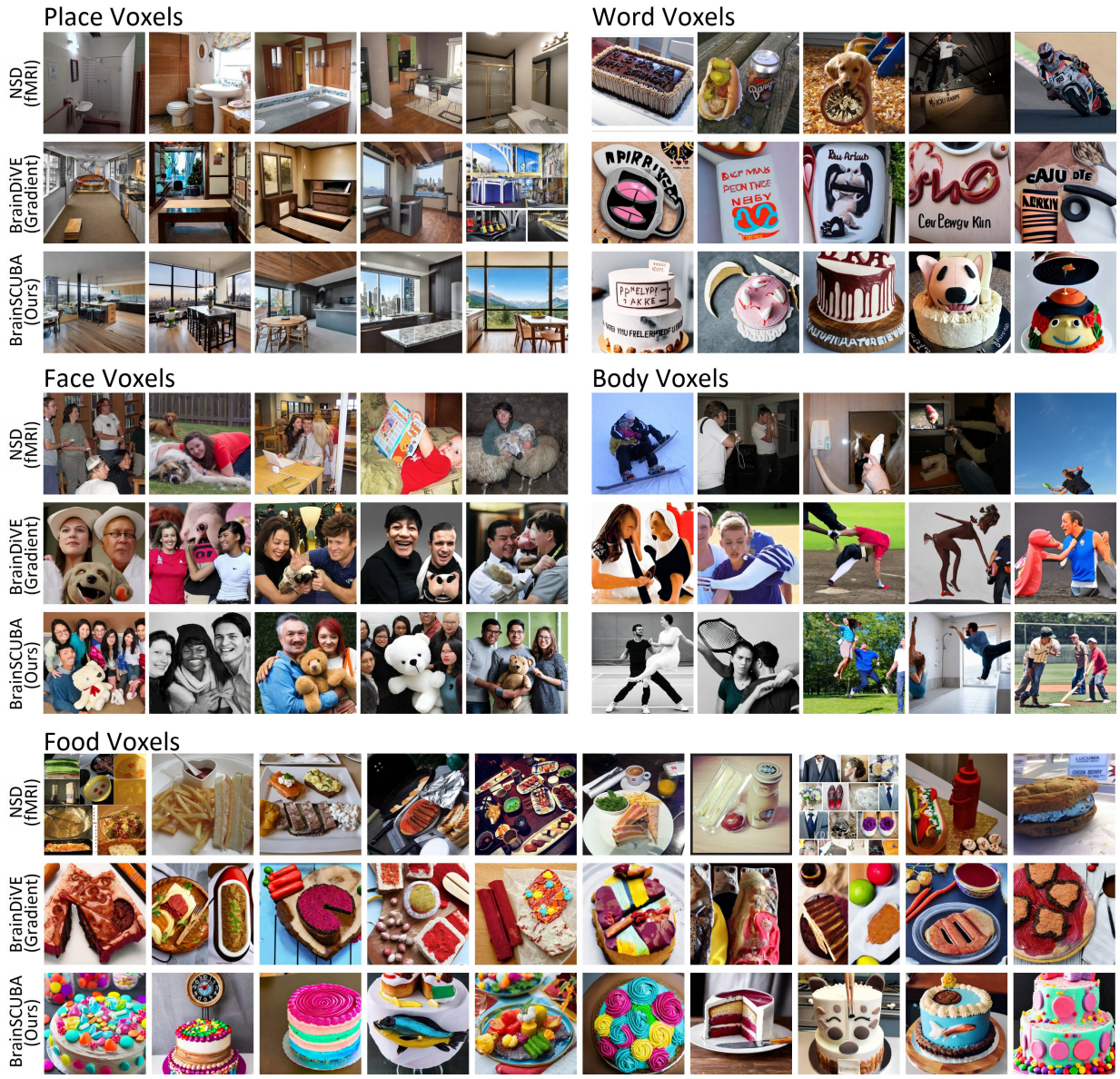


Figure 3.16: **Image generation for S5.** We visualize the top-5 for face/place/body/word categories, and the top-10 for food. NSD images are ranked by ground truth response. BrainDiVE and BrainSCUBA are ranked by their respective encoders. BrainSCUBA images have more recognizable objects and fewer artifacts, likely due to the use of captions rather than gradients as in BrainDiVE.

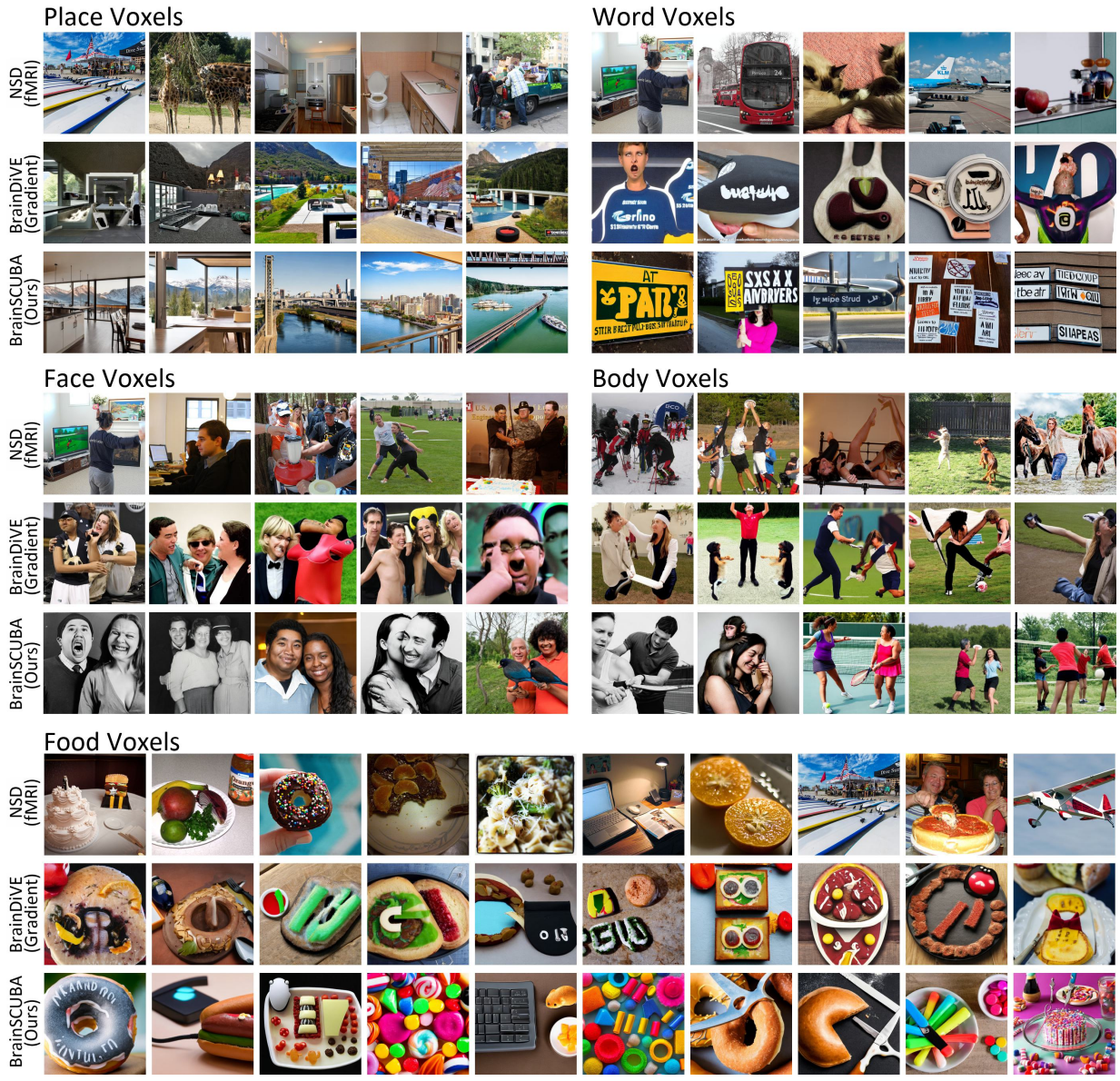


Figure 3.17: **Image generation for S6.** We visualize the top-5 for face/place/body/word categories, and the top-10 for food. NSD images are ranked by ground truth response. BrainDiVE and BrainSCUBA are ranked by their respective encoders. BrainSCUBA images have more recognizable objects and fewer artifacts, likely due to the use of captions rather than gradients as in BrainDiVE.

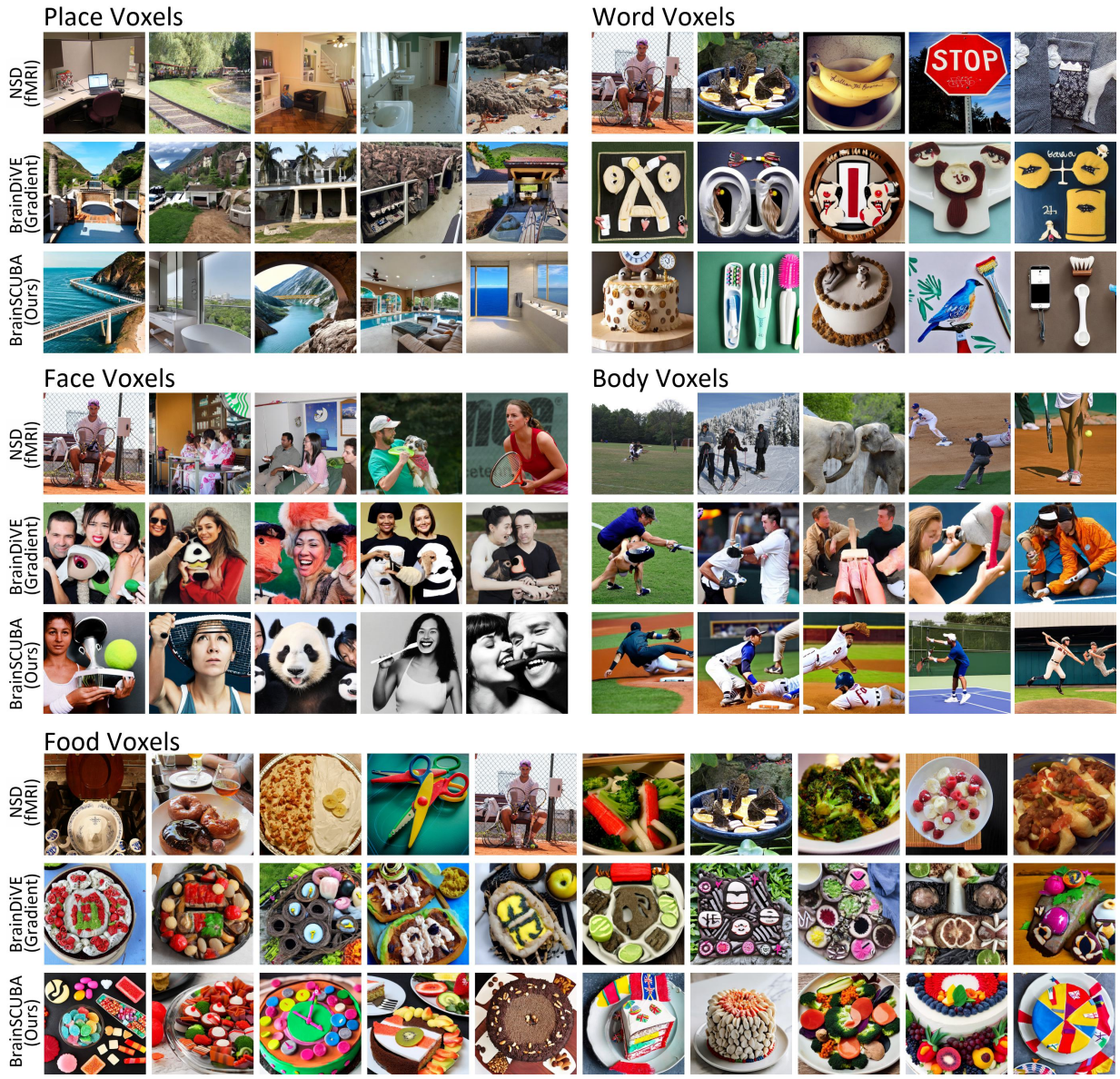


Figure 3.18: **Image generation for S7.** We visualize the top-5 for face/place/body/word categories, and the top-10 for food. NSD images are ranked by ground truth response. BrainDiVE and BrainSCUBA are ranked by their respective encoders. BrainSCUBA images have more recognizable objects and fewer artifacts, likely due to the use of captions rather than gradients as in BrainDiVE.

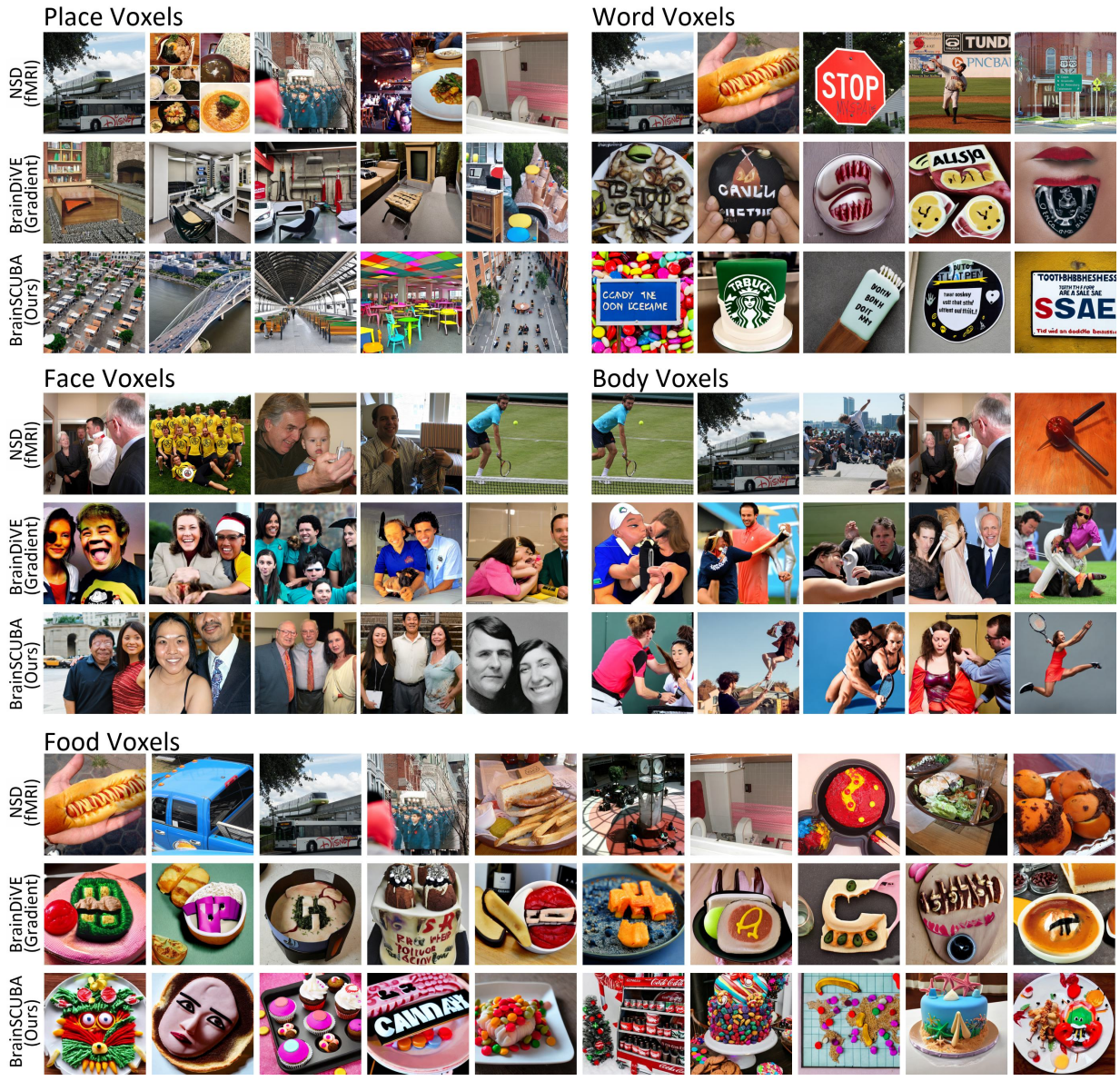


Figure 3.19: **Image generation for S8.** We visualize the top-5 for face/place/body/word categories, and the top-10 for food. NSD images are ranked by ground truth response. BrainDiVE and BrainSCUBA are ranked by their respective encoders. BrainSCUBA images have more recognizable objects and fewer artifacts, likely due to the use of captions rather than gradients as in BrainDiVE.

3.6.4 Distribution of "person" representations across the brain for all subjects

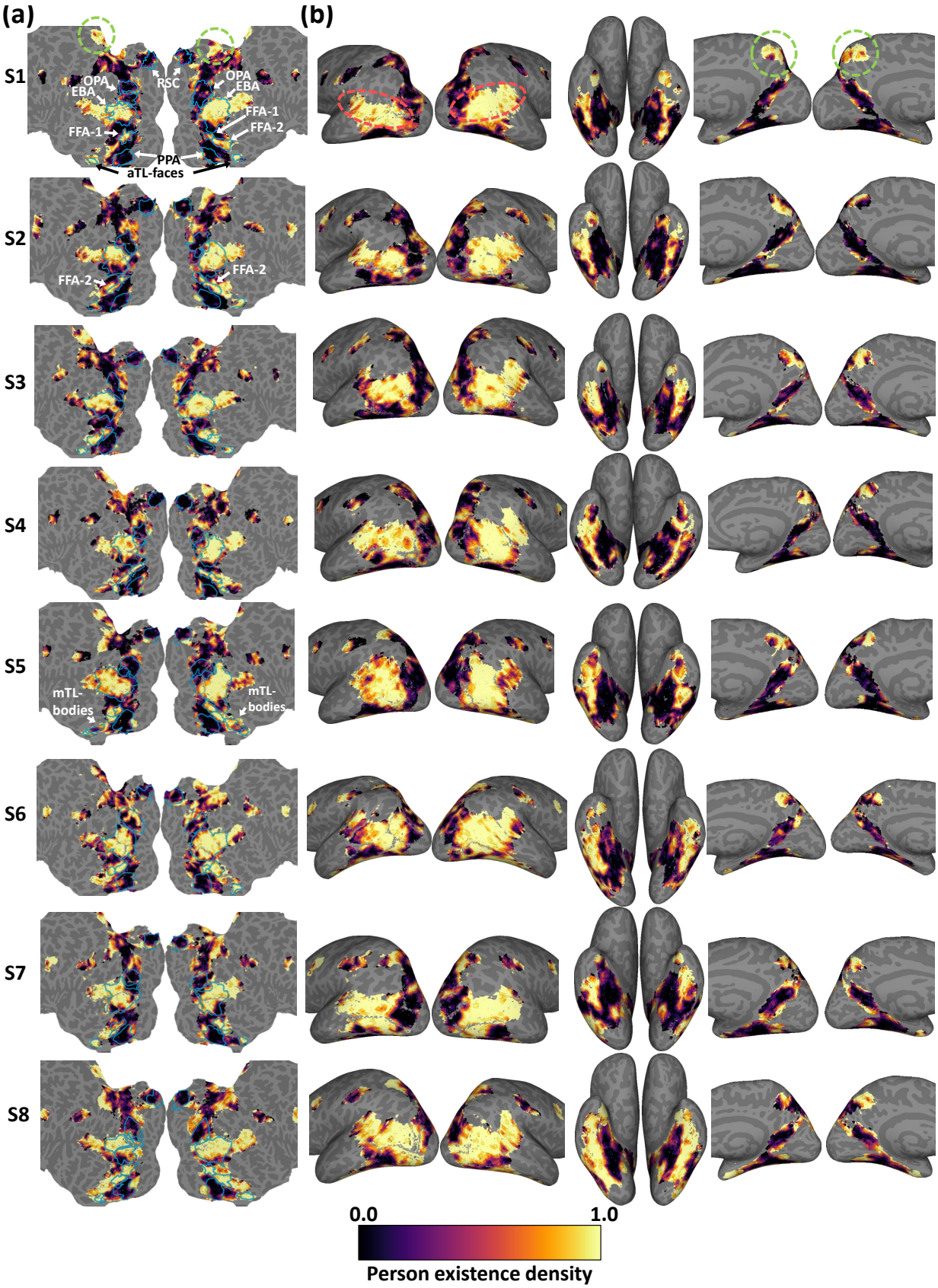


Figure 3.20: **Presence of people in captions for S1-S8.** **(a)** Flatmap of cortex. **(b)** Inflated map of cortex. Dotted orange oval shows approximate location of TPJ, which is linked to theory of mind; green circle shows location of PCV, associated with third-person perspective of social interactions. For S5 alone we additionally label the mTL-bodies area.

	Non-Person					Person		Other	
	RSC	OPA	PPA	Food	Word	EBA	FFA	PCV	TPJ
S1	12.9	17.3	10.6	11.5	32.0	87.2	88.5	89.7	92.1
S2	5.58	8.15	2.70	20.0	34.8	81.4	87.2	70.8	89.1
S3	6.57	16.9	4.49	24.4	33.9	84.7	90.3	75.3	83.2
S4	4.40	14.7	4.47	20.0	37.8	78.9	90.3	66.5	88.9
S5	9.31	6.43	1.95	17.8	38.4	79.5	89.4	78.5	79.9
S6	16.7	28.2	6.93	27.1	48.8	91.8	97.8	75.4	79.1
S7	7.14	9.87	5.99	10.7	36.9	84.3	89.5	84.2	90.3
S8	15.7	30.9	9.84	42.6	57.5	86.7	96.2	71.2	89.7
Mean	9.78	16.5	5.86	21.8	40.0	84.3	91.2	76.4	86.5

Table 3.4: **Percentage of captions in each region that contain people for S1-S8.** We observe a sharp difference between non-person regions (Scene RSC/OPA/PPA, Food, Word), and regions that are believed to be person selective (body EBA, face FFA). We also observe extremely high person density in PCV — a region involved in third-person social interactions, and TPJ — a region involved in social self-other distinction.

3.6.5 Additional extrastriate body area (EBA) clustering results

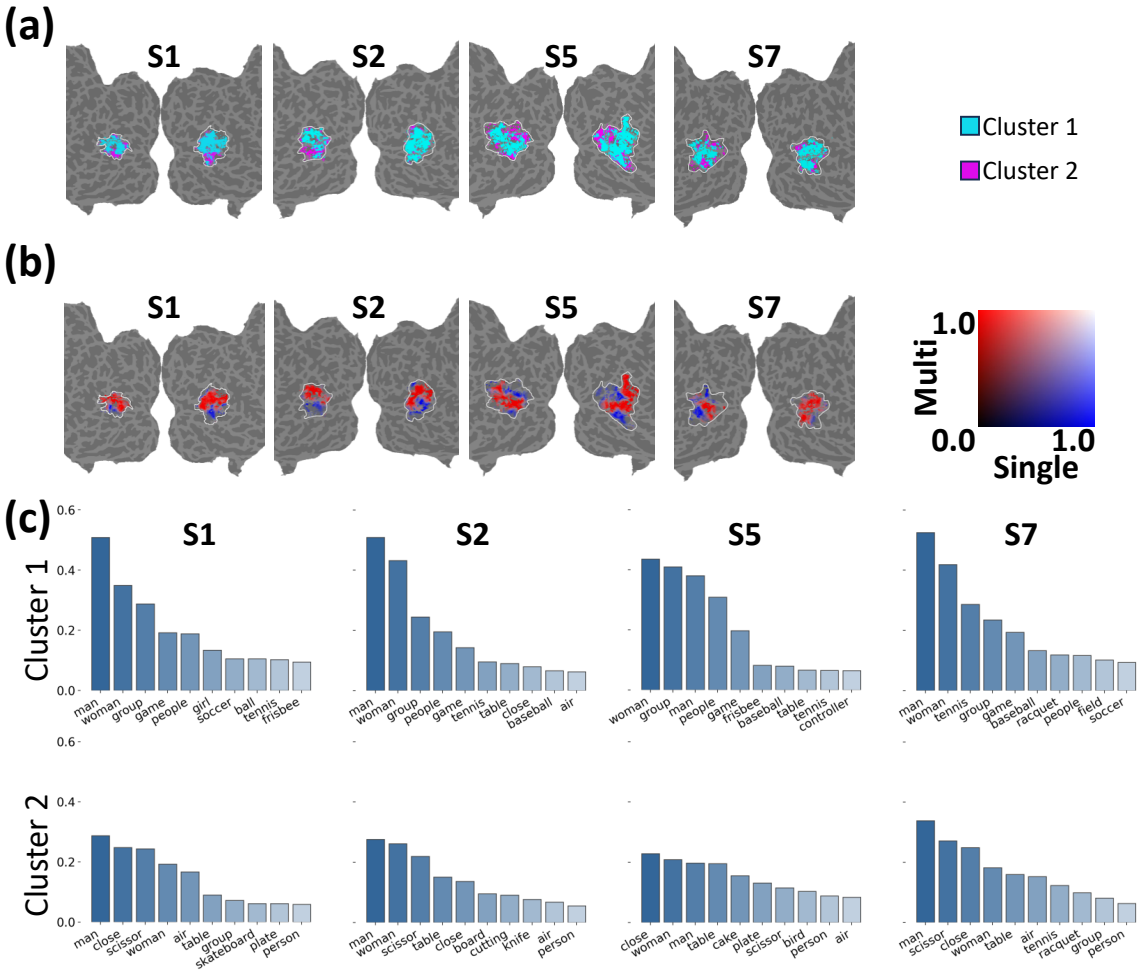


Figure 3.21: **EBA clustering for S1/S2/S5/S7.** **(a)** EBA clusters. **(b)** Voxels which mention just a single person and those that mention multiple people. **(c)** Top nouns. Note that clustering was performed jointly on S1/S2/S5/S7.

	Single		Multiple	
	EBA-1	EBA-2	EBA-1	EBA-2
S1	21.8	68.5	78.2	31.5
S2	31.5	69.5	68.6	30.5
S5	28.8	75.2	71.2	24.8
S7	29.0	63.8	71.0	36.2
Mean	27.8	69.3	72.3	30.8

Table 3.5: **Distribution of single/multi-person voxels within each EBA cluster.** After parsing each voxel's caption, we compute the single/multi voxels as a percentage of all voxels in the cluster that mention "person" class. We observe that EBA cluster 1 (EBA-1) has a higher ratio of voxels that mention multiple people. This is reflected in both the visualization, the nouns, and the human study on ground truth top NSD images for each cluster.

3.6.6 Human study details

Ten subjects were recruited via prolific.co. These subjects are aged $20 \sim 48$; 2 asian, 2 black, 6 white; 5 men, 5 women. For each NSD subject (S1/S2/S5/S7), we select the top-100 images for each cluster as ranked by the real average fMRI response. Each of the 100 images were randomly split into 10 non-overlapping subgroups.

Questions were posed in two formats. In the first format, subjects were simultaneously presented with images from the two clusters, and select the set where an attribute was ***more prominent***, possible answers include cluster-1/cluster-2/same. The second format asked subjects to evaluate a set of image from a single cluster, and answer yes/no on if an attribute/object-type was ***present in most*** of the images.

For the human study results in section 4.4, a human evaluator would perform 40 comparisons, from 10 splits and the 4 NSD subjects; with 10 human evaluators per question. We collected 1600 total responses for the four questions in the main text.

For the human study results below, a human evaluator would perform 16 judgements, from 4 splits and the 4 NSD subjects; with 10 human evaluators per question; across the 2 clusters of images. We collected 2560 total responses for the eight questions below.

Due to space constraints, we present the single set attribute evaluation (second format described above) results here in the appendix. We divide the results into two tables for presentation purposes.

Are most images...	social				sports				large-scale scene				animals			
	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7
EBA-1	88	80	85	85	90	85	88	100	80	85	83	85	20	20	18	23
EBA-2	28	23	35	45	28	25	30	50	38	28	33	60	30	30	30	28

Table 3.6: **Human study on EBA clustering, first set of image attributes.** Each human study subject was asked to evaluate groups of 10 images, and answer yes/no on if an attribute was present in most images. Units are in %.

Are most images...	artificial objs				body parts				human faces				multi person			
	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7
EBA-1	78	78	80	75	73	80	78	83	85	78	75	75	100	60	100	85
EBA-2	85	75	83	80	35	30	40	55	28	20	15	45	23	8	18	25

Table 3.7: **Human study on EBA clustering, second set of image attributes.** Each human study subject was asked to evaluate groups of 10 images, and answer yes/no on if an attribute was present in most images. Units are in %.

3.6.7 Training and inference details

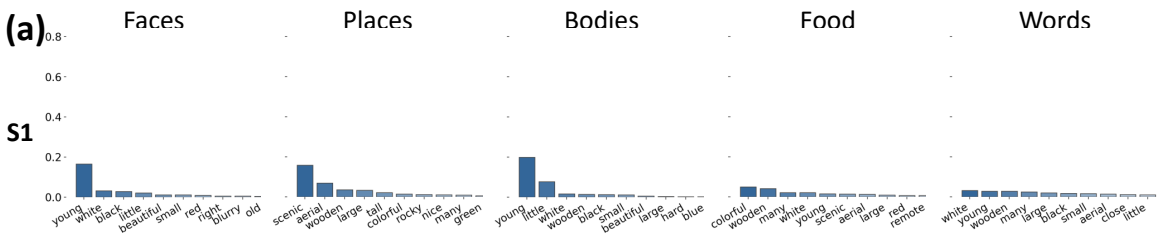
We perform our experiments on a mixture of Nvidia V100 (16GB and 32GB variants), 4090, and 2080 Ti cards. Network training code was implemented using pytorch. Generating one caption for every voxel in higher visual cortex ($20,000+$ voxels) in a single subject can be completed in less than an hour on a 4090. Compared to brainDiVE on the same V100 GPU type, caption based image synthesis with 50 diffusion steps can be done in < 3 seconds, compared to their gradient based approach of $25 \sim 30$ seconds.

For the encoder training, we use the Adam optimizer with decoupled weight decay set to $2e-2$. Initial learning rate is set to $3e-4$ and decays exponentially to $1.5e-4$ over the 100 training epochs. We train each subject independently. The CLIP ViT-B/32 backbone is executed in half-precision (fp16) mode.

During training, we resize the image to 224×224. Images are augmented by randomly scaling the pixel values between $[0.95, 1.05]$, followed by normalization using CLIP image mean and variance. Prior to input to the network, the image is randomly offset by up to 4 pixels along either axis, with the empty pixels filled in with edge padding. A small amount of normal noise with $\mu = 0, \sigma^2 = 0.05$ is independely added to each pixel.

During softmax projection, we set the temperature parameter to $1/150$. We observe higher cosine similarity between pre- and post- projection vectors with lower temperatures, but going even lower causes numerical issues. Captions are generated using beam search with a beam width of 5. A set of 2 million images are used for the projection, and we repeat this with 5 sets. We select the best out of 5 by measuing the CLIP similarity between the caption and the fMRI weights using the original encoder. Sentences are converted to lower case, and further stripped of leading and trailing spaces for analysis.

3.6.8 Top adjectives and more sentences



(b)

Faces	Places	Bodies	Food	Words
A painting of a woman with a tie	A view of a kitchen through a window	A man and woman dancing on the floor	A close up of sliced fruit on a plate	A close up of a clock on a building
A woman in a hat holding up a trophy	A scenic view of a lake and a bench	A woman flying a kite next to a man	A table topped with different types of food	A man holding a cup and a sign
A man in a tuxedo poses for a picture	A view of a kitchen and a dining room	A group of people standing in a room	A colorful display of food is shown on a table	A close up of a logo for a company

Figure 3.22: **Top adjectives. (a)** We extract the most frequent adjectives in each category selective region, note how the adjectives are related to the semantic category in the brain. **(b)** Additional example sentences for each region. The category selective regions are identified via official NSD functional localizer experiments with a different stimulus set.

3.6.9 Encoder fitting stability

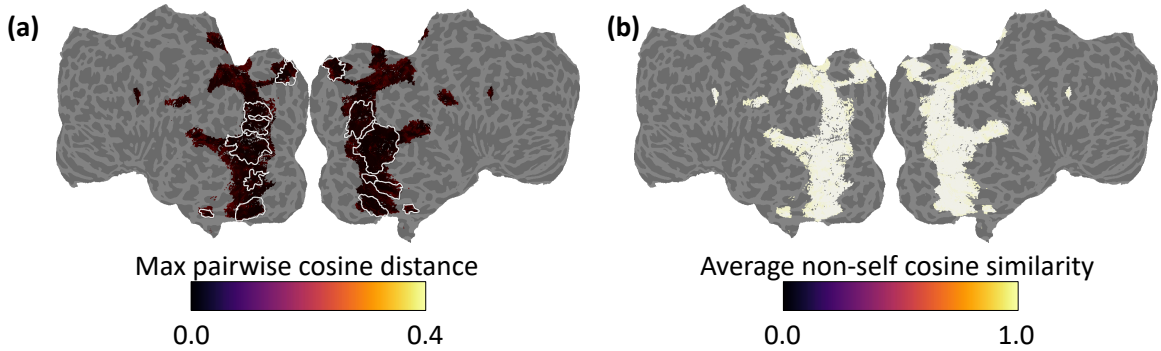


Figure 3.23: **Result on 10-fold cross-validation.** **(a)** We measure the cosine distance of the voxel-wise weights across 10-folds. Visualized is the maximum any-pair voxel-wise distance. We find the average maximum any-pair across voxels is 0.02. **(b)** Average non-self pair-wise cosine similarity across the 10-folds. Note that for each fold, we randomly initialize the weights with kaiming uniform. We find that the fitting process is stable across repeats with an average non-self cosine similarity of 0.98.

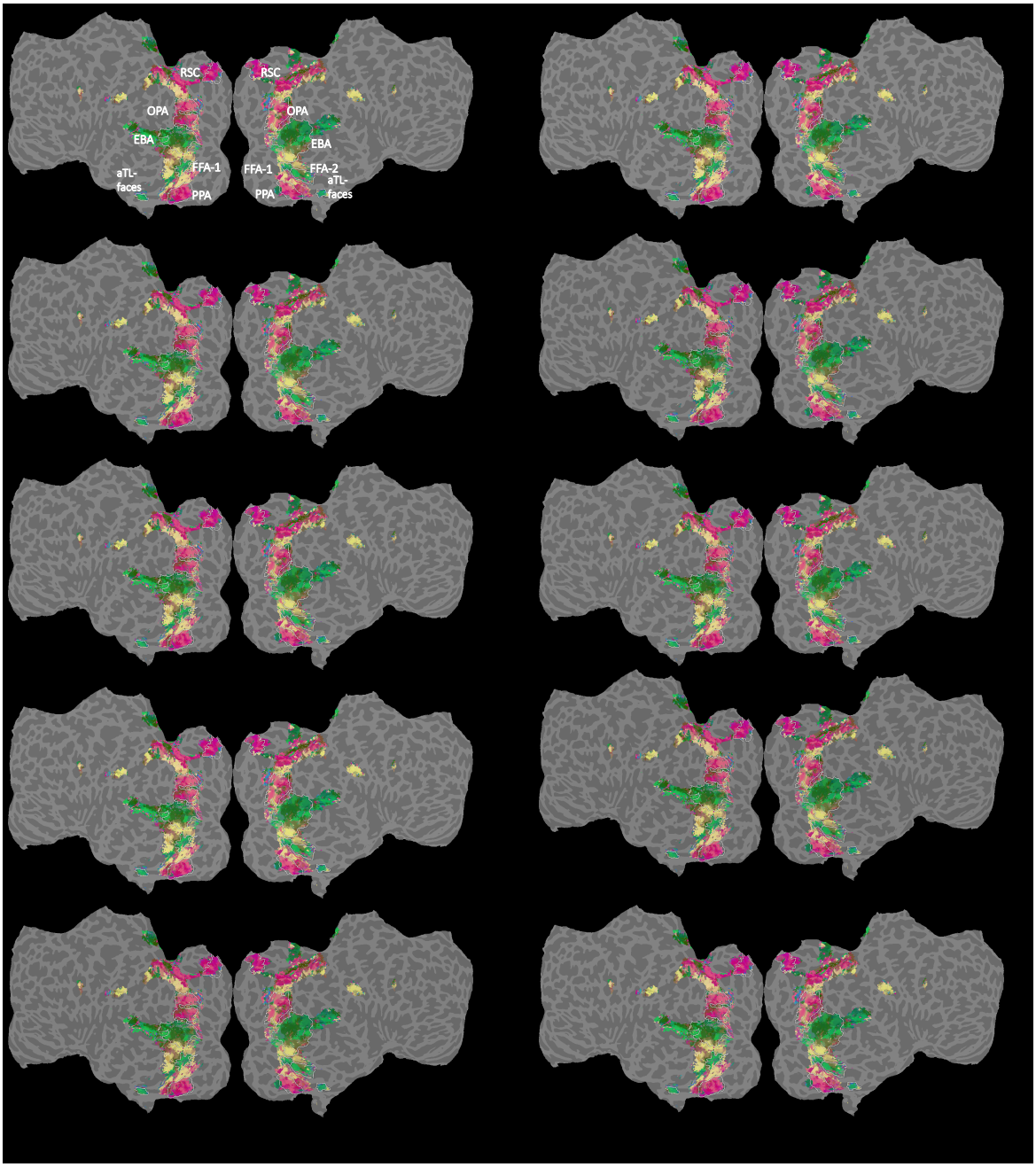


Figure 3.24: **Projection of the 10-fold cross-validation encoders.** We perform UMAP projection using the basis from the main paper on each of the 10 encoder weights. We find that aside from minor differences in the FFA/food intersection on the right hemisphere, the large-scale distribution is similar.

3.6.10 Ground truth functional localizer category distribution

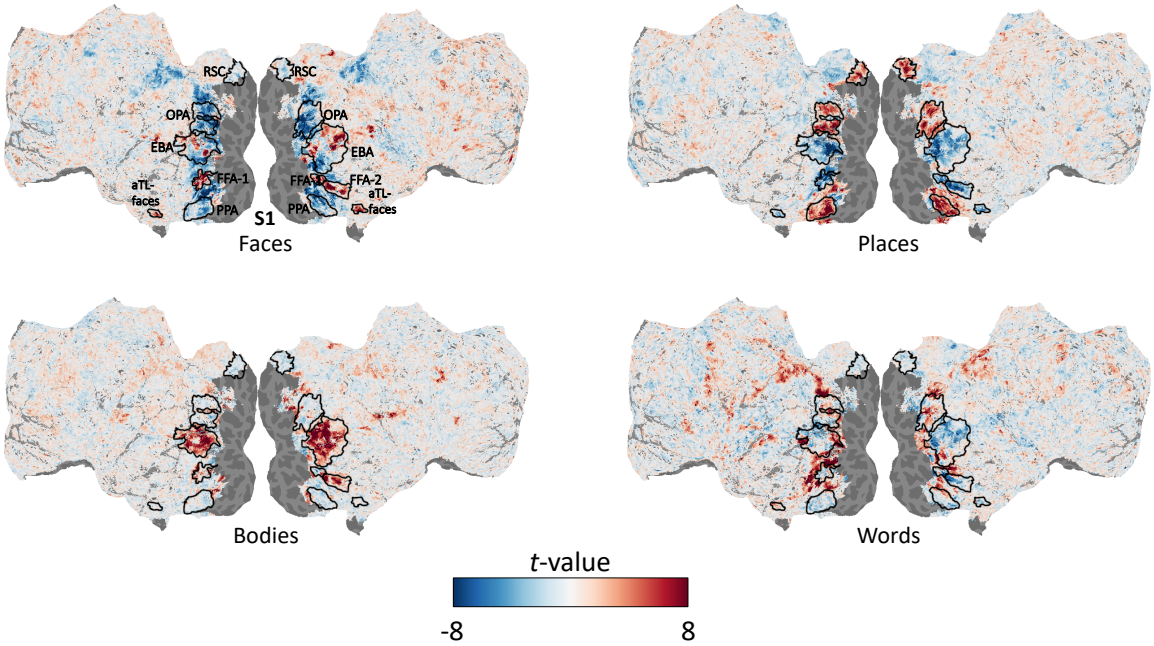


Figure 3.25: **Ground truth t-statistic from functional localizer experiments.** We plot the ground truth functional localizer result t-statistics. The official functional localizer results are provided by NSD, and are collected using the Stanford VPNL fLoc dataset. Here red indicates a region which is activated by images from a category. This plot shows the broad category selectivity present in the high order visual areas.

3.6.11 Fine-grained concept distribution outside EBA

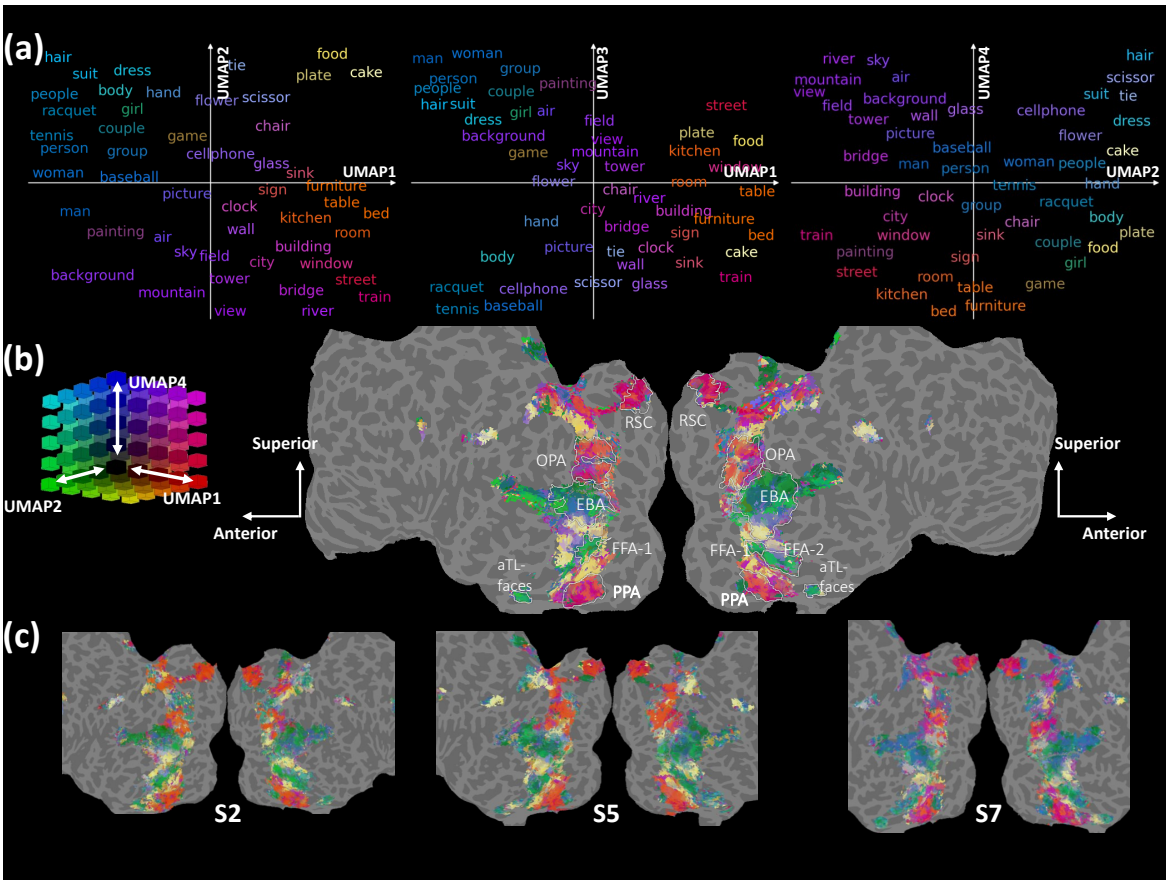


Figure 3.26: **Additional UMAP visualizations.** Here we plot the UMAP dimensionality reduction, and identify the indoor/outdoor concept split in OPA using the 4th UMAP component. Note the Indoor (orange) and Outdoor (purple) gradient along the anterior-posterior axis.

3.6.12 Norm of the embeddings with and without decoupled projection

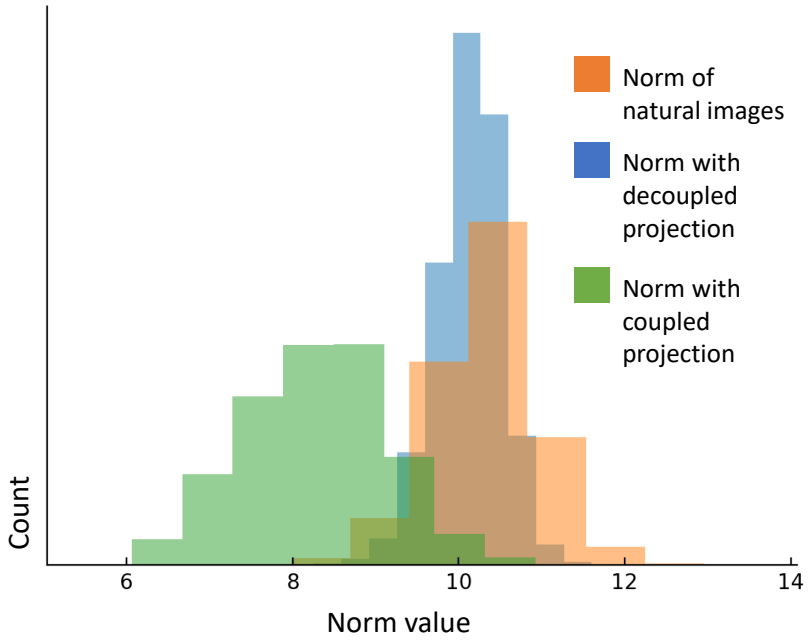


Figure 3.27: **Norm of the embedding vectors.** In the main paper we decouple the projection of the norm and direction. Here we visualize the norm of natural image embeddings in orange, the norm of the post-projection weights using decoupled projection in blue, and the norm of the post-projection weights using coupled norm/direction projection in green. As vectors can cancel each other out, the use of decoupled projection in the main paper yields a better distribution alignment.

Chapter 4

BrainSAIL – Semantic Attribution and Image Localization

4.1 Introduction

Understanding how the human brain processes and represents visual information from natural experience is a fundamental challenge in neuroscience. The vast majority of our knowledge of the visual system comes from tightly controlled experiments using simplified, hand-crafted images or, at best, real-world photographs of objects against noise backgrounds. Although this paradigm has revealed a pattern of preferential neural responses to semantic categories such as faces, places, bodies, words, objects, and food [Aguirre et al., 1996, Allison et al., 1994, Downing et al., 2001, Epstein and Kanwisher, 1998, Grill-Spector, 2003, Jain et al., 2023, Kanwisher et al., 1997, Khosla et al., 2022a, Malach et al., 1995, McCarthy et al., 1997, Pennock et al., 2023b, Sergent et al., 1992b], the visual world we actually experience consists of rich, complex scenes containing many co-occurring objects, textures, and contextual associations [Simoncelli and Olshausen, 2001, Torralba and Oliva, 2003]. As such, using minimal or single-object stimuli narrows the space of hypothesis testing and limits the ecological relevance of any conclusions, leaving us with an incomplete characterization of how the brain represents and processes real-world visual stimuli.

Recent developments in computer vision models trained on web-scale datasets have enabled learning rich multimodal representations that capture semantic concepts in a human-aligned manner [Conwell et al., 2022b, Wang et al., 2022]. In this work, we introduce a novel methodology that leverages the power of such models to decompose selectivity patterns in visual cortex by analyzing responses to dense, localized semantic features present in naturalistic images: **S**emantic **A**ttribution and **I**mage **L**ocalization ("BrainSAIL"). BrainSAIL allows us to isolate the specific image regions that activate different cortical areas when viewing naturalistic scenes. This method advances our prior work by focusing on selectivity in single-object images at the broad category level, thereby enabling a richer decomposition grounded in the full semantic complexity of natural visual experiences.

The core of BrainSAIL involves extracting spatially dense semantic embeddings from images using state-of-the-art models such as CLIP, DINO, or SigLIP [Caron et al., 2021, Radford et al., 2021, Zhai et al., 2023]. These embeddings bridge the traditionally disparate domains of raw vision data, dense deep semantic features, and measured neural responses. Within this rich embedding space, we can isolate and identify the specific visual features and corresponding image regions that drive selectivity effects in different cortical areas during perception of naturalistic visual scenes. By concurrently modeling localized semantic information, high-level semantic categories, and observed brain activity patterns, BrainSAIL can tease apart the image-level visual drivers of neural tuning preferences across higher visual areas. We validate this dense feature mapping method on a large-scale fMRI dataset consisting of human participants viewing many thousands of diverse natural images that span a wide range of semantic categories and visual statistics [Allen et al., 2022].

BrainSAIL's dense embedding framework offers an interpretable view of feature representations across visual regions of the brain. Critically, this view explicitly grounds neural selectivity to localized semantic characteristics inherent in real-world visual experiences. First, we demonstrate the utility of our model for natural images applied to known category-selective regions of the cortex. Second, we show that our model can be used to identify the preference of brain

regions sensitive to scene statistics. Finally, we use our model to compare and contrast the feature selectivity for different vision foundation models. In sum, the dense semantic grounding realized in BrainSAIL enables exciting new directions towards understanding and modeling high-level visual representation in humans.

4.2 Related Work

A growing body of work leveraging computational modeling and machine learning has explored semantic representation in the higher visual cortex. Approaches include generative image models [Gu et al., 2022, Luo et al., 2024, 2023, Pierzchlewicz et al., 2023, Ratan Murty et al., 2021] and the decoding of visual stimuli [Chen et al., 2022, Doerig et al., 2022, Ferrante et al., 2023, Liu et al., 2023, Scotti et al., 2024, Takagi and Nishimoto, 2022]. These diverse studies are united by their consideration of the stimulus image as a whole, primarily focusing on the global information contained within the image rather than the individual scene components. In contrast, the method we introduce explicitly decomposes an image into its semantic components, enabling the identification of individual, semantically meaningful activating concepts within complex natural images.

Semantic Representation in the Visual Cortex. Using hand-crafted image stimuli, functional mapping studies have identified regions in the human brain that respond preferentially to stimuli representing distinct semantic concepts such as faces, places, bodies, words, objects, and food [Aguirre et al., 1996, 1998, Allison et al., 1994, Aminoff et al., 2007, Cohen et al., 2000, Desimone et al., 1984, Downing et al., 2001, Epstein and Kanwisher, 1998, Gauthier and Tarr, 1997, Grill-Spector, 2003, Jain et al., 2023, Kanwisher et al., 1997, Khosla et al., 2022a, McCarthy et al., 1997, Nakamura et al., 2000, O'Craven and Kanwisher, 2000, Pennock et al., 2023b, Sergent et al., 1992b]. One limitation of this simplified approach is that it may not fully capture the contextual complexity of natural vision [Gallant et al., 1998, Mahon, 2022]. Addressing this concern, recent work on image-computable encoders has enabled computational tests of visual selectivity using

naturalistic images [Conwell et al., 2022b, Efird et al., 2024, Eickenberg et al., 2017, Huth et al., 2012, Kubilius et al., 2019, Luo et al., 2024, 2023, Naselaris et al., 2011, Popham et al., 2021, Prince et al., 2023, Wang et al., 2022, Wen et al., 2018, Yamins et al., 2014, Yang et al., 2024a,b]. Building on this work, our method leverages state-of-the-art brain encoding backbones based on vision transformers [Dosovitskiy et al., 2020, Wang et al., 2022] to further explore finer-grained semantic representation in visual cortex.

Visual Contrastive Representation Learning. Self- or weakly-supervised vision models that use contrastive [Chopra et al., 2005, Musgrave et al., 2020, Schultz and Joachims, 2003, Sohn, 2016, Wu et al., 2018, Xing et al., 2002] and masked prediction objectives [Chen et al., 2020, Kolesnikov et al., 2019, Li et al., 2021, Pathak et al., 2016, Zhao et al., 2021, Zhou et al., 2021] are scalable and can be trained on massive, diverse datasets to achieve high zero-shot performance on downstream tasks. Contrastive models such as CLIP, DINO, and SigLIP demonstrate strong classification performance without further fine-tuning [Caron et al., 2021, Oquab et al., 2023, Radford et al., 2021, Zhai et al., 2023]. Models that jointly train on language and vision (CLIP/SigLIP) can also classify images using text-based descriptions without fine-tuning. Interestingly, this high level of performance is mirrored in the fact that contrastive models show high performance for predicting neural responses in visual cortex when paired with linear probes [Conwell et al., 2022b, Wang et al., 2022].

Exploring the Brain with Foundation Models. There has been strong interest in leveraging generative models for decoding (reconstructing) visual stimuli conditioned on brain activations either directly or via intermediate language-based captions [Chen et al., 2023, Doerig et al., 2022, Ferrante et al., 2023, Han et al., 2019, Kamitani and Tong, 2005, Liu et al., 2023, Lu et al., 2023, Mai and Zhang, 2023, Ozcelik and VanRullen, 2023, Ren et al., 2021, Scotti et al., 2024, Seeliger et al., 2018, Shen et al., 2019, Takagi and Nishimoto, 2022]. A related approach generates novel stimuli that are posited to best to activate a target brain region (as opposed to reconstructing the

original stimulus) [Bashivan et al., 2019, Walker et al., 2019] with recent attempts utilizing GANs or Diffusion models to constrain the synthesized output [Gu et al., 2022, Luo et al., 2024, 2023, Ponce et al., 2019, Ratan Murty et al., 2021]. While these models have shown positive results, they all rely on images as a whole, whereas BrainSAIL seeks to disentangle complex images into their semantically meaningful components and localize those parts of the image that elicit activation for different brain voxels or regions.

4.3 Methods

Our aim is to generate spatial attribution maps for arbitrary voxels in the higher visual cortex. Unlike the early visual cortex, which is believed to be primarily selective for "simple features" [Stork and Wilson, 1990], the higher visual cortex exhibits semantic selectivity – a pattern that, at present, is best predicted by deep networks [Conwell et al., 2022b, Wang et al., 2022]. As illustrated in Figure 3.1, to create spatial attributions maps for brain voxels, we first train voxel-wise fMRI encoders to map images to brain activations. Second, we derive dense features from pre-trained vision transformers (ViT) used as the backbone for these encoders. Third, we demonstrate that an artifact-free dense feature map can be derived for high-throughput exploration of selectivity with the visual cortex.

4.3.1 Image-to-Brain Encoders for the Higher Visual Cortex

A voxel-wise image-computable fMRI encoder is a model F_ϕ that predicts fMRI activations (betas) for $B \in \mathbb{R}^{1\times N}$ where N represents the number of voxels in the brain. The encoder is conditioned on image input $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, where $F_\phi(\mathcal{I}) \Rightarrow B$. Recent work has demonstrated that encoders that rely on features extracted from large vision foundation models achieve excellent predictive performance, where higher visual cortex is best predicted by deeper layers in the model [Wang et al., 2022]. In this setting, the backbone model is usually frozen, while a per-voxel adapter typically parameterized as a linear layer is trained to map from network features to voxel

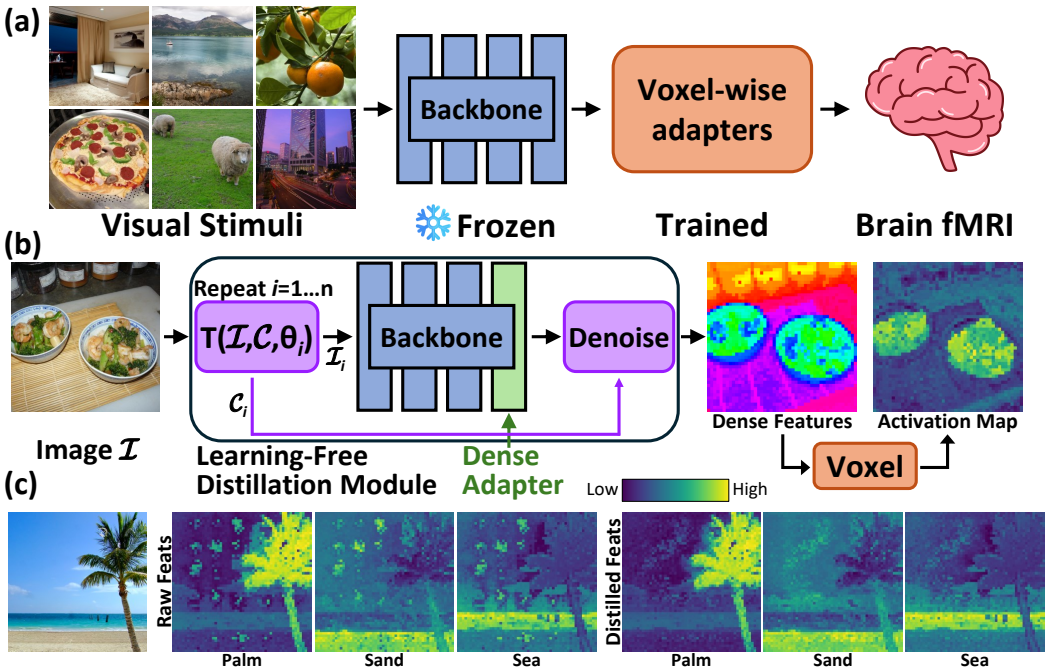


Figure 4.1: **The BrainSAIL framework leverages dense visual features**. **(a)** An fMRI encoder learns a map from images to voxel-wise activations in the brain. Encoders leveraging frozen foundation models based on vision transformers (ViTs) with voxel-wise adapters are currently the highest accuracy models for brain prediction [Conwell et al., 2022b, Wang et al., 2022]. **(b)** Given an image and a ViT backbone for the fMRI encoder, we modify the backbone to output dense features. The dense backbone is wrapped inside of a **Learning-Free Distillation Module**. This module takes an image $\mathcal{I}$ and $2D$ image coordinates $\mathcal{C}$, and generates transformed images and coordinates $(\mathcal{I}_i, \mathcal{C}_i)$ for a given transform θ_i. The dense features and transformed coordinates are provided to a denoising module to generate clean dense features. A voxel-wise adapter then generates dense relevance maps which highlight the image regions activating the voxel. **(c)** Using `CLIP ViT-B/16` with the latest `NACLIP` adapter, we show relevance maps using the CLIP text encoder. The original, raw features are highly noisy and contain artifacts, while the distilled features are localized to the relevant semantic components with high accuracy. Note that we achieve state-of-the-art open vocabulary CLIP-based segmentation results using our method.

activations. In that we focus on the higher visual cortex exclusively, we utilize a two component design for our encoder: (1) a frozen vision foundation model backbone $\mathrm{G}(\mathcal{I})$ which outputs a $R^{1\times M}$ dimension embedding vector for each image; (2) a per-voxel adapter parameterized as a linear probe with weight $W \in \mathcal{R}^{M\times N}$ and bias $b \in \mathcal{R}^{1\times N}$, which takes as input a unit-norm

image embedding.

$$\left[\frac{\mathrm{G}_{\mathrm{img}}(\mathcal{I})}{\|\mathrm{G}_{\mathrm{img}}(\mathcal{I})\|_2} \times W + b\right] \Rightarrow B \tag{4.1}$$

It should be noted that BrainSAIL is not restricted to linear probes, and can work with arbitrary voxel-wise parameterizations, including MLPs. Linear probes are used here as they are widely adopted in fMRI encoder literature and empirically achieve good performance. BrainSAIL is compatible with any Vision Transformer (ViT)-based model, making it readily applicable to the vast majority of modern visual foundation models which predominantly employ ViT architectures. Additional results are presented in the supplemental. We train our model with MSE loss, and evaluate the encoder on the test set. In Figure 4.7 we show that our encoder achieves state-of-the-art R^2.

4.3.2 Deriving Dense Features from ViT backbones

The emergence of vision models trained on a contrastive image-text objective has fueled interest in zero-shot open-vocabulary image classification methods. For example, CLIP has shown that images can be classified without foreknowledge of the test time classes during training; instead the category of interest can be described using language during test time. Of late, this capability has been extended from classification to segmentation. Compared to methods that require human annotation [Li et al., 2022a] and perform poorly on out-of-distribution images [Jatavallabhula et al., 2023, Kerr et al., 2023], these new methods require no further training and directly extract dense features that lie in the same space as the image/text embedding. These dense feature extraction methods operate by modifying the last self-attention (SA) block within the typical ViT architecture (MaskCLIP, Zhou et al. [2022]; SCLIP, Wang et al. [2023]; NACLIP, Hajimiri et al. [2024]). For vision models of this sort trained on a contrastive objective, the output is composed of a single `[CLS]` token, which is supervised using a contrastive loss; and numerous patch tokens which correspond to specific spatial locations. Let (q_i, k_i, v_i) be the query, key,

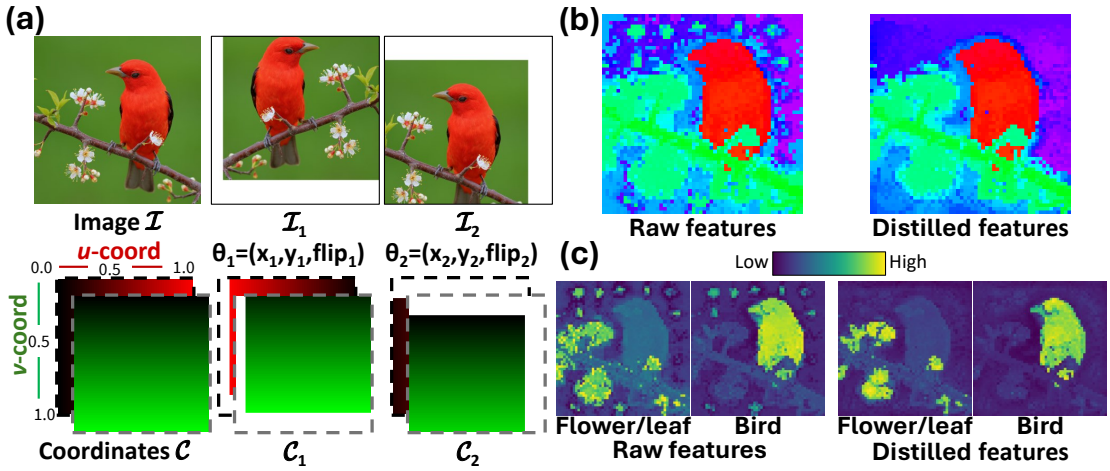


Figure 4.2: **The Learning-Free Distillation Module**. **(a)** Given an image, we generate image-space coordinates (u, v) for each pixel. We then randomly sample from $\theta_{1...n}$, where θ_i has vertical/horizontal offset, and left-right flips. The augmented images are provided to a frozen backbone with dense adapter. The features are projected to the original image space via an inverse transform $\mathcal{T}^{-1}(\theta_i)$. **(b)** UMAP visualization of the dense features. The same fitted basis is used for both visualizations. **(c)** With CLIP, we can perform zero-shot text queries. Note the artifacts above the bird's head. In practice artifact location is different for each image. The distilled results are significantly better.

value features respectively for a single image patch i, with a total of m spatial patches. For a given patch j at the final layer, where f denotes any function applied to the `[CLS]` after the last self-attention, the `[CLS]` token and each dense token is a convex combination of v features:

$$\text{Out\_Orig}_j = f\left(\sum_{k=1}^{m}\left[\text{softmax}(\frac{q_j k^T}{C})_j \cdot v_k\right]\right) \quad \text{Out\_Mask}_j = f(v_j) \quad \text{Out\_NA}_j = f\left(\sum_{k=1}^{m}\left[\text{softmax}(\frac{q_j q^T + \omega_j}{C})_j \cdot v_k\right]\right) \quad (2)$$

MaskCLIP proposes to directly remove the convex re-weighting and output the value feature for each patch token directly. SCLIP and NACLIP reintroduce the weighting to reduce output artifacts, but modify it with correlative self-attention (CSA); or by using CSA with a spatial attentive bias ω. Here, we utilize NACLIP as the dense adaptor for CLIP. The other two backbones in Section 4.4.4 use an updated ViT architecture with "register tokens" [Darcet et al., 2023]. As these have not been explored in the context of CSA, we utilize MaskCLIP as the dense adaptor.

4.3.3 Learning-Free Feature Distillation

As only the [CLS] is supervised in these contrastive models, as shown in Figures 4.1 and 4.2, the extracted dense embeddings often have artifacts – even when using the latest NACLIP method which seeks to reduce artifacts. While methods such as Darcet et al. [2023] improve spatial consistency via architectural improvements, they require training the model with architecture modifications that are computationally costly. Consequently, in order to facilitate high-throughput characterization of the visual cortex over large datasets, we propose an efficient learning-free distillation module. Given an image $\mathcal{I}$, we first generate n augmentation parameters $\theta_{1...n}$, where θ_i consists of a horizontal/vertical offset (u_i, v_i) and horizontal flip$_i \in \{0, 1\}$. We further generate the image space coordinates $\mathcal{C} = (\text{u-coord}, \text{v-coord})$, where $u \in [0, 1]$ goes from top-to-bottom, while $v \in [0, 1]$ goes left-to-right. We describe our full transform in Algorithm 4.3. Our method distills a clean semantic map, as visual semantics are equivariant to shift and horizontal flips. We note that averaging over the number of augmentation is extracting an *optimal* embedding under mean squared error (squared euclidean). Let $\vec{p^*}$ be the optimal embedding under MSE for a given patch, and $\vec{p_i}$ with $i \in \{1...n\}$ be the feature candidates under image augmentation:

Input: Image $\mathcal{I}$;
Image space coordinates $\mathcal{C}$;
Augmentation parameters $\theta_{1...n}$;
Augmentation function $\mathcal{T}$;
ViT model with dense adapter M;
1. Zero init clean feature tensor Q
2. Zero init count tensor K
3: **For** i in {1...n}:
4. $\theta_i = (u_i, v_i, \text{flip}_i)$
5. $(\mathcal{I}_i, \mathcal{C}_i) = \mathcal{T}(\mathcal{I}, \mathcal{C}, \theta_i)$
6. Dense feature $F_i = M(\mathcal{I}_i)$
7. $(F_i^{\text{valid}}, \mathcal{C}_i^{\text{valid}}) = \mathcal{T}^{-1}(F_i, \mathcal{C}_i, \theta_i)$
8. $Q[\mathcal{C}_i^{\text{valid}}] = Q[\mathcal{C}_i^{\text{valid}}] + F_i^{\text{valid}}$
9. $K[\mathcal{C}_i^{\text{valid}}] = K[\mathcal{C}_i^{\text{valid}}] + 1$
10. **return** Q/K

Algo 4.3: **Learning-Free Feature Distillation**

$$\vec{p^*} = \min_{\hat{p}} \left(\sum_{i=1}^{n} \|\vec{p_i} - \hat{p}\|_2^2 \right) = \min_{\hat{p}} \left(\|\vec{p_1} - \hat{p}\|_2^2 + ... + \|\vec{p_n} - \hat{p}\|_2^2 \right) \tag{3}$$

$$= \min_{\hat{p}} \left(\vec{p_1}^T \vec{p_1} - 2\vec{p_1}^T \hat{p} + \hat{p}^T \hat{p} + ... + \vec{p_n}^T \vec{p_n} - 2\vec{p_n}^T \hat{p} + \hat{p}^T \hat{p} \right) \quad \text{omitting } \vec{p_i}^T \vec{p_i} \tag{4}$$

$$= \min_{\hat{p}} \left(n \cdot \hat{p}^T \hat{p} - 2 \sum_{i=1}^{n} (\vec{p_i}^T \hat{p}) \right) = \min_{\hat{p}} \left(n \cdot \hat{p}^T \hat{p} - 2n \sum_{i=1}^{n} ((1/n) \cdot \vec{p_i}^T \hat{p}) \right) \tag{5}$$

	ADE20k		COCO Object		COCO Stuff		VOC20	
	mIoU↑	Pearson↑	mIoU↑	Pearson↑	mIoU↑	Pearson↑	mIoU↑	Pearson↑
SCLIP	16.45	0.308	33.52	0.353	21.95	0.309	**81.54**	**0.551**
NACLIP	17.69	0.425	33.14	0.418	22.58	0.393	77.09	0.473
+Distilled	**18.19**	**0.443**	**34.15**	**0.435**	**23.08**	**0.405**	79.09	0.489

Table 4.1: **Open-vocabulary segmentation with dense CLIP features.** We validate the effectiveness of our learning-free smoothing approach on segmentation datasets in a zero-shot setting (without any training). This performance is state-of-the-art for open vocabulary segmentation. Note mIoU scores are multiplied by 100. Our smoothing improves the results.

The objective can be expressed as $\|\hat{p} - (1/n)\sum \vec{p_i}\|_2^2 \geq 0$, then $\vec{p^*} = (1/n)\sum \vec{p_i}$.

In Table 4.1, we compare pre- and post- smoothing results. Under the Pearson metric, which does not assume prior category knowledge, smoothing yields the best performance in three of four datasets. We apply the voxel-wise adapters to the per patch dense features to derive the final relevance map.

4.4 Results

We utilize BrainSAIL to localize the semantic selectivity of different brain regions and demonstrate that the relevance maps are interpretable throughout the brain and correlate well with the known category-selective regions. We then explore the selectivity of higher visual cortex with respect to localized scene structure and image properties. Finally, we compare and contrast the localization results from three different vision foundation models. These results establish BrainSAIL as a novel technique for mapping and understanding the semantics of visual representations in the brain.

4.4.1 Setup

We use the Natural Scenes Dataset (NSD; Allen et al. [2022]), the largest 7T fMRI dataset of human visual responses, focusing on four subjects (S1, S2, S5, S7) who viewed the full 10,000 image set (a subset of COCO images) three times each. fMRI activations (betas) were derived

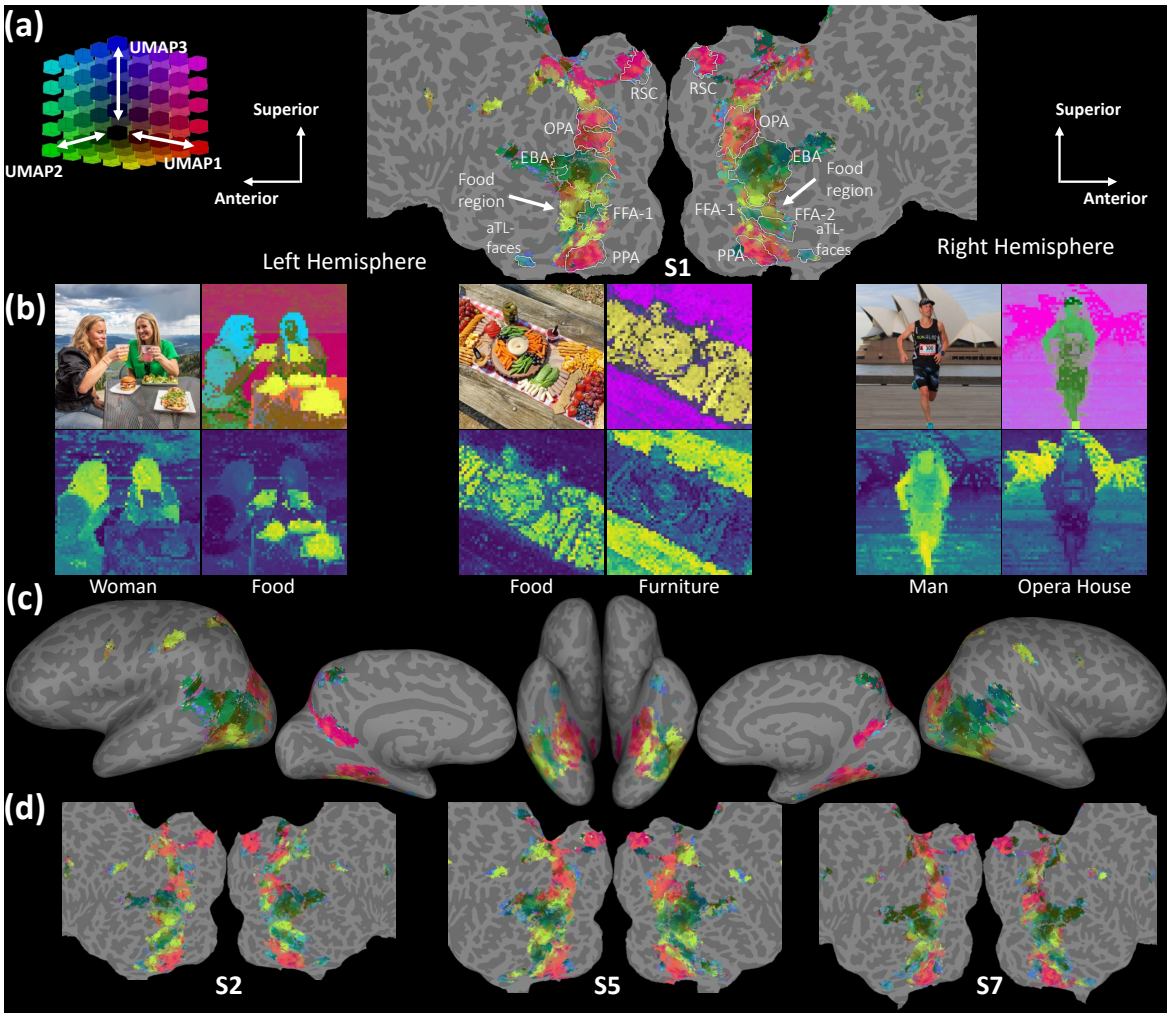


Figure 4.4: **Joint dimensional reduction of higher visual cortex encoder weights and images using BrainSAIL**. We use a UMAP to perform visualization of the encoder weights. This same UMAP basis is reused for images. **(a)** Cortical flatmap of **S1**. Note that the overlaid white region outlines and labels were derived from *functional localizer data collected independently from the visualized UMAP results*. **(b)** Embeddings from novel images are computed with BrainSAIL and transformed using the fMRI UMAP. For each quartet of images, the content is as follows. Top left: Original RGB image; Top right: Dimension reduction of BrainSAIL embeddings for the image; Bottom: Two text queries using CLIP text branch showing language-indicated relevance results. **(c)** UMAP results on an inflated view of the brain for **S1**. **(d)** UMAP results on cortical flatmaps for **S2**, **S5** and **S7**. These results demonstrate that BrainSAIL can effectively localize semantically meaningful components of natural images and map them to appropriate brain regions. The cortical maps show color-coded mappings that align well with functionally-defined regions: body regions (EBA), face regions (FFA/aTL-faces), place regions (RSC/OPA/PPA), and food regions (yellow). Note that the food regions have been identified as flanking FFA by Jain et al. [2023], but we do not have independent functional localizer data for food for these subjects.

Region	Faces				Places				Bodies				Words				Food			
	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7	S1	S2	S5	S7
Face	**46**	**48**	**54**	**40**	1	1	2	2	12	11	12	13	9	11	12	11	1	2	1	5
Places	1	1	1	3	**76**	**80**	**89**	**75**	0	2	3	2	9	12	11	9	8	7	10	7
Bodies	26	16	21	27	5	1	0	1	**50**	**41**	**55**	**48**	15	13	21	**30**	9	5	1	5
Words	1	7	3	8	6	3	2	9	8	6	7	7	**38**	28	26	23	16	13	17	35
Food	26	28	21	22	12	15	7	13	30	40	23	30	29	**36**	**30**	27	**66**	**73**	**71**	**48**

Table 4.2: **CLIP text alignment for each category selective brain region.** For each category selective brain region, we take the top-100 images from the NSD test set that elicit the highest fMRI response for each region. We then use BrainSAIL to compute the relevance maps for the top-100 images for each region. For each image, its relevance map is computed using the CLIP text encoder with text prompts from the five relevant categories. The text prompt with the highest Pearson correlation to the BrainSAIL relevance map is recorded as the category for that image. Units are in %.

using GLMSingle [Prince et al., 2022] and normalized per session ($\mu = 0, \sigma^2 = 1$). Responses to repeated images were averaged. A brain encoder for each subject was trained on ~ 9000 unique images per subject, with the remaining ~ 1000 images viewed by all subjects being used for R^2 validation as the test set. Supplementary results for other subjects are included in the appendix. Face, place, body, and word regions were defined using independent category localizer data from NSD with a threshold of $t > 2$ [Stigliani et al., 2015]. Food regions were defined using masks provided by Jain et al. [2023].

We train three encoders based on different neural network backbones. For all three, we utilize the `ViT-Base` model size. **(1)** For CLIP, we utilize OpenAI's official `ViT-B/16` weights. This is a network trained on an infoNCE contrastive image-text objective. **(2)** For DINO, we utilize the latest official `DINOv2 ViT-B/14+reg`, and is a network trained on image-only self-supervision [Darcet et al., 2023]. **(3)** For SigLIP, we utilize NVIDIA's implementation based on `RADIOv2.5 ViT-B/16` [Ranzinger et al., 2024], as the original Google variant used a non-standard architecture. SigLIP utilizes a pairwise non-contrastive image-text objective [Zhai et al., 2023]. All fMRI encoders are trained using MSE loss, with the backbone frozen. We validate the test time R^2 in Figure 4.7 and find that we achieve state-of-the-art results similar to Wang et al. [2022] and Luo et al. [2024]. We use CLIP for Sections 4.4.2 and 4.4.3, as it is the most widely used backbone in fMRI literature. We use 51 augmentation steps unless otherwise

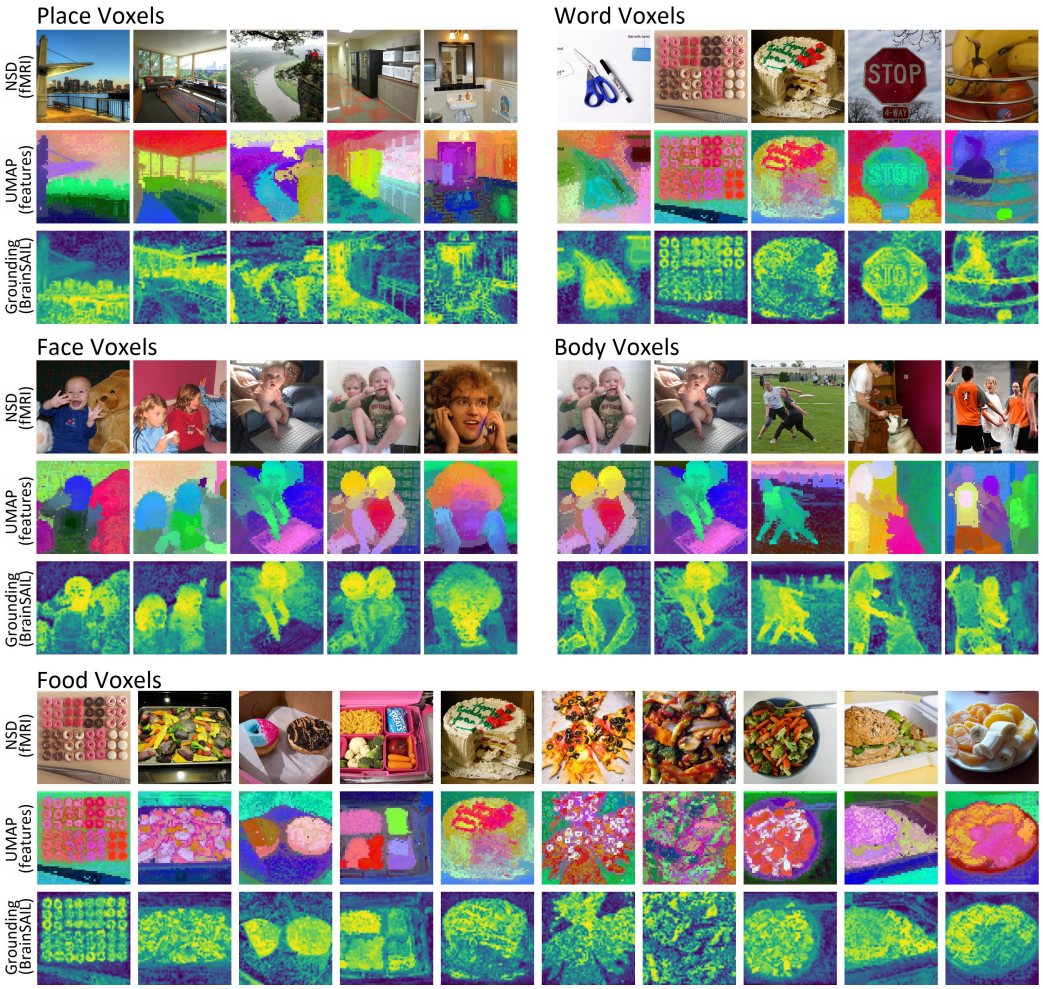


Figure 4.5: **Grounding results using BrainSAIL**. We visualize the top test set images as predicted by the CLIP fMRI encoder for each category selective region. For each image, we also visualize the image-wise UMAP for the distilled dense features. Note the UMAP basis here is computed imagewise, and not shared with Figure 3.3. For each image, we further visualize the feature relevancy map for the category selective voxels illustrating that this method extracts the semantically relevant regions in complex compositional images.

noted.

4.4.2 Image Factorization using the Brain

To explore how different areas in higher visual cortex align to different image parts we apply UMAP [McInnes et al., 2018] with an angular metric to linear brain weights and apply the same UMAP basis to dense features as produced by BrainSAIL. Note that during dimensionality reduction **we do not utilize any cortex category masks from NSD** – the region of interest outlines on the cortex in Figure 4.4a are for **visualization purposes only** and are derived from

independent NSD functional localizers. As shown in Figure 4.4, we find that the factorization of the brain is well aligned to pre-identified functional regions, and broadly segments the cortex into axes along "people", "scenes" and "food". In particular, place regions, including the retrosplenial cortex (RSC), occipital place area (OPA), and parahippocampal place area (PPA), show selectivity for scene components (magenta). People regions, including the extrastriate body area (EBA), fusiform face area (FFA), occipital face area (OFA), show selectivity for face and body parts in the image (Green-Blue). Finally, we find that the recently identified food region that roughly surrounds FFA (Yellow) Jain et al. [2023], Khosla et al. [2022a], Pennock et al. [2023b] strongly corresponds to food in images. These results establish that BrainSAIL can be used to characterize higher-level selectivity to individual semantic categories in complex natural images without prior knowledge of their semantic selectivity.

We further quantify the feature relevance maps for broad category selective regions in Figure 4.5 and Table 4.2. We use the brain encoder to predict the top-5 images for the place/word/face/body regions, and the top-10 images for the food region. We find that our method can effectively localize the objects relevant to each category- selective brain region. Note that the word region is known to have cross-selectivity to faces [Mei et al., 2010] and food [Khosla and Wehbe, 2022].

4.4.3 Cortex Selectivity to Image Features

Going beyond semantic categories, we seek to explore the low- and mid-level image feature correlates that correspond to different brain regions. Prior work explored this by training a convolutional encoder on each NSD subject, which is limited to $\sim 10,000$ images each [Sarch et al., 2023]. One concern is that using a small dataset with a convolutional backbone can lead to overfitting to the dataset's specific features and exacerbate the inherent biases of convolutional networks. To address this limitation, our method leverages vision transformers trained on massive datasets of hundreds of millions of images, thereby avoiding the hard-coded inductive biases present in CNNs [Raghu et al., 2021]. We visualize BrainSAIL feature dissection results in Figure 4.6. Our method can successfully identify the known scene selective regions (RSC/OPA/PPA) as preferring

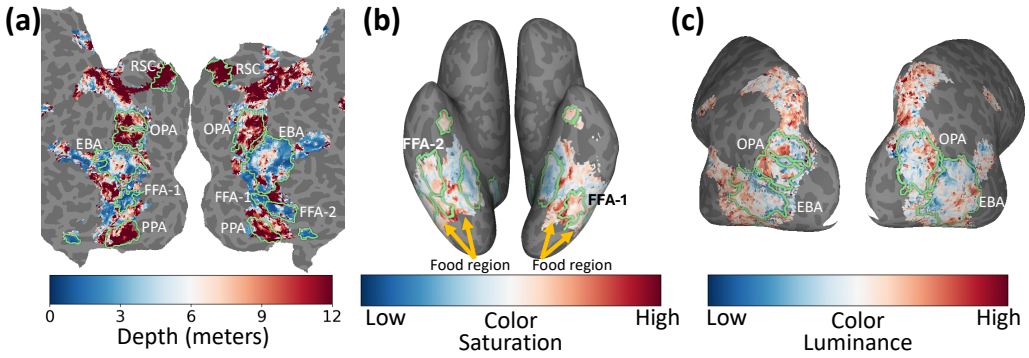


Figure 4.6: **Feature correlates with BrainSAIL**. We visualize the depth, color saturation, and color luminance (brightness) correlates for each brain region using BrainSAIL . **(a)** The scene selective regions, retrosplenial cortex (**RSC**), parahippocampal place area (**PPA**), and occipital place area (**OPA**) are all identified as having a preference for high depth. **(b)** On the ventral surface, we identify two stripes on each hemisphere, surrounding FFA with high saturation preference. These are the same brain regions identified by Jain et al. [2023] as being food selective. **(c)** In OPA, we identify an anterior/posterior split, where one region has high color luminance preference, and the other has low color luminance preference. This are the same regions identified by Luo et al. [2023] as being **outdoor/indoor** selective.

high depth, and is successful even in OPA where Sarch et al. [2023] fails. We believe this is likely because OPA processes higher-level associative content and affordances [Aminoff and Tarr, 2021, Bonner and Epstein, 2017]. Similarly, we identify the region surrounding FFA as being selective to high color saturation, which correspond to the food regions identified by Jain et al. [2023] and others. In OPA, we identify a split in color luminance preference, which is similar to the indoor/outdoor preferring regions identified by Lescroart and Gallant [2019], Peer et al. [2019], and Luo et al. [2023]. These results demonstrate that our method can identify fine-grained selectivity with more broadly characterized brain regions.

4.4.4 Are Brain Encoders Equivalent?

Recent high-performing models such as CLIP, DINO, and SigLIP differ in their training objectives, architectures, and datasets: CLIP employs a contrastive image-language objective, DINO utilizes a self-supervised image loss without explicit linguistic guidance, and SigLIP leverages a non-contrastive pairwise image-language loss. Despite these differences, when employed as the backbone for fMRI encoders, these models exhibit similar performance in predicting brain responses, achieving comparable R^2 values on the test set as shown in Figure 4.7. This observation

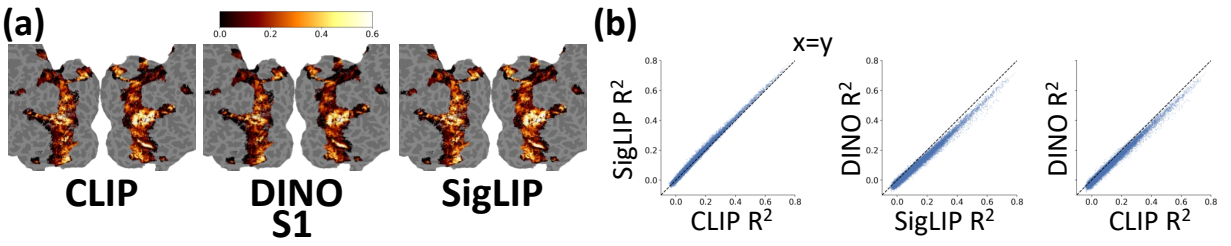


Figure 4.7: **Comparing the brain prediction performance for different encoder backbones**. **(a)** We validate each encoder R^2 on a test set and find that all three models achieve very high performance (comparable to Wang et al. [2022]). **(b)** The voxel-wise correlation of test set R^2 for the three models. CLIP and SigLIP, which rely on language supervision, achieve higher performance than DINO (which trained via self-supervision with images).

raises an important question about the nature of each model's learned features and their alignment with one another: Do these models converge upon similar feature representations for category selective brain regions despite their varied training paradigms?

To investigate the representational differences between the models, we perform BrainSAIL analysis for scene, face, and food-selective brain regions and qualitatively visualize the results in Figure 4.9. While all three models exhibit broad similarities in their grounding maps, DINO, trained without language supervision, demonstrates a stronger sensitivity to low-level visual features compared to CLIP and SigLIP (for comparisons between CLIP, ResNet, and simCLR see Wang et al. [2022]). This is evident in the food region (Figure 4.9), where DINO's grounding map for a pizza image excludes the toppings and assigns lower relevance to non-orange elements in a fruit bowl, suggesting a focus on color and texture rather than the concept of "food" itself. Similarly, in the face region, DINO's grounding map exhibits less reliance on semantically relevant features such as eyes, nose, and mouth. We hypothesize that this greater sensitivity to visual features in DINO stems from its lack of language guidance during training, preventing it from learning the higher-level semantic correlations that link visually disparate parts and objects within a category. As high-performing "proxy models" of visual brain representation [Leeds et al., 2013], these and other underlying model characteristics – architecture, training objective, training dataset, etc. – are important considerations for developing more robust encoding models that can bridge the gap between artificial and biological vision systems.

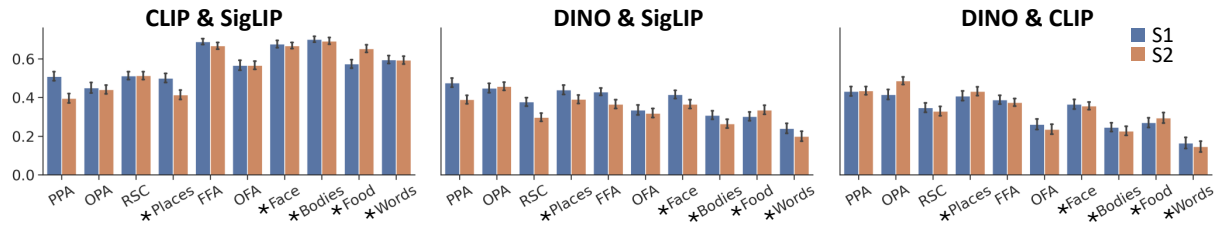


Figure 4.8: **Model similarity across ROIs**. Brain encoder backbone spatial similarity for the ground truth top-100 images from the test set for each category-selective brain region. A $\star$ denotes a domain-defined network of regions encompassing multiple ROIs. CLIP and SigLIP relevance maps are more similar to one another than either is to DINO. Error bars indicate standard error across the 100 images.

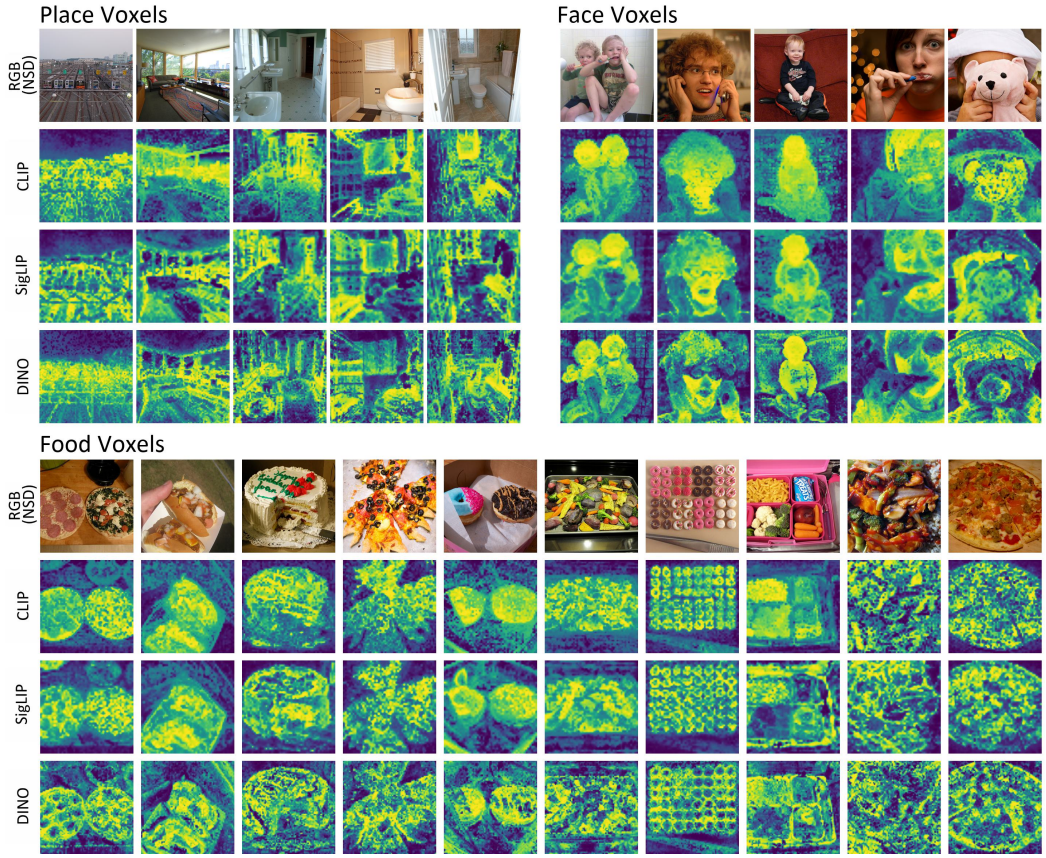


Figure 4.9: **Comparing different brain encoder backbones with BrainSAIL**. Visualization of the top test set images for the place, face, and food category-selective brain regions as predicted by CLIP, SigLIP, and DINO. While all models show broadly similar feature relevance for a given brain area, there are important differences. DINO, with no language supervision, exhibits greater sensitivity to visual similarity, at the cost of semantic coherence.

4.5 Discussion

Limitations and Future Work. BrainSAIL achieves strong localization performance and benefits from a pre-trained vision transformer, reducing reliance on the fMRI dataset for backbone training. However, it is still necessary to train the fMRI encoder on these data, and thus potential dataset biases in the human neural data and how it was collected can influence the learned representations and conclusions. Future work should explore training on larger and more diverse neural datasets to mitigate this limitation and enhance the generalizability of our findings.

Conclusion. We propose BrainSAIL, a method that leverages vision foundation models to interrogate which semantic components of complex natural images lead to the neural activation of specific regions of the brain. Based on the vision transformer architecture, we: (1) semantically attribute and localize relevant objects in complex compositional images; (2) jointly factorize images and semantically selective regions in the human brain; (3) identify the feature correlates of depth, saturation, and luminance that underlie semantic selectivity; (4) explicate differences in fMRI encoders that achieve similar overall brain prediction performance. *In toto,* these results establish that BrainSAIL is a powerful new approach to data-driven explorations of the human higher visual cortex.

Chapter 5

Conclusion

This thesis presents three novel methods that leverage recent advancements in computer vision to study the semantic selectivity of the human visual cortex in a data-driven and ecologically valid manner. The methods I propose leverage powerful vision models trained on massive image datasets. These models, capable of extracting semantically rich representations from natural images, enabled us to develop techniques for image synthesis, voxel-wise semantic captioning, and spatial attribution of semantic selectivity within the visual cortex.

My findings demonstrate the utility of these methods in uncovering previously unobserved aspects of visual cortex organization. These include fine-grained functional distinctions within established regions, novel functional subdivisions, and sensitivity to both high-level semantic categories and lower-level visual features. The ability to synthesize images specifically designed to activate targeted brain regions provides a powerful tool for exploring the underlying feature preferences and generating novel stimuli for future experiments. Additionally, generating natural language descriptions of voxel-wise preferred stimuli facilitates a deeper understanding of feature representations across the visual cortex, going beyond traditional category-level analyses. Finally, by spatially attributing selectivity within natural images, we gain valuable insights into the neural mechanisms underlying semantic processing in real-world contexts. By grounding my analyses in naturalistic images and generating interpretable outputs, this work paves the way for a more

nuanced and data-driven understanding of the human visual system. The developed techniques have broad implications for future research: 1. The data-driven nature of these methods can lead to the formulation of novel hypotheses regarding functional organization and selectivity within the visual cortex. 2. Synthesized images and captions can be used to create more effective and targeted stimuli for future fMRI experiments, testing specific hypotheses generated by our findings. 3. The framework allows for a direct comparison of different vision models in terms of their alignment with human brain representations, aiding in the development of more biologically plausible artificial vision systems. This thesis underscores the power of integrating advancements in computer vision with neuroscience research. By leveraging these powerful tools, we can move beyond the limitations of traditional approaches, ultimately gaining a more complete and ecologically valid understanding of visual cognition.

Bibliography

Geoffrey K Aguirre, John A Detre, David C Alsop, and Mark D'Esposito. The parahippocampus subserves topographical learning in man. *Cerebral cortex*, 6(6):823–829, 1996. 4.1, 4.2

Geoffrey K Aguirre, Eric Zarahn, and Mark D'Esposito. An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron*, 21(2):373–383, 1998. 4.2

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022. 2.1, 2.2, 2.4.1, 3.1, 3.4.1, 4.1, 4.4.1

Truett Allison, Gregory McCarthy, Anna Nobre, Aina Puce, and Aysenil Belger. Human extrastriate visual cortex and the perception of faces, words, numbers, and colors. *Cerebral cortex*, 4 (5):544–554, 1994. 4.1, 4.2

Elissa Aminoff, Nurit Gronau, and Moshe Bar. The parahippocampal cortex mediates spatial and nonspatial associations. *Cerebral cortex*, 17(7):1493–1503, 2007. 4.2

Elissa M Aminoff and Michael J Tarr. Functional context affects scene processing. *Journal of cognitive neuroscience*, 33(5):933–945, 2021. 4.4.3

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005. 2.4.4

Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image

synthesis. *Science*, 364(6439):eaav9436, 2019. 2.2, 3.2, 4.2

Michael F Bonner and Russell A Epstein. Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, 114(18):4793–4798, 2017. 4.4.3

Judy Borowski, Roland S Zimmermann, Judith Schepers, Robert Geirhos, Thomas SA Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain cnn activations better than state-of-the-art feature visualization. *arXiv preprint arXiv:2010.12606*, 2020. 3.2

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2.2

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 4.1, 4.2

Andrea E Cavanna and Michael R Trimble. The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129(3):564–583, 2006. 3.4.4

Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, 6(1):1–18, 2019. 2.1

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 4.2

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. *arXiv preprint arXiv:2211.06956*, 1(2):4, 2022. 2.2, 3.2, 4.2

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages

22710–22720, 2023. 3.2, 4.2

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 3.2

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005. 4.2

Laurent Cohen, Stanislas Dehaene, Lionel Naccache, Stéphane Lehéricy, Ghislaine Dehaene-Lambertz, Marie-Anne Hénaff, and François Michel. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2):291–307, 2000. 2.2, 3.2, 4.2

Colin Conwell, Jacob S Prince, George Alvarez, and Talia Konkle. Large-scale benchmarking of diverse artificial vision models in prediction of 7t human neuroimaging data. *bioRxiv*, pages 2022–03, 2022a. 2.2, 2.3.2

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, pages 2022–03, 2022b. 4.1, 4.2, 4.2, 4.3, 4.1

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, 2023. doi: 10.1101/2022.03.28.485868. 3.2, 3.3.1

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 4.3.2, 4.3.3, 4.4.1

Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant

to stimulus modality. *Journal of Neuroscience*, 39(39):7722–7736, 2019. 3.2

Robert Desimone, Thomas D Albright, Charles G Gross, and Charles Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4(8):2051–2062, 1984. 2.2, 3.2, 4.2

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2.3.3

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 2.2

Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*, 2022. 3.2, 3.2, 4.2, 4.2

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2.6.13, 4.2

Paul E Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001. 2.2, 3.1, 3.2, 3.4.4, 4.1, 4.2

Cory Efird, Alex Murphy, Joel Zylberberg, and Alona Fyshe. What's the opposite of a face? finding shared decodable concepts and their negations in the brain. *arXiv preprint arXiv:2405.17663*, 2024. 4.2

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152: 184–194, 2017. 2.2, 3.2, 4.2

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment.

Nature, 392(6676):598–601, 1998. 2.1, 2.2, 3.1, 3.2, 4.1, 4.2

Gidon Felsen and Yang Dan. A natural approach to studying vision. *Nature neuroscience*, 8(12): 1643–1646, 2005. 3.2

Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain captioning: Decoding human brain activity into images and text. *arXiv preprint arXiv:2305.11560*, 2023. 3.2, 3.2, 4.2, 4.2

Jack L Gallant, Charles E Connor, and David C Van Essen. Neural activity in areas v1, v2 and v4 during free viewing of natural scenes compared to controlled viewing. *Neuroreport*, 9(7): 1673–1678, 1998. 2.2, 3.2, 4.2

James S Gao, Alexander G Huth, Mark D Lescroart, and Jack L Gallant. Pycortex: an interactive surface visualizer for fmri. *Frontiers in neuroinformatics*, page 23, 2015. 2.6.13

Isabel Gauthier and Michael J Tarr. Becoming a "greeble" expert: Exploring mechanisms for face recognition. *Vision research*, 37(12):1673–1682, 1997. 4.2

Isabel Gauthier, Marlene Behrmann, and Michael J Tarr. Can face recognition really be dissociated from object recognition? *Journal of cognitive neuroscience*, 11(4):349–370, 1999. 3.1

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016. 3.7

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2.2

Kalanit Grill-Spector. The neural basis of object perception. *Current opinion in neurobiology*, 13 (2):159–166, 2003. 3.1, 3.2, 4.1, 4.2

Kalanit Grill-Spector and Rafael Malach. The human visual cortex. *Annual Review of Neuroscience*, 27:649–677, 2004. ISSN 0147006X. doi: 10.1146/ANNUREV.NEURO.27.070203.

144220. 2.1, 2.2

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. 2.1, 2.2, 2.4.3, 3.2, 3.2, 4.2, 4.2

Zijin Gu, Keith Jamison, Mert R Sabuncu, and Amy Kuceyeski. Human brain responses are modulated when exposed to optimized natural images or synthetically generated images. *Communications Biology*, 6(1):1076, 2023. 3.4.3

Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. *arXiv preprint arXiv:2404.08181*, 2024. 4.3.2

Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198:125–136, 2019. 3.2, 4.2

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022. 2.2

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. *arXiv preprint arXiv:2210.00939*, 2022. 2.4.1

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. 3.2, 4.2

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, 2016. 3.2, 3.3

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score

matching. *Journal of Machine Learning Research*, 6(4), 2005. 2.2

Kajsa M Igelström and Michael SA Graziano. The inferior parietal lobule and temporoparietal junction: a network perspective. *Neuropsychologia*, 105:70–83, 2017. 3.7

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. 2.4.1, 2.6.13

A Ishai, L G Ungerleider, A Martin, J L Schouten, and J V Haxby. Distributed representation of objects in the human ventral visual pathway. *Proc Natl Acad Sci U S A*, 96(16):9379–9384, 1999. 2.1

Nidhi Jain, Aria Wang, Margaret M. Henderson, Ruogu Lin, Jacob S. Prince, Michael J. Tarr, and Leila Wehbe. Selectivity for food in human ventral visual cortex. *Communications Biology 2023 6:1*, 6:1–14, 2 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04546-2. 2.1, 2.2, 2.4.2, 2.4.5, 2.6.12, 3.1, 3.2, 3.3, 3.4.1, 3.4.2, 3.4.3, 4.1, 4.2, 4.4, 4.4.1, 4.4.2, 4.6, 4.4.3

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023. 4.3.2

Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685, 2005. 3.2, 4.2

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11): 4302–4311, 1997. 2.1, 2.2, 3.1, 3.2, 3.4.4, 4.1, 4.2

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 4.3.2

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised,

models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014. 2.2

Meenakshi Khosla and Leila Wehbe. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*, pages 2022–03, 2022. 2.2, 3.4.2, 4.4.2

Meenakshi Khosla, N. Apurva Ratan Murty, and Nancy Kanwisher. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, 32:1–13, 2022a. 2.1, 3.1, 3.2, 4.1, 4.2, 4.4.2

Meenakshi Khosla, N Apurva Ratan Murty, and Nancy Kanwisher. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, 32(19):4159–4171, 2022b. 2.2

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2.6.13

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2.2

Reese Kneeland, Jordyn Ojeda, Ghislain St-Yves, and Thomas Naselaris. Second sight: Using brain-optimized encoding models to align image distributions with human brain activity. *ArXiv*, 2023. 2.2

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1920–1929, 2019. 4.2

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019. 2.2, 3.2, 4.2

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset,

Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 3.4.1

Tom Dupré la Tour, Michael Eickenberg, Anwar O Nunez-Elizalde, and Jack L Gallant. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728, 2022. 2.2

Daniel D Leeds, Darren A Seibert, John A Pyles, and Michael J Tarr. Comparing visual representations across human fMRI and computational vision. *J Vis*, 13(13)(25):1–27, 2013. 4.4.4

Mark D Lescroart and Jack L Gallant. Human scene-selective areas represent 3d configurations of surfaces. *Neuron*, 101(1):178–192, 2019. 4.4.3

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022a. 4.3.2

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a. 3.2

Wei Li, Xue Xu, Xinyan Xiao, Jiachen Liu, Hu Yang, Guohao Li, Zhanpeng Wang, Zhifan Feng, Qiaoqiao She, Yajuan Lyu, et al. Upainting: Unified text-to-image diffusion generation with cross-modal guidance. *arXiv preprint arXiv:2210.16031*, 2022b. 2.3.3, 2.6.13

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023b. 3.2, 3.3.2

Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34:13165–13176, 2021. 4.2

Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. MagicMix: Semantic mixing with

diffusion models, 2022. 2.6.3

Jia Liu, Alison Harris, and Nancy Kanwisher. Perception of face parts and face configurations: An fmri study. *Journal of cognitive neuroscience*, 22:203, 1 2010. ISSN 0898929X. doi: 10.1162/JOCN.2009.21203. 2.4.3

Yulong Liu, Yongqiang Ma, Wei Zhou, Guibo Zhu, and Nanning Zheng. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding from fmri. *arXiv preprint arXiv:2302.12971*, 2023. 3.2, 3.2, 4.2, 4.2

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2.6.13

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 2.4.1

Yizhuo Lu, Changde Du, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. *arXiv preprint arXiv:2303.14139*, 2023. 2.2, 3.2, 4.2

Andrew Luo, Margaret Marie Henderson, Michael J. Tarr, and Leila Wehbe. Brainscuba: Fine-grained natural language captions of visual cortex selectivity. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=mQYHXUUTkU`. 4.2, 4.2, 4.2, 4.4.1

Andrew F Luo, Margaret M Henderson, Leila Wehbe, and Michael J Tarr. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *arXiv preprint arXiv:2306.03089*, 2023. 3.2, 3.2, 3.4.1, 3.4.3, 4.2, 4.2, 4.2, 4.6, 4.4.3

Eleanor Maguire. The retrosplenial contribution to human navigation: a review of lesion and neuroimaging findings. *Scandinavian journal of psychology*, 42(3):225–238, 2001. 3.1, 3.2

Bradford Z. Mahon. Domain-specific connectivity drives the organization of object knowl-

edge in the brain. *Handbook of clinical neurology*, 187:221–244, 2022. doi: 10.1016/B978-0-12-823493-8.00028-6. Place: Netherlands. 4.2

Weijian Mai and Zhijun Zhang. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. *arXiv preprint arXiv:2308.07428*, 2023. 3.2, 4.2

Rafael Malach, JB Reppas, RR Benson, KK Kwong, H Jiang, WA Kennedy, PJ Ledden, TJ Brady, BR Rosen, and RB Tootell. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, 92(18): 8135–8139, 1995. 4.1

Bruce D McCandliss, Laurent Cohen, and Stanislas Dehaene. The visual word form area: expertise for reading in the fusiform gyrus. *Trends in cognitive sciences*, 7(7):293–299, 2003. 2.2

Gregory McCarthy, Aina Puce, John C Gore, and Truett Allison. Face-specific processing in the human fusiform gyrus. *Journal of cognitive neuroscience*, 9(5):605–610, 1997. 3.1, 3.2, 4.1, 4.2

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 3.3.2, 4.4.2

Leilei Mei, Gui Xue, Chuansheng Chen, Feng Xue, Mingxia Zhang, and Qi Dong. The "visual word form area" is involved in successful memory encoding of both words and faces. *Neuroimage*, 52(1):371–378, 2010. 3.4.2, 4.4.2

Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2.6.3

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2.2

Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 3.2, 3.3.2

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 681–699. Springer, 2020. 4.2

Katsuki Nakamura, R Kawashima, Nobuya Sato, A Nakamura, Motoaki Sugiura, T Kato, Kentaro Hatano, K Ito, H Fukuda, T Schormann, et al. Functional delineation of the human occipito-temporal areas related to face and scene processing: a pet study. *Brain*, 123(9):1903–1912, 2000. 4.2

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011. 2.2, 3.2, 4.2

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2.1, 2.3.3, 2.6.13

Kathleen M O'Craven and Nancy Kanwisher. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of cognitive neuroscience*, 12(6):1013–1023, 2000. 4.2

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4.2

Furkan Ozcelik and Rufin VanRullen. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion. *arXiv preprint arXiv:2303.05334*, 2023. 2.2, 2.3.3, 3.2, 3.2, 4.2

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 4.2

Michael Peer, Yorai Ron, Rotem Monsa, and Shahar Arzy. Processing of different spatial scales

in the human brain. *elife*, 8:e47492, 2019. 4.4.3

Ian M L Pennock, Chris Racey, Emily J Allen, Yihan Wu, Thomas Naselaris, Kendrick N Kay, Anna Franklin, and Jenny M Bosten. Color-biased regions in the ventral visual pathway are food selective. *Curr. Biol.*, 33(1):134–146.e4, 2023a. 2.1

Ian ML Pennock, Chris Racey, Emily J Allen, Yihan Wu, Thomas Naselaris, Kendrick N Kay, Anna Franklin, and Jenny M Bosten. Color-biased regions in the ventral visual pathway are food selective. *Current Biology*, 33(1):134–146, 2023b. 2.2, 3.1, 3.2, 4.1, 4.2, 4.4.2

Karin Petrini, Lukasz Piwek, Frances Crabbe, Frank E Pollick, and Simon Garrod. Look at those two!: The precuneus role in unattended third-person perspective of social interactions. *Human Brain Mapping*, 35(10):5190–5203, 2014. 3.4.4

Paweł A Pierzchlewicz, Konstantin F Willeke, Arne F Nix, Pavithra Elumalai, Kelli Restivo, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Katrin Franke, et al. Energy guided diffusion for generating neurally exciting images, 2023. 2.2, 4.2

David Pitcher and Leslie G. Ungerleider. Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, 25:100–110, 2 2021. ISSN 1364-6613. doi: 10.1016/J.TICS.2020.11.006. 3.4.4

David Pitcher, Vincent Walsh, and Bradley Duchaine. The role of the occipital face area in the cortical face perception network. *Experimental brain research*, 209:481–493, 4 2011. ISSN 1432-1106. doi: 10.1007/S00221-011-2579-1. 2.4.3

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, 2019. 2.1, 2.2, 3.2, 4.2

Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature neuroscience*, 24(11):1628–1636, 2021. 3.2, 4.2

Jacob S Prince, Ian Charest, Jan W Kurzawski, John A Pyles, Michael J Tarr, and Kendrick N Kay. Improving the accuracy of single-trial fmri response estimates using glmsingle. *eLife*, 11: e77599, nov 2022. ISSN 2050-084X. doi: 10.7554/eLife.77599. 2.4.1, 3.4.1, 4.4.1

Jacob S Prince, George A Alvarez, and Talia Konkle. A contrastive coding account of category selectivity in the ventral visual stream. *bioRxiv*, pages 2023–08, 2023. 4.2

Aina Puce, Truett Allison, Maryam Asgari, John C Gore, and Gregory McCarthy. Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *Journal of neuroscience*, 16(16):5205–5215, 1996. 3.1, 3.2

Alec Radford, Jong Wook Kim, Chris Hallacy, A. 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*, 2021. 2.4.1, 2.6.2, 3.2, 3.3.1, 4.1, 4.2

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021. 4.4.3

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2.2

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12490–12500, 2024. 4.4.1

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. 2.1, 2.2, 3.2, 3.2, 3.4.3, 4.2, 4.2

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine*

learning, pages 1060–1069. PMLR, 2016. 2.2

Ziqi Ren, Jie Li, Xuetong Xue, Xin Li, Fan Yang, Zhicheng Jiao, and Xinbo Gao. Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, 228:117602, 2021. 3.2, 4.2

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 2.2

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2.2, 2.2, 2.3.1

Gabriel H. Sarch, Michael J. Tarr, Katerina Fragkiadaki, and Leila Wehbe. Brain dissection: fmri-trained networks reveal spatial selectivity in the processing of natural images. *bioRxiv*, 2023. doi: 10.1101/2023.05.29.542635. URL `https://www.biorxiv.org/content/early/2023/11/20/2023.05.29.542635`. 4.4.3

Rebecca Saxe and Nancy Kanwisher. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". In *Social neuroscience*, pages 171–182. Psychology Press, 2013. 3.4.4

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022a. 2.2, 3.4.1

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation

image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022b. 2.4.1, 2.6.13

Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. *Advances in neural information processing systems*, 16, 2003. 4.2

Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024. 3.2, 4.2, 4.2

Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel AJ van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, 2018. 3.2, 4.2

J Sergent, S Ohta, and B MacDonald. Functional neuroanatomy of face and object processing: A positron emission tomography study. *Brain*, 115:15–36, 1992a. 2.1

Justine Sergent, Shinsuke Ohta, and Brennan Macdonald. Functional neuroanatomy of face and object processing: a positron emission tomography study. *Brain*, 115(1):15–36, 1992b. 4.1, 4.2

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019. 3.2, 4.2

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 3.2

Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001. 4.1

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine*

Learning, pages 2256–2265. PMLR, 2015. 2.2

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016. 4.2

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a. 2.2, 2.3.3

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b. 2.2

A. Stigliani, K. S. Weiner, and K. Grill-Spector. Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, 35:12412–12424, 2015. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.4822-14.2015. 2.4.2, 2.4.5, 4.4.1

David G Stork and Hugh R Wilson. Do gabor functions provide appropriate descriptions of visual cortical receptive fields? *JOSA A*, 7(8):1362–1373, 1990. 4.3

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2.3.2, 3.2

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*, pages 2022–11, 2022. 2.2, 3.2, 3.2, 4.2, 4.2

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928, 2022. 3.2, 3.5

Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: computation in neural systems*, 14(3):391, 2003. 4.1

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3.5

Maria Tsantani, Nikolaus Kriegeskorte, Katherine Storrs, Adrian Lloyd Williams, Carolyn McGettigan, and Lúcia Garrido. Ffa and ofa encode distinct types of face identity information. *Journal of Neuroscience*, 41:1952–1969, 3 2021. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1449-20.2020. 2.4.3

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2.2

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3.3.2

Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22(12): 2060–2065, 2019. 2.2, 3.2, 4.2

Aria Y. Wang, Ruogu Lin, Michael J. Tarr, and Leila Wehbe. Joint interpretation of representations in neural network and the brain. In *'How Can Findings About The Brain Improve AI Systems?' Workshop @ ICLR 2021*, 2021. 2.3.2

Aria Yuan Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *BioRxiv*, pages 2022–09, 2022. 2.2, 2.3.2, 3.2, 3.3.1, 4.1, 4.2, 4.2, 4.3, 4.3.1, 4.1, 4.4.1, 4.7, 4.4.4

Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. *arXiv preprint arXiv:2312.01597*, 2023. 4.3.2

Haiguang Wen, Junxing Shi, Wei Chen, and Zhongming Liu. Deep residual network predicts cortical representation and organization of visual features for rapid categorization. *Scientific reports*, 8(1):3752, 2018. 2.2, 3.2, 4.2

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 4.2

Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, 15, 2002. 4.2

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014. 2.2, 3.2, 4.2

Huzheng Yang, James Gee, and Jianbo Shi. Alignedcut: Visual concepts discovery on brain-guided universal feature space. *arXiv preprint arXiv:2406.18344*, 2024a. 4.2

Huzheng Yang, James Gee, and Jianbo Shi. Brain decodes deep nets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23030–23040, 2024b. 4.2

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2.4.1

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 4.1, 4.2, 4.4.1

Yucheng Zhao, Guangting Wang, Chong Luo, Wenjun Zeng, and Zheng-Jun Zha. Self-supervised visual representations learning by contrastive mask prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10160–10169, 2021. 4.2

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European*

Conference on Computer Vision, pages 696–712. Springer, 2022. 4.3.2

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 4.2

Roland S Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas Wallis, and Wieland Brendel. How well do feature visualizations support causal understanding of cnn activations? *Advances in Neural Information Processing Systems*, 34:11730–11744, 2021. 3.2