Contrastive Decoding for Concepts in the Brain

by

Cory Efird

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

This thesis presents a novel data-driven approach for identifying category-selective regions in the human brain that are consistent across multiple participants. By leveraging a massive fMRI dataset and a multi-modal (language and image) neural network (CLIP), we trained a highly accurate contrastive brain decoder to predict neural responses to naturalistic images in the human visual cortex. We then applied a novel adaptation of the DBSCAN clustering algorithm to identify clusters of voxels across multiple brains that decode similar concepts, which we term shared decodable concepts (SDCs). The SDCs are interpreted by identifying the closest embeddings to each cluster centroid and analyzing the associated images and text. In contrast to other methods, ours does not require registration to a template space, allowing us to maintain the unique functional layout of each participants brain. It also uncovers both activating and deactivating stimuli, highlighting the importance of both in understanding brain function. Our approach allowed us to uncover category-selective areas for food, subcategories of bodies and places, color, numerosity, object size, softness, lighting conditions, and more, demonstrating the versatility and potential of our approach for exploring brain functions.

Preface

This thesis was adapted from a paper submitted to the 2024 Conference on Neural Information Processing Systems (NeurIPS), and is a result of a collaborative effort with co-authors Alex Murphy, Joel Zylberberg, and Alona Fyshe. I would like to thank Alex Murphy for his work in preparing figures 3.1 and 3.3, which have significantly contributed to the clarity of this thesis.

Acknowledgements

First, I would like to thank my supervisor, Alona Fyshe, for accepting me as her Master's student and for providing invaluable guidance, expertise, and encouragement throughout my research.

A special thanks to Joel Zylberberg, Alex Murphy, and Richard Gerum for their continuous feedback and contributions to this project.

Next, I am very grateful to all of my lab-mates and peers at the university for their stimulating discussions and shared passion for discovery, which have inspired and motivated me throughout this journey.

Finally, to my friends and family, thank you for your unwavering support and for providing much-needed balance during this phase of my life.

Contents

List of Figures

Chapter 1

Introduction

The visual cortex is one of the most thoroughly studied regions in the human brain. Previous research has largely focused on identifying the types of visual stimuli that drive responses in distinct areas of the brain. It is well-established that early visual system constructs low level representations of texture, depth, color, and shape. This information is then projected to higher visual areas, some of which selectively respond to broad categories of stimuli such as faces, bodies, places, words, and food. These category-selective areas typically correspond to large patches of cortex, and many visual regions remain poorly characterized, leaving gaps in our understanding of their functional organization. Motivated by these gaps, our work seeks to explore and map these lesser-known regions using a novel methodology.

In this work, we introduce a data-driven approach to discover new category-selective areas in the visual cortex and to identify sub-categories within existing ones. Our method leverages a massive fMRI dataset, the Natural Scenes Dataset (NSD), and a multimodal (natural language and image) neural network architecture, CLIP, to train a contrastive decoder that maps brain responses to CLIP embeddings. We then apply a novel adaptation of the DBSCAN clustering algorithm to the decoder model parameters in order to find areas of high similarity across participants. We refer to these clusters as shared decodable concepts (SDCs)—dimensions of CLIP space that can be decoded from multiple participants in the NSD study. To interpret these SDCs, we identify sets of CLIP embeddings closest to each cluster centroid and examine

the corresponding images and text associated with those embeddings. Additionally, we investigate the negated cluster centroids, enabling us to explore stimuli that both activate and deactivate specific cortical regions.

Our method for identifying category-selective brain areas offers unique advantages over prior approaches. Typically, multi-participant data is warped to a template space, and further analysis proceeds with the assumption that functionally similar areas will also overlap anatomically. In contrast, we preserve the native space of each participant's brain, maintaining the unique cortical layout and avoiding potential misalignment issues. Despite the absence of spatial alignment constraints, our method consistently uncovers SDCs localized to similar cortical regions across participants. Moreover, the use of a contrastive loss function mitigates the decoder model's tendency to ignore categories that are under-represented in the stimulus set, and thus might otherwise go undiscovered. This is evidenced by the contrastive-trained decoder's significantly higher accuracy compared to a baseline ridge-regression decoder. Lastly, our examination of stimuli that both activate and deactivate each cluster provides compelling evidence for their potential functions, aligning with recent research on "offsembles" of neurons that are selectively inhibited by visual stimuli (Pérez-Ortega et al., 2024). This suggests that deactivation may be just as important as activation when interpreting the function of brain areas.

Our approach has enabled us to uncover areas selective for a diverse range of stimuli, including faces, food, subcategories of bodies (e.g., legs, hands, motion, sitting, groups), subcategories of places (e.g., indoors, outdoors, nature), color, numerosity, object size, softness/hardness, lighting conditions, and more. These findings not only expand our understanding of the visual cortex but also demonstrate the potential of our method to uncover the diverse functions of brain areas, making it a versatile tool that can be applied to studying various types of stimuli beyond just visual processing.

1.1 Objectives

The contributions of this master's thesis can be summarized as follows:

- We demonstrate that the use of a contrastive loss function significantly improves the accuracy of the decoder model, improving the decodability of under-represented categories in the stimulus set.

- We applied a novel adaptation of the DBSCAN clustering algorithm to the decoder model parameters, enabling the identification of SDC clusters across participants without relying on spatial alignment.

- We investigate the stimuli that both activate and deactivate the SDC clusters, providing deeper insights into the potential functions of brain areas.

1.2 Outline

In chapter 2, we review previous research that is relevant to this master's thesis. First, we highlight the core studies that have revealed the function of areas in the human visual system. Then we describe the fMRI recordings from the natural scenes dataset (NSD) that we use in this study. Next, contrastive learning techniques are discussed including the contrastive language-image pre-training (CLIP) model that is utilized in this work. Following this, we discuss the DBSCAN clustering algorithm which we modify to discover areas that are used to decode similar concepts across participants. Finally we outline previous works that have investigated the similarity between deep learning representations and neural data.

Chapter 3 describes our methodology for training decoder models that predict deep learning representations from CLIP using neural recordings from NSD. The full pipeline including the data split, pre-processing, decoding model, hyperparameter selection, and evaluation are described. Then we present our novel adaptation of the DBSCAN clustering algorithm. This algorithm is applied to the decoder models to find brain areas of high similarity across multiple participants. Next we provide the details of how we select the representative images and text that are used to interpret the discovered brain areas.

Chapter 4 displays the discovered brain areas on flattened cortical surface

maps, along with corresponding representative images and text for interpretation.

In the final chapter 5 we discuss our conclusions and contributions of this master's thesis. Additionally, we consider some potential limitations of this work, broader societal impacts, and directions for future work.

Chapter 2

Related Work

In this section, we present a review of the relevant literature that supports our research question. First, we provide an overview of the primary brain areas in visual cortex and their known functions. Following this, we introduce the Natural Scenes Dataset (NSD), which provides the fMRI recordings utilized in this thesis. We then briefly examine contrastive learning techniques and the CLIP language-vision model, which we used to generate numerical representations of the stimulus images from NSD. Subsequently, we discuss the DBSCAN clustering algorithm, which we utilize in our analysis to identify novel category-selective brain regions. In the final section we highlight other studies similar to ours that use representations generated by deep neural networks to analyze neural recordings in human visual cortex.

2.1 Functional Localization in Visual Cortex

When light enters the eye, visual information travels from the retina to the lateral geniculate nucleus through the optic nerve, and then projects to the primary visual cortex (area V1) in the occipital lobe. Hubel and Wiesel, 1962 were the first to discover that neurons in V1 are tuned to respond to edges that are oriented at particular angles. V1 then projects to many areas including V2, V3, and V4, which build more complex representations of visual properties such as texture, depth, color, and shape.

Many previous studies have discovered higher visual areas that selectively respond to particular types of stimuli. Kanwisher et al., 1997 found an area

in the fusiform gyrus that responds much stronger to faces than any other stimuli. Multiple tests were conducted to confirm that the FFA does not respond to low-level features of human faces or any other human body part. This provided convincing evidence that the FFA will only respond strongly to an image of a face.

Soon after, Downing et al., 2001 discovered a region that they named the extrastriate body area (EBA). The EBA was shown to have a large response to images of human bodies, or to cutouts of individual body parts. Notably, the EBA will respond mildly to animal bodies and to cutout parts of human faces (i.e. eyes, ears, or a mouth), but responds very weakly to whole human faces. Later, the visual word form area (VWFA) was discovered to reside in the left fusiform gyrus (McCandliss et al., 2003). This area responds specifically to visually presented words, but not to auditory word stimuli.

Another key functional area in visual cortex is the parahippocampal place area (PPA Epstein and Kanwisher, 1998. They showed that this area has a strong response to images of indoor and outdoor scenes, a mild response to cutouts of objects, and no response to images of faces. Interestingly, the PPA will respond just as strongly to an empty room as a furnished room, or to a room where the component pieces have been cutout and separated. However if the cutout pieces are rearranged and scrambled, the response is significantly diminished. This suggests is that the geometry of the scene is what drives a response in PPA.

In addition to the PPA, the occipital place area (OPA) and retrosplenial cortex (RSC) have been shown to be involved in scene processing. Dilks et al., 2013 provided evidence that the OPA has a functional purpose in processing scenes. When participants were subject to transcranial magnetic stimulation (TMS) [1] of the OPA, they struggled to discriminate between matching scenes, and also to categorize types of scenes. However, the perception of other types of images such as faces and objects was not disrupted. Furthermore, perception

[1]Transcranial magnetic stimulation (TMS) is a non-invasive method for stimulating neurons in the brain. An electromagnetic coil is placed on the participants scalp which creates a varying magnetic field. This field induces an electric current in a target neural population.

of scenes was not disrupted when TMS was applied to other control brain areas. Maguire, 2001 highlighted 10 cases in the literature where a lesion in RSC resulted in difficulties in navigation. In all cases, the patients could recognize landmarks, but they struggled to place them on a map or navigate in familiar environments. This suggests that navigation is an important function of RSC.

With the onset of the Natural Scenes Dataset (NSD), three studies have uncovered another region in the visual cortex that reliably responds to images of food. Pennock et al., 2022 and Jain et al., 2022 both use a hypothesis-driven approach, where they designed their analysis specifically to test whether food selective cortex exists. Pennock et al., 2022 localized visual cortex that were highly correlated with the color saturation of presented stimulus images. This area of cortex was further investigated to determine the contribution of image saturation, luminance, warmth, presence of circular objects (due to overlap of shape-selective cortex), and presence of food in the presented images to brain responses in different ROIs. The number of food pixels was determined to have the greatest contribution to brain responses. Jain et al., 2022 found food-selective cortex by fitting voxel-wise encoding models that predict brain responses as a function of a small set of binary labels that indicated the location, perspective, and content of each image. The encoder models that had significantly high weights for the food-related labels were used to identify the voxels that had a specific response to food stimuli. Khosla et al., 2022 used a purely data driven approach to locate the food areas. Bayesian non-negative matrix factorization was applied to brain responses to find approximately 20 components for each participant. An inter-subject spatial consistency metric was applied in the MNI template space to identify 5 components that were consistent across participants. These components corresponded to the FFA, PPA, EBA, VWFA, and the unexpected food area.

2.2 Natural Scenes Dataset

The natural scenes dataset (NSD) is a massive fMRI dataset acquired to study the underpinnings of natural human vision (Allen et al., 2022). Eight participants were presented with 30,000 images (10,000 unique images over 3 repetitions) from the Common Objects in Context (COCO) naturalistic image dataset (Lin et al., 2014). Out of the 10,000 images shown to each participant, 9,000 were unique images that were not shown to any other participant in the study, while the remaining 1,000 were shared images that were shown to all participants. The brain images were acquired with a 7-Tesla scanner, which enabled very high 1.8 mm spatial and 1.6 second temporal resolutions, with full brain coverage. The participants were instructed to fixate on the center of the screen while images were presented for a 3 second duration with a 1 second gap between each image. A total of 40 1 hour scanning sessions were administered over the course of a year with each participant to complete the study. The participants were additionally challenged to perform a continuous recognition task during scanning, where they were provided with a button and instructed to press it on the second and third presentations of each image. The three presentations of each image were randomly distributed to any of the 40 scanning sessions, and could be presented many months apart, making this task fairly challenging. Some participants did not complete all fMRI recording sessions and three sessions were held out by the NSD team for the Algonauts challenge. Further details can be found in Allen et al., 2022.

As a first level of analysis, the raw fMRI data was pre-processed using a general linear model (GLM). The GLM is typically implemented as a regularlized linear regression that predicts the fMRI time series from a design matrix that describes the timing of stimulus presentations. The design matrix is constructed as a $N \times T$ array where N is the number of image presentations, and T is the number of brain images acquired in a scanning run. The matrix is initialized to zeros and a value of 1 is inserted on each row to indicate which image is currently presented. To account for delay in the blood oxygen level dependant (BOLD) response, the columns of the design matrix are convolved

with a haemodynamic response function (HRF), that mimics the typical observed response profile of neural populations when recorded with fMRI. The resulting parameters of this model are referred to as the beta-weights and are used in our subsequent analyses. Each beta-weight represents a voxel's percent signal change from baseline when a particular image was presented.

2.3 Contrastive Learning

Contrastive learning is a powerful representation learning technique that has applications in many deep learning domains including computer vision, natural language processing, and reinforcement learning. The core idea is to maximize the similarity between similar datapoints (positive pairs), and minimize the similarity of dissimilar datapoints (negative pairs) in a latent space. One of the earliest examples of contrastive learning is the Siamese network (Bromley et al., 1993) that consisted of a twin network with shared weights and a contrastive loss function. This model was applied to distinguish authentic hand-written signatures from forgeries. Much later, Chopra et al., 2005 formulated one of the first supervised contrastive loss functions for applications in deep learning. Their loss function takes pairs of data points (x_i, x_j) and corresponding categorical labels (y_i, y_j), and minimizes the distance between their embeddings $||f(x_i) - f(x_j)||_2^2$ if they have the same class i.e. $y_i = y_j$. The full loss function is defined as follows:

$$\mathcal{L}_{\text{contrastive}}(x_i, x_j) = \begin{cases} ||f(x_i) - f(x_j)||_2^2 & y_i = y_j \\ \max(0, \epsilon - ||f(x_i) - f(x_j)||_2^2) & y_i \neq y_j \end{cases}$$

where ϵ is a hyperparameter that defines a lower bound on the distance between negative samples.

For unsupervised datasets, augmentation strategies can be applied. The SimCLR technique (Chen et al., 2020) applies visual augmentations (random cropping, color distortions, and gaussian blur) to create two views of the same image as positive pairs, while other random images are used to create negative pairs.

The contrastive language-image pre-training (CLIP) (Radford et al., 2021) model is critical to this masters thesis. CLIP consists of a jointly-trained image and text encoder that share a representation space. These encoders are trained on a dataset of over 400 million image and text pairs with a massive $N = 32,768$ minibatch size. The InfoNCE contrastive loss function (Oord et al., 2018) is used to maximize the similarity between images and their corresponding text captions, while minimizing the similarity of mismatched images and captions. At each training iteration, the N matching image and text pairs are used as the positive samples, and the $N^2 - N$ mismatched image and text pairs are used as negative samples. The main power of CLIP comes from the use of natural language as a training signal. While most computer vision models are limited to a finite set of image classes, CLIP has the potential to learn any image property that can be explained with natural language. This allows CLIP to function as a general purpose vision model that reliably transfers to many computer vision tasks.

2.4 DBSCAN Clustering Algorithm

This work uses a modification of a clustering algorithm known as density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996), to discover category-selective regions in human visual cortex. The DBSCAN algorithm works by finding points that are densely packed together and groups them into clusters, while points that are in sparse regions are marked as outliers. The model is parameterized by a neighborhood size ε and a point threshold `minPts`. Each point is classified as either a core point, a border point, or an outlier. The algorithm can be summarized using the following steps:

1. Points that have at least `minPts` points in their ε-neighborhood are marked as *core* points.

2. A graph $\mathcal{G}$ is constructed where the vertices are core points, with edges between core points that have a distance less than ε.

3. Clusters are formed by finding the connected components of $\mathcal{G}$

4. The remaining *non-core* points are added to clusters as border points if they are within the ε-neighborhood of a core point, otherwise, they are marked as outlier points that do not belonging to any cluster.

The main advantages of DBSCAN is that it does not require the user to pre-specify the number of clusters in the dataset, it can find clusters of any shape, and it can robustly handle datasets with many outlier points. However, the hyperparameters ε and `minPts` require careful tuning to extract meaningful clusters from the data, and the algorithm can struggle to deal with datasets where the clusters have varying densities. A common issue if for all of the points to merge into a single cluster if ε is too high or `minPts` is too low. This can be resolved by grid searching different combinations of ε and `minPts` until there is a reasonable amount of clusters and points per cluster.

2.5 Decoding Deep Representations from Brain Activity

A widely used technique in neuroscience involves training models to establish a mapping between neural recordings, such as fMRI data, and numerical representations of the corresponding stimuli. When these models map brain responses to stimulus representations, they are referred to as decoders. Conversely, when they map stimulus representations to brain responses, they are known as encoders. In recent years, it has become common to use deep learning models to generate the stimulus representations that are then used to train brain decoding and encoding models. Khaligh-Razavi and Kriegeskorte, 2014 laid the foundations for this approach by investigating the similarity between convolutional neural network (CNN) representations and neural recordings from inferior temporal (IT) cortex in monkeys and humans. In a comparison of 37 models, the representations in a deep CNN trained to classify images in the ImageNet (Deng et al., 2009) dataset had the highest similarity to IT. Later, Horikawa and Kamitani, 2016 more thoroughly investigated the similar-

ities between early and deep representations in a CNN, and brain recordings from early and later stages of visual cortex. They found that early features in the CNN had the highest correlations to early visual areas such as V1 and V2, but had low correlation to areas such as PPA and FFA. In contrast, the deep representations in the CNN had very high correlations to PPA and FFA, with slightly lower correlations to V1 and V2. This provided evidence that the hierarchy of features in deep models somewhat mirrors the hierarchy of representations in the human visual cortex.

In parallel with the preparation of this maters thesis, numerous studies have trained encoding and decoding models using neural recordings from NSD and stimulus representations generated with CLIP. Wang et al., 2023 compared the voxel-wise encoding performance using representations from CLIP and a CNN trained on ImageNet. The encoders trained with CLIP representations were able to explain far more variance in almost every region in the visual cortex, suggesting that CLIP generates representations that are significantly more brain-like compared to previous models. Further evidence that CLIP representations have strong brain-like representations comes from studies that aim to fully reconstruct the stimulus image from held-out brain responses (Q. Liu et al., 2024; Y. Liu et al., 2023; Ozcelik and VanRullen, 2023; Scotti et al., 2023). These methods typically train a decoder that predicts CLIP representations from brain responses, and then the brain-decoded CLIP vectors are input to diffusion based image-generation models. The resulting brain-reconstructed images have many striking similarities in structure, composition, and semantic detail to the original stimulus images. This motivates the hypothesis that the brain responses recorded in NSD contain semantic information that is not limited to the presence of faces, places, words, bodies, and food.

2.6 Chapter Summary

In summary, this review has outlined the literature relevant to our research. We have detailed the key brain areas in the visual cortex and their functions, introduced the Natural Scenes Dataset (NSD) that provides the fMRI data for

our study, and briefly discussed contrastive learning techniques along with the CLIP language-vision model, that we use to create numerical representations of stimulus images. Additionally, we have described the DBSCAN clustering algorithm that is modified in our analysis to discover novel category-selective brain regions. Finally, we highlighted similar studies that utilize deep neural network representations for analyzing neural recordings.

In the next chapter, we will focus on the process of decoding CLIP representations from brain responses within the NSD.

Chapter 3

Methods

To identify shared decodable concepts (SDCs) in the brain, we derived a mapping from anatomical brain space to a representational space in which we can explore semantic sensitivity to visual input. In this chapter we describe the components of this mapping: a dataset of fMRI recordings (NSD), a multimodal image-text embedding model (CLIP), and our method to map from per-image brain responses to their associated multimodal embeddings. We consider two models and verify that our proposed contrastive decoder outperforms a baseline ridge regression model. In the final section we introduce our clustering method for identifying concepts that are decodable from the brains of multiple participants in NSD.

3.1 fMRI Data

As described in section 2.2, this work utilizes neural recordings from the Natural Scenes Dataset (NSD) that were pre-processed with a general linear model (GLM). We denote the brain responses as $\boldsymbol{X}^{(k)}$ for participants $k \in \{1 \dots 8\}$ (for brevity, we sometimes drop the superscript (k) in the following sections). Each value in these matrices represents the signal change of a voxel in response to the presentation of a particular image in the NSD experiment. The dimensionality of these matrices varies as some participants did not fully complete all 40 fMRI recording sessions.

3.2 Representational Space for Visual Stimuli

To generate representations for each stimulus image, we use CLIP (Radford et al., 2021), a model trained on over 400 million text-image pairs with a contrastive language-image pretraining objective. CLIP consists of a text-encoder and image-encoder that jointly learns a shared low-dimensional space trained to maximize the cosine similarity of corresponding text and image embeddings. We use the 32-bit Transformer model (ViT-B/32) implementation of CLIP to create a 512-dimensional representation for each stimulus image used in the NSD experiment. We train a decoder model to map from fMRI responses during image viewing to the associated CLIP vector for that same image. Notably, because CLIP is a joint image-language model, these CLIP vectors also correspond to text captions in the pretraining stage, which can be used to describe the images presented to the participants.

3.3 Data Preparation

Data Split We split the per-image brain responses $\boldsymbol{X}$ and CLIP embeddings $\boldsymbol{Y}$ into training $(\boldsymbol{X}_{\text{Train}}, \boldsymbol{Y}_{\text{Train}})$, validation $(\boldsymbol{X}_{\text{Val}}, \boldsymbol{Y}_{\text{Val}})$, and test $(\boldsymbol{X}_{\text{Test}}, \boldsymbol{Y}_{\text{Test}})$ folds for each of the 8 NSD participants. For each participant, the validation and test folds were chosen to have exactly 1,000 images with three presentations. Some participants in the NSD did not complete all scanning sessions and only viewed certain images once or twice. The images that were not viewed 3 times are assigned to the training set, which varies in size across participants. Of the shared 1,000 images that all participants saw, 413 images were shown three times to every participant across the sessions released by NSD. These 413 images appear in each participant's testing fold.

Voxel Selection The NSD fMRI data comes with voxelwise noise ceiling estimates that can be used to identify reliable voxels. However, the noise ceiling is calculated using the *full dataset*. Therefore, these noise ceiling estimates should not be used to extract a subset of voxels for decoding analyses because they are calculated using images in the test set, which is a form of double-

dipping (Kriegeskorte et al., 2009). We therefore re-calculated the per-voxel noise ceiling estimates specifically on our designated training data only. We selected voxels with noise ceiling estimates above 8% variance explainable to use as inputs to the decoding model. The exact number of voxels input to the model was $[17883, 18358, 13476, 11899, 17693, 18692, 9608, 6772]$ for subjects 1 through 8 respectively. In our visualizations, regions of the flattened brain surface in black represent voxels that have passed this voxel selection threshold.

Voxel Normalization Brain responses were per-voxel normalized to have zero mean and unit standard deviation within each scanning session. This standardizes the data across scanning sessions, ensuring that variation due to external factors does not influence our analysis. With this normalization, a value of zero corresponds to the mean response across all stimulus images, while positive and negative values represent above average and below average activation, respectively. This approach allows our decoder model to utilize both the positive and negative aspects of stimulus driven brain responses.

3.4 Decoding Methodology

Decoding Model The decoding model $\boldsymbol{g}(\boldsymbol{X};\boldsymbol{\theta}) = \hat{\boldsymbol{Y}}$ is a linear model trained to map brain responses $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n], \boldsymbol{x}_i \in \mathbb{R}^v$ to brain-decoded CLIP embeddings $\hat{\boldsymbol{Y}} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n], \boldsymbol{y}_i \in \mathbb{R}^{512}$. Here n is the number of training instances, and v is the participant-specific number of voxels that pass the noise ceiling threshold. We optimize the brain decoder using the InfoNCE definition of contrastive loss (Oord et al., 2018), which is defined below in Equations 3.1 and 3.2.

$$\text{Contrast}(\boldsymbol{A}, \boldsymbol{B}) = -\frac{1}{M}\sum_{i=1}^{M} log\left(\frac{exp(\boldsymbol{a}_i \cdot \boldsymbol{b}_i/\tau)}{\sum_{j=1}^{M} exp(\boldsymbol{a}_i \cdot \boldsymbol{b}_j/\tau)}\right) \tag{3.1}$$

$$\mathcal{L}_{\text{InfoNCE}}(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{2}[\text{Contrast}(\boldsymbol{A}, \boldsymbol{B}) + \text{Contrast}(\boldsymbol{B}, \boldsymbol{A})] \tag{3.2}$$

In this definition $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$ and $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n]$ are embeddings for two modalities representing the same data points, the $\cdot$ operator represents co-

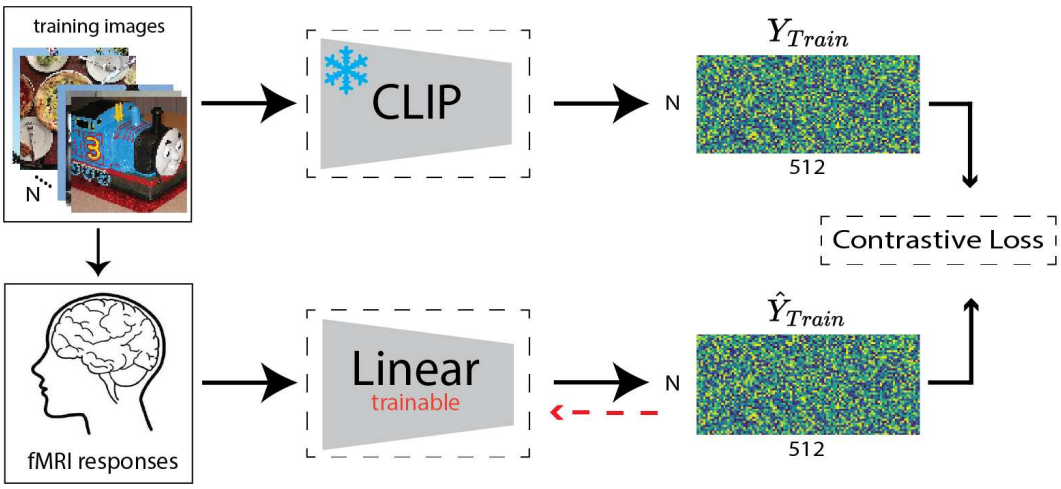

Figure 3.1: Decoding CLIP representations from the brain. We create CLIP representations by passing NSD training set stimuli to CLIP (frozen). The NSD team showed these same images to human participants during fMRI scans. We train a linear decoder with contrastive loss (Eq. 3.2) to predict CLIP embeddings from the fMRI responses to the corresponding images. Black arrows represent the flow of data through the procedure and dashed red lines represent gradient updates used to train the model.

sine similarity, and τ is a temperature hyper-parameter. The loss is minimized when the distance between matching embeddings is small, and the distance between mismatched embeddings is large. In the original CLIP setting, $\boldsymbol{A}$ and $\boldsymbol{B}$ represent embeddings for images and corresponding text captions. In our implementation, we apply the contrastive loss to image embeddings computed by a pretrained frozen CLIP model and CLIP embeddings predicted from fMRI, i.e. we optimize $\min_{\boldsymbol{\theta}} \mathcal{L}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}_{\text{CLIP}}) = \mathcal{L}(g(\boldsymbol{X}; \boldsymbol{\theta}), \boldsymbol{Y})$. An illustration of the decoding procedure is given in Figure 3.1.

The decoding model is trained for 5000 iterations (29 to 45 epochs depending on the participant's training set size) with the Adam optimizer, a batch size of 128, and a fixed learning rate of $1e^{-4}$. Data augmentation is applied to help slow overfitting by adding random noise to training samples $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i + \boldsymbol{z}$ sampled from a normal distribution $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2)$ where the noise standard deviation σ is a hyper-parameter. We set $\tau = 0.03$ and $\sigma = 0.1$ in our implementation. Hyper-parameters were selected based on performance on the validation set. We compare our contrastive decoder to a baseline ridge

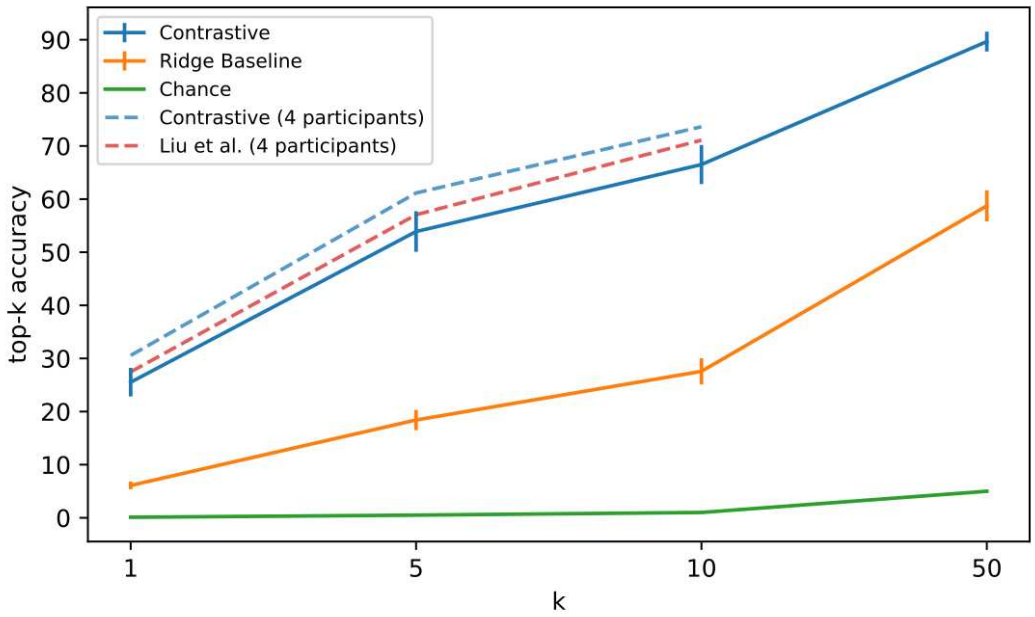

Figure 3.2: Top-k accuracy for CLIP decoding using the proposed contrastive decoder and a ridge regression baseline. The contrastive decoder implementation outperforms the ridge regression baseline across various values of k. Chance performance is given in green. Accuracy calculated on held-out data. Bars indicate the standard error (SE) across the 8 NSD participants. Additionally, we include a comparison to Y. Liu et al., 2023 who report the mean top-k accuracy across 4 subjects (1, 2, 5, 7) using brain responses to 982 test images (dashed red line). We average the results of our own contrastive decoder across these participants for a fair comparison (dashed blue line). Despite a slightly larger test set, our contrastive decoder has a higher accuracy for all reported values of k.

regression model trained on the same data. We used grid search to select the best ridge regularization parameter $\lambda \in \{0.1, 1, 10, 100, 1000, 10000, 100000\}$ using the validation data. The optimal λ was 10000 for participants 1, 2, 5, 6, and 1000 for participants 3, 4, 7, 8. The decoder models were trained on an NVIDIA GeForce RTX 2060. Training time was approximately 2 minutes for each model.

Evaluation We evaluate our models using top-k accuracy, which is computed by sorting in ascending order all true representations $\{y_1 \dots y_n\}$ by their cosine distance to a predicted representation $\hat{y}_i$. Top-k Accuracy is the percentage of instances for which the true representation y_i is amongst the top-k

items in the sorted list. Chance top-k accuracy is $\frac{100 \cdot k}{n}\%$ where n is the number of held-out data points used for evaluation. Figure 3.2 shows the results of this evaluation. The contrastive decoder outperforms ridge regression across all values of k. Recall that our end goal is to identify shared decodable concepts (SDC) in the brain. Our methodology for this task relies on the *predicted* CLIP vectors, and so our subsequent analyses require an accurate decoding model.

3.5 Finding Shared Brain-Decodable Concepts and their Representative Images

In this section, we describe our methods for finding and interpreting areas of the brain that respond to semantically similar images, and that are consistent across participants. We approach this problem by analyzing the weight matrices $\boldsymbol{W}^{(k)}$ from the optimized linear decoders described in section 3.4 $g^{(k)}(\boldsymbol{x}^{(k)}) = \boldsymbol{W}^{(k)}\boldsymbol{x}^{(k)} = \boldsymbol{y}$ for all subjects $k \in \{1 \dots 8\}$. First, we observe that the decoder's linear transformation $\boldsymbol{W}^{(k)}\boldsymbol{x}^{(k)} = \sum_{i=1}^{v} \boldsymbol{w}_i^{(k)} \cdot {x_i}^{(k)}$ is simply a summation of parameter vectors $\boldsymbol{w}_i^{(k)} \in \mathbb{R}^{512}$ that are scaled by brain response values $x_i^{(k)} \in \mathbb{R}$. This means that when $x_i^{(k)}$ is above or below baseline activation, the CLIP dimension represented by $\boldsymbol{w}_i^{(k)}$ is increased or decreased in the brain-decoded embedding. This motivates the key to our analysis: we view the parameter vector $\boldsymbol{w}_i^{(k)}$ as a CLIP vector that represents a brain-decodable concept for voxel i. This allows us to use cosine distance $d(\boldsymbol{w}_i^{(k)}, \boldsymbol{w}_j^{(r)})$ between the weight vectors for voxels i, j from participants k, r as a measure of similarity between the decodable concepts of brain voxels across participants. Our objective is to apply a clustering algorithm using this metric in order to find areas in the brains of multiple participants that have a high conceptual similarity. We refer to these underlying concepts as shared decodable concepts (SDCs). We interpret the SDCs by retrieving a set of representative images associated with the brain-decoded CLIP vectors that are closest to the centroid of each SDC cluster. A schematic of the algorithm we apply is given in Figure 3.3.

Cross Participant Clustering To discover SDCs, we apply a novel clustering method to the per-voxel model parameter vectors $\boldsymbol{w}_i^{(k)}$ across all participants. We base our clustering method on the density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996) algorithm, which is described in detail in the related works section 2.4. Compared to the original algorithm, we redefine the core point criteria in step (1) and rename the threshold variable `minPts` as `minNeighbors`. Our second modification is to add a final expansion step (5) where each cluster can grow slightly larger within each participant. Steps 2-4 are unchanged. Figure 3.4 outlines the application of our modified DBSCAN algorithm to modeling fMRI data. Our modified DBSCAN algorithm is summarized as follows:

1. A point $\boldsymbol{w}_i^{(k)}$ is marked as a core point if there are points from at least `minNeighbors` other participants within its ε neighborhood.

2. A graph $\mathcal{G}$ is constructed where the vertices are core points, with edges between core points that have a distance less than ε.

3. Clusters are formed by finding the connected components of $\mathcal{G}$

4. The remaining *non-core* points are added to clusters as border points if they are within the ε-neighborhood of a core point, otherwise, they are marked as outlier points that do not belonging to any cluster.

5. All points inside the $\varepsilon_{\text{expansion}}$ neighborhood of a point in a cluster become members of that cluster with the constraint that they must be from the *same* participant.

With the modification to step 1, a core point now identifies voxels that represent a brain-decodable concept that is shared with at least `minNeighbors` participants. The addition of step 5 was motivated by our early experiments, where we noticed that there were sometimes only one or two voxels belonging to certain participants for each cluster. This was helped by applying a within-participant expansion of clusters. We introduced a new hyperparameter $\varepsilon_{\text{expansion}}$ that controls the degree of cluster expansion. Unlike steps 1-4,

it is possible for a point to be assigned to multiple clusters in step (5). This allows the cluster boundaries to grow slightly larger for each participant, and we found that this did not significantly change the cluster location or semantic interpretations.

We present results using a fixed value for `minNeighbors` = 3. This means that each cluster must include at least 4 out of 8 participants in the study. This allows for the discovery of SDCs that may show up less reliably in some participants. Furthermore, we use a range of values for $\varepsilon \in \{0.5, 0.55, 0.6, 0.65\}$ to explore clusters at different density scales. We set the expansion neighborhood size $\varepsilon_{\text{expansion}} = \min(\varepsilon + 0.05, 0.65)$ so that the cluster boundaries can grow slightly larger than the baseline neighborhood size ε, but not exceeding 0.65 as we noticed neighborhoods become over-connected as ε approaches 0.7. The number of cross-participant clusters found was 9, 14, 12, 11 for ε values of 0.5, 0.55, 0.6, 0.65 respectively. This variant of the DBSCAN clustering algorithm uses minimal CPU resources and executes within a few minutes.

To reduce the impact of random initialization and combat the noise inherent in fMRI, we re-trained the decoder model 50 times for every participant and compute the average of the resulting parameter vectors. In other words, each averaged parameter vector $\boldsymbol{w}_i^{(k)}$ used in our analysis represents the average of 50 concepts that a voxel could represent depending on the random initialization of the decoder. We found that this significantly improved the quality and consistency of the clusters revealed by our modified DBSCAN algorithm.

Selecting Representative Images To interpret the semantic meaning of an SDC cluster, we first computed the cluster centroid by taking the mean of all parameter vectors within the cluster $\bar{\boldsymbol{w}} = \frac{1}{|I|} \sum_{(i,k)\in I} \boldsymbol{w}_i^{(k)}$ where I is a set of tuples identifying the voxel and subject indices for an SDC cluster. When the cluster voxels are above or below baseline activation, the CLIP dimension represented by $\bar{\boldsymbol{w}}$ is increased or decreased in the brain-decoded embedding, respectively. We note that DBSCAN has the ability to create non-globular clusters for which the centroid might not be a good representative. In our

experience the centroid is within ε of a core point within the cluster, implying that it is likely fairly representative of the cluster.

In order to find the representative images for an SDC cluster, we can consider the nearby image-embeddings to the cluster centroid $\bar{\boldsymbol{w}}$. Instead of using the original CLIP embeddings for each image, we instead used the subject-specific *brain-decoded* embeddings $\hat{\boldsymbol{Y}}^{(k)}_{\text{Test}}$. This allows us to focus specifically on the information that can be decoded from fMRI, which could be a subset of the information represented by CLIP space. For images that were viewed more than once we average the embeddings within and across participants. These averaged brain-decoded embeddings from held-out test data are denoted by $\hat{\boldsymbol{Y}}^{\text{AVG}}_{\text{Test}}$. To select nearby images we define $D(\boldsymbol{w}, \boldsymbol{Y}) = \{d(\boldsymbol{w}, \boldsymbol{y}) | \boldsymbol{y} \in \boldsymbol{Y}\}$ which is the set of distances between a CLIP vector $\boldsymbol{w}$ to a set of CLIP embeddings $\boldsymbol{Y}$. We take the cosine distances between the SDC cluster centroid $D(\bar{\boldsymbol{w}}, \hat{\boldsymbol{Y}}^{\text{AVG}}_{\text{Test}})$ and the *negated* centroid $D(-\bar{\boldsymbol{w}}, \hat{\boldsymbol{Y}}^{\text{AVG}}_{\text{Test}})$ and retrieve the images corresponding to the smallest values in these sets as the positive and negative representative images respectively.

Selecting Representative Words Similarly, we can find sets of representative words for the SDC clusters. To accomplish this we first embed all 5 captions for each of the 73000 images in the NSD stimulus set to obtain $\boldsymbol{Y}_{\text{text}} \in \mathbb{R}^{365000 \times 512}$. Unlike the representative images, these embeddings are not brain-decoded and are generated solely by the CLIP text encoder. To select the best representative captions we compute $D(\bar{\boldsymbol{w}}, \boldsymbol{Y}_{\text{text}})$ and $D(-\bar{\boldsymbol{w}}, \boldsymbol{Y}_{\text{text}})$ and select the captions corresponding to the 50 smallest distances in each set as the positive and negative captions respectively. We then utilize the word-cloud python package to create visualizations of the most frequently occuring words in the captions, which we present in the following chapter.

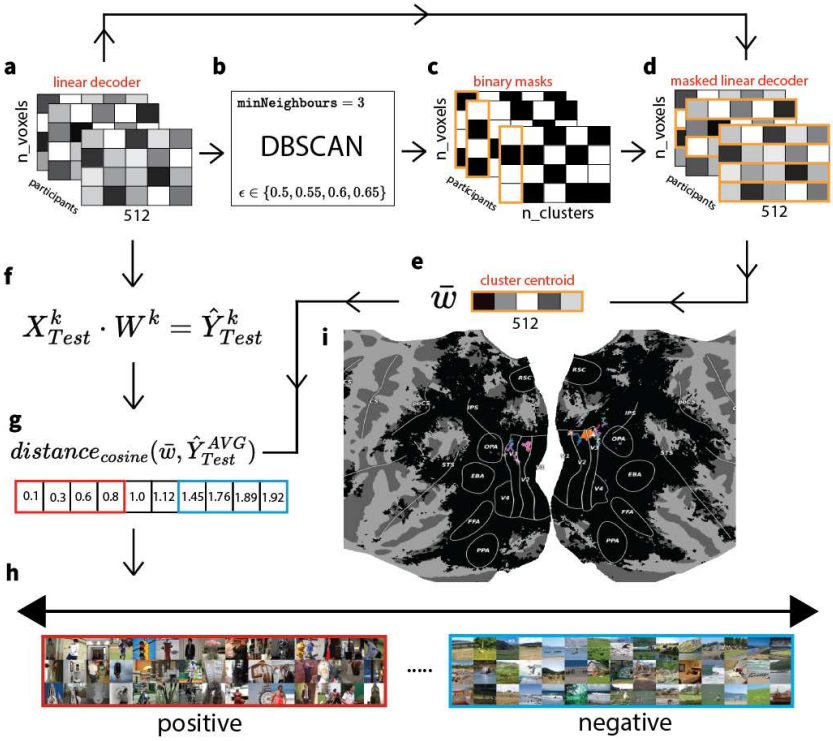

Figure 3.3: **Deriving SDC clusters.** **(a):** The participant-specific linear decoder matrices. **(b):** Our modified DBSCAN clustering procedure is applied to the linear decoders (See Figure 3.4 for details). **(c):** Our DBSCAN procedure derives binary masks over the voxels in the linear decoders for a specified number of clusters (of which one is highlighted in orange) **(d):** The rows corresponding to the selected voxels in the binary masks are extracted from the linear decoder matrices. **(e):** The 512-dimensional representations from the previous step are averaged over voxels and participants to derive a cluster centroid for each cluster derived from DBSCAN. We visualize the cluster centroid for a particular DBSCAN cluster. **(f):** The linear decoders are used to predict brain-decoded embeddings $\hat{Y}^k_{Test}$ for the held-out test data. The predicted embeddings for repeated images are averaged within and across participants to give $\hat{Y}^{AVG}_{Test}$. **(g):** Cosine distance is calculated between the cluster centroid and the brain-decoded embeddings. **(h):** The images most associated with the cluster centroids (positive images) and most negatively associated with the cluster centroids (negative images) are identified. Positive / negative images for the SDC cluster pictured here appears to correspond to global vertical/horizontal orientation in the associated images. **(i):** Color-coded participant-specific voxel clusters are displayed on a flatmap of the brain's cortical surface in common *fsaverage* space (overlapping areas are displayed in white). Regions of interest labels are highlighted on the flatmap image in white outlines. For the specified cluster (e), whose positive / negative images are associated with orientation, the flatmap indicates bilateral shared voxel clusters in early visual cortex.

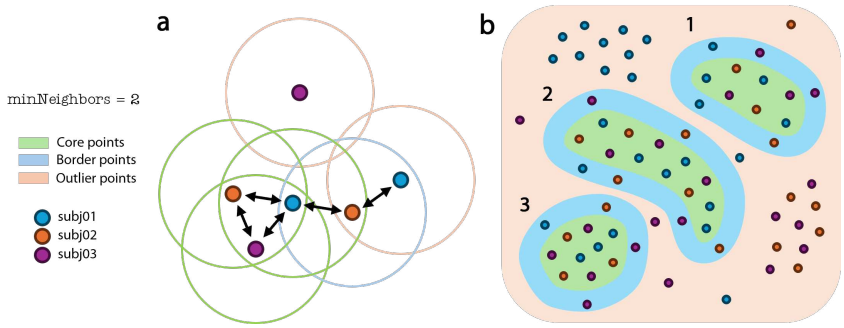

Figure 3.4: **Illustration of our DBSCAN variant applied to multi-participant fMRI data.** There are 3 participants in this example and `minNieghbors` $= 2$. **(a):** A zoomed-in view of a cluster with three core points. The outer ring around each point shows its ε-neighborhood and whether it is a core, border, or outlier point. Black arrows emphasize points that are neighbors. The points with neighboring points from at least 2 other distinct participants are marked as core points. Non-core points that neighbor core points are added to the cluster as border points. The remaining points are marked as outliers. **(b):** A zoomed-out sketch of a set of points that form 3 clusters. Since `minNieghbors` $= 2$, a high-density region will form clusters if and only if it contains points from at least 3 participants.

3.6 Chapter Summary

In this chapter we described the NSD dataset and the CLIP model that was used to generate numerical representations of the stimulus images. Using NSD and CLIP, we trained a contrastive decoder that maps brain responses to CLIP embeddings, and demonstrated that a contrastive approach outperforms a baseline ridge regression model. Finally, we needed a methodology to interpret our contrastive model. A novel DBSCAN variant supported our interpretation efforts, allowing us to discover SDC clusters. We explored these clusters using representative images and words. In the next section, we will showcase a selection of the SDC clusters, the corresponding images and words, and offer our interpretations.

Chapter 4

Results

The clustering method described in chapter 3 not only identifies previously discovered functional areas but additionally identifies new ones, both of which we explore in this section. We emphasize that there is no constraint that voxels within a shared visual semantic cluster be spatially adjacent in the brain, yet during visualization we consistently find contiguous patches both within and across participants. In order to identify the locations of shared voxel clusters in our visualizations, we overlaid region of interest boundaries onto the flatmap visualizations in Freesurfer's *fsaverage* space. For functional ROIs, we retraced the ROIs given in the NSD dataset during the functional localization experiments (fLoc, Stigliani et al., 2015).

4.1 Faces

One of the first reported functionally localized areas for higher-order vision was the fusiform face area (FFA) (Kanwisher et al., 1997). Our method identifies a face-related concept (Figure 4.1 $\varepsilon = 0.55$, cluster 7) that is localized to FFA and includes voxels from all 8 participants. This cluset also has some voxels in the extrastriate body area (EBA), likely because many face images include part or all of a person's body. We note that the positive representative images are not exclusively human and include a range of animal faces. The images often depict people eating and handling food, or holding other objects such as toothbrushes or cell phones.

Interestingly, the negative representative images often display bodies, but

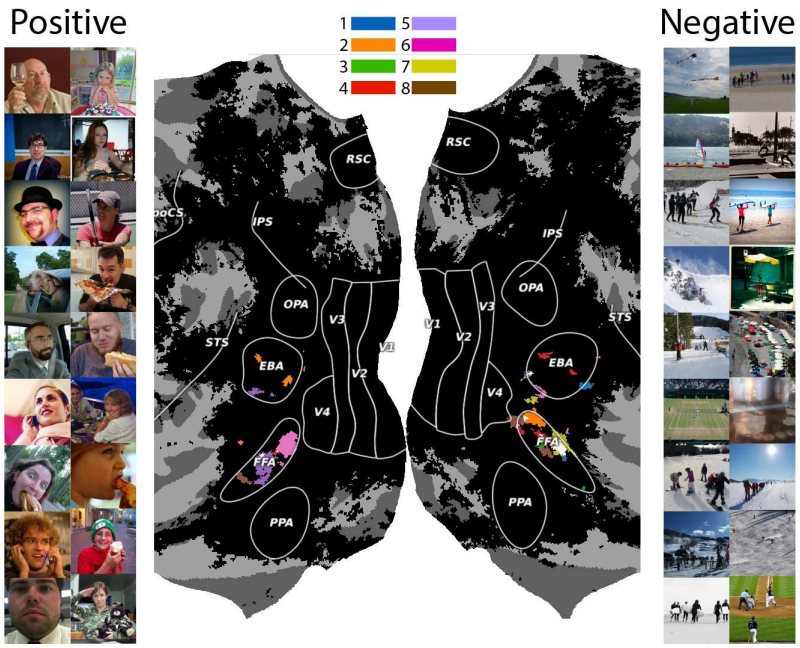

Figure 4.1: Cluster 7 ($\varepsilon = 0.55$). Positive images are strongly associated with faces, while the negative images represent depictions of people whose faces are not visible. Voxel clusters are primarily found in bilateral FFA and EBA.

there is a striking *lack* of clearly visible faces. There are several examples where people are visible, but their faces are obscured or they are facing away from the camera. This suggests that there may be a dip in FFA activity for images where a person's face might be expected (and thus FFA is primed to activate) but not clearly visible. Thus the brain's representation for the opposite of a face is a scene with the conspicuous absence of faces.

4.2 Food and Color

Previous work has specifically noted the correlation of very colorful images with food-related images, and attempted to define food areas in the absence of color (Jain et al., 2023). Our method also identifies a large possibly food-related cluster that spans FFA, PPA, and V4 (cluster 0, $\varepsilon = 0.55$, Figure 4.2). At first inspection most of the positive representative images are food-related. However, we also observed many vibrant and colorful positive images that contained no food. Strikingly, the negative representative images are *entirely gray-scale*, suggesting that this cluster may be related to color. Thus, we

speculate that this cluster is not specifically related to the identification of food, but might rather correspond to any colorful image.

We notice two other food related clusters when $\varepsilon = 0.65$. The positive images for cluster 3 (Figure 4.4) are mostly images of food. However, we observed themes of softness, round objects, and animals shared across the food and non-food images. Meanwhile, the negative images depicted many rigid and boxy objects such as trains, busses, and city buildings. This supported the idea that this cluster responds to a soft versus hard concept and is again not entirely food-specific.

Cluster 5 (Figure 4.3) is shared across seven participants and is localized to frontal brain regions. Voxels are located in the orbital sulci, as well as the boundary between the triangular part of the inferior frontal gyrus and the inferior frontal sulcus. These frontal regions are consistent with a network of voxels reported by Pennock et al., 2022 to be activated by food images. We noted that the positive images to contain images of food on plates, as well as a few non-food images (bear, skiing, playing baseball). We were not able to discern a strong core concept in the negatively associated images, aside from a distinct lack of food.

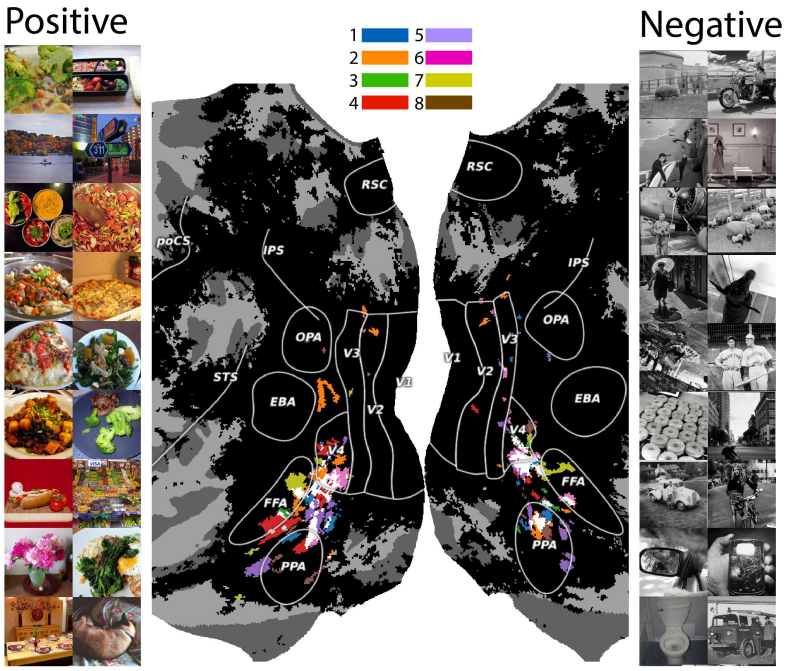

Figure 4.2: Cluster 0 ($\varepsilon = 0.55$). Positive images are associated with food and color. Negative images are entirely grayscale. Voxel clusters span bilateral FFA, V4, and PPA.

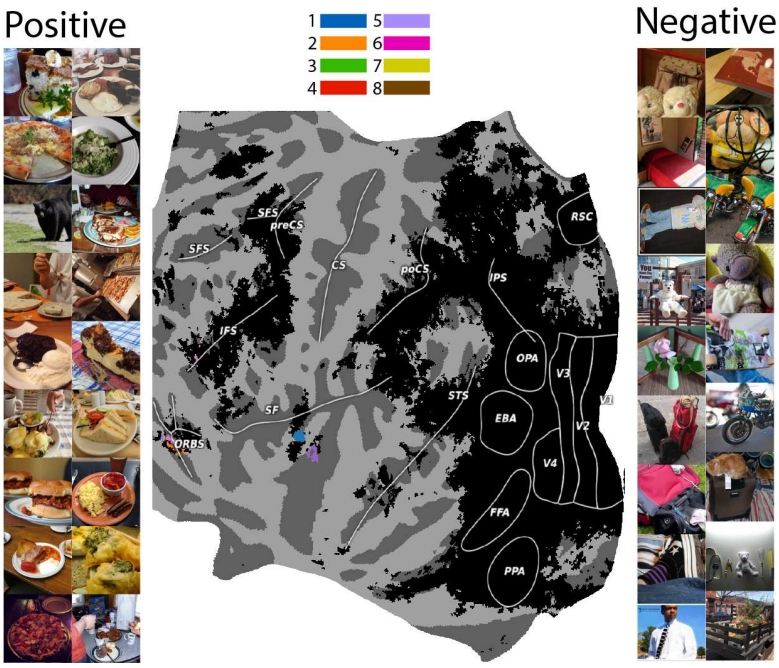

Figure 4.3: Cluster 5 ($\varepsilon = 0.65$). Food-related images with shared voxel clusters in anterior regions of the brain.

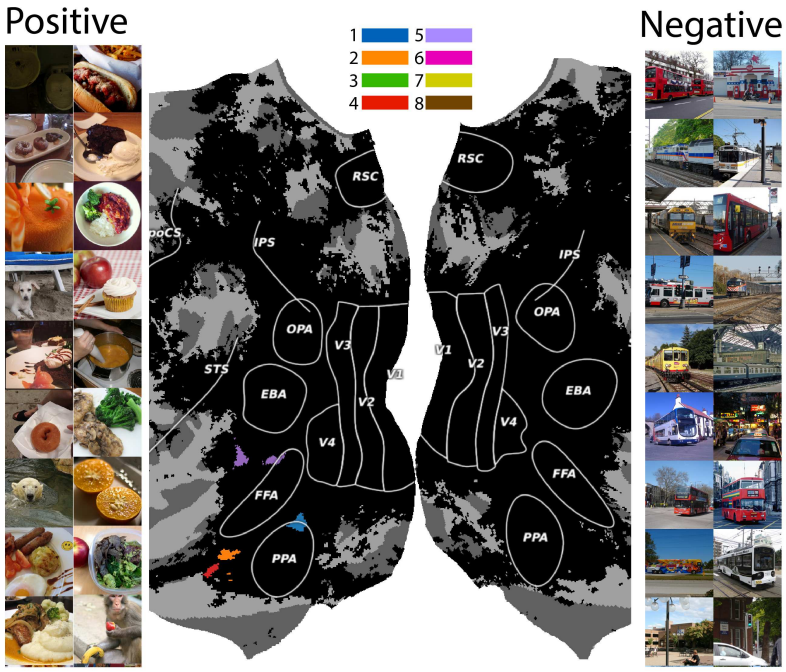

Figure 4.4: Cluster 3 ($\varepsilon = 0.65$). Positive images are of soft, often circular food. Negative images depict hard objects (buses or buildings). Voxel clusters are in left hemisphere around PPA.

4.3 Bodies

Our method identifies five notable body-related areas at $\varepsilon = 0.55$ in and around EBA (Downing et al., 2001). The positive representative images for cluster 2 (Figure 4.5) show people and animals outside with an emphasis on legs and active movement. The negative images typically depict people indoors sitting with their legs obscured by tables. Cluster 11 (Figure 4.6) has a similar emphasis on hands instead of legs. People are displayed in a variety of contexts with their hands clearly visible, while the negative images are exclusively non-primate animals without hands. This cluster also shows indoor/outdoor contrast in the representative image groups and so there is some PPA activation. Cluster 3 (Figure 4.7) shows groups of three or more people in the positive images, with a strong focus on an individual person, animal, or object in the negative images. Cluster 6 (Figure 4.8) appears to be related to full-body leaping motions, with the negative images showing people sitting or standing still. In contrast, cluster 5 (Figure 4.9) shows a mixture of images

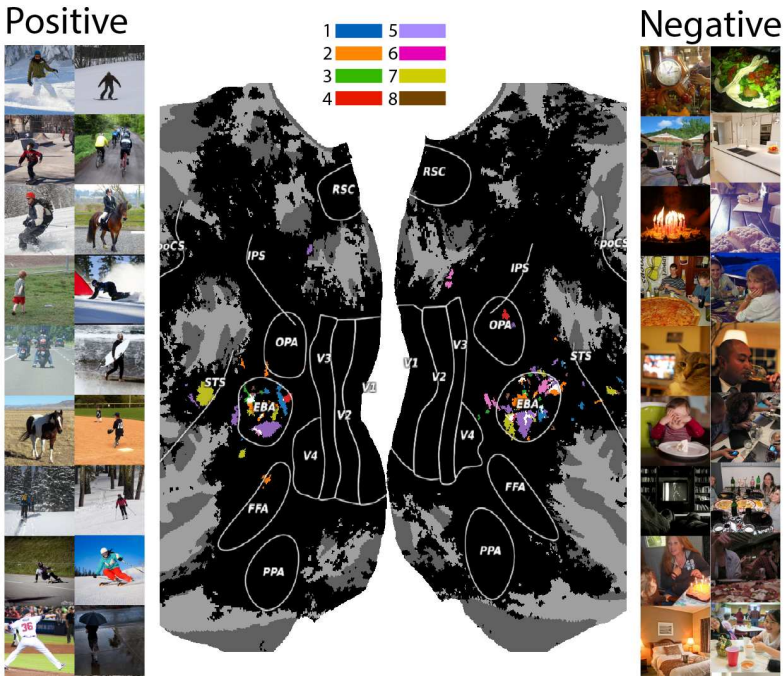

Figure 4.5: Cluster 2 ($\varepsilon = 0.55$). Positive images are strongly associated with presence of legs, while negative images are typically people at tables whose legs are obscured. Voxel clusters are primarily in bilateral EBA.

of people who are crouched or sitting, along with images of pet cats and dogs laying down.

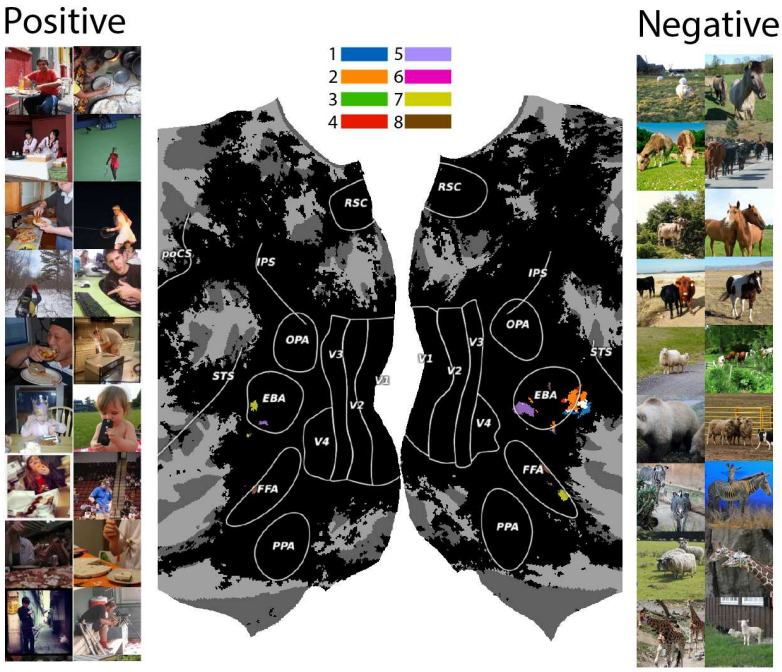

Figure 4.6: Cluster 11 ($\varepsilon = 0.55$). Positive images are strongly associated with hands and hand motion, while negative images are associated with animals (no hands). Voxel clusters are primarily in bilateral EBA and FFA.

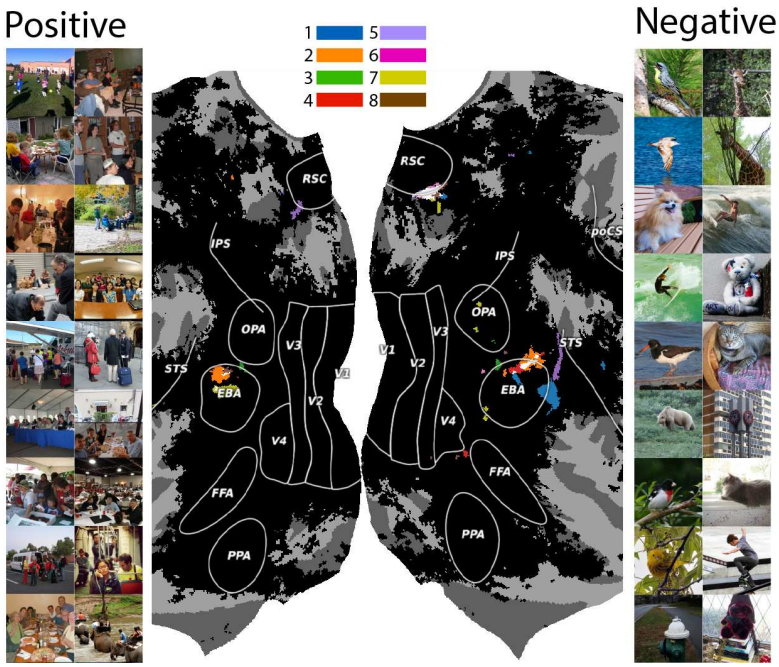

Figure 4.7: Cluster 3 ($\varepsilon = 0.55$). Positive images display crowds, while negative images are of individuals. Voxel clusters are in bilateral EBA.

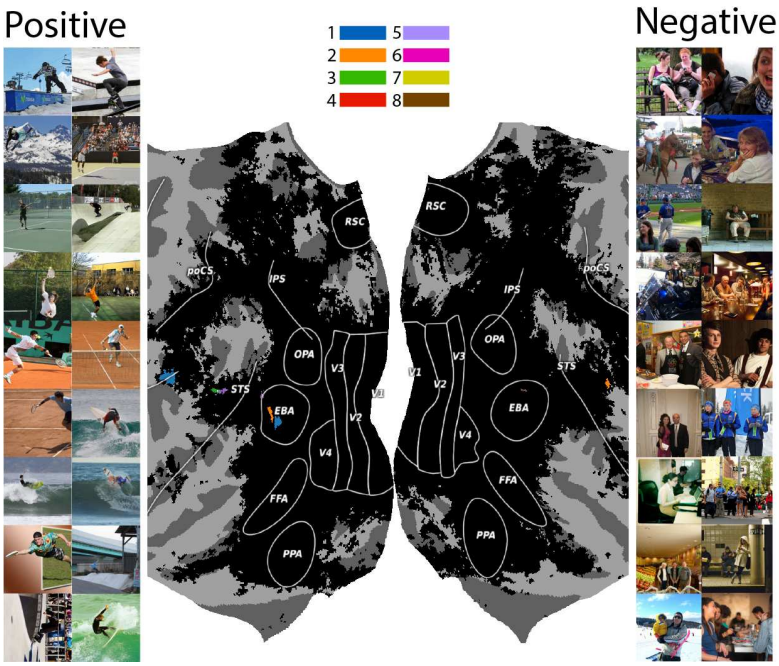

Figure 4.8: Cluster 6 ($\varepsilon = 0.55$). Positive images show people jumping or leaping, while negative images are of people standing still or sitting. Voxel clusters are in and around left hemisphere EBA.

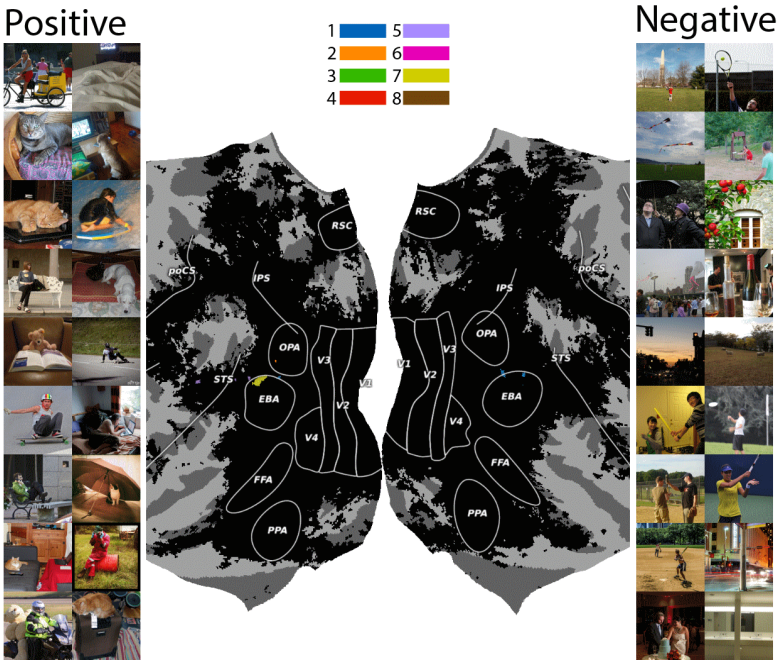

Figure 4.9: Cluster 5 ($\varepsilon = 0.55$). Positive images show people and animals who are crouched, sitting, or laying down, while the negative images show people standing. Voxel clusters are primarily found in bilateral FFA and EBA.

4.4 Places

When $\varepsilon = 0.55$, cluster 1 (Figure 4.10) is very large and overlaps strongly with PPA, OPA, and RSC. The positive representative images usually display an urban area where the camera is looking forward down a long path or road. In contrast, the negative images contain many close-up images of objects on tables where the surrounding environment is not visible. This suggests that this cluster may be related to scene geometry, navigation, or the reachability of locations in a scene (Dilks et al., 2013; Epstein and Kanwisher, 1998; Maguire, 2001). Cluster 9 (Figure 4.11) strongly overlaps with PPA and displays outdoor scenes with consistent vegetation for the positive images. The negative images show indoor scenes where human-made objects are prominent, with a distinct lack of vegetation.

At $\varepsilon = 0.6$, cluster 4 (Figure 4.12) display indoor scenes (the positive representative images). This cluster is mostly localized in OPA with some voxels in PPA. The positive images depict cluttered indoor scenes such as kitchens, work spaces, and living rooms. There is typically a flat surface such a desk, counter-top, or table at the focal point of the image. The negative images depict outdoor scenes with animals.

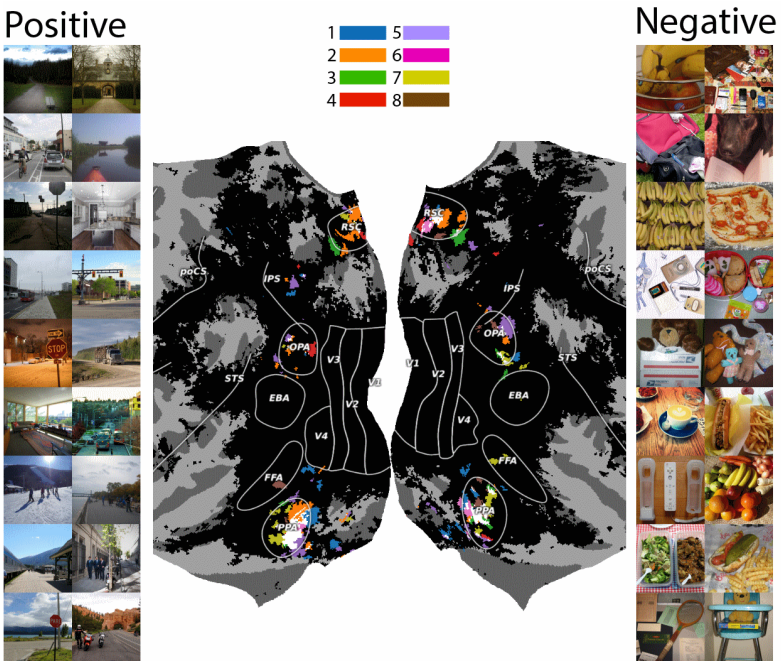

Figure 4.10: Cluster 1 ($\varepsilon = 0.55$). Positive images show scenes with large open areas. Negative images are close-ups of objects that have no scene geometry. Clusters span bilateral RSC, OPA, and PPA.

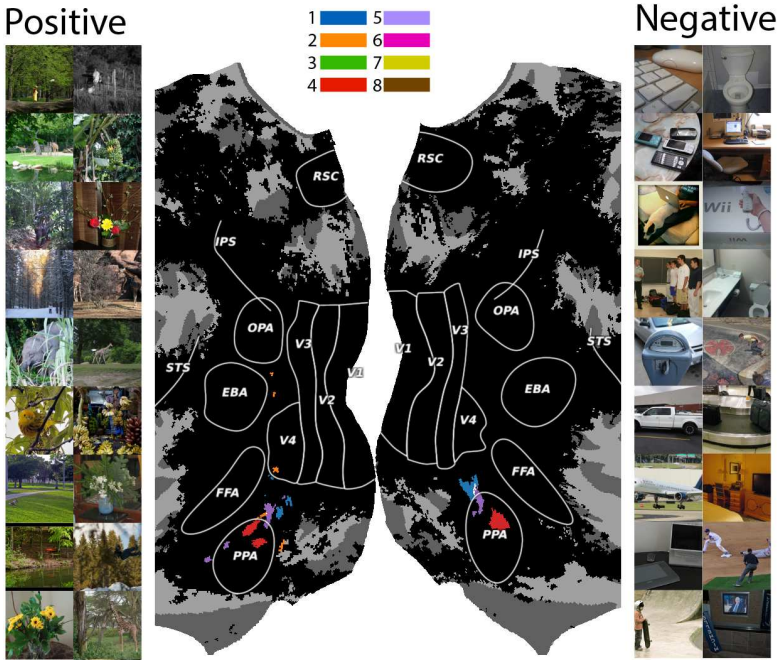

Figure 4.11: Cluster 9 ($\varepsilon = 0.55$). Positive images display scenes with plants and foliage, while negative images show human-made objects. Voxel clusters are in bilateral PPA.

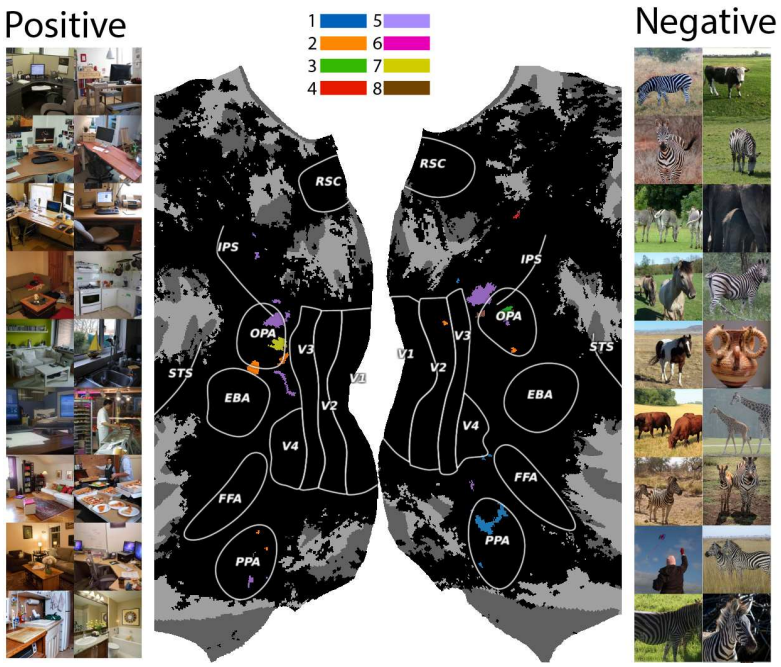

Figure 4.12: Cluster 4 ($\varepsilon = 0.60$). Positive images show indoor scenes with desks and tables. Negative images show outdoor scenes with animals. Voxel clusters are primarily in bilateral OPA and PPA.

4.5 Horizontal/Vertical Cluster

Very early work on the visual system discovered the tuning of the early visual system for lines of a particular orientation (Hubel and Wiesel, 1962). Two clusters emerge that reflect this tuning, both at $\varepsilon = 0.65$. Cluster 1 (Figure 4.13) has a strong horizontal component with strong horizon lines or large objects spanning the middle of the visual field creating a horizon-like line. Notably, the negative images are images with a strong vertical component. The inverse is true for cluster 7 (Figure 4.14) which shows strong verticality in the positive images, and horizontal in the negative. Interestingly some of the negative images from cluster 1 appear as positive images in cluster 7, and vice versa. The localization of both clusters is V2/V3.

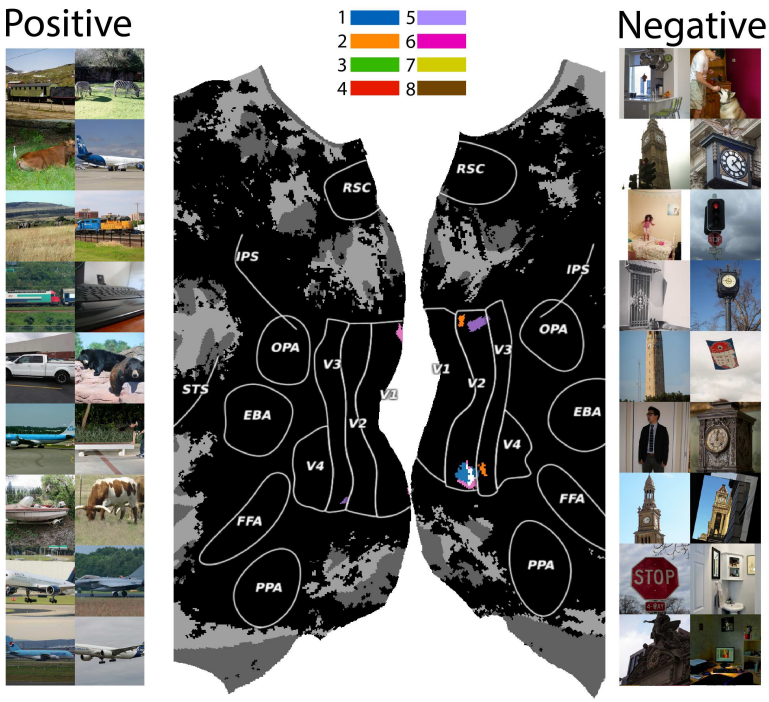

Figure 4.13: Cluster 1 ($\varepsilon = 0.65$). Positive images are associated with a horizontal mid-line, while negative images display a vertical mid-line. Voxel clusters are primarily in right hemisphere V2.

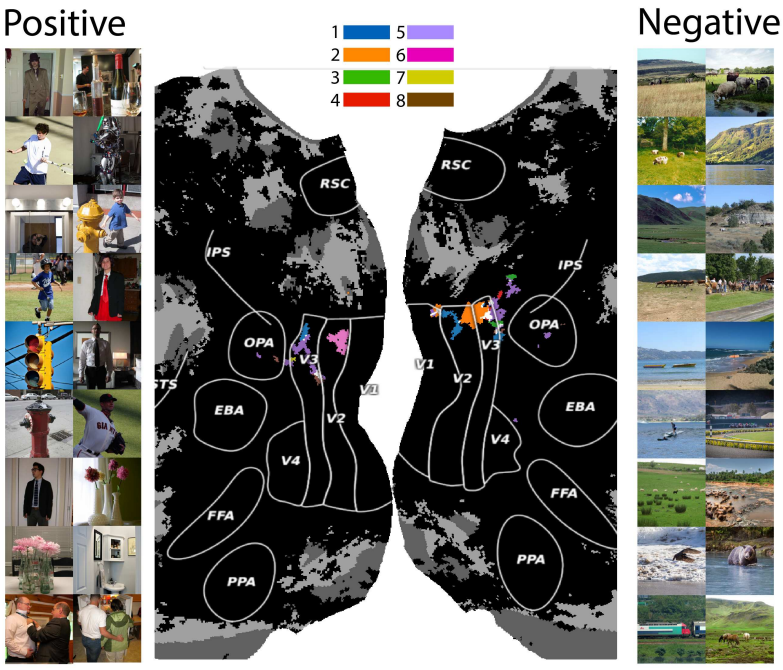

Figure 4.14: Cluster 7 ($\varepsilon = 0.65$). Positive images are associated with a vertical mid-line, while negative images display a horizontal mid-line. Voxel clusters are primarily in bilateral V3.

4.6 Repeated Elements (Numerosity) Cluster

We observed that cluster 12 ($\varepsilon = 0.55$, Figure 4.15) has positively associated images typically with repeated elements of related items, while negative images typically contained singular instances of items. We suggest that this SDC could be related to quantity processing and numerosity. We observed voxel clusters primarily in right OPA and IPS regions. As with other clusters, naturalistic images often implicates functional areas associated with place (OPA). The IPS has been widely reported in prior work to be associated with quantity processing, e.g. grammatical number processing Carreiras et al., 2010, mathematical processing deficits Ganor-Stern et al., 2020 and numerical processing Koch et al., 2023). Our observations in the visual domain provide converging evidence that the right IPS is associated with quantity processing across multiple modalities.

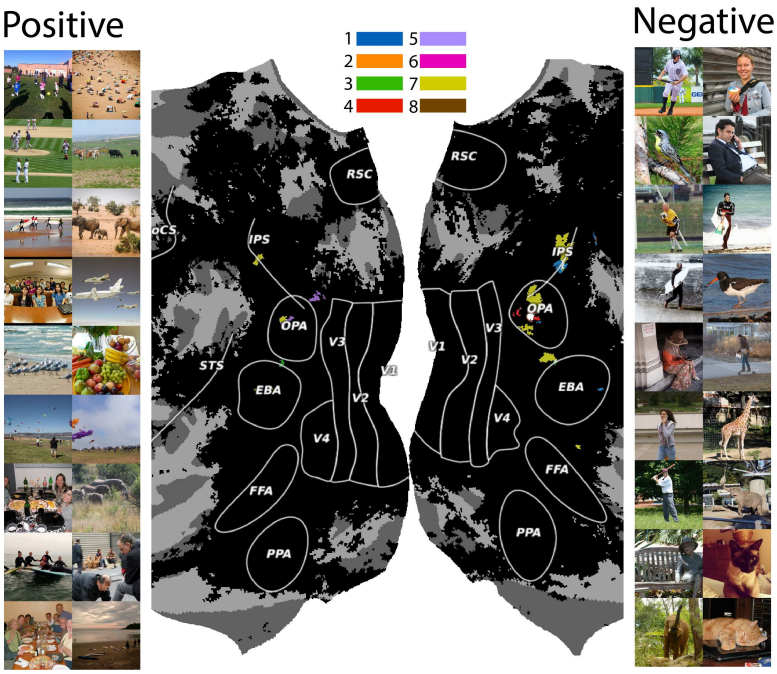

Figure 4.15: Cluster 12 ($\varepsilon = 0.55$). Positive images are associated with repeated elements, while negative images are single objects or individuals. Voxel clusters are primarily located in OPA and IPS.

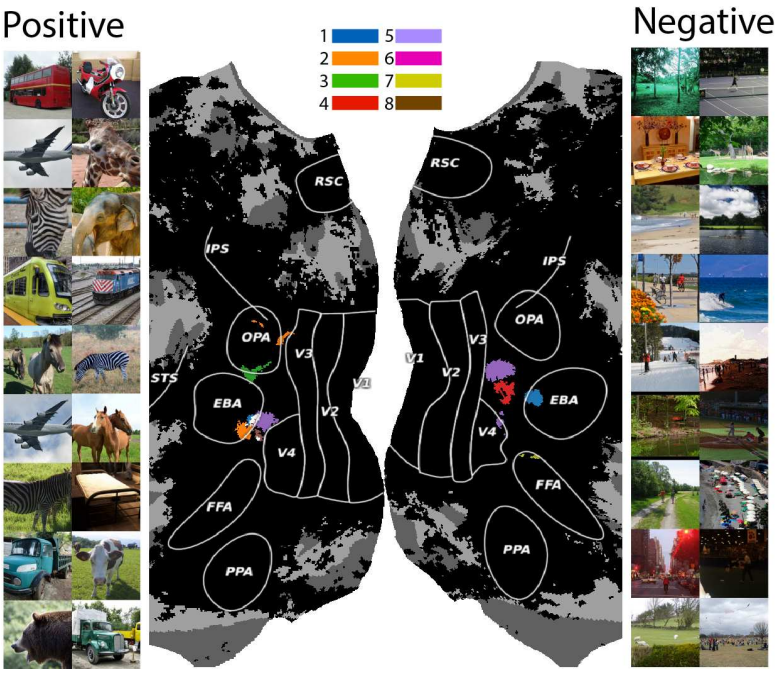

Figure 4.16: Cluster 2 ($\varepsilon = 0.65$). Positive images display close-ups of large animals or objects. Negative image show objects at a distance. Voxel clusters are located between EBA and V4.

4.7 Close vs Far Cluster

With $\varepsilon = 0.65$, cluster 2 (Figure 4.16) seems to represent big things (airplanes, busses) and close up pictures of larger animals (zebras, elephants). The negatively associated images are scenes that usually depict significant depth, including paths leading into the hills with animals or people in the distance. Sarch et al., 2023 explored the representation of image depth in cortex and also found that similar areas of cortex respond strongly to images with objects very close to the camera. In addition, Luo et al., 2023 reported that this general area is sensitive to relatively large objects.

4.8 Soft vs Hard Clusters

We observe two clusters that relate to soft versus hard objects. When $\varepsilon = 0.60$, cluster 8 shows images with a strong focus on clothing, bedding, and other textiles. The negative images show trains, and concrete-dominant architecture. The voxels in this cluster are mostly found in the right hemisphere bordering the area between FFA and EBA. As discussed in section 4.2, cluster 3, $\varepsilon = 0.65$ displays many soft foods in the positive representative images (Figure 4.4), with some instances of soft non-food objects. Additionally, we observe many hard and rigid objects in the negative representative images. This suggests that this cluster also responds to soft versus hard objects and is not entirely food related.

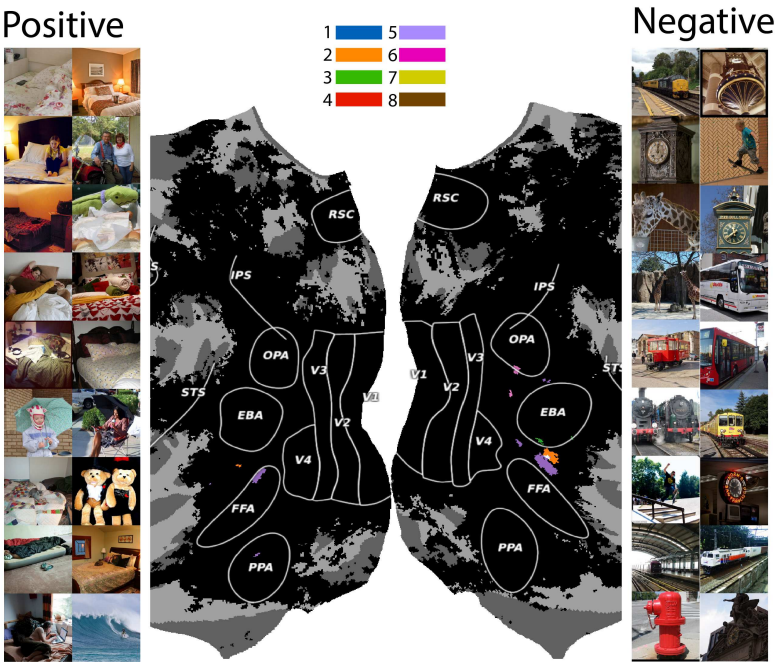

Figure 4.17: Cluster 8 ($\varepsilon = 0.60$). Positive images contain textiles, bedding, and other soft things. Negative images show hard objects. Voxel clusters are primarily in the right hemisphere between EBA and FFA.

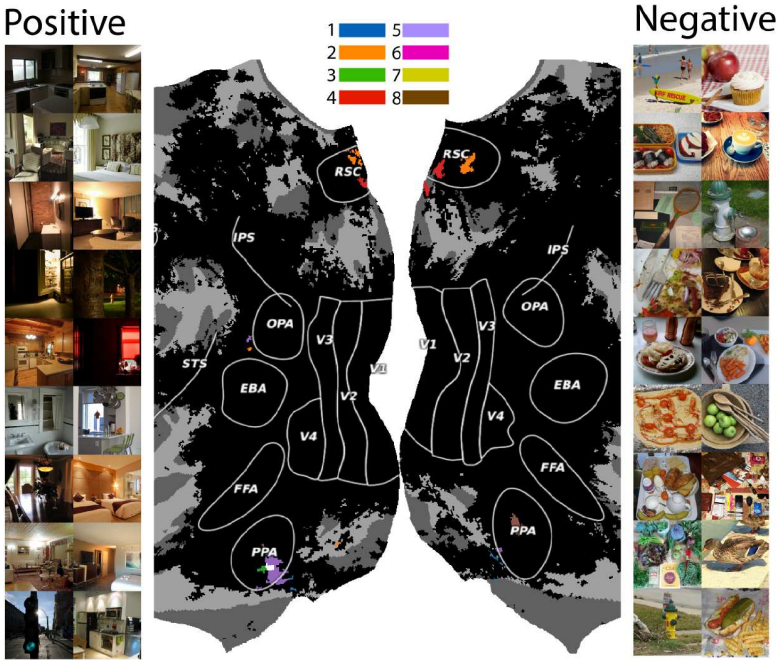

Figure 4.18: Cluster 7 ($\varepsilon = 0.50$). Positive images display a contrast between a light source and a dark environment. Negative images show uniform ambient lighting. Voxel clusters are primarily in bilateral RSC and PPA.

4.9 Lighting-Related Cluster

When $\varepsilon = 0.5$, most clusters are smaller or disconnected versions of clusters seen at higher ε values. The exception is cluster 7 (Figure 4.18), which is unique to $\varepsilon = 0.5$ and includes voxels in RSC and PPA. The positive images depict scenes with a high contrast in lighting. There is typically a dark environment with a bright light or a window that is partially illuminating the room. Conversely, the negative images depict close-up pictures of objects with uniform ambient lighting. This strongly suggests that this cluster of voxels responds to images that display a high contrast in lighting.

4.10 Words

Words and signs appear in NSD, and so we were interested to see if a word-related concept would emerge from our technique. When $\varepsilon = 0.60$ we found cluster 11 to have images with objects with characters like signs and clocks. The negation of this concept is perhaps best characterized as having strong visual contrast over large areas of the image (e.g. a backlit traffic light or a white desk). This cluster is bilaterally localized to the borders of V4. We note that this cluster is adjacent but not strongly overlapping with the visual word form area (VWFA) (McCandliss et al., 2003).

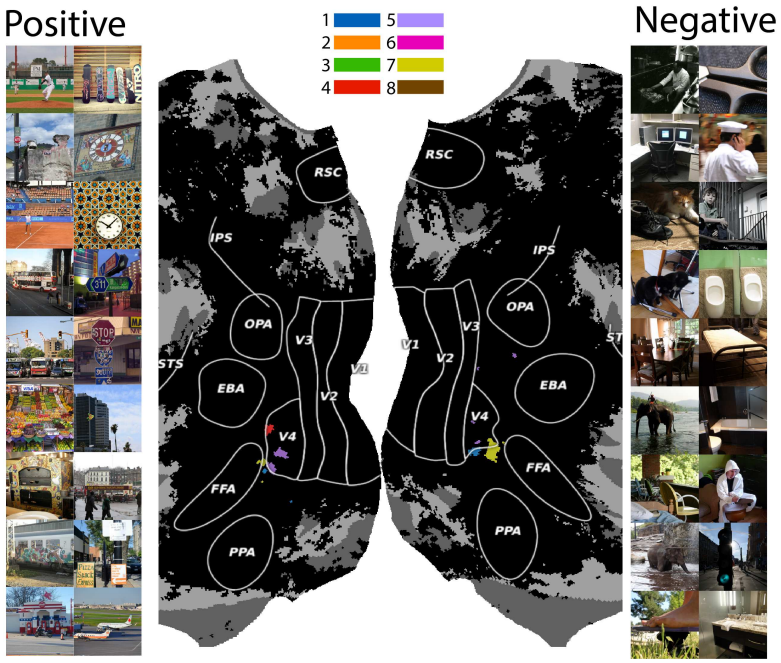

Figure 4.19: Cluster 11 ($\varepsilon = 0.60$). Positive images display words and symbols, while negative images have no discernible theme. Voxel clusters are primarily around bilateral V4.

4.11 Representative Word Clouds

In Figure 4.20, we present word clouds containing the most frequent words in the representative captions for each image. In most cases the word clouds support our interpretation of each underlying cluster concept. For example the word cloud for the lighting cluster (Figure 4.18) containing the words lit, lamp, sun, illuminated, dark, dimly, brightly, all of which are indicative of lighting conditions. Similarly, the repeated elements cluster (Figure 4.15) is well-characterized by words like group, together, formation, squadron, herd, which emphasize the concept of repetition or grouping.

In some cases, while the word clouds provide partial support for our descriptions, the thematic patterns become more apparent when examining the representative images. For example the positive representative words for the hands cluster (Figure 4.6) contains action-oriented words like as cutting and slicing, and the legs cluster (Figure 4.5) features words such as riding and running. However it is difficult to identify the overall theme of hands or legs

without the representative images.

The word clouds for the food clusters offer additional insights. The food and color cluster (Figure 4.2) is strongly indicative of favoring the color versus grayscale concept. In contrast, the anterior food cluster (Figure 4.3) has a strong association to food-related terms like food, eating, plate, avocado, and feeding. However, the food and softness cluster (Figure 4.4) lacks food-specific words despite an evident theme of food in the representative images. Although there are some softness-related words such as sheep, fur, slush, and snowy, there is not a strong theme of softness in the word cloud. These observations highlight the complexity of interpreting semantic representations in data-driven analysis, motivating the importance of targeted, hypothesis-driven experiments. For example, an fMRI experiment could be designed to test the neural responses to food stimuli compared to alternative categories like softness or color.

Figure 4.20: Word clouds for all SDC clusters. The positive and negative representative words are displayed side-by-side along with a subset of the representative images. The text to the left identifies the corresponding flatmap figure and our interpretation of the SDC. The vertical line SDC is omitted for space.

4.12 Chapter Summary

In this section we showcased a selection of SDC concept clusters that were uncovered using our contrastive decoder and modified DBSCAN clustering algorithm. Our interpretation of each cluster was based off selecting the closest CLIP embeddings to each cluster centroid. As expected, we identified clusters corresponding to familiar categories such as faces, places, bodies, food, color, and orientation. Interestingly, we also discovered clusters representing object size, softness, lighting conditions, and object repetition. These interpretations were further supported by generating representative word clouds using CLIP text representations for each cluster (Figure 4.20). Overall, our findings demonstrate the effectiveness of our method for identifying both expected and novel category-selective areas in the human brain.

Chapter 5

Conclusion

In this master's thesis, we present a novel, data-driven approach designed to uncover new category-selective areas in the brain and refine our understanding of known ones. Leveraging the Natural Scenes Dataset (NSD) and the CLIP neural network, we trained a contrastive decoder to map brain responses to CLIP embeddings. We demonstrated that a contrastive decoder outperforms a ridge-regression baseline, improving the decodability of under-represented image categories in NSD. By applying a modified DBSCAN clustering algorithm, we identified SDC clusters, representing dimensions of CLIP space that are decodable across multiple participants. Analyzing the stimuli that both activate and deactivate these SDC clusters provided deeper insights into their functions. Negative representations were particularly helpful in our understanding of SDCs for faces, softness, food and color, object size, and numerosity. Our method serves as a powerful hypothesis-generation technique with broad applicability to new datasets.

5.1 Broader Impacts

Methods that decode neural activity could have substantial impacts on both neuroscience and broader society. Identifying new category-selective brain regions could assist in diagnosing neurological and psychiatric conditions, surgical planning, development of brain-computer interfaces, and potentially helping individuals with locked-in syndrome or related disabilities to better communicate.

However, there are also significant ethical considerations with technologies that decode neural responses. If these techniques are used outside controlled research environments, there is a risk that they could be used for intrusive monitoring of mental states. As with all emerging technologies, responsible deployment is needed in order for society as a whole to maximize benefit while mitigating risks.

5.2 Limitations and Future Work

While our approach offers many advantages, it is not without drawbacks. Firstly, we are limited by the stimulus images that were chosen for the NSD experiment. Although the use of a contrastive loss function helps mitigate this, our model may still be biased toward over-represented categories in the stimulus set. Additionally, the use of CLIP may introduce its own biases, as some signal in the brain responses may not be fully captured by the CLIP embeddings. The use of fMRI also presents certain limitations. Despite its high spatial resolution, its limited temporal resolution restricts the ability to capture rapid neural activity that may play a crucial role in visual processing. Future work could address these limitations by utilizing larger and more diverse fMRI datasets, incorporating higher temporal imaging techniques such as electroencephalogram (EEG) or electrocorticography (ECoG), and exploring alternative stimulus representations beyond CLIP.

Additionally, our clustering approach might merge concepts within regions of CLIP space that have relatively uniform densities, potentially missing interesting SDCs. This issue could be addressed by adapting the heirarichcal version of DBSCAN, or by using different criteria to merge the cross-participant core points into clusters.

Furthermore, we acknowledge that our method is not a replacement for traditional hypothesis-driven experiments. It will be important for future work to test for alternative hypotheses for what might be driving responses in SDC clusters. Another interesting research direction is the application of this method to other sensory modalities beyond vision, such as auditory or

somatosensory stimuli.

References

Allen, Emily J. et al. (2022). "A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence." In: *Nature Neuroscience* 25.1. DOI: 10.1038/s41593-021-00962-x.

Bromley, Jane et al. (Aug. 1993). "Signature Verification using a "Siamese" Time Delay Neural Network." In: *International Journal of Pattern Recognition and Artificial Intelligence* 7, p. 25. DOI: 10.1142/S0218001493000339.

Carreiras, Manuel et al. (2010). "Where syntax meets math: Right intraparietal sulcus activation in response to grammatical number agreement violations." In: *NeuroImage* 49.2, pp. 1741–1749. URL: http://dblp.uni-trier.de/db/journals/neuroimage/neuroimage49.html#CarreirasCBH10.

Chen, Ting et al. (2020). *A Simple Framework for Contrastive Learning of Visual Representations*. arXiv: 2002.05709 [cs.LG]. URL: https://arxiv.org/abs/2002.05709.

Chopra, S., R. Hadsell, and Y. LeCun (2005). "Learning a similarity metric discriminatively, with application to face verification." In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1, 539–546 vol. 1. DOI: 10.1109/CVPR.2005.202.

Deng, Jia et al. (2009). "ImageNet: A large-scale hierarchical image database." In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

Dilks, Daniel D. et al. (2013). "The Occipital Place Area Is Causally and Selectively Involved in Scene Perception." In: *Journal of Neuroscience* 33.4, pp. 1331–1336. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.4081-12.2013. eprint: https://www.jneurosci.org/content/33/4/1331.full.pdf. URL: https://www.jneurosci.org/content/33/4/1331.

Downing, Paul et al. (Oct. 2001). "A Cortical Area Selective for Visual Processing of the Human Body." In: *Science (New York, N.Y.)* 293, pp. 2470–3. DOI: 10.1126/science.1063414.

Epstein, Russell and Nancy Kanwisher (1998). "A cortical representation of the local visual environment." In: *Nature* 392.6676, pp. 598–601.

Ester, Martin et al. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, pp. 226–231.

Ganor-Stern, Dana et al. (2020). "Damage to the Intraparietal Sulcus Impairs Magnitude Representations of Results of Complex Arithmetic Problems." In: *Neuroscience* 438, pp. 137–144. ISSN: 0306-4522. DOI: https://doi.org/10.1016/j.neuroscience.2020.05.006.

Horikawa, Tomoyasu and Yukiyasu Kamitani (2016). *Generic decoding of seen and imagined objects using hierarchical visual features*. arXiv: 1510.06479 [q-bio.NC]. URL: https://arxiv.org/abs/1510.06479.

Hubel, D. H. and T. N. Wiesel (Jan. 1962). "Receptive fields, binocular interaction and functional architecture in the Cat's visual cortex." In: *The Journal of Physiology* 160.1, pp. 106–154. DOI: 10.1113/jphysiol.1962.sp006837.

Jain, Nidhi et al. (2022). "Food for thought: selectivity for food in human ventral visual cortex." In: *bioRxiv*. DOI: 10.1101/2022.05.22.492983. eprint: https://www.biorxiv.org/content/early/2022/09/08/2022.05.22.492983.full.pdf. URL: https://www.biorxiv.org/content/early/2022/09/08/2022.05.22.492983.

— (2023). "Selectivity for food in human ventral visual cortex." In: *Communications Biology* 6.1. DOI: 10.1038/s42003-023-04546-2.

Kanwisher, Nancy, Josh McDermott, and Marvin M. Chun (1997). "The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception." In: *Journal of Neuroscience* 17.11, pp. 4302–4311. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.17-11-04302.1997. eprint: https://www.jneurosci.org/content/17/11/4302.full.pdf. URL: https://www.jneurosci.org/content/17/11/4302.

Khaligh-Razavi, Seyed-Mahdi and Nikolaus Kriegeskorte (2014). "Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation." In: *PLoS Computational Biology* 10. URL: https://api.semanticscholar.org/CorpusID:14942477.

Khosla, Meenakshi, N. Apurva Ratan Murty, and Nancy Kanwisher (2022). "A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition." In: *Current Biology* 32.19, 4159–4171.e9. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2022.08.009. URL: https://www.sciencedirect.com/science/article/pii/S0960982222012866.

Koch, Griffin E. et al. (Feb. 2023). "Representations within the Intraparietal Sulcus Distinguish Numerical Tasks and Formats." In: *Journal of Cognitive Neuroscience* 35.2, pp. 226–240. ISSN: 0898-929X. DOI: 10.1162/jocn_a_01933.

Kriegeskorte, Nikolaus et al. (May 2009). "Circular analysis in systems neuroscience: The dangers of double dipping." In: *Nature neuroscience* 12, pp. 535–40.

Lin, Tsung-Yi et al. (2014). *Microsoft COCO: Common Objects in Context*. cite arxiv:1405.0312Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list.

Liu, Qing et al. (May 2024). "Mind-bridge: Reconstructing visual images based on diffusion model from human brain activity." In: *Signal, Image and Video Processing* 18.S1, pp. 953–963. DOI: 10.1007/s11760-024-03207-z.

Liu, Yulong et al. (2023). *BrainCLIP: Bridging Brain and Visual-Linguistic Representation Via CLIP for Generic Natural Visual Stimulus Decoding*. arXiv: 2302.12971 [cs.CV]. URL: https://arxiv.org/abs/2302.12971.

Luo, Andrew F. et al. (2023). "BrainSCUBA: Fine-Grained Natural Language Captions of Visual Cortex Selectivity." In: *CoRR* abs/2310.04420. DOI: 10.48550/ARXIV.2310.04420. arXiv: 2310.04420. URL: https://doi.org/10.48550/arXiv.2310.04420.

Maguire, Eleanor (2001). "The retrosplenial contribution to human navigation: A review of lesion and neuroimaging findings." In: *Scandinavian Journal of Psychology* 42.3, pp. 225–238. DOI: https://doi.org/10.1111/1467-9450.00233. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9450.00233. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9450.00233.

McCandliss, Bruce D., Laurent Cohen, and Stanislas Dehaene (2003). "The visual word form area: expertise for reading in the fusiform gyrus." In: *Trends in Cognitive Sciences* 7.7, pp. 293–299. ISSN: 1364-6613. DOI: https://doi.org/10.1016/S1364-6613(03)00134-7. URL: https://www.sciencedirect.com/science/article/pii/S1364661303001347.

Oord, Aäron van den, Yazhe Li, and Oriol Vinyals (2018). "Representation Learning with Contrastive Predictive Coding." In: *CoRR* abs/1807.03748. arXiv: 1807.03748.

Ozcelik, Furkan and Rufin VanRullen (Sept. 2023). "Natural scene reconstruction from fmri signals using Generative Latent Diffusion." In: *Scientific Reports* 13.1. DOI: 10.1038/s41598-023-42891-8.

Pennock, Ian Morgan Leo et al. (2022). "Color-biased regions in the ventral visual pathway are food-selective." In: *bioRxiv*. DOI: 10.1101/2022.05.25.493425. eprint: https://www.biorxiv.org/content/early/2022/05/26/2022.05.25.493425.full.pdf.

Pérez-Ortega, Jesús, Alejandro Akrouh, and Rafael Yuste (Apr. 2024). "Stimulus encoding by specific inactivation of cortical neurons." en. In: *Nature Communications* 15.1, p. 3192. ISSN: 2041-1723. DOI: 10.1038/s41467-024-47515-x.

Radford, Alec et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision." In: *CoRR* abs/2103.00020. arXiv: 2103.00020.

Sarch, Gabriel et al. (2023). "Brain Dissection: fMRI-trained Networks Reveal Spatial Selectivity in the Processing of Natural Images." In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., pp. 46255–46283. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/90e06fe49254204248cb12562528b952-Paper-Conference.pdf.

Scotti, Paul S. et al. (2023). *Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors*. arXiv: 2305.18274 [cs.CV]. URL: https://arxiv.org/abs/2305.18274.

Stigliani, A., K. S. Weiner, and K. Grill-Spector (Sept. 2015). "Temporal Processing Capacity in High-Level Visual Cortex Is Domain Specific." en. In: *Journal of Neuroscience* 35.36, pp. 12412–12424. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.4822-14.2015.

Wang, Aria Y. et al. (2023). "Natural language supervision with a large and diverse dataset builds better models of human high-level visual cortex." In: *bioRxiv*. DOI: 10.1101/2022.09.27.508760. eprint: https://www.biorxiv.org/content/early/2023/07/11/2022.09.27.508760.full.pdf. URL: https://www.biorxiv.org/content/early/2023/07/11/2022.09.27.508760.