# Using Task Driven Methods to Uncover Representations of Human Vision and Semantics

Aria Yuan Wang

July 27, 2023

Joint Ph.D Program in Neural Computation and Machine Learning
Carnegie Mellon University
Pittsburgh, PA 15213

**Thesis Committee:**
Leila Wehbe, Co-chair
Michael J. Tarr, Co-chair
Bradford Mahon
Surya Ganguli (Stanford)

*Submitted in partial fulfillment of the requirements*
*for the degree of Doctor of Philosophy.*

**Keywords:** human vision, semantic processing, scene understanding, taskonomy, fMRI

# Abstract

Humans are multimodal and multitasking agents – a fact reflected in the complexity of our visual system. Robust behavior is supported by multiple visual processing pathways in the human brain, each capable of facilitating a wide variety of downstream tasks, ranging from recognizing familiar objects, navigating a new scene, and inferring ongoing events from a picture, etc. Slow progress in unraveling the representational basis and mechanisms of these pathways has been a challenge for vision scientists for many decades. Excitingly, our ability to account for neural responses in visual brain pathways has recently advanced much more rapidly due to the use of representations derived from task optimized neural networks. Previously unaccounted for high-level tasks in both visual and semantic processing can now be "explained" by state-of-the-art deep neural networks. While such explanations are promising, they are based on black box-like models that achieve much better accuracy than before, but still suffer from a lack of meaningful interpretations for understanding neural processing. My work advances our understanding of visual and semantic processing in the human brain by: 1) leveraging relationships among modern computer vision tasks to reveal the task-specific architecture of the human visual system; and 2) using multi-modal networks trained on both vision and language to investigate representational basis of visual semantic processing. I also address the limitations of task driven methods as well as potential for applying these methods in a more dynamic, cross-modal setting to model the interplay of language and visual representations.

# Acknowledgments

First and foremost, I would like to express my gratitude to my advisors, Leila Wehbe and Michael Tarr — I could not have completed this journey without their support. In addition to showing me how to properly tackle research questions and communicate my findings, your constant drive for curiosity, patience and optimism in research are what truly fueled me on this journey. You both taught me that doing research, like most things in the world, is full of compromises and imperfections but it is more important to carry on despite setbacks. Thank you both for creating supportive and open lab environments which I will forever cherish.

I never considered myself a good student growing up amongst intense competition in China, let alone someone who would earn a PhD. I am grateful for the time I spent in Berkeley exploring, and for the people I met that inspired me to take on this path: Walter Freeman III and Weishun Zhong. I still remember the excitement I felt in our discussions about the boundary-less unknown that led me to find purpose in pursuing science. And thank you to Jack Gallant and Mark Lescroart who showed me how research is done and heavily influenced the flavor of research that I ended up doing. I would also like to thank my committee members Brad Mahon and Surya Ganguli; your great works and useful feedback helped in forming this thesis. I am also grateful to my collaborators, and everyone in Tarr lab and Wehbe lab for useful discussions and feedback. You all made me a better researcher. I would like to especially thank Jayanth Koushik and Nadine Chang, for all the time spent together debugging code, rushing for deadlines, watching trash TV, and being there for each other during all the ups and downs in both work and personal life. You made working in the office fun. Lastly, I would like to thank my parents, who made this path possible by supporting me through these years of education.

The years I spent in getting this degree also taught me how to forgive myself at times, how to love others and how to try to live a more meaningful life. For that I would like to thank Roie Levin (and his jokes) for creating probably the happiest time I have had during my PhD amid the pandemic. And also Ruizhe Liu and Jenny Sun for their company over the years, and for being the strong women they are. You inspired me with your pursuits in academia and in life. I would also like to thank my cat Màn-Màn for her company. Dealing with her cancer in the last year of my PhD constantly reminded me there was more to life than publishing papers. For maintaining my mental and physical health, I would like to thank Ascend climbing gym for introducing me to rock climbing, and the climbing community for providing me an outlet and an alternative home when I needed a break from academia.

Lastly I would like to thank Arjun Teh for supporting me through this journey and in starting the next chapter of my life. Thank you for your company and patience, and for being a rock in my life. I look forward to more adventures together, on and off the mountains. Thank you.

# Contents

# Chapter 1

# Introduction

Human vision supports a wide range of high-level tasks, ranging from object classification, navigation, and scene interpretation, etc. Since Hubel and Wiesel [1959a] discovered the orientations receptive fields of neurons in 1959, and Kanwisher et al. [1997a] identified the patch in the brain consistently responds to face images in 1997, for many decades, progress has been slow in unraveling the representational basis of these neural pathways that support these high-level tasks. One of the main reason why this is a persistent challenge for vision scientists is that, manually curating these mid to high level features that the brain represents are simply beyond our ability. It is for the same reason why computational models for vision did not work well before deep learning and big datasets. We simply cannot come up with layers of perfect descriptors to make up a model to recognize a dog, for example.

Excitingly, our ability to account for neural responses in the visual brain has recently advanced much more rapidly due to the use of representations derived from task optimized deep neural networks. Previously unaccounted for high-level tasks in both visual and semantic processing can now be "explained" by state-of-the-art deep neural networks [Yamins et al., 2014a, Kell et al., 2018]. While not created as models of the brain *per se*, these deep neural network models' dramatic increases in prediction performance appear to be driven by the fact that common task goals between artificial and natural systems lead to similar representations [Yamins and

DiCarlo, 2016a, Agrawal et al., 2014, Güçlü and van Gerven, 2015]. While the exact reason why these networks are so good at mapping brain representation is not well studied, it is reasonable to hypothesize that the "engineering solution" that these networks land at solving a task, can be useful to probe the "biological solution" that the brain comes up with. One thing to note here is that this approach simply maps representations from task-optimized network to the brain, instead of optimizing these neural networks directly to predict brain responses, therefore the success of this approach cannot simply be explained by the fact that neural networks are simply powerful approximators for any functions.

Works presented in this thesis further explores the efficacy and limitation of this approach. First of all, this high-level correspondence between "engineering solution" and "biological solution" has held particularly in the study of vision, where computer vision models of object classification are very effective at accounting for neural responses in human ventral-temporal cortex – the neural pathway that supports visual object processing and recognition[Güçlü and van Gerven, 2015, Yamins and DiCarlo, 2016b, Toneva and Wehbe, 2019]. Until recent years, almost all neural-network models deployed in visual neuroscience are trained on a single *visual* task – object classification. The long focus on object-centered vision is not surprising since it is itself a better defined task compared to scene vision – Image of a object can have a clear label. However, human visual perception is a multifaceted process that incorporates both a wide variety of task objectives and interactions between visual and non-visual knowledge [Aminoff and Tarr, 2021]. Vision allows us to navigate everyday life, understand the things around us and to interact with the environment. To learn about the world, we constantly think and reason about what we see. Therefore high-level visual representations are thought to reflect both the structure of the visual world, relevant information for potential task (i.e. navigation) and *semantics* – non-perceptual associations such as object function or linguistic meaning [**?**]. As such, relying on neural-network models optimized for visual tasks such as object classification necessarily limits prediction to purely visual components of the perceptual process.

In my thesis, I show my work of extending the method of using task driven representations for brain mapping to a multi-task and multi-modality approach. The thesis is organized as follows:

- Section 2 introduces related works.
- Section 3 describes challenges in modeling scene understanding in human with fMRI data, and explains why we adopted specific approaches and datasets.
- Section 4 introduces our work on using a pool of 21 task-trained networks (Taskonomy[Zamir et al., 2018]) to quantify task relevant information in the brain. This work aims at using task driven networks and their relations as interpretation tools to disentangle high level visual representations in the brain.
- Section 5 details our attempts in modeling affordances (i.e. actions an objects affords to an observer) with task-driven neural networks, where we designed a affordances datasets, trained a neural network to predict affordances to pictures.
- Section 6 introduces our work on using representations from a multimodal model with language and vision pre-training to map out relevant semantic dimensions in the visual semantic processing in the brain. This works aims at exploring the fact that humans learn to see with language learning and shows that language grounded visual representation is a better model and explains unique variances in the brain.
- Section 7 describes our ongoing attempts on applying task-driven models on modeling attention switching across modality. We extended the use of visual and language models to probe the attended modality while humans subjects are watching a famous TV sitcom.
- Section 8 describes our attempt to look at the potential problems and limitations in using different task driven representations for brain mapping and propose a potential solution with network sparsification.
- Section 9 contains conclusions and discussion of the works described in this thesis.

# Chapter 2

# Background and Related Work

As one of the most primitive systems in animals, the visual systems has been the forefront topic of neuroscientists' endeavors in understanding how the brain works. One of the reasons is that vision is essential for primal animal behaviors such as food gathering, prey and recognising enemies and basically anything related to interaction with the environment. Another reason is that visual processing is relatively straightforward and representative of other perceptive systems. It is an mostly unconscious process where animals do effortlessly. Unlike more complex and conscious process such as decision making where modeling the internal states is crucial, without the manipulation of attention, we can treat vision as a faithful system where if you give it the same input, it will respond with the same output.

In one of the most important early experiments in visual neuroscience, Hubel and Wiesel showed light bars to cats and recorded cells that consistently responded to different orientations of edges [Hubel and Wiesel, 1959a]. This experiment allowed the pinpointing of a cell's receptive field, and led researchers to use cellular recording to eventually identify properties such as shape preferences in V4 neurons [Roe et al., 2012, Gallant et al., 1993] and direction preference in motor cortex by populations of neurons [Georgopoulos and Carpenter, 2015]. The difficulties of going down this path lies in coming up with exactly what the neuron is coding for to present it to the system and record the neuronal firing.

It took a couple decades for the data collection to catch up from single cell recordings to calcium imaging, which allows us to record populations of neuron at a time, and as well as to fMRI, which uses blood flow as a proxy for neuronal activity at the whole brain level. The opportunity to observe the system in a noninvasive way at the whole brain level allowed us to model various levels of processing across the human visual system and fMRI became a popular method among visual neuroscientists. This in turn led to the discovery of specific semantically defined functional brain regions such as the fusiform face areas (FFA) [Kanwisher et al., 1997a, 2002], parahippocampal place area (PPA) [Epstein and Kanwisher, 1998a], extrastriate body area (EBA) [Downing et al., 2001a], etc. At this stage, most of the findings were driven by univariate methods where images of a few categories or different properties are presented and the contrast of voxels responses are recorded as selective responses for the presented categories. The hypothesis space are largely limited with this method.

More than a decade later, a new framework, voxelwise encoding models, was developed by Naselaris et al. [2011] to describe neural processing at the voxel level across the whole brain. In this framework, hypotheses about input properties that the neurons care about (e.g., 30 degree line, human face) are parameterized as a function of the input. The encoded properties are then mapped linearly to voxel responses from seeing an image. Prediction accuracy of the encoding models are evaluated on held out datasets and successful prediction indicates that a voxel shares similar representations as the hypothesis incorporated in the encoding model. The encoding model framework allows researchers to use highly complex naturalistic stimuli (such as a scene in real life) and test a wide range of hypothesis at the same time. After the model is learned, it provides a basis of latent visual features that the brain represents from the highly complex visual environments we live in. With this framework, neuroscientists were able to map out low to high level visual features the brain uses and even decode what a subject sees in the scanner [Kay et al., 2008, Nishimoto et al., 2011]. It was also widely applied in other neural functions such semantic and language, where representations of the semantic concepts were mapped out

in the brain [Huth et al., 2016], and we have a better picture of how the brain represent language with contexts, syntax as well as meaning [Wehbe et al., 2014a, Deniz et al., 2019]. Encoding models thus allow the test of a large number of hypotheses from the same naturalistic data. The difficulties now lie in finding good hypotheses about the input that capture how visual systems process image for downstream tasks.

Almost in parallel to the quest for biological intelligence, computer vision has been one of the main driver that pushes artificial intelligence forward. Large scale image datasets starting with ImageNet [Deng et al., 2009a] have helped to scale up the training of artificial networks. Increasingly large-scale naturalistic dataset allowed for networks trained to do different tasks[Zamir et al., 2018], and bridge across modalities (such as CLIP [Radford et al., 2021]). An interesting point for researchers using encoding models is that artificial networks that solve tasks that human can do, such as recognizing an object from a picture and navigating through a space, can be used as proxy models as well as hypotheses for what we think the brain is doing for solving the same tasks. This is not merely an coincidence but the fact that human visual systems likely use similar efficient coding that extract the statistical regularities from the visual world for use in downstream task. Properties we can already pinpoint to be important for the visual system, such as hierarchies of edges and shapes, can be found in networks learned with sparse coding [Olshausen and Field, 1996], as well as deep neural network trained in an end-to-end fashion. What's more, these networks also provides us other visual features that are hard to be described by language or designed by hand, and can be used as hypotheses to test with the visual system. Indeed, the use of the activations from task trained networks to model representations in the brain has provided a leap of prediction of brain responses for visual stimuli in recent years [Agrawal et al., 2014, Güçlü and van Gerven, 2015, Yamins and DiCarlo, 2016b, Yamins et al., 2014b].

Advances in dataset scale in neuroscience also opened up a new chapter in modeling of brain representations. Datasets such as BOLD5000 [Chang et al., 2019] with around 5000 images per subject and Natural Scene Dataset (NSD) [Allen et al., 2021] with 10,000 images per subject,

rely on images from benchmark computer vision datasets. These images are taken from real life scenes, and span much larger visual and semantics space than the images from most other controlled visual experiments. Neural datasets that are extensive in both stimuli space and recording time, with working end-to-end trained image-computable models, opens up a brand new frontier for data driven neuroscience research.

Data driven methods have shortcomings as well, especially in their poor interpretability. As features extracted from these proxy model become more and more complex, they become harder to interpret. Dimensionality reduction methods such as principal component analysis (PCA) are handy to make sense of the low dimension latent space of the model. Unit visualization tool such as Net Dissect [Bau et al., 2017], and attribution analysis through pruning [Tanaka et al., 2019] are useful as well to tease apart representations of specific units in a network.

My thesis focuses on modeling the processing of natural scenes in the human visual system. Scene understanding requires the integration of space perception, visual object recognition, and the extraction of semantic meaning. The human brain's solution to this challenge has been elucidated in recent years by the identification of scene-selective brain areas via comparisons between images of places and common objects [Epstein and Kanwisher, 1998a]. This simple contrast has been extended across a wide variety of image manipulations that have provided evidence for the neural coding of scene-relevant properties such as relative openness [Kravitz et al., 2011, Park et al., 2014, Harel et al., 2012], the distance of scenes to the viewer [Kravitz et al., 2011, Park et al., 2014, Lescroart et al., 2015], 3D spatial layout [Ferrara and Park, 2016, Kamps et al., 2016, Kornblith et al., 2013] and navigational affordances [Bonner and Epstein, 2017]. Independent findings in how the human brain process features related to scene understanding are hard to organically built on top of of each together since each experiment might use different but correlated hypothesis. Recently, in 2019, to help with this, Lescroart and Gallant [2019] developed an encoding model using a feature space that parametrizes 3D scene structures along the distance and orientation dimensions and provides a computational framework to account for human scene

processing. Intriguingly, Lescroart and Gallant [2019] were able to identify distance and openness within scenes as the dimensions that best account for neural responses in scene-selective brain areas. At a higher, semantic, level, Stansbury et al. [2013] found that neural responses in scene-selective brain areas can be predicted using scene categories that were learned from object co-occurrence statistics. Such findings demonstrate that human scene-selective areas represent both visual and semantic scene features.

# Chapter 3

# Motivations and Methods

## 3.1 Motivations: Challenges in scene understanding modeling

Natural scenes are complex, high dimensional in nature and ill-defined by language. It is hard to answer the question: what is a natural scene? It encompasses any visual scene we encounter in the wild. Indoor, outdoor, close-up, or further away, natural scenes contain a wide variety of different visual and semantic contents. Furthermore, building models of scene understanding is complicated by the fact that visual and semantic features are correlated with one another. For example, man-made objects such as buildings have more rectilinear lines as compared to natural objects such as animals (cite), or, for example, that horses and grasses tend to co-occur. For these and many other reasons, scene perception in humans entails a model that encapsulates a broad space in terms of both visual and semantics features. Only through such a model will we have a fuller picture of how the human brain processes the complex information carried in natural scenes.

At the same time, there is strong evidence that the neural representation of visual concepts and features is distributive in nature. From past work we learned that, even though there are regions of the brain that can be shown to be associated with important semantic categories such as faces [Kanwisher et al., 1997a] or places [Epstein et al., 1999], most other semantic categories

are distributively coded in the human brain [Huth et al., 2016] and even nominally category-selective regions also encode information about non-domain categories [McGugin et al., 2012, Grill-Spector et al., 2006].

These challenges directly inform our chosen methodologies for modeling scene understanding in human:

1. We build an **efficient voxelwise encoding model** pipeline that utilizes linear algebra tricks such as the Woodbury matrix identity[Woodbury, 1950] and is implemented on PyTorch for running on GPUs. This pipeline could fit regression models for all voxels from the whole brain fMRI with efficient hyperparameter tuning.
2. We use **naturalistic stimuli** in human fMRI experiments to obtain realistic neural responses to complex scene images.
3. We combine naturalistic scene images together with **ecological tasks as training feedback to build an image computable proxy model** for scene understanding in the human brain.
4. We extract activations from the trained models as features to build a **voxelwise encoding model** that informs us as to where in the brain visual and semantic information is represented.
5. We build **interpretable tools** with variance partitioning and principal component analysis (PCA) to further interrogate the feature spaces that are represented across the brain.

The following sections provide details on each of these methods.

## 3.2 Our approaches

### 3.2.1 Encoding Models

Encoding models – predictive models of brain activity that are able to generalize and predict brain responses to novel stimuli [Naselaris et al., 2011] – are widely used in understanding feed-

forward information processing in human perception, including scene perception. Researchers have also used encoding models to infer which dimensions are critical for prediction by comparing the weights learned by the model [Huth et al., 2016, Lescroart and Gallant, 2019]. One of the successes of encoding models lies in predicting low- to mid-level visual cortex responses in humans and primates using features that were learned via a convolutional neural network trained on object recognition [Agrawal et al., 2014, Güçlü and van Gerven, 2015, Yamins et al., 2014b, Eickenberg et al., 2017]. Most interestingly, these studies demonstrate a correspondence between human neural representation and learned representations within CNN models along the perceptual hierarchy: early layers tend to predict early visual processing regions, whereas later layers tend to predict later visual processing regions. Similarly, researchers have found that network representations from other task-driven networks, including networks trained on speech or music related tasks, are able to explain neural responses in human auditory pathways [Kell et al., 2018]. Many researchers hold that such successes are not mere coincidences but rather indications of how fundamental task-driven representations are to both task training and to information processing in the brain [Yamins and DiCarlo, 2016b].

In the works I present here, I apply encoding models across multiple works. The procedure is as follows. I first parameterize each image in the training set into values along different feature dimensions in a feature space. For example, if the feature space of interest is an intermediate layer in a task-driven network, we simply feed the image into the network and extracted its layer activation. The activations are then used as regressors in a ridge regression model (implemented in PyTorch; see [Koushik, 2017]) to predict each voxels' response to that image. The fMRI dataset to images are split into training and testing sets, usually with a ratio of 4-to-1 ratio. Training data are further split into train and validation sets with 4-to-1 ratio. Performance from the validation data is used to choose the regularization parameter in the ridge regression model. We choose to use a ridge regression model instead of more complicated models in order to retain the interpretability of model weights, which may provide insights into the underlying

dimensions of the brain responses. For each subject, each voxel's regularization parameter was chosen independently via 7-fold cross-validation based on the prediction performance of the validation data. We swept through 100 regularization parameters spaced evenly on a log scale from $10^{-8}$ to $10^{10}$, i.e. np.logspace(-8, 10, 100).

Ridge solutions for multiple regularization values are computed efficiently by using the Woodbury identity[Woodbury, 1950]. With X (n x d) representing the feature matrix, and y (n x 1) the outcomes, the ridge solution is given by

$$\beta = (X^T X + lI)^{-1} X^T y$$

where $l$ is the regularization coefficient. This can be reduced to

$$(1/l)(X^T y - X^T V(e + lI)^{-1}(X^T V)^T X^T y)$$

where $Ue^{1/2}V^T$ is the singular-value decomposition of $X^T$. Since $(e + lI)$ is a diagonal matrix, its inverse can be computed efficiently simply by taking the reciprocal of the diagonal elements. Then, $(X^T V)^T X^T y$ is a vector; so it can be multiplied by $(e + lI)^{-1}$ just by scalar multiplication.

After the regularization parameters are selected, ridge models are retrained with the training and validation data. Final model performance was evaluated on the test data using both Pearson's correlation and coefficient of determination ($R^2$). To determine the significance of the predictions, we ran permutation tests where we shuffle the stimuli evoked brain responses 5000 times, re-computed the performance metrics such as correlation score and $R^2$, and obtained FDR corrected $p$-values for both ROI and whole brain results.

### 3.2.2 Variance Partitioning

To obtain unique variance by two model A and B, as shown in 3.1, we first create joint model of A and B by concatenating features from these two models. We then fit voxelwise ridge regression

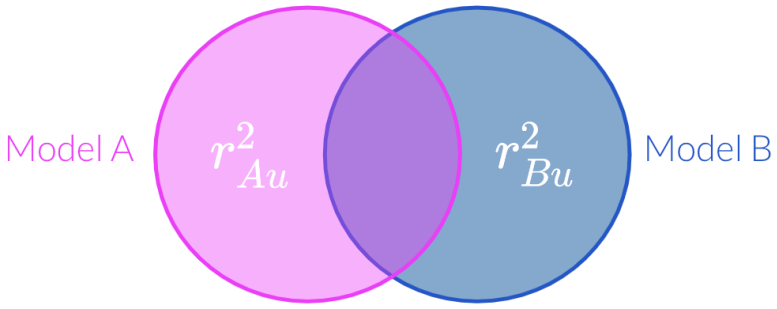



Figure 3.1: Schematic of variance partitioning between two models. The goal is to identify variance that is explained by one model and not the other (outer crescents).

model to the joint model and obtain $R^2_{A\&B}$. The variance explained by individual model A and B are denoted as $R^2_A$ and $R^2_B$, respectively. We then calculate the unique variance for model A and B:

$$R^2_A = R^2_{\&B} - R^2_B$$

$$R^2_B = R^2_{A\&B} - R^2_A$$

.

### 3.2.3 PCA analysis

Principal component analysis (PCA), or singular vector decomposition (SVD) are widely used as tools to recover basis of latent space in encoding models trained to predict brain responses [Huth et al., 2016, Lescroart and Gallant, 2019]. More specifically, after the encoding model is learned, we applied PCA to the weight matrix that has dimension of #-of-features by #-of-voxel. This could be applied on subject specific weight matrix, or a group weight matrix from concatenation of subject weight matrices. For group models, we selected usually 20,000 best predicted voxels (out of roughly 100k) for each individual subject based on the noise corrected model performance.

## 3.3 fMRI Datasets with naturalistic stimuli

### 3.3.1 BOLD5000

BOLD5000 is a publicly available large-scale fMRI dataset [Chang et al., 2019]. In the BOLD5000 study, participants' brains were scanned while they fixated at real-world images and judged how much they liked the image using a button press. Images in the BOLD5000 dataset were chosen from standard computer vision datasets (ImageNet [Russakovsky et al., 2015], COCO [Lin et al., 2014a] and SUN [Xiao et al., 2010]). The experiment was run in a slow-event setting where trials are separated by 10 seconds. From BOLD5000, we used data from three participants viewing 4916 unique images. These 4916 image trials are separated into random training, validation, and testing sets during model fitting. Average of TR 3 and 4 of each slow-event trial is used for model fitting and testing. Region of interest (ROI) boundaries that identify category-selective brain regions in the whole-brain map presented in our results were generated directly from the ROI masks provided with the BOLD5000 dataset.

### 3.3.2 Natural Scene Dataset (NSD)

Natural Scenes Dataset (NSD) [Allen et al., 2022] is an open dataset of 7T whole brain high-resolution fMRI responses from eight subjects (S1-S8) who each viewed $\sim$10,000 unique images of natural scenes, each image repeated 3 times. These scene images were a subset of the images in the annotated Microsoft Common Objects in Context (COCO) dataset [Lin et al., 2014b]. COCO is unique among large-scale image datasets in that COCO images contain contextual relationships and non-iconic (or non-canonical) object views. In comparison to ImageNet [Deng et al., 2009b], COCO contains fewer labeled categories (91), but includes more examples for each category ($> 5,000$ for 82 of the categories). Note, however, that many labeled categories in ImageNet are at the subordinate level – COCO likely contains at least as many *unlabeled* subordinate categories. The complete set of COCO images and additional details can be found

on the COCO website: `https://cocodataset.org`.

Of the 70,566 total images presented across subjects, ∼1,000 images were viewed by all subjects. fMRI data were collected during 30-40 scan sessions. Stimulus images were square cropped, presented for 3 s at a size of $8.4^\circ \times 8.4^\circ$ with 1 s gaps in between image presentations. Subjects were instructed to fixate on a central point and to press a button after each image if they had seen that image previously.

The functional MRI data were acquired at 7T using whole-brain gradient-echo EPI at 1.8-mm resolution and 1.6-s repetition time. Preprocessing steps included a temporal interpolation (correcting for slice time differences) and a spatial interpolation (correcting for head motion). Single-trial beta weights were estimated with a general linear model. In this paper we used the betas_fithrf_GLMdenoise_RR preparation of the betas. FreeSurfer [Dale et al., 1999, Fischl et al., 1999] was used to generate cortical surface reconstructions to which the beta weights were mapped. The beta weights were z-scored across run and were averaged across repetitions of the image (up to 3 repetitions of each image), resulting in one averaged fMRI response to each image per voxel, in each subject. NSD also includes several visual ROIs that were identified using separate functional localization experiments. We drew the boundaries of those ROIs for each subject on their native surface for better visualization and interpretation of the results. All brain visualizations were produced using Pycortex software [Gao et al., 2015a].

### 3.3.3 Friends Dataset

Friends Dataset is provided by the Courtois NeuroMod group (data release cneuromod-2022)[Boyle et al., 2021]. This dataset contains functional data acquired while showing 6 participants episodes of the Friends TV show in English. It includes seasons 1-6 for all subjects, except sub-04 who only completed seasons 1-4 (and a few segments of season 5). Each episode is cut in two segments (a/b) to allow more flexible scanning and give participants opportunities for breaks. There is a small overlap between the segments to allow participants to catch up with the storyline. The

fMRI sampling rate (TR) was 1.49s. The data were prepossessed using fMRIPrep 20.1.0 (ref. 59). These data are available on request at https://docs.cneuromod.ca/en/latest/ACCESS.html.

# Chapter 4

# Neural Taskonomy: inferring the similarity of task-derived representations from brain activity

## 4.1 Introduction

Scene understanding requires the integration of space perception, visual object recognition, and the extraction of semantic meaning. The human brain's solution to this challenge has been elucidated in recent years by the identification of scene-selective brain areas via comparisons between images of places and common objects [Epstein and Kanwisher, 1998a]. This basic contrast has been extended across a wide variety of image manipulations that have provided evidence for the neural coding of scene-relevant properties such as relative openness [Kravitz et al., 2011, Park et al., 2014, Harel et al., 2012], the distance of scenes to the viewer [Kravitz et al., 2011, Park et al., 2014, Lescroart et al., 2015], 3D spatial layout [Ferrara and Park, 2016, Kamps et al., 2016, Kornblith et al., 2013] and navigational affordances [Bonner and Epstein, 2017]. Recently, to help explain such findings, Lescroart and Gallant [2019] developed an encoding model using a feature space that parametrizes 3D scene structures along the distance and orientation dimensions

and provides a computational framework to account for human scene processing. Intriguingly, Lescroart and Gallant [2019] were able to identify distance and openness within scenes as the dimensions that best account for neural responses in scene-selective brain areas. At a higher, semantic, level, Stansbury et al. [2013] found that neural responses in scene-selective brain areas can be predicted using scene categories that were learned from object co-occurrence statistics. Such findings demonstrate that human scene-selective areas represent both visual and semantic scene features. At the same time, there is still no robust model of how these different kinds of information are integrated both within and across brain regions.

Encoding models are widely used in understanding feedforward information processing in human perception, including scene perception. Encoding models are predictive models of brain activity that are able to generalize and predict brain responses to novel stimuli [Naselaris et al., 2011]. Researchers have also used encoding models to infer which dimensions are critical for prediction by comparing the weights learned by the model [Huth et al., 2016, Lescroart and Gallant, 2019]. One of the successes of encoding models lies in predicting low- to mid-level visual cortex responses in humans and primates using features that were learned via a convolutional neural network trained on object recognition [Agrawal et al., 2014, Güçlü and van Gerven, 2015, Yamins et al., 2014b, Eickenberg et al., 2017]. Most interestingly, these studies demonstrate a correspondence between human neural representation and learned representations within CNN models along the perceptual hierarchy: early layers tend to predict early visual processing regions, whereas later layers tend to predict later visual processing regions. Similarly, researchers have found that network representations from other task-driven networks, including networks trained on speech or music related tasks, are able to explain neural responses in human auditory pathways [Kell et al., 2018]. Such successes are not mere coincidences but rather indications of how fundamental task-driven representations are to both task training and to information processing in the brain.

Despite these advances, CNN features themselves are notoriously difficult to interpret. First,

activations from the convolutional layers lie in extremely high-dimensional spaces and it is difficult to interpret what each feature dimension signifies. Second, features from a CNN tailored for a particular visual task can represent any image information that is relevant to that task. As a consequence of these two issues, the feature representations learned by the network are not necessarily informative with respect to the nature of visual processing in the brain despite their good performance in predicting brain activity.

To better understand the specificity of the information represented in the human visual processing pathways, we adopted a different approach. Instead of choosing a generic object-classification CNN as a source of visual features, we built encoding models with individual feature spaces obtained from different task-specific networks. These tasks included mid-level features such as surface normal estimation, edge detection, scene classification, etc. In any task-driven network, the feature space learned to accomplish the task at hand should only represent information from input images that is task-relevant. Therefore we can use the predictive regions from each of the models to identify the brain regions where specific task-relevant information is localized. Independently, Dwivedi and Roig [2018] have shown that representation similarity analysis (RSA) performed between task representations and brain representations can differentiate scene-selective regions of interest (ROIs) by their preferred task. For example, representations in scene-selective occipital place area (OPA) are more highly correlated with representations from a network trained to predict navigational affordances. However, this study was limited to pre-defined regions of interest, while the task representations we identify span the entire brain. Consequently, the brain regions predicted by each model provide an atlas of neural representation of visual tasks and allow us to further study the representational relationships among tasks.

Independently of the brain, visual tasks have relationships among them. Task representations that are learned specifically for one task can be transferred to other tasks. Computer vision researchers commonly use transfer learning between tasks to save supervision and computational resources. In this vein, Zamir et al. [2018] recently showed that by standardizing model struc-

ture and measuring performance in transfer learning, one can generate a taxonomic map for task transfer learning ("Taskonomy"). This map provides an account of how much information is shared across different vision tasks. Given this global task structure, we can infer clusters of information defined by segregation of tasks, and then ask: does the brain represent visual information in the same task-relevant manner?

We compared the relationships between tasks using both brain representations and task learning. These comparisons reveal clustering of 2D tasks, 3D tasks, and semantic tasks. Compared to general encoding models, building individual encoding models and exploiting existing relationship among models has the potential to provide more in-depth understanding of the neural representation of visual information.

## 4.2 Methods

### 4.2.1 Encoding Model

To explore how and where visual features are represented in human scene processing, we extracted different features spaces describing each of the stimulus images and used them in an encoding model to predict brain responses. Our reasoning is as follows. If a feature is a good predictor of a specific brain region, information about that feature is likely encoded in that region. In this study, we first parameterized each image in the training set into values along different feature dimensions in a feature space. For example, if the feature space of interest is an intermediate layer in a task-driven network, we simply fed the image into the network and extracted its layer activation. These values are used as regressors in a ridge regression model (implemented in PyTorch; see [Koushik, 2017]) to predict brain responses to that image. Performance from the validation data is used to choose the regularization parameter in the ridge regression model. We chose to use a ridge regression model instead of more complicated models in order to retain the interpretability of model weights, which may provide insights into the underlying dimensions

of the brain responses. For each subject, each voxel's regularization parameter was chosen independently via 7-fold cross-validation based on the prediction performance of the validation data. Model performance was evaluated on the test data using both Pearson's correlation and coefficient of determination ($R^2$). To determine the significance of the predictions, we ran permutation tests where we shuffled responses 5000 times, computed the correlation scores, and obtained FDR corrected *p*-values for both ROI and whole brain results.

### 4.2.2 Feature Spaces

To simultaneously test representations from multiple 2D, and 3D vision tasks, we used the latent space features from each of the 21 tasks in Taskonomy [Zamir et al., 2018] model bank: autoencoding, colorization, curvature estimation, denoising, depth estimation, edge detection (2D), edge detection (3D) or occlusion edges detection, keypoint detection (2D), keypoint detection (3D), depth, reshading, room layout estimation, segmentation (2D), segmentation (2.5D), surface normal estimation, vanishing point estimation, semantic segmentation, jigsaw puzzle, inpainting, object classification and scene classification. In the Taskonomy training scheme, an intermediate latent space with fixed dimension ($16 \times 16 \times 8$) was enforced for each of these networks. We obtained these latent space activations by feeding our images into each pre-trained task-specific network in the task bank provided with the Taskonomy paper. Four of the 25 tasks were excluded from this analysis because these tasks take multiple images as input, while the brain responses we have are only to single images. Examples of these excluded tasks include camera pose estimation and egomotion estimation. We then built individual ridge regression models with the extracted latent features to predict brain responses and measured the correlation between the prediction and the true response in the held-out dataset.

### 4.2.3 Neural Data

The images used in this paper are from a publicly available large-scale fMRI dataset, BOLD5000 [Chang et al., 2019]. In the BOLD5000 study, participants' brains were scanned while they fixated at real-world images and judged how much they liked the image using a button press. Images in the BOLD5000 dataset were chosen from standard computer vision datasets (ImageNet [Russakovsky et al., 2015], COCO [Lin et al., 2014a] and SUN [Xiao et al., 2010]). The experiment was run in a slow-event setting where trials are separated by 10 seconds. From BOLD5000, we used data from three participants viewing 4916 unique images. These 4916 image trials are separated into random training, validation, and testing sets during model fitting. Average of TR 3 and 4 of each slow-event trial is used for model fitting and testing. Region of interest (ROI) boundaries that identify category-selective brain regions in the whole-brain map presented in our results were generated directly from the ROI masks provided with the BOLD5000 dataset.

### 4.2.4 Task Similarity Computation

For each task, we took prediction performance scores across all voxels ($n \approx 55,000$). We set the score of a voxel to zero if the *p*-value of the correlation obtained from permutation test is above significance threshold ($p > 0.05$, FDR corrected). This gave us a performance matrix of meaningful correlations of size $m \times v$, where $m$ is the number of tasks of interest and $v$ is the number of voxels. To analyze the relationship between tasks based on neural representations, we computed pairwise similarity across tasks in the performance matrix using cosine similarity. These pairwise similarities were then used to construct graphs and similarity trees among tasks. Other distance or similarity functions such as euclidean distance did not show substantial differences.

## 4.3 Results

### 4.3.1 Model Prediction on ROIs

In Figure 4.1 we show the prediction accuracy measured using the Pearson correlation coefficient. This was done for the 21 task-related feature spaces that were used to predict brain responses in predefined ROIs. Each bar shown in the figure represents the average correlation score across all voxels in that ROI. Overall, the predictions using these feature spaces—which come from mid-level computer vision tasks—show significant correlations with brain responses, except for the feature space from the curvature task. Among scene-selective regions, such as parahippocampal place area (PPA), retrosplenial complex (RSC), occipital place area (OPA), and lateral occipital complex (LOC), models with 3D features (e.g. keypoints, edges) show far better predictions than models with 2D features. This finding is consistent with the results of Lescroart and Gallant [2019]. In contrast, within early visual areas, the prediction results between 2D and 3D features are not differentiable. Across all ROIs, features from object and scene classification tasks provide the best predictions. For more scene specific tasks or semantic tasks such as 3D keypoints/edges, 2.5D/semantic segmentation, depth, distance, reshading, surface normal, room layout, vanishing points estimation, and object/scene classification, scene-selective regions are better predicted as compared to early visual areas. These patterns are consistent across all three participants. To quantify the consistency of results across subjects, we computed correlations of prediction accuracy for each pair of subjects: 0.7957 (S1 vs. S2), 0.9034 (S1 vs. S3) and 0.9345 (S2 vs. S3). These results provide evidence that scene-selective areas show selectivity for scene-specific task representations.

### 4.3.2 Model Prediction Across the Whole Brain

Prediction performance in pre-defined ROIs may omit relevant information arising in other brain regions. In Figure 4.2 and 4.3 we show prediction performance across the entire brain in a flat-

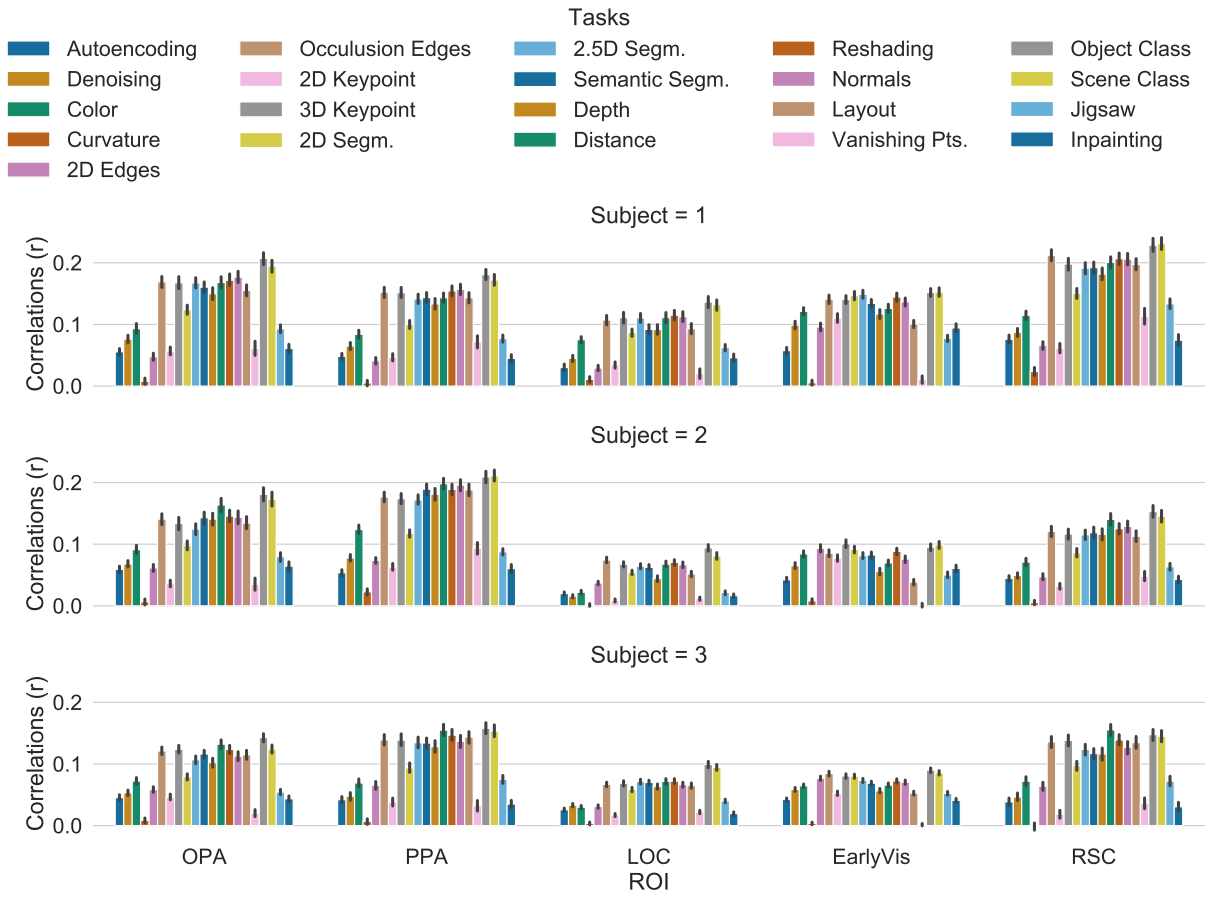



Figure 4.1: **Pearson correlation coefficient between predicted and true responses across tasks.** Each sub-figure corresponds to a particular participant. Colors in the legend are arranged by columns. Features from 3D tasks, compared to those from 2D tasks, predict better in OPA, PPA, RSC, and LOC.

tened view (generated using Pycortex [Gao et al., 2015b]). Figure 4.2 shows the raw prediction performance as correlation coefficients for each task feature space. Figure 4.3 shows a contrast in prediction between 3D and 2D keypoints as well as edges. In this figure, red-colored voxels are better predicted by 3D features than 2D features, and vice-versa for blue-colored voxels; white-colored voxels are well predicted by both features. We find that 3D features make better predictions for scene-selective regions—those delimited by ROI borders, while 3D and 2D features seem to predict early visual areas equally well. Figure 4.2 shows that prediction results are consistent across three participants despite anatomical differences in their brain structures

and 4.3 shows that the results are consistent across tasks.

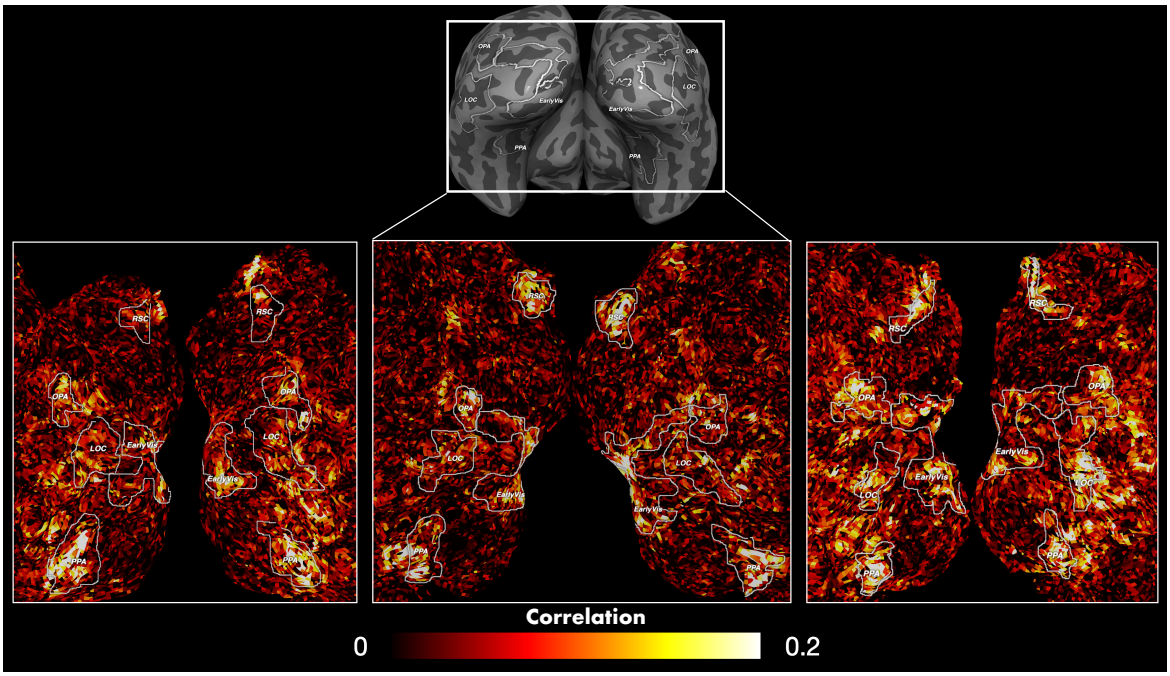



Figure 4.2: **Whole brain prediction correlation using task representation of scene classification network.** The flat maps are cropped from the occipital regions of the brain. The upper zoom-out view shows the relative locations of the flat maps. Lower colored figures are the prediction performance across 3 participants. Prediction results are consistent across subjects.

Model performance using feature spaces from other tasks are shown in Figure 4.4. Here we plot 6 of the 21 tasks, and the remaining figures for this sample subject (subject 1) are provided in the appendix. Voxels with insignificant predictions ($p \geq 0.05$, FDR corrected) are masked in these figures. Prediction performances of all tasks and all three subjects can be viewed at `https://cs.cmu.edu/~neural-taskonomy`.

To provide a better estimate of the variance ceiling, we ran ridge regression to predict responses of one subject from another. In Figure 4.5 we show the prediction correlation for each subject from the remaining two subjects. The average correlation between predictions and true responses across voxels for each subject are: 0.0931, 0.0932, 0.112, as shown by the black lines on each plot. The histogram includes low signal to noise ratio (SNR) voxels that are not engaged by the task. The histograms distribution indicates that the accuracy we obtained on significant

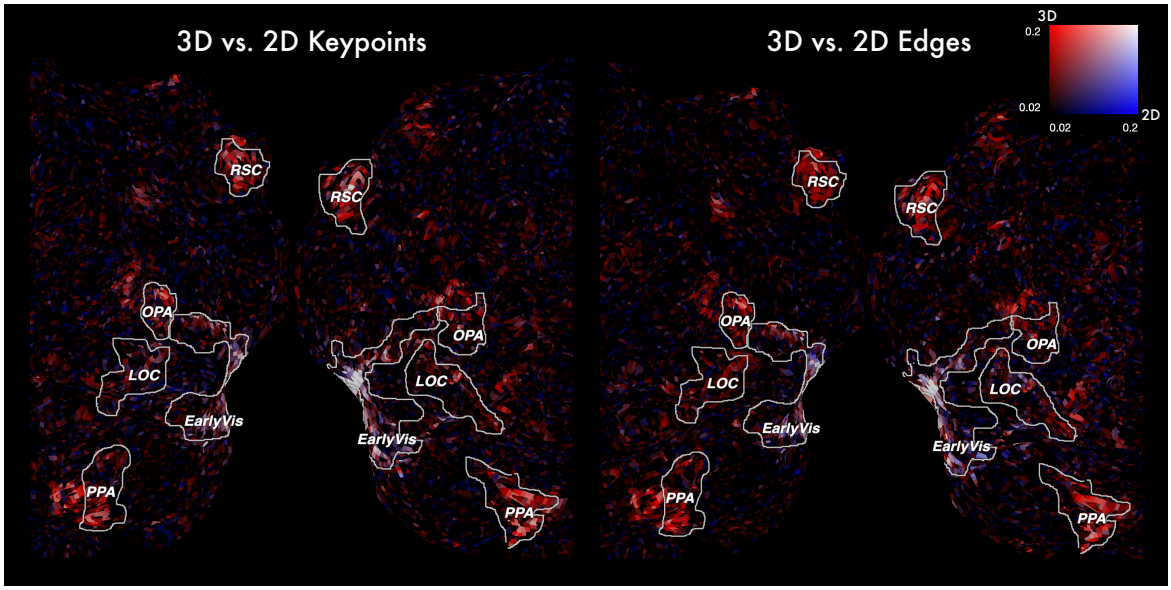
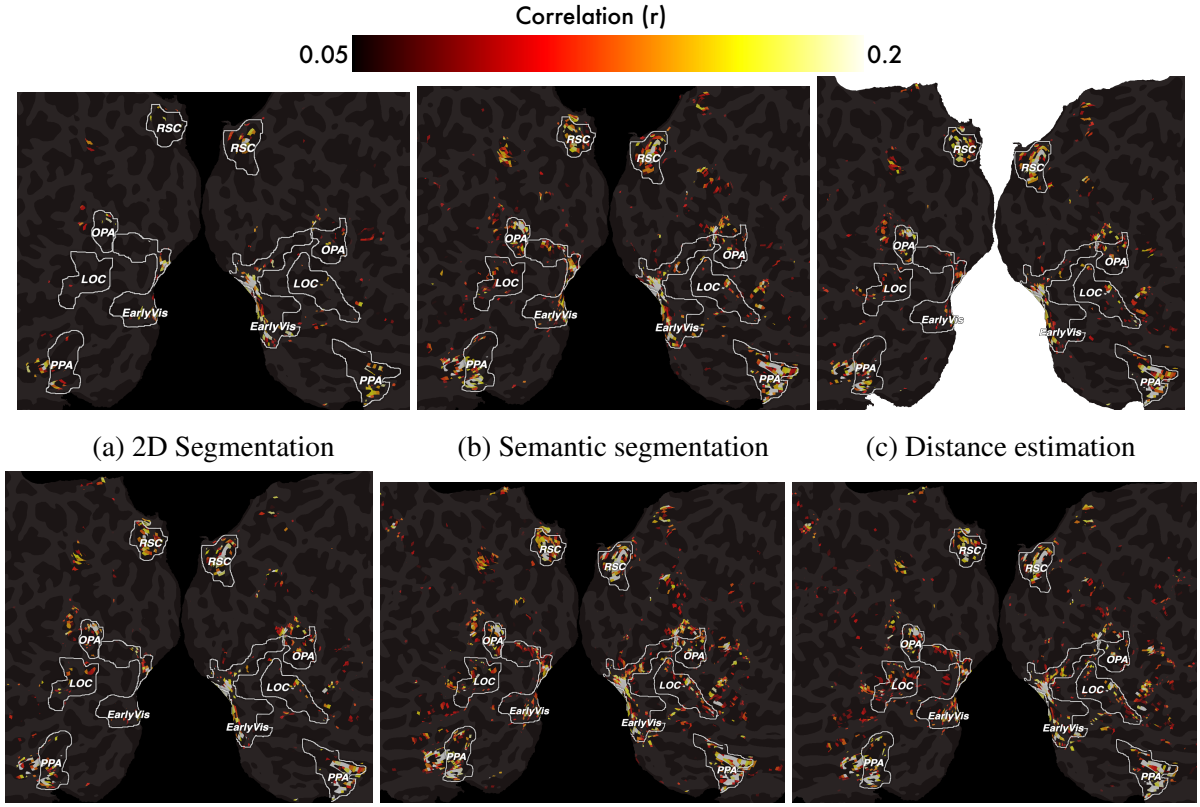



Figure 4.3: **Contrast of prediction performance (measured with Pearson correlation coefficients) between 2D and 3D features in one sample subject (subject 1).** The flat maps are cropped similar to as in Figure 4.2. The color map indicates the difference in correlation coefficients: red: 3D $>$ 2D; blue: 2D $>$ 3D. 3D task features predict better in scene selective regions and in more anterior parts of the brain.

voxels across the whole brain using features from various task is close to the ceiling. Cross subject prediction results from each pair of subjects is provided in the appendix. Note that we are predicting single-trial fMRI data with no repetitions, which leads to a lower signal to noise ratio (and therefore lower variance ceiling) than other fMRI studies that average repetitions.

### 4.3.3 Evaluation of Neural Representation Similarity

To this point we have shown that the neural prediction maps across tasks differ from one another; at the same time, there are many overlapping voxels across the predicted regions. Importantly, this pattern of voxels as predicted by the tasks can be exploited and used to infer task relationships in the brain. We computed task similarity averaged across 3 subjects using the methods discussed in 4.2.4 (Figure 4.6). The individual patterns of task similarity are almost identical across 3 subjects. More specifically, correlations (Pearson's r) between similarity matrices for each pair




(a) 2D Segmentation (b) Semantic segmentation (c) Distance estimation

(d) Reshading (e) Object Classification (f) Scene Classification

Figure 4.4: **Predictive voxels using tasks features from Taskonomy [Zamir et al., 2018] in one sample subject (subject 1).** Predictive regions of different tasks differ from each other across tasks.

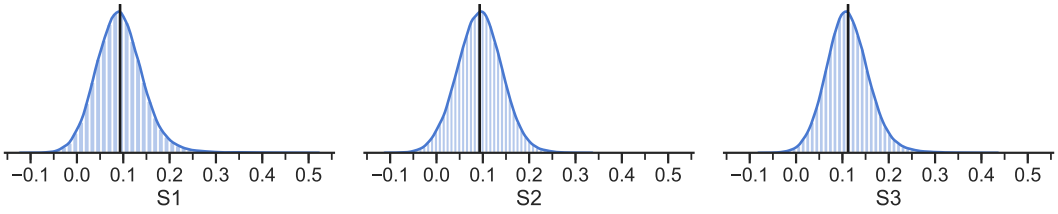



Figure 4.5: **Noise ceiling derived from cross subjection prediction.** Each subfigure is a histogram of correlation scores across voxels. The black lines on each subfigure indicate the average correlation values.

of subjects are: 0.9610 (S1 vs. S2), 0.9477 (S1 vs. S3) and 0.9407 (S2 vs. S3). In this comparison across the whole brain, tasks such as 2.5D segmentation, room layout estimation, surface normal estimation, scene classification etc. have similar predictions patterns.

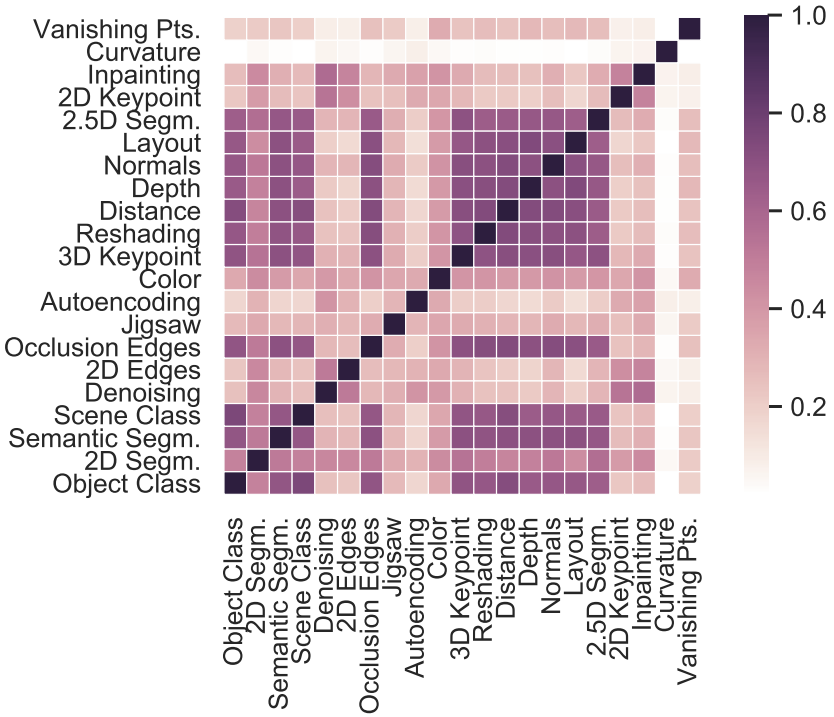
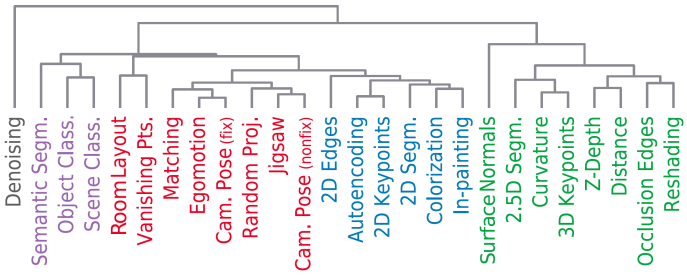


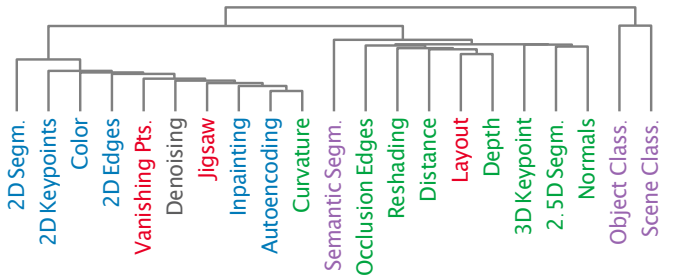

Figure 4.6: **Prediction similarity matrix across 21 tasks, averaged across 3 subjects.** A large similarity value between task X and Y indicates encoding models with features representation from task X and Y have similar predictions of brain responses.

### 4.3.4 Task Similarity Tree

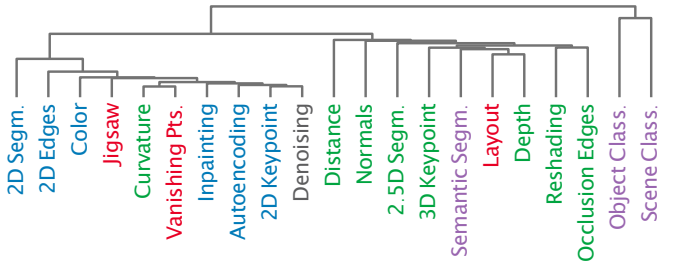
To further explore the relationship between tasks as represented in the brain, we ran hierarchical clustering on the prediction correlation results and visualized the clustering results as dendrograms. Figure 4.7 compares the task similarity tree based on transferring-out patterns in the original Taskonomy paper [Zamir et al., 2018], with the task similarity tree generated based on similarity in voxel prediction performance. Trees independently generated for each subject show great similarity. In the Taskonomy result, tasks are clustered into 3D (indicated in green), 2D (blue), low-dimensional geometric (red) and semantic (purple) tasks. Interestingly, the tree derived from brain representation also shows a similar structure: semantic, 2D and 3D tasks are clustered together. The differences between the two similarity trees may be due to low abso-



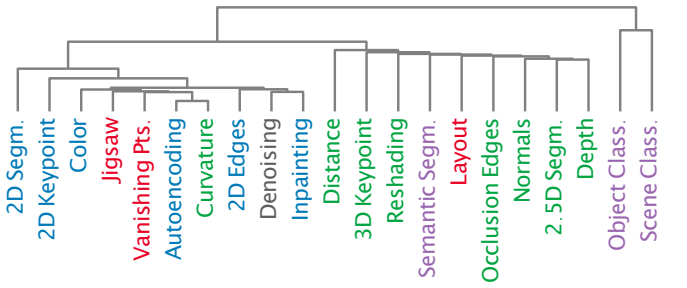

(a) Taskonomy from transfer learning




(b) Neural Taskonomy (subject 1)




(c) Neural Taskonomy (subject 2)




(d) Neural Taskonomy (subject 3)

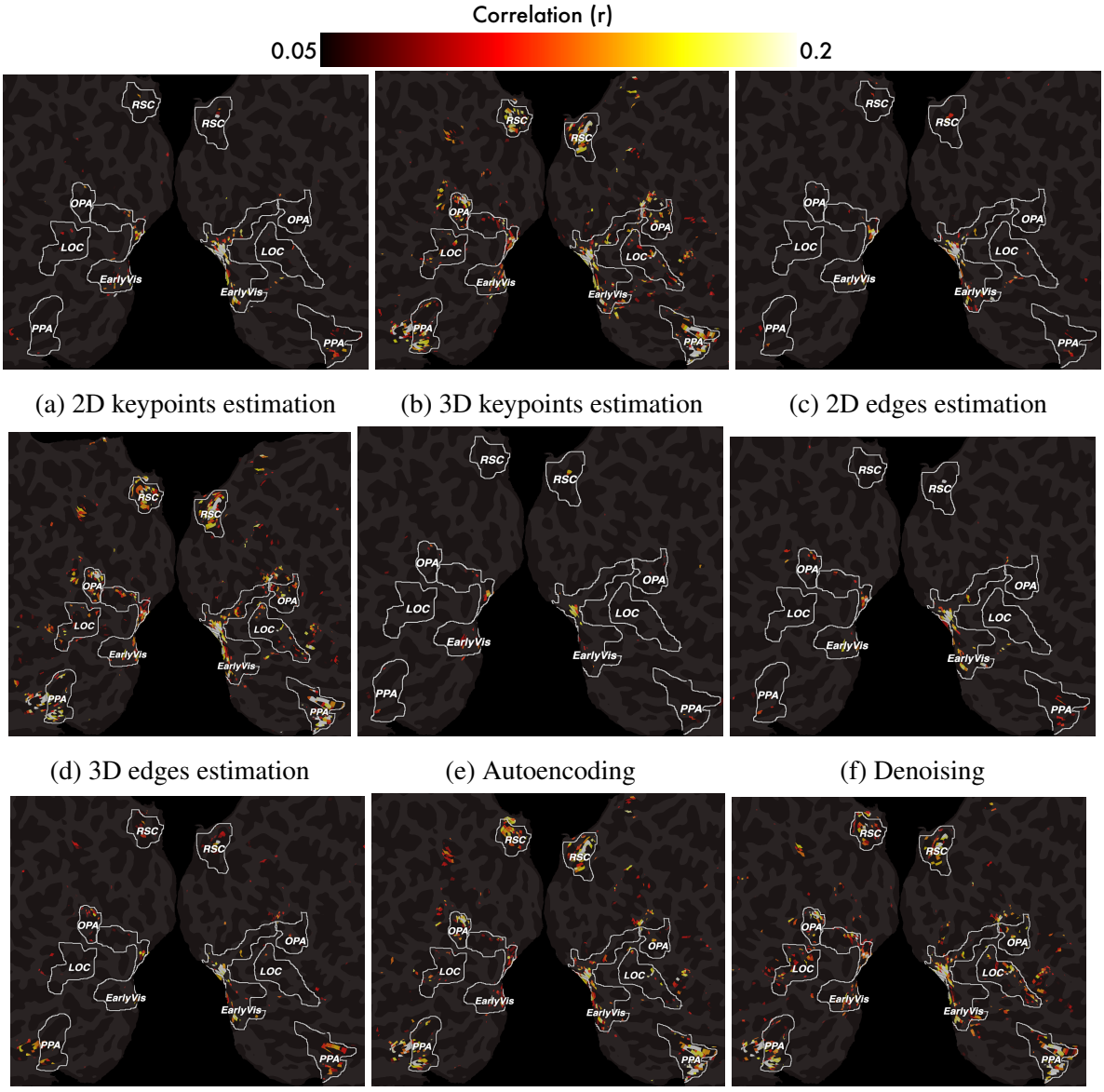
Figure 4.7: **Task Trees from (a) Taskonomy [Zamir et al., 2018] and (b-d) brain representation of tasks from 3 differnet subjects.** Tasks in (b-d) are colored according to colors in (a). Similar clusters of 2D (in blue), 3D (green) and semantic (purple) tasks are found among neural taskonomy trees. Clustering results are highly consistent across three subjects.

lute performance of the encoding model. For example, the model with features from curvature estimation task has less than 10 significant voxels in some subjects which may lead to bias in the representation of the task tree. Overall the similarity between two task trees shows that, at a coarse level, neural representation of task information is similar to that found through transfer learning. The clustering and dendrogram structures are stable across subjects and across different linkage criteria. Aside from using "average" linkage for clustering, as shown here, we also used "ward" linkage criterion (shown in the appendix) and obtained similar structures.

## 4.4 Discussion

The architecture of the primate visual system reflects a series of computational mechanisms that enable high performance for accomplishing evolutionarily adaptive tasks [Yamins and DiCarlo, 2016b]. However, the precise nature of these tasks remains unknown because of the limitations of neuroscience data collection methods and the lack of interpretability of intermediate visual representations. To address these issues we leveraged the space of vision tasks learned through transfer learning in Taskonomy [Zamir et al., 2018] and the recent availability of a larger-scale human functional imaging dataset, BOLD5000 [Chang et al., 2019]. One challenge we faced was the substantial difference between the image distributions of BOLD5000 (which contains general objects and scenes) and the Taskonomy dataset (which includes indoor scenes exclusively). As such, when we applied the pre-trained Taskonomy models to BOLD5000 images, we found that these models didn't perform as well as on the Taskonomy dataset, especially for the outdoor images used in BOLD5000. Such inconsistency in image distribution is unavoidably reflected in the encoding model performance and hinders us from making more specific claims about task spaces in the brain. One solution to this issue would be to use a more general computational model of visual tasks, as well as a larger brain dataset based on more images, both of which are outside of the scope of this paper.

In the future we would also like to investigate the unique and shared variance explained




(a) 2D keypoints estimation (b) 3D keypoints estimation (c) 2D edges estimation

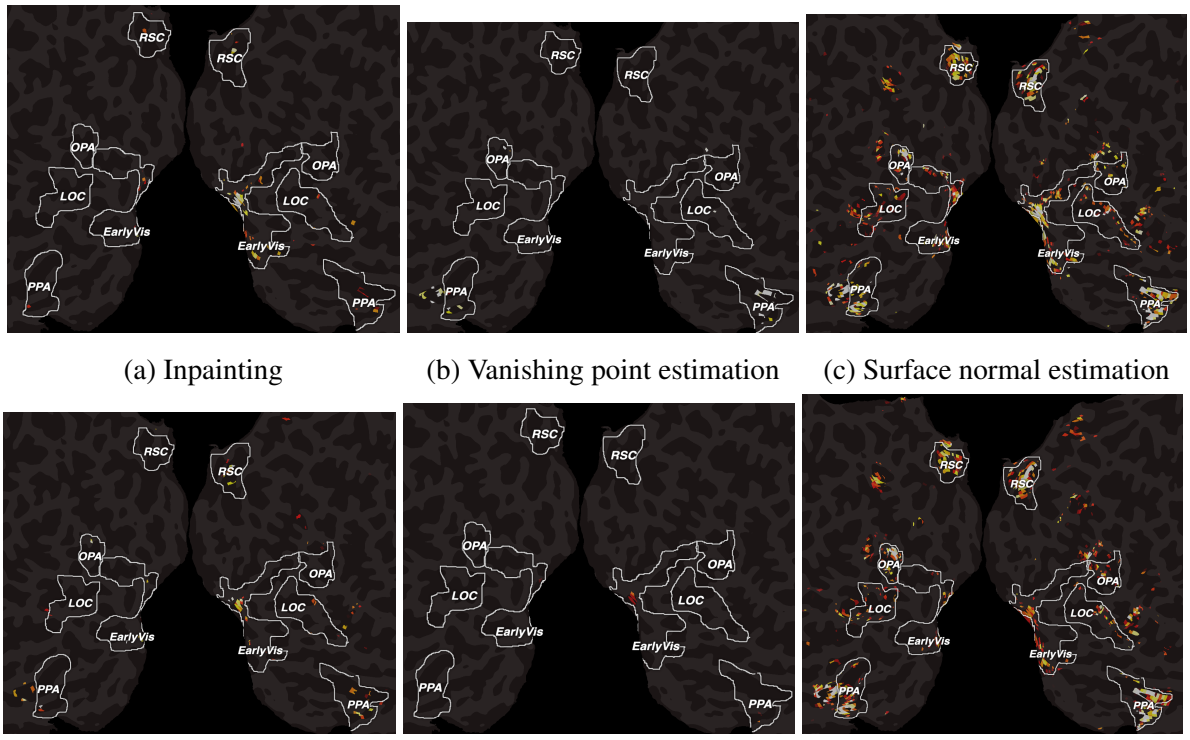
(d) 3D edges estimation (e) Autoencoding (f) Denoising

(g) Colorization (h) Depth estimation (i) 2.5D segmentation

Figure 4.8: Model predictions of tasks from Taskonomy (Part 1). Voxels below significance threshold ($p \geq 0.05$, FDR corrected) are masked.

by each task. At present we are still unclear as to what transferability between tasks within Taskonomy predicts for similarity in task representations within the brain.

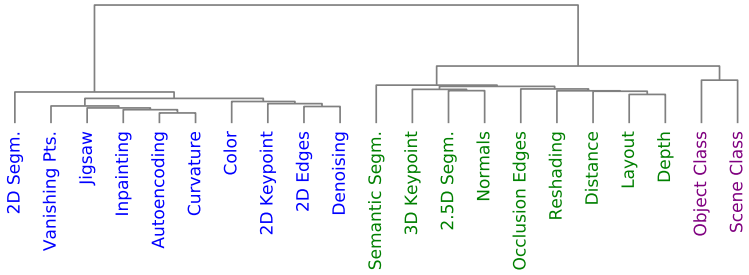
Finally, although our whole brain prediction maps do seem to suggest the involvement of additional functional brain areas beyond the pre-defined ROIs, we strongly feel that making claims about new functionally-defined brain areas would be premature given our current data



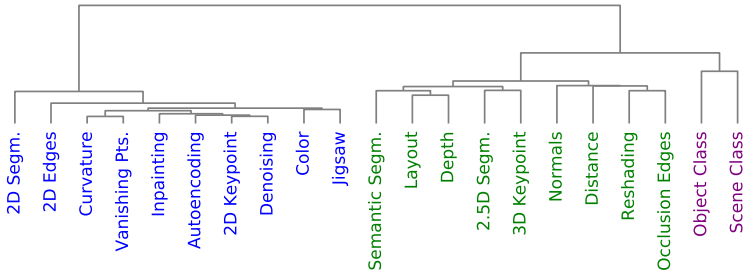

(a) Inpainting (b) Vanishing point estimation (c) Surface normal estimation

(d) Jigsaw puzzle (e) Curvature estimation (f) Room layout estimation

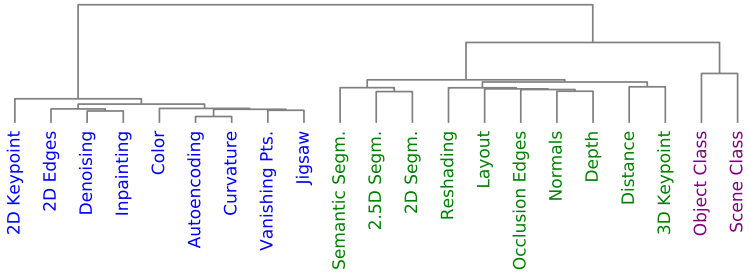
Figure 4.9: Model predictions of tasks from Taskonomy (Part 2). Voxels below significance threshold ($p \geq 0.05$, FDR corrected) are masked.




(a) Neural Taskonomy (Subject 1)




(b) Neural Taskonomy (Subject 2)




(c) Neural Taskonomy (Subject 3)

and analysis. We believe that to make robust claims about new "functional territories", we would first need to run additional validation experiments in which specific manipulations are used to establish that specific brain regions are sensitive to the tasks in question.

## 4.5 Conclusion

Our results reveal that task-specific representations in neural networks are useful in predicting brain responses and localizing task-related information in the brain. One of the main findings is that features from 3D tasks, compared to those from 2D tasks, predict a distinct part of visual cortex. In the future we will incorporate features from other tasks to obtain a more comprehensive picture of task representation in the brain.

For years neuroscientists have focused on recovering which parts of the brain represent a given type of information. However, what are the computational principles behind the encoding of information in the brain? We observe feedforward hierarchies in the visual pathways, but what are the stages of information processing? To date, we have few satisfying answers. The ultimate goal in studying task representation in the brain is to answer some of these questions. We exploited the task relationship found in transfer learning and used it as a ground truth of visual information space to study the neural representation of visual and semantic information. In sum, our paper provides an initial attempt in using task relationships to answer broader questions of neural information processing.

# Chapter 5

# Learning Intermediate Features of Object Affordances with a Convolutional Neural Network

While interacting with our environment, we naturally infer the functional properties of the objects around us. These properties, typically referred to as affordances, are defined by [Gibson, 1979], as all of the actions that an object in the environment offers to an observer. For example, "kick" for a ball and "drink" for water. Understanding affordances is critical for understanding how humans are able to interact with objects in the world.

In recent years, convolutional neural networks have been successful in preforming object recognition in large-scale image datasets [Krizhevsky et al., 2012]. At the same time, convolutional networks trained to recognize objects have been used as feature extractors and can successfully model neural responses as measured by fMRI in human visual cortex [Agrawal et al., 2014] or by electrodes in monkey IT cortex [Yamins and DiCarlo, 2016b]. To understand the relevant visual features in an object that are indicative of affordances, we trained a CNN to recognize affordable actions of objects in images.

## 5.1 Dataset Collection

Training deep CNNs is known to require large amounts of data. Available affordance datasets with images and semantic labels are largely limited at this moment. The only relevant dataset currently available to the public was created by [Chao et al., 2015], and only includes affordance labels for 20 objects from the PASCAL dataset and 90 objects from the COCO dataset. Here we built a large scale affordance dataset with affordances labels attached to all images in the ImageNet dataset [Deng et al., 2009a]. This dataset forms a more general representation of the affordance space and allows large scale end-to-end training from the image space and to this affordance space. The dataset collection process is shown in Figure 5.1. Human labelers were presented with object labels from ImageNet object categories and answered the question "What can you do with that object?". All answers were then co-registered with WordNet [Miller, 1995] action labels so that our labels could be extended to other datasets. The top five responses from labelers were used as canonical affordance labels for each object. 334 categories of actions were labeled for around 500 objects categories. When combined with image to object label mappings from ImageNet, these affordance labels provided us with the image to affordance label mappings that were used to train our CNN.

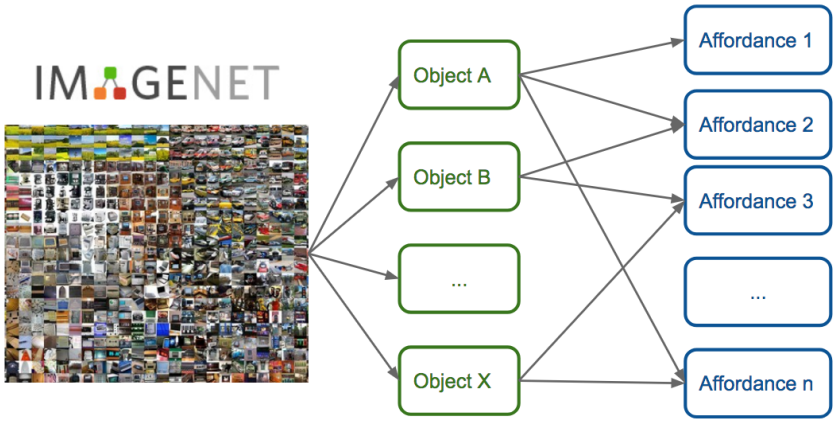



Figure 5.1: Dataset Collection. The labelers are given object labels, indicated in the green boxes here and assign to them affordances labels, indicated with blue boxes.

## 5.2 Visualization of Affordance Space

In our affordance dataset, each object was represented by a binary vector indicating whether each of the possible actions was available for this object or not. Each object can then be represented as a point in the affordance space. We used PCA to project these affordance vectors into a 3D space and plotted the object classes as illustrated in Figure 5.2. In the 3D space created for visualization, the objects appear to be well separated. More specifically, the majority of living things were organized along the top axis; the majority of small household items were organized along the left axis; and transportation tools and machines were organized along the right axis. Human-related categories such as dancer and queen do not belong to any axis and appear as flowing points in the space.

Figure 5.2: ImageNet images in the affordance space.

## 5.3 Results

### 5.3.1 Network Training

A CNN was trained to predict affordance categories from images. A total of 55 affordances were selected as potential actions after ensuring that each affordance label had at least 8 object categories associated with it (by removing affordances that were associated with too few object categories). Each object category was placed in the training, validation or testing sets. These sets were exclusive, such that, if one object category appeared in one set, it would not appear in the other two sets. Such separation ensures that the learning of affordances was not based on recognizing the same objects and learning linear mappings between objects and affordances.

We used the ResNet18 model [He et al., 2016] (other models such as VGG produced similar results), and trained it using the Adam optimizer [Kingma and Ba, 2014] by minimizing binary cross-entropy loss. Approximately 630,000 images from ImageNet were used in training, and approximately 71,000 images each were used for validation and testing. The trained CNN was evaluated by computing the average percentage of correctly predicted affordance labels, and the results are reported in Table 5.1. The trained networks showed significantly better performance compared to the baseline.

Table 5.1: Training Results. "Fine-tuning" indicates that the network was pre-trained to predict image categories, while "Training from Scratch" indicates that the network was initialized with random weights. Baseline accuracy was calculated by estimating the most frequent categories.

| | Baseline | Fine-tuning | Training from scratch | Fine-tuning + oversampling | Training from scratch + oversampling |
|---|---|---|---|---|---|
| Training Acc (%) | 7.61 | 80.39 | 71.42 | 87.60 | 85.05 |
| Testing Acc (%) | 6.86 | 44.62 | 37.47 | **55.42** | **53.43** |

### 5.3.2 Skewed Distribution and Oversampling

Since actions such as “hold” and “grab” would be used on objects much more often than actions such as “thrust”, we obtained an uneven distribution of affordance labels across image categories, as shown in Figure 5.3. In computer vision, oversampling is a commonly used solution for this problem. However, because of the multi-label nature of the affordance recognition problem, proper oversampling is challenging. Less frequently appearing classes need to be oversampled without over representing the more frequently appearing classes. We used Multi-label Best First Over-sampling (ML-BFO) [Ai et al., 2015], and re-trained the CNN with the resampled data. This produced a considerable increase in prediction performance, as seen in Table 5.1.

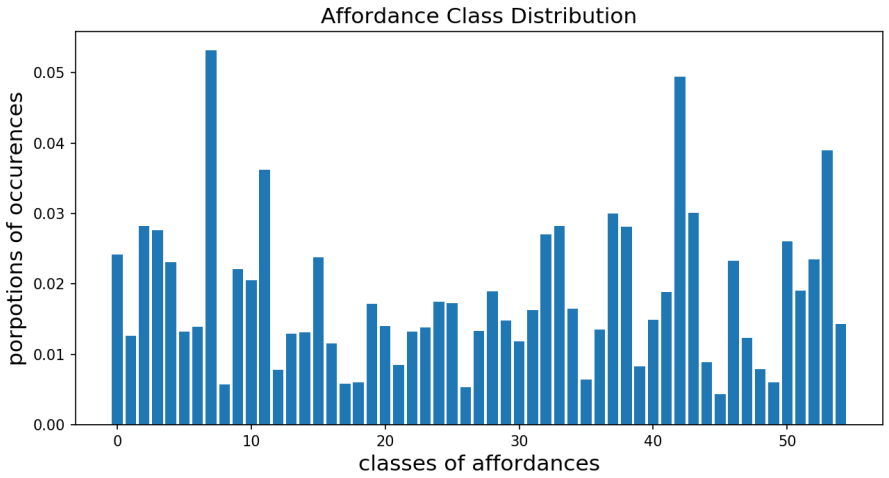



Figure 5.3: Percentages of objects classes assigned to each affordances categories.

### 5.3.3 Sample Predictions

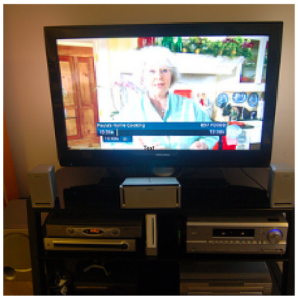
Figures 5.4(a)–(d) demonstrate images where the network was able to predict correctly. However, the presence of distinct features can mislead the network. For example, in Figure 5.4(e), where white bars stand out in the image, the network predicted “grab” and “drive”, potentially mistaking the image as a bar or a road. On the other hand, human labelers, knowing that it is a image of a wall, provided labels such as “walk” and “enter”. Since ImageNet contains natural scene images,

multiple objects are likely to appear in one image, even though each image is assigned only one object label. Such images confuse both the labelers and the network, and therefore can lead to incorrect affordance recognition as shown in Figures 5.4(f) and (g).

(a) care/feed

(b) hang/wear/grab

(c) eat/taste



(d) switch on/off

(e) P: grab/drive
GT: walk/exit

(f) P: empty/fill
GT: taste

(g) P: hunt/chase
GT: talk/meet

Figure 5.4: Sample predictions. (a)-(d): Examples of images with correct affordance predictions (correct label below each image). (e)-(g): Examples of images with incorrect affordance predictions (P: indicates the CNN prediction, while GT: indicates the ground truth based on human labeling.

## 5.4 Visualizing the Learned Representation Space

### 5.4.1 RDM across Layers

To visualize the representations learned by the network, we randomly sampled 10 images from each of 30 objects classes, and extracted activations from the network layers. Pairwise correlation distance between network activation across layers was computed for each pair of images, and is

shown in Figure 5.5. Pairwise distance between affordance labels is shown in the bottom-right matrix. This matrix denotes the ground truth distance in affordance space. Similar patterns begin to emerge in Layer 4 for both the fine-tuned network and the network trained from scratch. Critically, this pattern is not seen for the off-the-shelf network that was not trained on affordances. This demonstrates that our network learns representations that effectively separate different affordance categories.

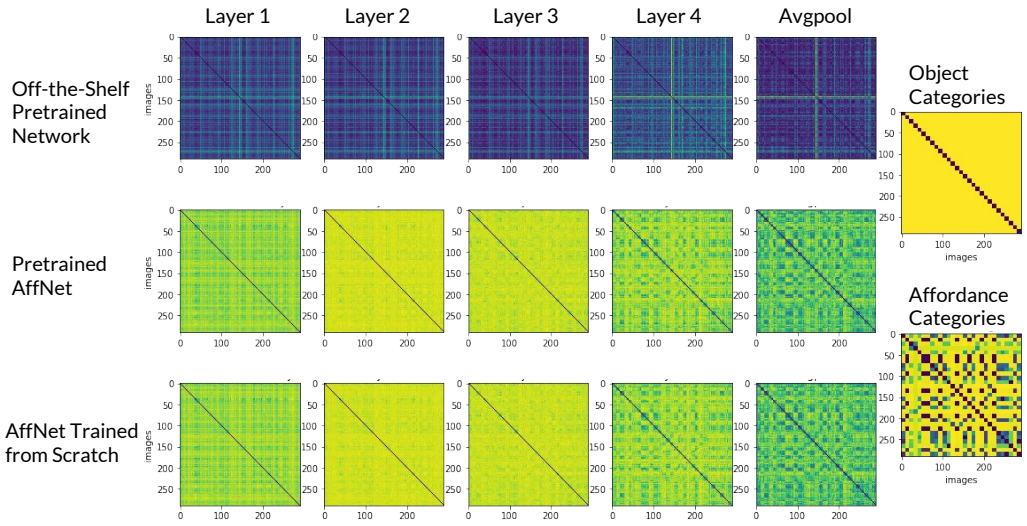



Figure 5.5: RDM matrix of layers from CNN from off-the-shelf pre-trained network, fine-tuned network and network trained from scratch for affordances.

### 5.4.2 t-SNE

Activations from the second to last layer in the network trained from scratch were visualized using t-SNE [Maaten and Hinton, 2008], as shown in Figure 5.6. Images are coarsely split into four groups based on their distinct affordances: living things, vehicles, physical spaces and small items. In the 2D t-SNE visualization, the representation of living things (in green), vehicles (in red) and physical spaces (in blue) are visibly separable. Small items (in yellow), in contrast, span the entire space. The category of small items does not appear well separated, which is likely due

to the visualization being limited to 2 dimensions.

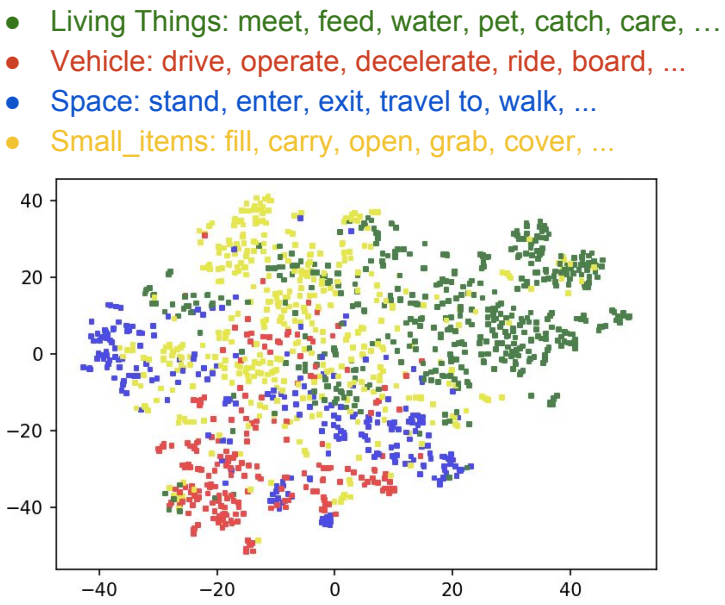
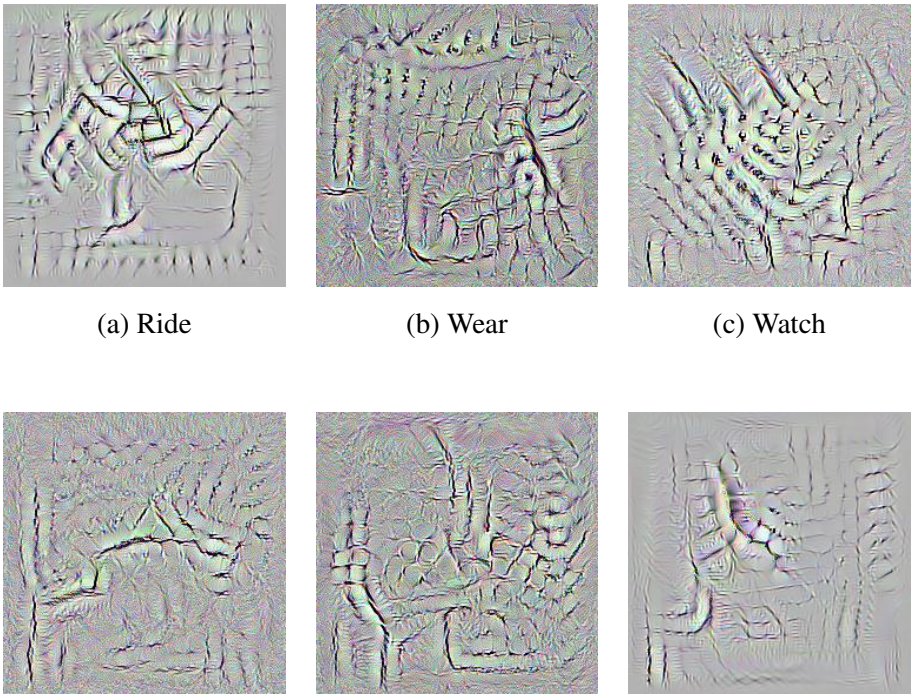



Figure 5.6: t-SNE visualization of the second to last layer in the CNN trained from scratch. Representations of images are coarsely split into four groups based on the distinct affordances of the images: living things (green), vehicles (red), physical spaces (blue) and small items (yellow).

### 5.4.3 Unit Visualization

We were able to visualize the output layer units of the CNN by optimizing in pixel space to determine which images maximally activated a specific unit. Figure 5.7 shows such visualization of 6 units from the output layer. The "ride" unit, for example, shows human- and horse-like structures; the "wear" unit shows a coarse clothing pattern and details of common textures often associated with clothing. Similarly, units "climb", "sit", and "fill" show stairs-like, chair-like, and container-like structures respectively. Interestingly, the "watch" unit shows preference for dense textures in the center of the image space, which may correlate with image characteristics from objects that are related to watching (e.g., TV). It should be noted that unit visualization is very limited for capturing the learned intermediate features. Interpreting features in a limited 2D space is inherently biased and subjective.



(a) Ride (b) Wear (c) Watch

(d) Climb (e) Fill (f) Sit

Figure 5.7: Visualization of 6 last layer units in the CNN.

## 5.5 Discussion

We successfully trained a CNN to predict affordances from images, as a means for learning the underlying dimensionality of object affordances. The intermediate features in the CNN constitute an underlying compositional structure for the representation of affordances.

## 5.6 Discussions

To ensure the objectivity of the affordance labeling, affordance labels for images – as opposed to just object categories labels – are being collected currently using Amazon Mechanical Turk. This dataset is availaby by request.

With a CNN trained for affordance recognition, weights from the intermediate layers can be extracted and used to featurize each image. A model can then be trained to predict the BOLD

responses to each image. Correlations between the predicted responses and the true responses can be used to measure model performance. If a linear model is built to perform this task, the model weights could then be used as a proxy to localize where information about affordances is represented in the human brain.

Finally, affordance categories can be split into two large groups: semantically relevant ones, such as “eat”, which requires past experience with the objects in question; and non-semantically relevant ones, such as “sit”, which may be inferred directly from the shapes of the objects. If semantic affordances are being processed in the brain, top-down information about the objects is potentially necessary in order to inform an observer about affordable actions, while the non-semantic ones would not require top-down information. Given such differences we may be able to differentiate between top-down and bottom-up visual processing in the human brain using our model; in particular, by distinguishing the different brain regions that are engaged in either or both of these two processes.

# Chapter 6

# Joint natural language and image pre-training builds better models of human higher visual cortex

## 6.1 Introduction

Advances in deep learning have sparked a revolution in machine intelligence over the past decade [LeCun et al., 2015a]. Somewhat surprisingly, these recent advances have also sparked a parallel revolution in how we explain brain recordings [Yamins et al., 2014a]. Heretofore unaccounted for brain responses associated with tasks in both visual and semantic processing can now be well predicted by deep neural networks [Yamins et al., 2014a, Toneva et al., 2022]. As suggested by Yamins and DiCarlo [Yamins and DiCarlo, 2016a], these dramatic improvements in prediction performance may be driven by the fact that artificial neural network models sharing task goals with natural systems also learn representations shared with those systems. This is true not only for complex behaviors, but also for mid-level tasks realized within our perceptual systems [Wang et al., 2019]. However, the use of modern neural networks to study biological intelligence has been limited by the fact that most artificial models learn only a single or low

dimensional task objective. Consequently, to date, models used to account for the visual behavior of biological systems have been trained using purely *visual* tasks. Moreover, much of this research has also been based on networks pre-trained on the relatively small and undiversified ImageNet dataset [Deng et al., 2009b]. In contrast, natural vision is an active process that evolved (i.e., was trained) over hundreds of millions of years in order to support complex behaviors such as scene interpretation and navigation and which incorporates information from diverse perceptual, conceptual, and language sources [Aminoff and Tarr, 2015, Gauthier et al., 2003, Schaffner et al., 2023, Lupyan et al., 2010]. In this context, a major challenge for understanding how biological systems process and represent visual information is to consider such sources, including incorporating complex training signals that capture human-relevant information and greater diversity in experience.

Of particular note, recent state-of-the-art models in machine intelligence capture complex human semantics by learning from multiple modalities simultaneously[Radford et al., 2021, Li et al., 2022, Yuan et al., 2021, Jia et al., 2021, WuD]. The dramatic performance improvements seen in both vision and language tasks for these models may be attributed to several factors, including the fact that the confluence of information from different sources can help delineate what is important in inputs and the fact that their training sets tend to be larger and more diverse than earlier models. For the former, this is especially true if one of the modalities is language, since human language is generated by humans and has evolved to highlight aspects of the world that are behaviorally relevant [Pinker, 2007]. For the latter, increasing training set size not only provides more (and possibly better) examples, but concomitantly, there may be an increase in the diversity of training inputs [Fang et al., 2022]. Parallel to the improvements these multimodal and large-scale models bring to machine intelligence, using these same models we find that we also obtain dramatic improvements in our ability to explain aspects of biological intelligence.

In one of the recent studies to explore whether joint vision-language models are effective for predicting brain data, Devereux et al.[Devereux et al., 2018] investigated the representation of

semantic information in anterior regions of the ventral visual stream by combining a deep neural network for vision with an attractor network model of semantics. Within this "visuo-semantic" model, concepts associated with visual inputs are activated due to co-occurrence between visual information and semantic features. This model was used to predict brain responses from fMRI arising from an object naming task. Consistent with the idea that higher level visual areas encode semantics, anterior regions of the ventral stream were best explained by high level layers of the attractor network representing semantics. Such results provide an important demonstration of how a joint vision-language model can account for patterns of activation at different stages of visual processing. However, there are a couple limitations in this work. For instance, the model used in this paper only takes in object centered image with white background and is not able to account for complex scene semantics. The use of human labeled concept property norm also has limited generalizblity to semantics on any new images. Another limitations is that the analyses in this paper are all based on representational similarity analysis (RSA) on coarsely defined regions in the brain, it is hard to pinpoint exactly where each semantic dimension is represented and for others to compare their findings against other visual-semantic models.

Building on this work and aiming to resolve aforementioned limitations, we took state-of-the-art models using "Contrastive Language-Image Pre-training" or "CLIP" [Radford et al., 2021] as representative of the class [Li et al., 2022, Yuan et al., 2021, Jia et al., 2021, Mu et al., 2022] in that models with CLIP successfully leverage supervision from natural language (image captions) for vision and supervision from vision (complex scene images) for language. In CLIP schemes, models are trained with real-world image/associated caption pairs, learn separate image and text encoders from scratch that encode each image/caption pair of training data with similar representations at the final layer. Different than most other previous multimodal models (e.g., VisualBERT [Li et al., 2019], LXMERT [Tan and Bansal, 2019]), multimodal loss signals in the final layer of CLIP are propagated through all earlier layers of both the visual and language encoders, and, therefore, model learning with CLIP may be more similar to human visual

learning, where top-down knowledge has been found to influence even the earliest layers of the visual pathway [Murray et al., 2006, Gilbert and Li, 2013]. Of note, CLIP can be used with different model architectures – here we tested two different model backbones with CLIP: visual transformer (ViT-32; "ViT-32$_{CLIP}$") and ResNet50 ("ResNet$_{CLIP}$"). The use of these two quite different backbones allowed us to rule out performance improvements due to network architecture. Of particular interest, multimodal models using CLIP excel at current zero shot benchmark tests in computer vision – outperforming vision models that do not include natural language supervision. When considered in light of the complex, multitask visual abilities of humans, we initially decided to explore whether brain prediction based on models using CLIP would be better than earlier models – our motivating assumption being that any improved performance might be attributable, in part, to the multimodal structure of CLIP.

To begin to address this possibility, we extracted network representations from standard neural network models trained with CLIP, such as ResNet$_{CLIP}$ (using each image or its associated caption) and from several single modality task-optimized models: ImageNet [Deng et al., 2009b] pre-trained ResNet50 [He et al., 2016] (which we refer to as "ResNet$_{ImageNet}$") and BERT [Devlin et al., 2019] (using the caption associated with each image). We then constructed voxelwise encoding models[Naselaris et al., 2011] (Figure 6.1a) to explain whole brain responses arising from *viewing* natural images from Allen et al.'s Natural Scenes Dataset (NSD) [Allen et al., 2022]. Our objective was to use this large-scale brain activity dataset to evaluate and quantify the contribution of multimodal pre-training in generating more brain-like visual representations.

As already alluded to, a variety of factors characteristic of CLIP and different from those of most prior models used for brain prediction, may be contributing to the improved prediction performance we obtained with CLIP. However, as a proprietary model, we are unable to vary and control for these different aspects of CLIP. Consequently, as a next step towards exploring these factors, we selected a series of recent models that allow for more direct comparisons between four important factors: model architecture, training feedback, dataset size, and data diversity.

Specifically, we extended our analyses to a self-supervised model, simCLR [Chen et al., 2020], a self-supervised model that included language feedback as in CLIP, SLIP [Mu et al., 2022], and an open source version of CLIP [Schuhmann et al., 2022a]. As shown in Figure 6.5a, these models were trained with datasets that included 15 million (YFCC [Thomee et al., 2016]), 400 million (as in the original CLIP model), or 2 billion examples from LAION [Schuhmann et al., 2022a]. As we did with CLIP, we extracted network representations from these models and constructed voxelwise encoding models. These encoding models were then used to explain responses from NSD, allowing us to evaluate and quantify, as compared to the performance of CLIP, the contributions of model architecture, multimodal pre-training, dataset size, and data diversity in generating brain-like visual representations. To preview our most important contributions, we find that:

1. Pre-training with CLIP enables encoding models that are much more accurate at predicting visual and semantic representations in the human brain as compared to single modality models that are pre-trained with ImageNet. These improvements are *not* due to architectural differences and beyond a certain training dataset size, appear to be related more to the quality and/or diversity of the data, as well as the joint image/caption training that this data affords. Critically, we also see consistent improvement due to language feedback when dataset factors are controlled.
2. Models using CLIP are able to predict the recorded activity in human visual cortex using image captions alone – indicating that these models learn a robust latent space bridging natural language and vision.
3. Models using CLIP account for more of the unique variance in high-level visual regions. We observe the greatest improvement from a latent dimension of CLIP that represents complex scenes of humans interacting with one another and their environment, which in turn allows us to better predict brain areas that process such information.

## 6.2 Results

### 6.2.1 Multimodal embeddings best predict high-level visual cortex

The central question of our study is whether models using CLIP is a better model for human high-level visual cortex as compared to previous, vision-only models. To address this question, we extracted representations from the last layer of the ResNet$_{CLIP}$ image encoder and ResNet$_{ImageNet}$. Recall that both these networks have the same architecture but are trained with different objectives.

We expect that images are represented differently by ResNet$_{CLIP}$ and ResNet$_{ImageNet}$, such that ResNet$_{CLIP}$ embeddings contain more semantic information, and ResNet$_{ImageNet}$ contain more visual properties. In Figure 6.1b, we show the similarity between pairs of images computed using the two embeddings. For each pair of 10000 randomly sampled images, a similarity score was computed (measured in correlation) between the representations of these two images extracted from ResNet$_{CLIP}$ and ResNet$_{ImageNet}$ (i.e. $Sim_{i,j}^{CLIP}$ and $Sim_{i,j}^{ResNet_I}$). Pairs of images were ranked according to the differences between the similarities. Namely, $S_{i,j} = Sim_{i,j}^{CLIP} - Sim_{i,j}^{ResNet_I}, \forall i,j \in \{1,...,10,000\}$, where $Sim_{i,j}^{CLIP}$ and $Sim_{i,j}^{ResNet_I}$ are correlations of representations between Image $i$ and Image $j$ in ResNet$_{CLIP}$ and ResNet$_{ImageNet}$, respectively. Figure 6.1b shows the pairs of images that are most similar in ResNet$_{CLIP}$ and dissimilar in ResNet$_{ImageNet}$ (ranked by $S_{ij}$) and vice versa. Images represented similarly in the ResNet model trained with CLIP are semantically related, but this is not so for the ResNet model trained with ImageNet. For example, within ResNet$_{CLIP}$, images of people surfing and skateboarding are more similar and images of giraffes and an elephant are more similar. In contrast, within ResNet$_{ImageNet}$, images with different contexts are more similarly represented according to their visual similarity. For example, people wearing dark suits with a white shirt and a contrasting tie. These "corner" images illustrate that with natural language as training feedback, representations within ResNet$_{CLIP}$ capture contextual similarities that are not present

in ResNet$_{ImageNet}$ (which appears to be much more anchored in visual similarity).

We used the stimuli representation from ResNet$_{CLIP}$ image encoder and ResNet$_{ImageNet}$ to predict fMRI voxelwise responses across the brain. In Figure 6.1c we show the $R^2$ performances in the held out data set across the whole brain. For visualization purpose, we only plotted in the flatmap the voxels that are predicted significantly higher than chance ($p < 0.05$, FDR-corrected [Benjamini and Hochberg, 1995]). The encoding model built with the last layer of ResNet$_{CLIP}$'s visual encoder explains variance close to the voxel noise ceiling (see Fig. 6.6 for performance measured in $r$ for Subject S5). As a reference, earlier papers using voxelwise encoding models for brain prediction report well below 0.7 in maximum correlation [Güçlü and van Gerven, 2015, Huth et al., 2016]. In Allen et al. [Allen et al., 2022] a brain optimized model of early visual cortex (V1-V4) explains up to 0.8 in $R^2$, similar to what we observe here in high-level visual cortex. However, directly comparing performance across wide range of models is challenging due to the fact that different studies are carried out with very distinct experimental designs and rely on different data preprocessing and fitting pipelines. Studies that report model performance in terms of averages within ROIs and representation similarity (RSA) scores are also difficult to compare to our present results. Importantly, the high level performance we observed was not idiosyncratic to a few subjects: both the overall level and the pattern of prediction performance were highly consistent across S1-S8 (results for S5 are shown in Fig. 6.1, results for S1-S8 are shown in Fig. 6.7 and Fig. 6.8).

The ResNet$_{CLIP}$ encoding model's superior prediction performance provides compelling evidence that joint supervision from text information leads to representations that are better predictive of high-level visual cortex. We discuss this further in the *Discussion* section. From a theoretical point of view, these results suggest that the semantic information summarized in the image captions plays an important role in the organization of high-level visual knowledge in the human brain.

Beyond overall performance metrics, performance peaks in the brain prediction maps were

aligned with common functionally-defined category-selective ROIs. In particular, peaks within regions implicated as scene-selective [Epstein and Baker, 2019], body-selective [Downing et al., 2001b], and face-selective [Sergent et al., 1992, Kanwisher et al., 1997b] were sufficiently well defined so as to allow localization of these ROIs based solely on the prediction performance of ResNet$_{CLIP}$. We speculate that these alignments signal the importance of semantic associations in scene understanding and person recognition.

In order to rule out performance improvements based on a specific network architecture, we extracted features from two available model backbones that are pre-trained with CLIP: visual transformer (ViT-32) and ResNet50. Differences in prediction performance were small (see Fig. 6.11), indicating that the improvement provided by models using CLIP is not due to any particular neural-net architecture.

To explore whether captions associated with images alone could predict the brain activity in response to viewing the corresponding image, representations extracted from the last layer of the CLIP text encoder were also used to predict voxelwise responses across the brain. To accomplish this we provided the text encoder with the captions of the images viewed in the scanner by each subject. The text encoder representation was then used to make voxelwise brain predictions. Somewhat surprisingly, in the absence of any image information, the model is still able to predict higher level visual cortex similar to that of the model based on ResNet$_{CLIP}$'s image encoder (Fig. 6.2), though the visual encoder still explains most of the unique variance throughout the cortex (see Fig. 6.12 for variance partitioning between visual and text encoder model). This result shows CLIP does enable models to learn meaningful latent space that bridges between vision and natural language as well as the efficacy of this latent space in capturing brain relevant visual-semantic information from the images and the captions. The fact that both the image and text encoders have similar patterns of high predictive performance indicates that the information encoded in these high-level visual areas is highly anchored in semantics.

### Visualization of ROIs with common English words

The trained encoding model with CLIP created a new tool for us to probe semantic representation of a brain region. From previous results, we show that the trained model reliable maps from the latent space of both visual and text encoder of CLIP to all voxels in the brain. Using the learned mapping from text encoder, we can now provide prediction of how each voxels respond to any text stimuli, in or out of the dataset. We then take a step further and ask, given a set of general words, what are the words or sentence that maximally activates a brain area. In table X, we show the set of the words that maximally activate brain areas that's well known to us. For the fusiform face area (FFA), for example, the words maximally activate it according to the trained model are: people, face, smile, etc, which is consistent with our previously knowledge of FFA. For the parahippocampal place area (PPA), we see a different set of words including: land, property, and stations, etc. This result is also consistent with our prior knowledge of this brain area. Together this shows that the visualizing the areas with text can consistently reveal semantic turning of brain areas, and encoding model with CLIP text encoder provides a simple visualization tool for semantic representation in the brain.

## 6.2.2 Embeddings learned with CLIP explain more unique variance than unimodal embeddings

As compared to the ImageNet trained ResNet50 (ResNet$_{ImageNet}$), ResNet$_{CLIP}$ explains more variances in individual voxels across the whole brain, as shown in 6.3a and Fig. 6.10. In order to measure the unique variance accounted for by ResNet$_{CLIP}$ as compared to unimodal models, we performed a variance partitioning analysis [Lescroart et al., 2015, de Heer et al., 2017] (Fig. 6.3). Only voxels with significantly higher than chance unique variance are plotted for both models ($p < 0.05$, FDR-corrected). We compared the unique variance accounted for by the last layer of the ResNet$_{CLIP}$ image encoder to that accounted for by last layer of ResNet$_{ImageNet}$ (which had the same ResNet50 architecture), ruling out potential performance differences arising from

model architecture.

Consistent with our results for prediction performance, ResNet$_{CLIP}$ accounts for the majority of the unique variance in areas anterior to primary visual cortex, particularly in OPA, PPA and EBA – all functional ROIs implicated in scene and person perception. To evaluate ROI-level improvement we also present a series of voxel scatter plots for a range of functional ROIs in Figure 6.3b. With the exception of early visual areas (e.g., V1v, h4v), ResNet$_{CLIP}$ accounts for a much larger portion of the unique variance for the majority of voxels in these high-level ROIs. Beyond category-selective ROIs that respond to faces, places, and bodies, we also identified ROIs such as TPOJ and Angular Gyrus (AG) that were much better explained by ResNet$_{CLIP}$. Interestingly, these two areas are held to be related to theory of mind and language [Saxe and Kanwisher, 2013].

Note that the last layer of ResNet$_{CLIP}$ explained less of the variance in early visual cortex as compared to ResNet$_{ImageNet}$; however, this does not imply that ResNet$_{CLIP}$ fails to capture information represented in these regions. The last layer of ResNet$_{CLIP}$ is the bottleneck layer that captures the image embeddings optimized to match in similarity with the text embeddings. As shown in Fig. 6.13, the entire visual pathway is best predicted by a progression of ResNet$_{CLIP}$ layers (including ones below the bottleneck layer). More generally, ResNet$_{CLIP}$ is the best predictive model for the whole of visual cortex.

### 6.2.3 Regions that benefit most from ResNet$_{CLIP}$ embeddings encode scenes of humans interacting with their environment

To better understand the semantic dimensions learned in the encoding model built with CLIP model representations, we performed principal component analysis (PCA) on the learned weight matrix concatenated across the 20,000 top predicted voxels from each of the eight subjects in NSD. We projected the concatenated voxels onto the principal component (PC) dimensions to understand the tuning of the entire voxel space, following previous work [Huth et al., 2016, Çukur

et al., 2013]. By visualizing each PC of the learned model and its corresponding voxel projection, we were able to uncover some of the semantic bases that underlie semantic organization in the brain. To interpret the information captured by different PCs, we visualized the top images that have the most similar representations to a given PC, for the top 5 PCs, which account for most of the explained variance (as shown in Fig. 6.15). These images, as shown in Fig. 6.16, were identified by computing the dot product similarity of ResNet$_{CLIP}$ image embeddings with the vector corresponding to the PC direction.

As illustrated in Figure 6.4d, we observed that animate and inanimate images are separated by PC1, and its brain projections correspond to functionally-defined body and face regions (e.g., FFA and EBA). As illustrated in the bottom row of Figure 6.17, we observed that scenes and food images are separated by PC2 when we split the functional areas identified from PC1 with PC2; its brain projections corresponded to functionally-defined place regions (e.g., PPA, RSC, OPA) and the food region [Jain et al., 2023, Pennock et al., 2023, Khosla et al., 2022]. Of note, we obtained interpretable PC dimensions up to PC10 (despite the relatively low explained variance from PC6 onwards), allowing us to identify more fined-grained semantic distinctions within high-level visual cortex. Images visualization of the rest of the PCs are shown in Fig. 6.16.

We directly compared the brain projection for PC1 and the unique variance map for ResNet$_{CLIP}$. We found that voxels that have large negative values on the PC1 overlay the majority of the time with voxels where ResNet$_{CLIP}$ has the largest unique variance (Figs. 6.4a and 6.4b). These voxels clustered in ventral EBA, FFA-1, FFA-2, as well as ventral RSC. Figure 6.4c further validates this finding by showing a strong negative correlation for the voxels with a negative projection between the magnitude of this projection and the unique variance explained by ResNet$_{CLIP}$. (Note that the sign of the PC is arbitrary and can be flipped; we use "negative" here to refer to one of the sides of PC1.) Thus, PC1 appears to separate the regions of high-level visual cortex that benefit the most when ResNet$_{CLIP}$ is used to predict performance.

Figure 6.4d shows the top 10 images for both ends of PC1. Top negative images are people

participating in sports, whereas the top positive images are indoor scenes. This separation is consistent with the location of the best predicted voxels from ResNet$_{CLIP}$ being centered on the extrastriate body area (EBA). Distribution of object categories present in the images that are on the two sides of the PC1 further validate this finding (Fig. 6.4d). We leveraged the known category and super-category labels of images in COCO and found that images that lie on the negative end of the PC1 are more likely to contain people, animals, and sports items. These observations suggest that the representation of people in ResNet$_{CLIP}$ is the domain for which the model provided the most leverage in terms of predicting brain responses (i.e., as compared to ResNet$_{ImageNet}$). From an ecological standpoint this finding appears to capture high-level semantic statistics regarding the world around us: scenes of people and human interactions are heavily present in our daily life. Returning to our original hypothesis, by including natural language as input (image captions) along with complex scenes, ResNet$_{CLIP}$ is more effective at capturing the rich semantics of scenes as compared to models trained with image/label pairs pre-training (e.g., ImageNet). At the same time, it is important to recognize that the process of how CLIP is trained has other major differences from earlier models. These differences, for example, larger training datasets and greater diversity in dataset distribution, may also contribute to the ability of CLIP to predict brain responses in high-level visual areas. We explore these factors in the next section.

### 6.2.4 Disentangling the effects of language feedback, model architecture, dataset size, and data diversity

As noted, beyond model architecture, the training data distribution and size may both impact model representations and how predictive they are in voxelwise encoding models of the brain. To better understand, in a controlled manner, the contributions of each of these factors towards the high brain prediction performance we observe with CLIP, we included three further variance partitioning analyses. The additional models we compared to one another in these analyses are

listed with their relevant characteristics in Figure 6.5a. Critically, these models also allow us to use the publicly available YFCC ("Yahoo Flickr Creative Commons") [Thomee et al., 2016] and LAION [Schuhmann et al., 2022a] datasets to better control for dataset size and diversity. YFCC is a 100 million example dataset comprised of multimedia "objects" which includes 15 million photos with captions selected from Flickr, while LAION is a large-scale dataset that contains 5.85 billion multilingual CLIP-filtered image-text pairs. Both the YFCC and LAION datasets provide sufficient multimodal data to retrain CLIP with different dataset parameters that allow for better control of dataset parameters.

We visualized averaged model performance across all models in Figure 6.5b for several well-characterized ROIs within each general anatomical and semantic categories (EarlyVis: V1v, h4v; Scene: PPA, OPA, RSC; Body: EBA; Face: FFA-1, FFA-2; TPOJ: TPOJ-1, TPOJ-2). Each point in the figure is a region's averaged performance, averaged across 8 subjects and error bar indicates standard error across 8 subjects. We observe that all CLIP models and the SSL models explained brain responses in higher visual cortex significantly better than the $\text{ResNet}_{ImageNet}$ model, while differences among SSL and CLIP models are small. An important point about these summary results is that they describe *average* responses across all voxels in a given ROI. Consequently, they do not reflect any spatial patterns of unique variance within that ROI. In that some ROIs as defined in NSD are large in terms of number of voxels, ROI average analyses are likely to mask meaningful spatial prediction patterns across models. In the flattened cortical maps of the unique variance shown in Figure 6.5c-e, and as discussed in detail in the next section, model comparisons that are similar on the average, as described in [Conwell et al.], actually also carry fine-grained information about the effects of model architecture, data distribution, and dataset size.

To take a closer look on how model feedback, dataset size and diversity affects predictions of each individual voxels, we present three voxelwise analyses in which we extracted features from models to build encoding models to predict voxel responses and analyzed voxelwise unique

variances explained by each model: 1) the effect of language feedback when controlling for the dataset parameters of distribution and size; 2) the effect of dataset size when controlling for the data distribution, feedback type, and model architecture; and 3) the effect of data distribution when controlling for feedback type and training dataset size. Note that while other comparisons between models are possible, all of them would involve models that vary from one another on more than one factor. As such, these three included comparisons are the ones that are maximally informative in terms of isolating single factors with respect to their impact on brain prediction performance.

First, we evaluated the effect of language feedback while controlling for dataset size and data distribution by comparing the simCLR and SLIP (which combines simCLR and CLIP losses, meaning that language is the only varying factor) models trained on 15 million photos captions pairs from the YFCC dataset. Figure 6.5c shows the spatial brain map of the unique variance comparison, thresholded by statistical significance ($p < 0.05$, FDR-corrected [Benjamini and Hochberg, 1995]). Using the exact same dataset, we see unique variance explained by SLIP in EBA, FFA and adjacent to the boundary of RSC, while simCLR shows more unique variance explained in early visual cortex and posterior EBA. Below the brain maps, we further exam model preferences using a histogram of unique variance across all EBA voxels (for all 8 subjects). We observe a bimodal distribution of voxels preferring one model or the other, with more voxels skewing towards YFCC SLIP. Flatmaps of significant unique variance explained by SLIP are also visualized across 8 subjects in the common MNI space in Figure 6.18. In MNI space, significant voxels show consistent patterns across subjects. More generally, these data visualizations indicate that interpreting brain prediction across models requires analysis at the voxel, rather than the ROI, level, in order to move beyond the broad functional roles associated with different ROIs.

Second, we evaluated the effect of dataset size while controlling for data distribution by comparing two CLIP models trained on 400 million or 2 billion image/caption pairs from the LAION

dataset [Schuhmann et al., 2022a]. Figure 6.5d shows the spatial brain map of the unique variance comparison, thresholded by statistical significance ($p < 0.05$, FDR-corrected [Benjamini and Hochberg, 1995]). The representations from CLIP trained using the larger dataset explained more unique variance than CLIP trained using the smaller dataset in EBA, FFA, and areas outside of RSC. However, this improvement in prediction performance due to dataset size was small. The scale of the unique variance accounted by both versions of CLIP is shown in the histogram of EBA voxels below the brain maps. The narrow spread indicates that dataset size, after reaching a critical level for model training, does not seem to be a critical factor in the improvements in brain prediction we obtained using CLIP.

Third, we evaluated the effect of data distribution while controlling both feedback type and dataset size by comparing two CLIP models trained on 400 million image/caption pairs from OpenAI [Radford et al., 2021] or from LAION [Schuhmann et al., 2022a]. Figure 6.5e shows the spatial brain map of the unique variance comparison, thresholded by statistical significance ($p < 0.05$, FDR-corrected [Benjamini and Hochberg, 1995]). The representations from CLIP trained using OpenAI's dataset explained more unique variance than CLIP trained using LAION's dataset in regions including the EBA, FFA and areas outside of RSC. This result aligns with a previous study [Fang et al., 2022] that argues that data diversity in the training dataset contributes significantly to the robustness of the OpenAI CLIP model. As shown in the histogram of EBA voxels below the brain maps, the scale of difference arising from data distribution is larger than that arising from dataset size, indicating that data diversity is an important factor in the improvements in brain prediction we obtained using OpenAI CLIP.

To summarize, we argue that the inclusion of language feedback and the diversity/quality of training data, given some baseline amount of data, are important model characteristics for achieving improved brain prediction in high level visual cortex. At this point in time, however, these factors cannot easily be disentangled further in that natural language labeling (i.e., image captions), compared to category labels, may introduce large variations that are concomitant with

dataset size and data diversity. That is, if language is the training feedback that distinguishes a good proxy model of the brain [Leeds et al., 2013], then it logically follows that the quality of those language annotations also make a difference in the brain prediction performance of these models.

We also performed a variance partitioning analysis comparing ViT-$32_{CLIP}$ and BERT, and likewise found that ViT-$32_{CLIP}$ accounts for almost all unique variance, again ruling out improvements in prediction due to architectural differences (Fig. 6.14). Thus, the advances we observed in brain prediction using models using CLIP do not appear to arise from incorporating complex semantics alone, but rather, can be attributed to a meaningful mapping between visual and semantic representations.

## 6.3 Discussion

Do higher performing models using more human-like training feedback, and in particular, natural language, as well as more diverse, ecologically-valid training sets, also perform better at accounting for brain data in response to complex, real-world scenes? To address this question, we evaluated and quantified the contributions of the multimodal pre-training as provided by CLIP for generating semantically-grounded, behaviorally-relevant representations of natural scenes. We find that models using CLIP as a pre-training task are extraordinarily good at predicting voxelwise responses to viewing real-world scenes in the Natural Scenes Dataset [Allen et al., 2022]. Validating this finding, a second study also finds that models using CLIP as a pre-training task can better predict responses in NSD as compared to other deep network models (likewise controlling for training data and model architectures)[Conwell et al., 2022a,b]. However, while it is appealing to attribute the improved prediction performance of CLIP to its inclusion of language feedback during training, several other factors may also contribute to the high level of brain prediction performance using CLIP. More specifically, CLIP models, as compared to prior models used for brain prediction [Yamins et al., 2014a, Güçlü and van Gerven, 2015], often have

different model architectures and were pre-trained with many more examples. Consequently, model architecture, multimodal pre-training, dataset size, or data diversity (or some combination therein) may underlie some or all of the improved brain prediction performance we obtained.

To address the potential contributions of these factors, we examined the prediction performance for several other models that enabled us to control for each factor. Using models for brain prediction that differ from one another along only a single dimension, we found that: 1) Models trained with natural language feedback show a consistent advantage in prediction performance in certain high-level brain regions, including the EBA and TPOJ; 2) The quality, diversity, and quantity of the training data may set a ceiling for improvements in prediction arising from adding language feedback; 3) The size of the training dataset for the CLIP model appears to be both less consequential for improved prediction performance and shows diminishing returns as compared to other data characteristics (i.e., data diversity). As such, we conclude that models using CLIP (along with sufficient data diversity) are better candidate models for understanding representation in high-level human visual cortex.

More broadly, in contrast to simply quantifying overall brain prediction performance across a range of models, we focus on human-like training in the form of natural language feedback and larger, more diverse datasets. We not only identify the brain areas that benefit most from natural language feedback, but also provide analyses that help us to better understand the ways in which CLIP pre-training facilitates learning brain-like representations. Visualizations of the representations from ResNet$_{CLIP}$ and a unimodal network reveal that ResNet$_{CLIP}$ better captures semantic information. This observation is consistent with our finding that natural language feedback is central to CLIP's improved performance and our hypothesis that this improvement is associated with the encoding of complex semantic representations in high-level visual cortex. We then used PCA analyses to explore our results, finding that, withing ResNet$_{CLIP}$, the fine-grained representation of scenes depicting human interaction drives the largest gains in brain prediction, which in turn illuminates some of the underlying reasons why models using CLIP yield such excellent

performance particularly in EBA. Our conjecture is that ResNet$_{CLIP}$ is capturing information about humans interacting with the world around them, and that such complex semantic information is predictive of this fundamental aspect of higher level brain regions [Bracci and Op de Beeck, 2023].

*In toto*, our results support the theory that human higher-level visual representations reflect semantics and the relational structure of the visual world beyond object identity; for example, non-perceptual associations such as function or linguistic meaning [Gauthier et al., 2003, Bracci and Op de Beeck, 2023, Maier and Abdel Rahman, 2019]. Supporting this point, in concurrent work [Charest et al., 2020], an embedding model based on text captions for viewed images also suggests that higher-level visual cortex represents semantic information related to those images. Indeed, while it is well established that language plays a critical role in the acquisition of semantics [Nappa et al., 2009, Waxman and Markow, 1995], language also influences the acquisition of visual categories during development, where visual learning occurs concurrently with language and conceptual learning [Waxman and Markow, 1995, Lupyan et al., 2007, Shusterman and Spelke, 2005a]. In this context, in conjunction with our results, it seems clear that language and semantics strongly influence the high-level organization of visual information encoded in the human brain.

Returning to the question of what it is about models pre-trained with CLIP that enables them to excel not just at visual tasks such as few-shot learning, but also at brain prediction, we note that this is a question that is still being debated within the field of computer vision [Fang et al., 2022]. In our view, the same natural language feedback, along with data diversity, helps prompt higher performance in both domains - another instance where higher performing models also perform better at brain prediction [Yamins and DiCarlo, 2016a]. Interestingly, beyond a certain size, the size of the training dataset appears to have little impact on performance. This is in line with the intuition that even if we were able to re-train ResNet on a much bigger dataset, but continued to include only category labels, the resultant model would be unlikely to learn fine-grained

representations of semantically complex scenes. Such nuanced information regarding human interactions in real-world scenes is not typically carried by category labels. That is, while some labels do contain semantic information beyond the category (e.g., "party"), language feedback provides context and a broader understanding, for example, the semantic relationships between different real-world scenes. Supporting this point, we found that, given equivalent training data, models with natural language feedback outperformed self-supervised and unimodal models in higher level visual areas. As such, the natural language feedback present in models pre-trained with CLIP appears crucial to their excellent performance in tasks related to both machine and biological intelligence.

In sum, the ability of CLIP pre-trained models to predict brain responses opens up new possibilities for developing a deeper understanding of the functional architecture of the human brain. Further exploring the implications of this finding will require new ways of thinking about both machine and biological systems. Future large-scale efforts should incorporate stimuli, tasks, representations, and models that reflect the natural complexity of how we interact with the world around us.

## 6.4 Materials and Methods

### 6.4.1 Datasets

**fMRI data.** Brain recordings were obtained from the the Natural Scenes Dataset (NSD) [Allen et al., 2022], an open dataset of 7T whole brain high-resolution fMRI responses from eight subjects (S1-S8) who each viewed $\sim$10,000 unique images of natural scenes, each image repeated 3 times. These scene images were a subset of the images in the annotated Microsoft Common Objects in Context (COCO) dataset [Lin et al., 2014b]. Of the 70,566 total images presented across subjects, $\sim$1,000 images were viewed by all subjects. fMRI data were collected during 30-40 scan sessions. Stimulus images were square cropped, presented for 3 s at a size of $8.4^\circ \times$

$8.4^{\circ}$ with 1 s gaps in between image presentations. Subjects were instructed to fixate on a central point and to press a button after each image if they had seen that image previously.

The functional MRI data were acquired at 7T using whole-brain gradient-echo EPI at 1.8-mm resolution and 1.6-s repetition time. Preprocessing steps included a temporal interpolation (correcting for slice time differences) and a spatial interpolation (correcting for head motion). Single-trial beta weights were estimated with a general linear model. In this paper we used the betas_fithrf_GLMdenoise_RR preparation of the betas. FreeSurfer [Dale et al., 1999, Fischl et al., 1999] was used to generate cortical surface reconstructions to which the beta weights were mapped. The beta weights were z-scored across run and were averaged across repetitions of the image (up to 3 repetitions of each image), resulting in one averaged fMRI response to each image per voxel, in each subject. NSD also includes several visual ROIs that were identified using separate functional localization experiments. We drew the boundaries of those ROIs for each subject on their native surface for better visualization and interpretation of the results (e.g., Fig. 6.1). All brain visualizations were produced using Pycortex software [Gao et al., 2015a].

**Natural scene images.** All stimulus images used in NSD and in our experiments were drawn from the COCO dataset [Lin et al., 2014b]. COCO is unique among large-scale image datasets in that COCO images contain contextual relationships and non-iconic (or non-canonical) object views. In comparison to ImageNet [Deng et al., 2009b], COCO contains fewer labeled categories (91), but includes more examples for each category ($> 5,000$ for 82 of the categories). Note, however, that many labeled categories in ImageNet are at the subordinate level – COCO likely contains at least as many *unlabeled* subordinate categories. The complete set of COCO images and additional details can be found on the COCO website: `https://cocodataset.org`.

### 6.4.2 Model details and feature extraction

Models used in the analysis includes: 1) OpenAI trained CLIP (with ViT-32 transformer and ResNet50 backbones); 2) YFCC trained SLIP, CLIP, simCLR; 3) Open CLIP models trained on LAION 400m and LAION 2B; 4) ImageNet pretrained ResNet50. All NSD stimuli images were input to the these models.

For model comparison, we use the output of the "image encoder" in CLIP models and the second to the last layer in ImageNet trained models as feature spaces for the encoding models. The feature dimensions for each of the model feature spaces are as follows: ImageNet trained ResNet50: 2048; OpenAI CLIP with ViT-32 backbone: 512; OpenAI CLIP with ResNet50 backbone: 1024; YFCC simCLR: 768; YFCC SLIP: 512; YFCC CLIP: 512; LAION 400m CLIP: 512; LAION 2B CLIP: 512.

For image captions, we use the human generated captions for each of the NSD images provided by the COCO dataset and input them into both BERT and CLIP-based models' text encoders for their layerwise activations. On average, COCO provides 5-6 captions for each image. Caption embeddings for a image are extracted individually and the average is used in the encoding models.

### 6.4.3 Voxelwise encoding models

We built ridge regression model (implemented in PyTorch; see [Koushik, 2017]) to predict one averaged fMRI response to each image per voxel, in each subject. We chose to use a ridge regression model instead of more complicated models in order to retain the interpretability of model weights, which may provide insights into the underlying dimensions of the brain responses. We randomly split the total number of images a subject sees into training and test set with a 4-to-1 ratio. For each subject, each voxel's regularization parameter was chosen independently via 7-fold cross-validation across the training set. We swept through 100 regularization parameters spaced evenly on a log scale from $10^{-8}$ to $10^{10}$, i.e. np.logspace(-8, 10, 100). Cross-validation was han-

dled by `sklearn.model_selection.KFold`, where data are split into consecutive folds without shuffling. Each fold is used once as validation while the rest of the set are used for training. Model performance was evaluated on the test data using both Pearson's correlation and coefficient of determination ($R^2$). To determine the significance of the predictions, we perform a bootstrap test where we resample the test set with replacement for 2000 times and compute the FDR corrected *p*-values threshold for various performance statistics.

### 6.4.4 Variance Partitioning

To obtain unique variance by two model A and B, we first create joint model of A and B by concatenating features from these two models. We then fit voxelwise ridge regression model to the joint model and obtain $R^2_{A\&B}$. The variance explained by individual model A and B are denoted as $R^2_A$ and $R^2_B$, respectively. We then calculate the unique variance for model A and B, where $R^2_A = R^2_{\&B} - R^2_B$, $R^2_B = R^2_{A\&B} - R^2_A$.

### 6.4.5 PCA analysis

We performed principal component analysis (PCA) on the learned matrix to recover the semantic basis of the learned model. We selected the 20,000 best predicted voxels for each individual subject based on the noise corrected model performance of the ResNet$_{CLIP}$. We then concatenated the weight matrices (used in encoding model with ResNet$_{CLIP}$) corresponding to these voxels from all eight subjects along the voxel dimension. We then centered the matrix, applied PCA, and obtained the first 20 PCs. Explained variance by these PCs are plotted in Figure 6.15.

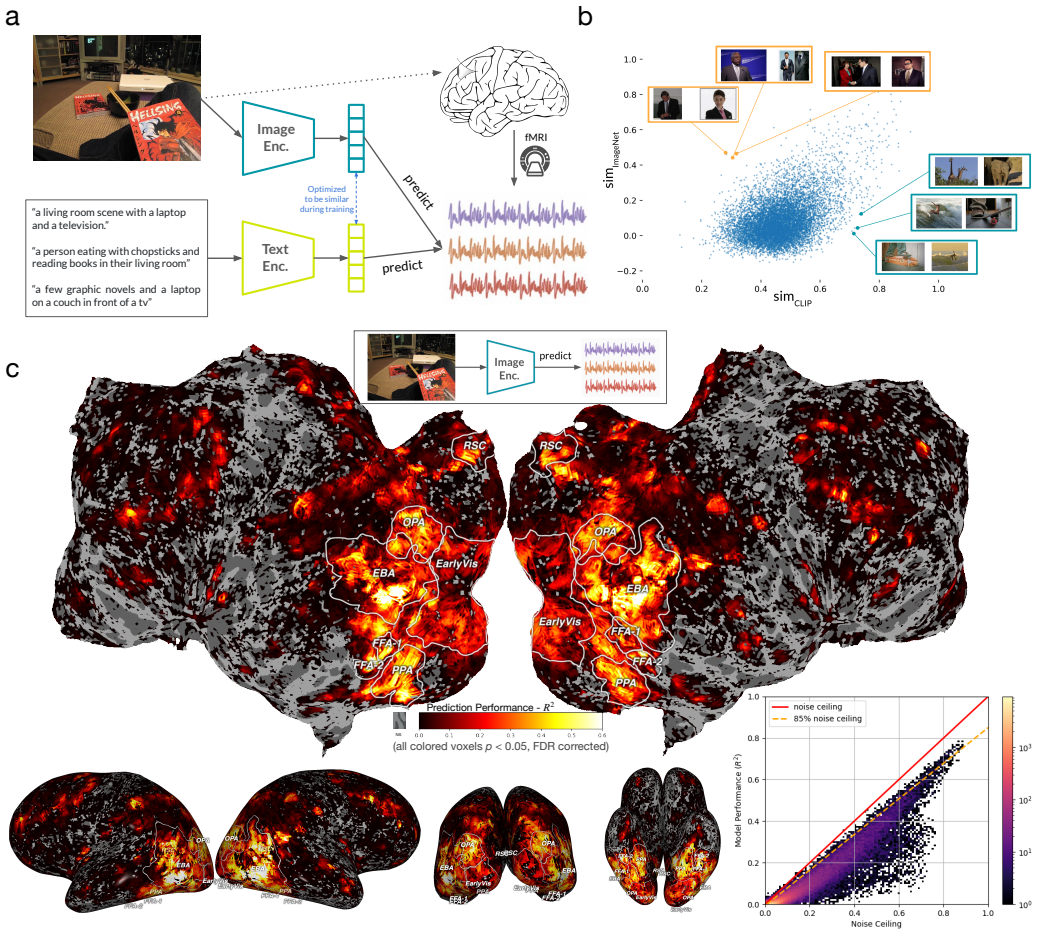



Figure 6.1: **Model pipeline, motivation and prediction performance for the ResNet$_{CLIP}$ visual encoder**. **(a)** Last-layer representations from the CLIP image and text encoders are extracted from images and captions, respectively. These representations are used in voxelwise encoding models to predict brain responses to each image. **(b)** The similarities of pairs of images when using embeddings from ResNet$_{CLIP}$ and ResNet$_{ImageNet}$ are compared. The position of each dot in the scatter plot is determined by similarity scores for the same pair of images in ResNet$_{CLIP}$ and ResNet$_{ImageNet}$ model spaces. Pairs of images in the bottom right corner are those most similar in ResNet$_{CLIP}$ and most dissimilar in ResNet$_{ImageNet}$; for example, images of people surfing and skateboarding and images of giraffes and an elephant. In contrast, pairs of images in the top left corner are those most similar in ResNet$_{ImageNet}$ and least similar in ResNet$_{CLIP}$; for example, visually similar pictures of people wearing dark suits with a white shirt and a contrasting tie. **(c)** Voxelwise prediction performance (measured in $R^2$) on a held-out test set is shown for Subject S5 in a flattened view of the brain with overlays for functionally-defined, category-selective ROIs (top), as well as in lateral, posterior and bottom views (bottom row, left-to-right). (Bottom-right) A 2D histogram of model performance in $R^2$ against noise ceiling and 85% noise ceiling across all voxels in the whole brain. Density of voxels are shown in a log scale. Most voxels are predicted close to its noise ceiling and some are above the 85% noise ceiling.

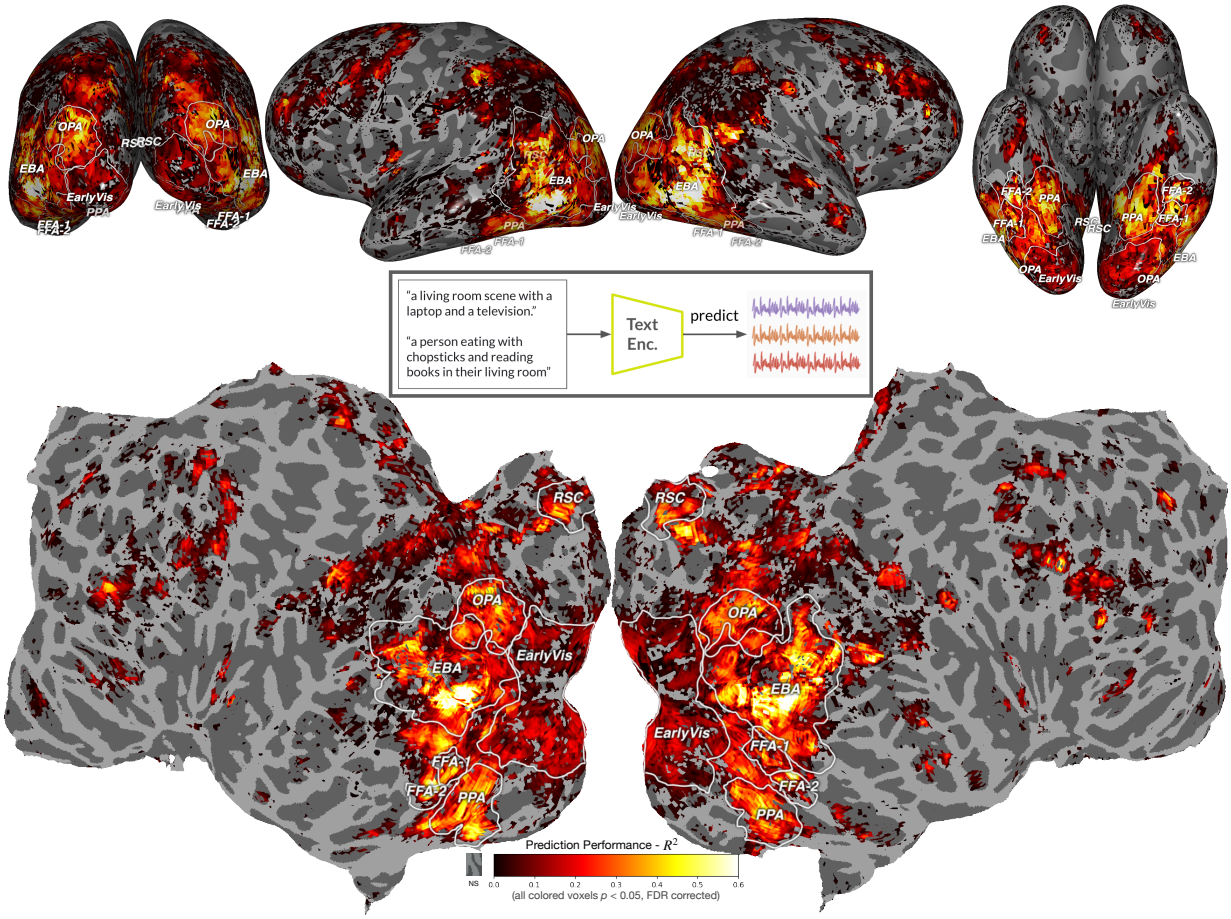



Figure 6.2: **Prediction performance for the CLIP text encoder.** Prediction performance for voxelwise responses – $R^2$ – in held out data for the CLIP text encoding model for S5 with overlays for functionally-defined, category-selective ROIs. Although only having access to the captions of the images that the subjects viewed, the CLIP text encoder is still able to predict fMRI data in many functionally-defined ROIs (e.g., EBA, PPA, RSC, FFA).

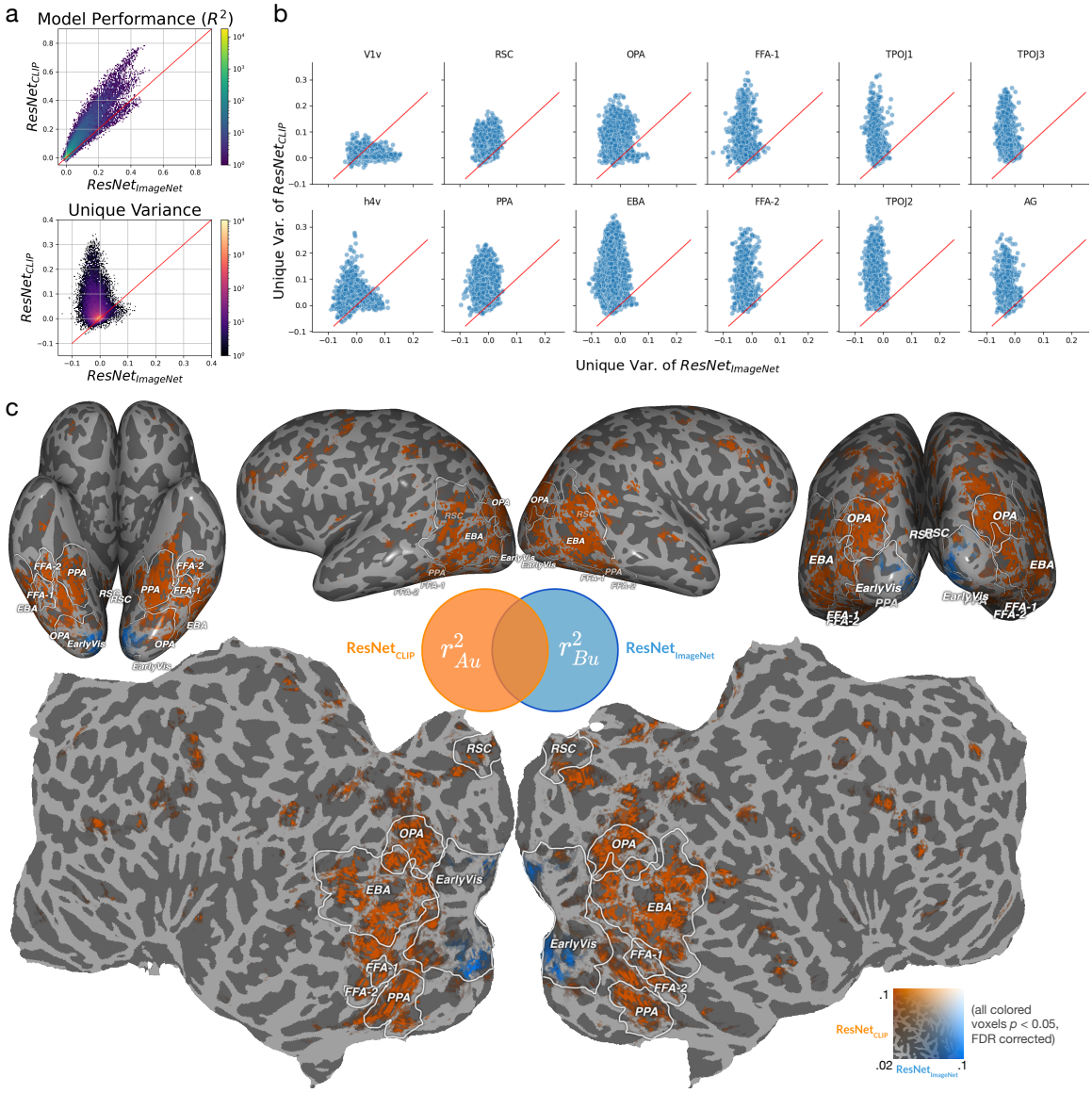



Figure 6.3: **Performance for the CLIP visual encoder using a ResNet backbone as compared to ResNet$_{ImageNet}$ (a)** 2D distribution plots of voxels from the whole brain in S5 in model performance (in $R^2$) and unique variance comparing between ResNet$_{CLIP}$ and ResNet$_{ImageNet}$. The red lines indicates equal performance for the two models. ResNet$_{CLIP}$ predicts much better in terms of total variance and unique variance. **(b)** Unique variance accounted for by ResNet$_{CLIP}$ as compared to ResNet$_{ImageNet}$ for 12 different ROIs for all eight subjects. Individual voxels are plotted as blue points. The red lines indicate iso-variance, that is, $(y = x)$. ResNet$_{CLIP}$ accounts for overwhelmingly more variance than ResNet$_{ImageNet}$ in higher-level visual cortex. In contrast, ResNet$_{ImageNet}$ only accounts for more variance in ventral V1 and a reasonable proportion of the variance in ventral V4. **(c)** Unique variance accounted for by ResNet$_{CLIP}$ as compared to ResNet$_{ImageNet}$ for S5 – obtained by subtracting $R^2$ for each model from that of the joint model (with concatenated feature spaces). Voxels where ResNet$_{CLIP}$ accounts for greater unique variance are orange and voxels where ResNet$_{ImageNet}$ accounts for greater unique variance are blue.

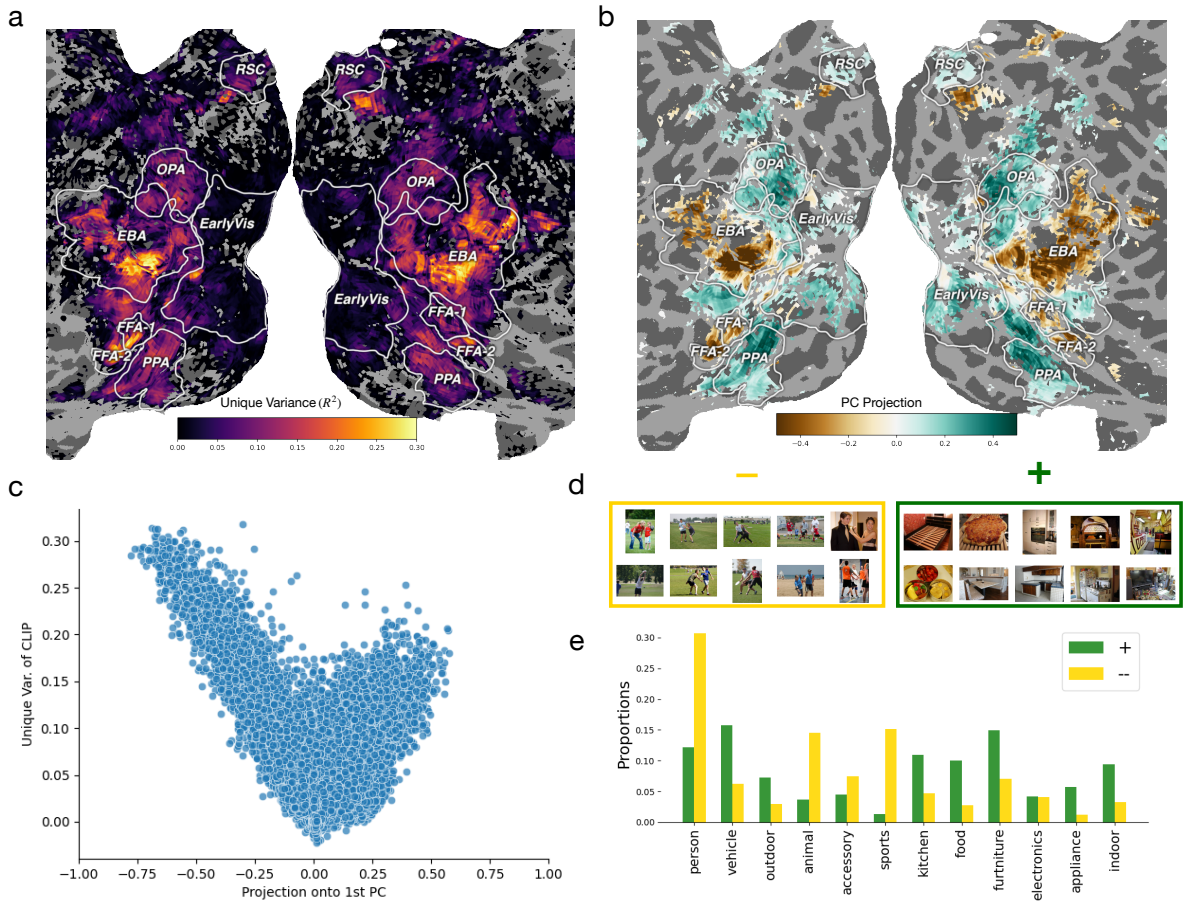



Figure 6.4: **Better representations of scenes with people in a model trained with CLIP can account for gains in unique variance. (a)** Unique variance explained by ResNet$_{CLIP}$ plotted on a flatmap from S5. **(b)** Projection of voxels onto PC1 of ResNet$_{CLIP}$ for S5. Voxels that are best explained by ResNet$_{CLIP}$ overlap largely with the voxels that lie on positive side when projected onto the 1st PC. **(c)** Voxelwise scatter plot illustrating that for voxels lying on the negative side of 1st PC projection, the further down the voxel lies on the projection, the better it is explained by ResNet$_{CLIP}$. **(d)** Images are grouped in to "+" and "−" depending on which side the image lies on when projected onto the PC1. The top 10 images that best align with either end of the PC1 are shown in the yellow and green boxes respectively. For the positive projection we observe images of indoor scenes, whereas for the negative projection we observe images of people participating in outdoor sports. **(e)** Category distribution of two groups of images validates that images on the negative side consist more of people, animal, and sports, relative to images on the positive side.

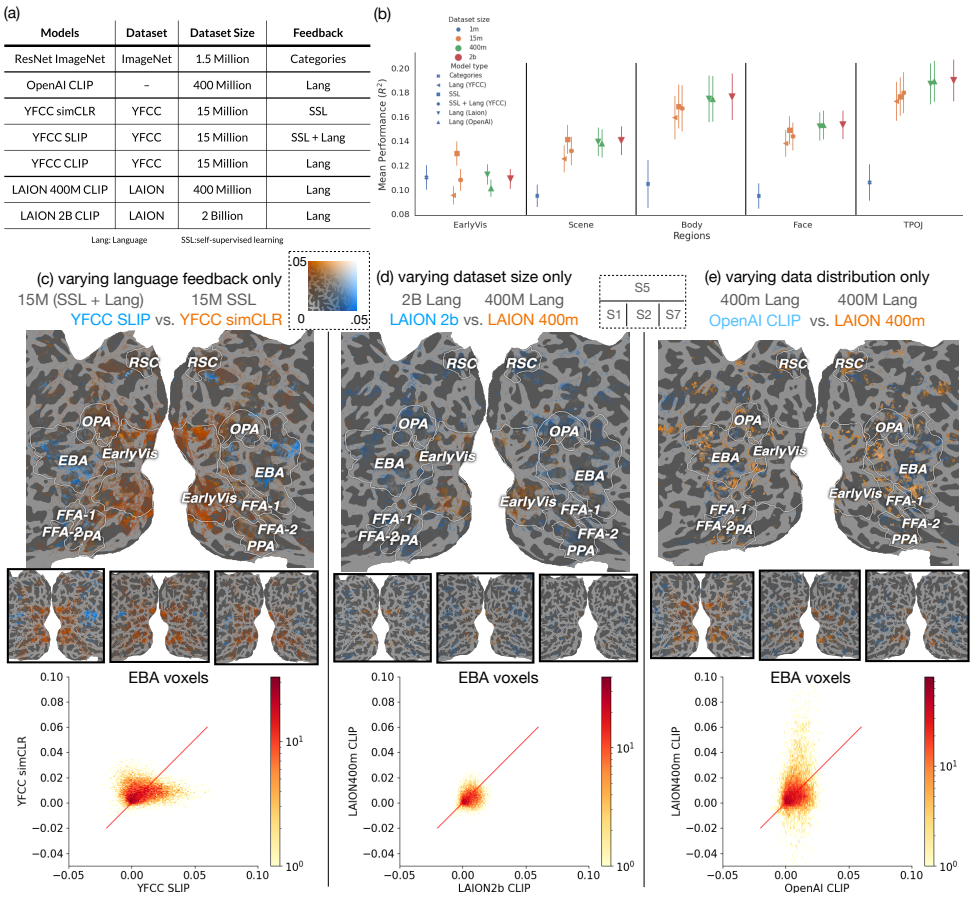



Figure 6.5: Variance partitioning analyses controlling for model architecture, data distribution, and dataset size indicate that dataset size and diversity have comparatively smaller effects on voxel prediction than language input does. (a) The models we consider with their relevant characteristics; (b) Brain prediction performance averaged across all voxels in a given brain region for each model+dataset combination ("SSL" denotes self-supervised learning; "Lang" denotes natural language feedback for a given model). Error bars reflects standard error across 8 subjects. When looking at average brain prediction performance with an ROI, all three CLIP pre-trained models and the SSL model perform significantly better than ImageNet trained ResNet50, while differences between all three CLIP models and the SSL model are relatively small. (c-e) Cortical flatmaps showing the fine-grained, spatial distribution of unique variance for model comparisons varying a single factor while controlling for the others. For each comparison, the first row shows brain maps from S5, while the second row shows unique variance brain maps from S1, S2, and S7, respectively. The third row of each comparison shows a 2D histogram of unique variance for individual voxels in EBA for all 8 subjects. The red line indicates identical unique variance $(y = x)$. Notably, as shown in (c), when the same dataset is used for training across models, SLIP, a model that includes language feedback, accounts for more unique variance in high-level brain areas such as EBA and some parts of FFA, relative to simCLR, an otherwise identical model that does not include language feedback. Beyond language feedback, as shown in (e), a good data distribution appears to also account for unique variance in some high-level visual areas, while, as shown in (d), dataset size *per se* appears to account for very small improvements in unique variance past a certain size.

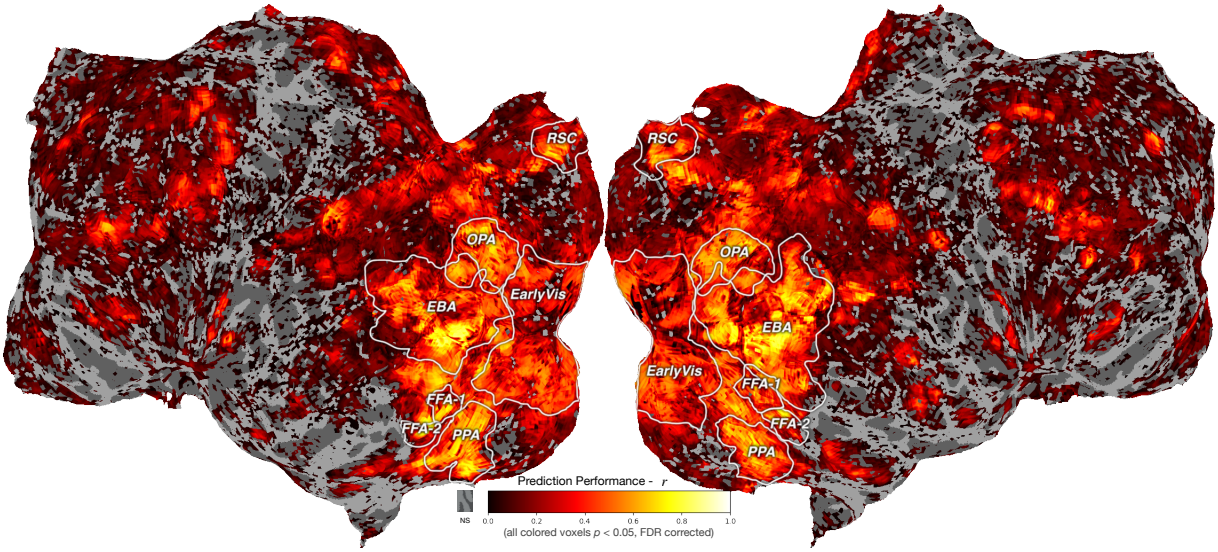



Figure 6.6: **Prediction performance meansured in corrlation using the CLIP visual encoder**. Voxelwise prediction performance (measured in $r$) on a held-out test set is shown for S5 in a flattened view of the brain.

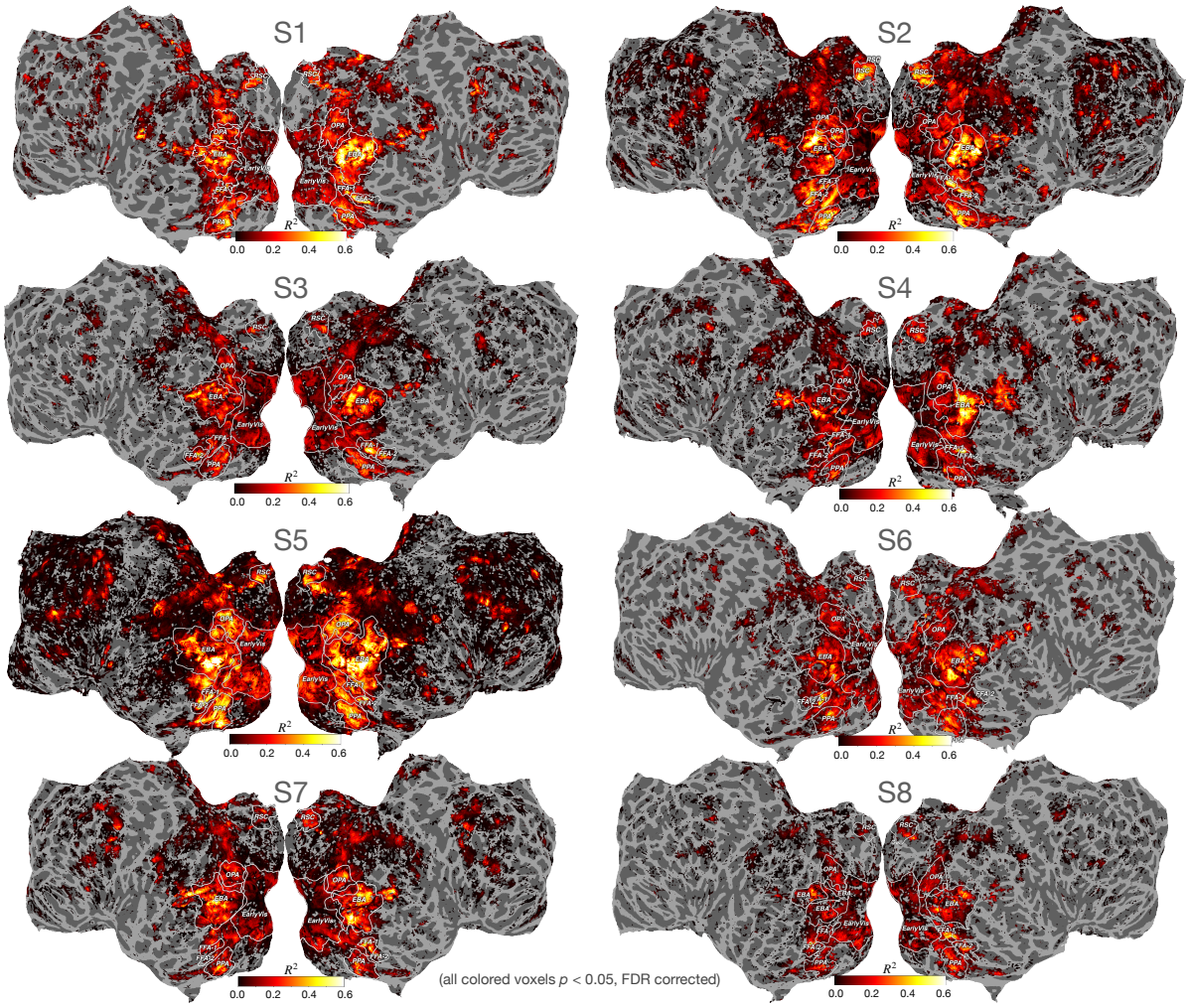



Figure 6.7: **Prediction performance with CLIP visual encoder for all eight subjects**. Voxelwise prediction performance (measured in $R^2$) on a held-out test set is shown for S1-S8 in a flattened view of the brain.

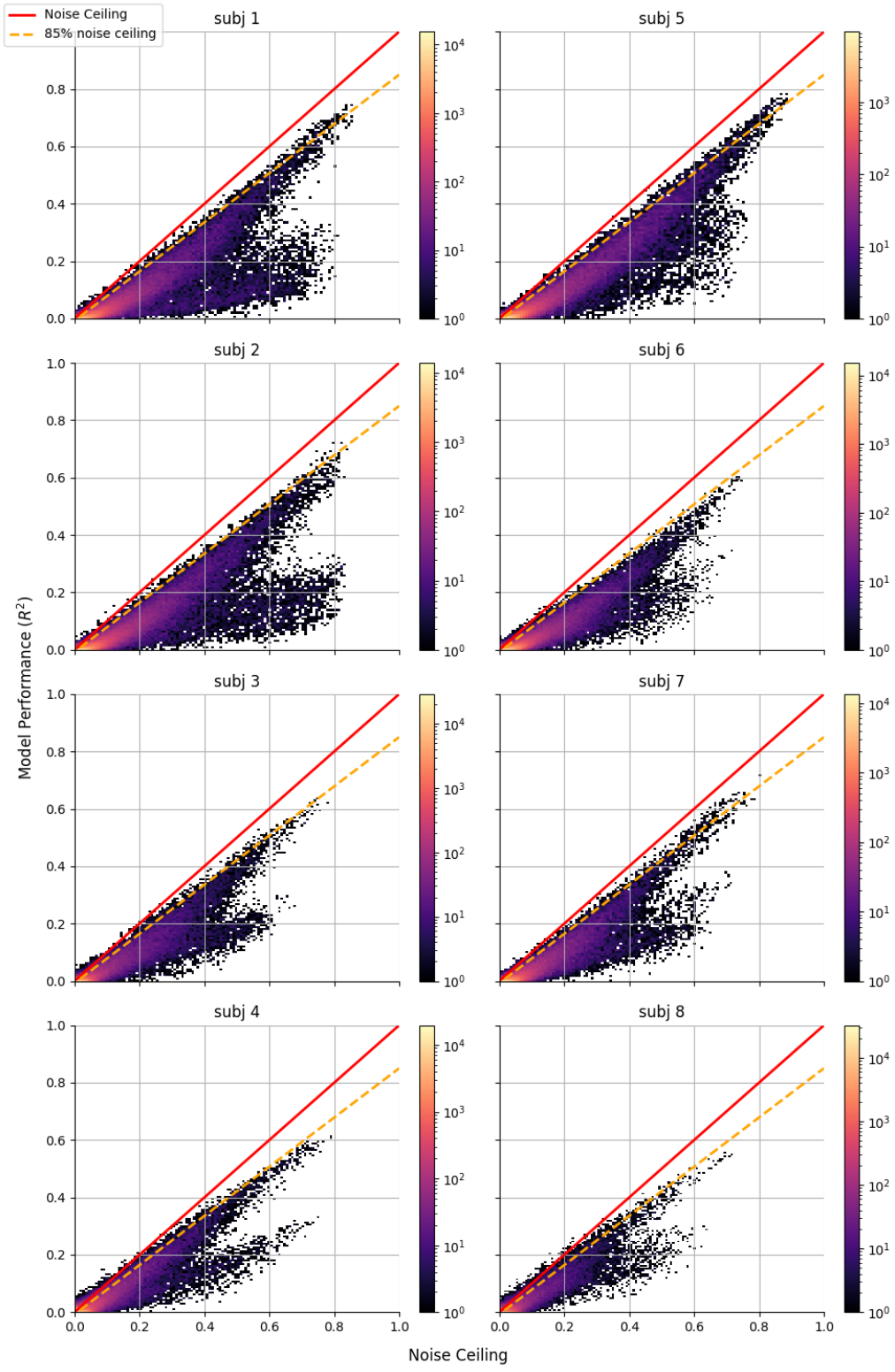



Figure 6.8: **Scatterplots of noise ceiling against model performance in $R^2$ for all subjects**.

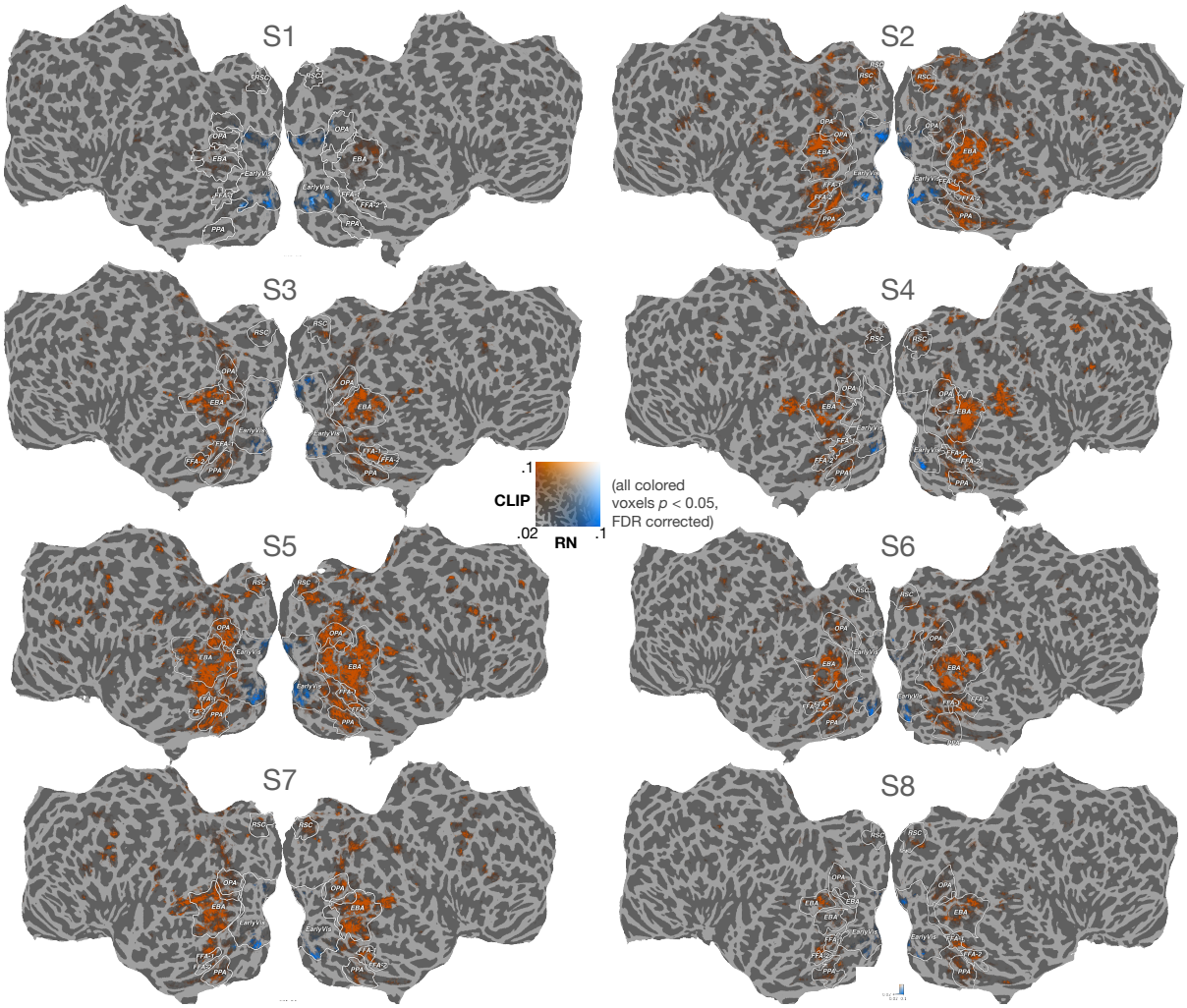



Figure 6.9: **Unique variance accounted for by ResNet$_{CLIP}$ as compared to ResNet$_{ImageNet}$ (noted as RN in the figure) for all eight subjects.** Voxels where ResNet$_{CLIP}$ accounts for greater unique variance are orange and voxels where ResNet$_{ImageNet}$ accounts for greater unique variance are blue.

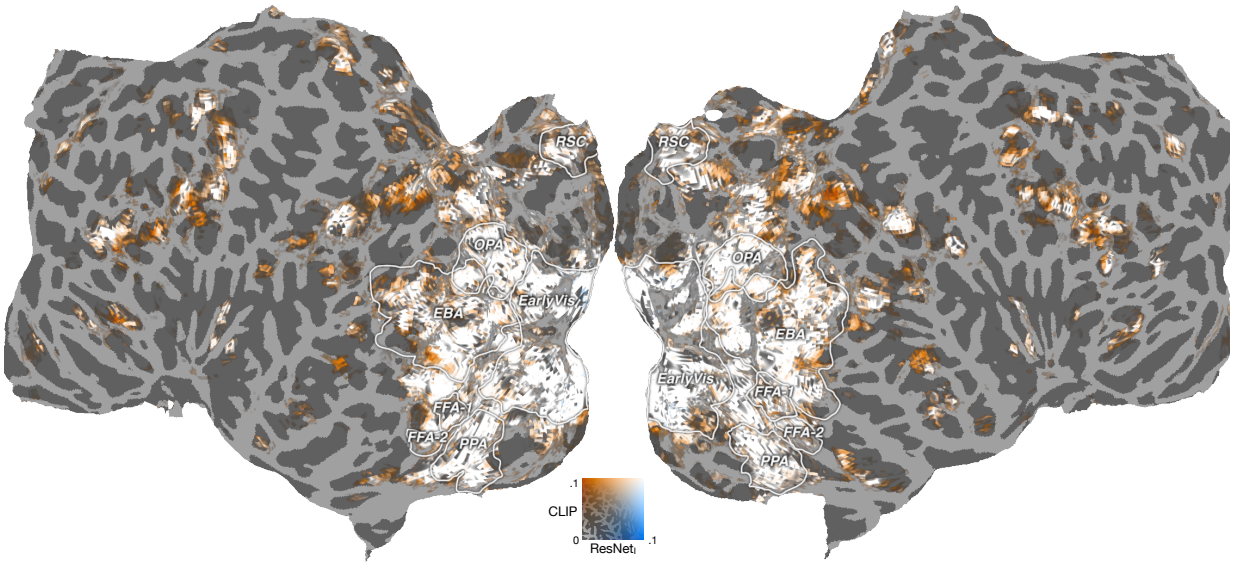



Figure 6.10: **Total variance accounted for by ResNet$_{CLIP}$ as compared to ResNet$_{ImageNet}$ for S5** Voxels where ResNet$_{CLIP}$ accounts for greater variance are orange and voxels where ResNet$_{ImageNet}$ accounts for greater variance are blue. White voxels are where both models explain well.

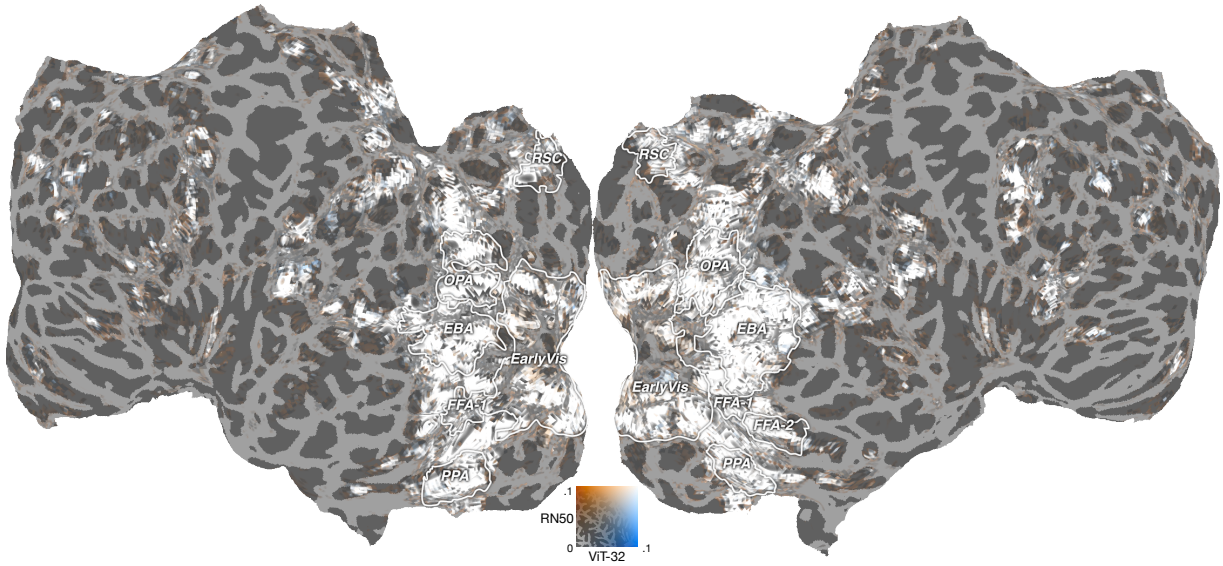



Figure 6.11: **Performance 2D map between ResNet$_{CLIP}$ and ViT-32$_{CLIP}$.**

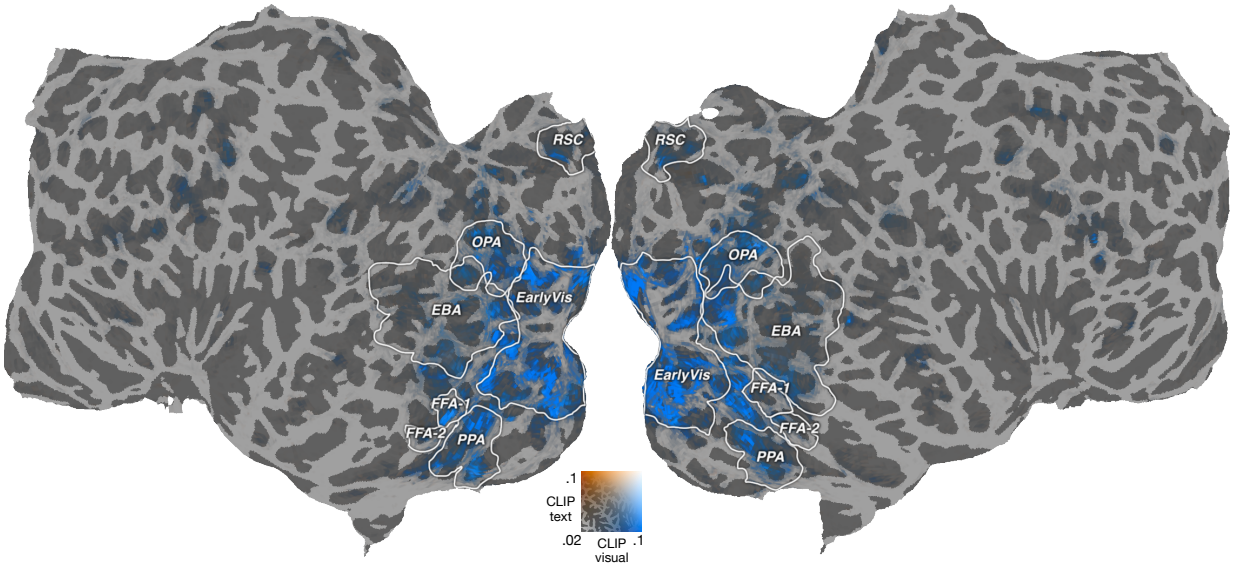



Figure 6.12: **Unique variance by CLIP visual encoder and CLIP text encoder**.

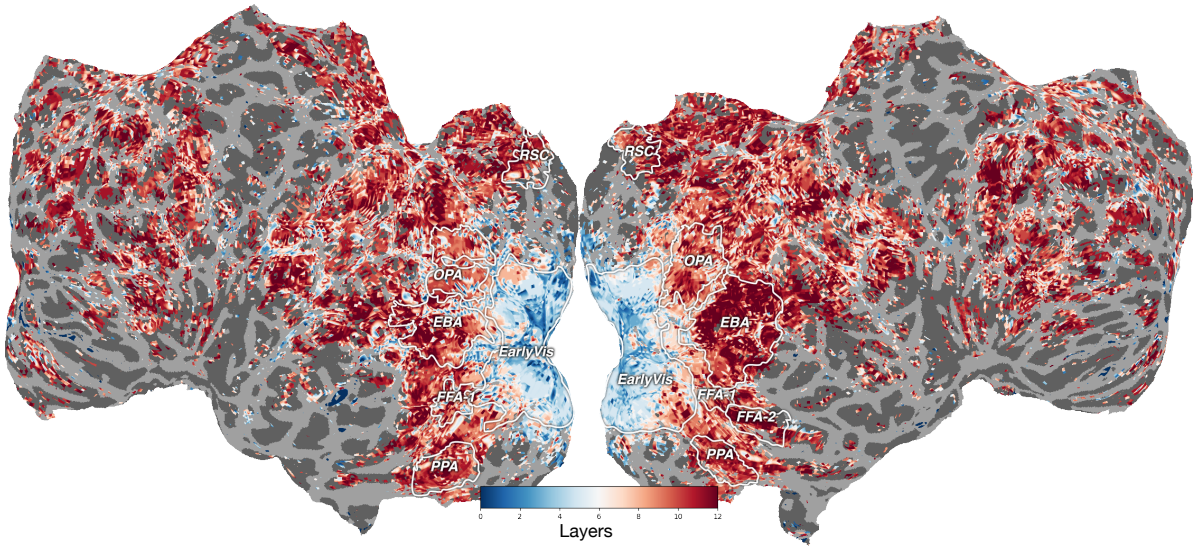



Figure 6.13: **Layer preference by voxels across the brain**.

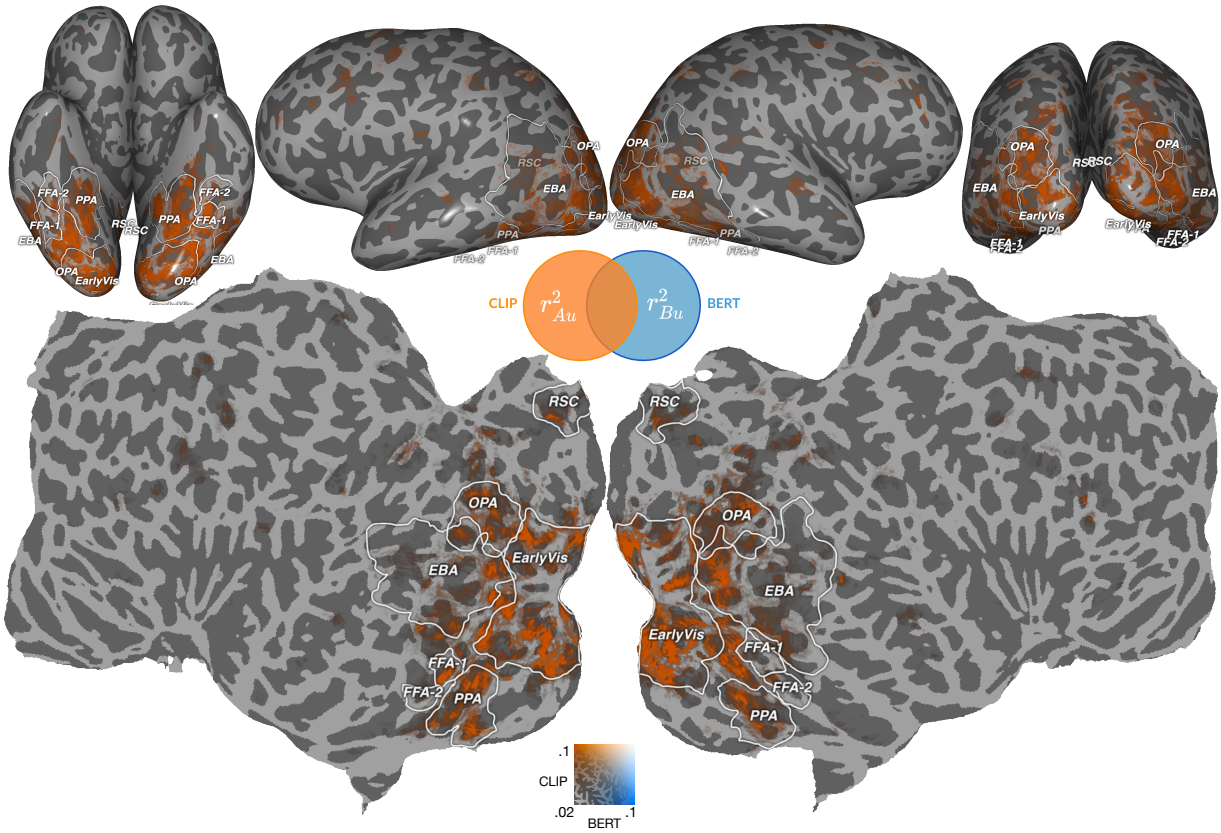



Figure 6.14: **Performance comparison between CLIP text encoder with BERT** Unique variance accounted for by CLIP as compared to BERT for S5 – obtained by subtracting $R^2$ for each model from that of the the concatenated model. Voxels where CLIP accounts for greater unique variance are orange and voxels where BERT accounts for greater unique variance are blue.

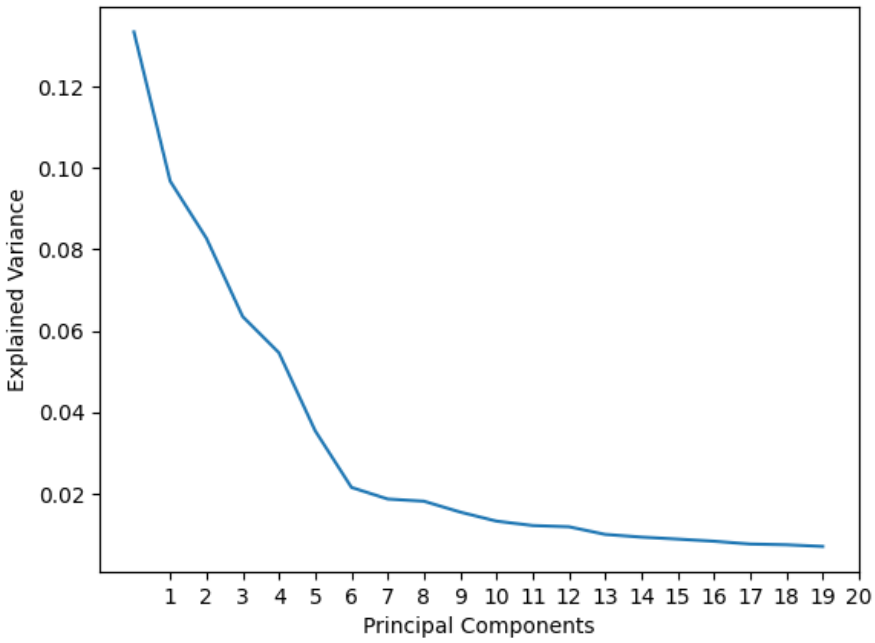



Figure 6.15: **Explainable variances across 20 PCs**


Figure 6.16: **Top 15 images for top 5 PCs**

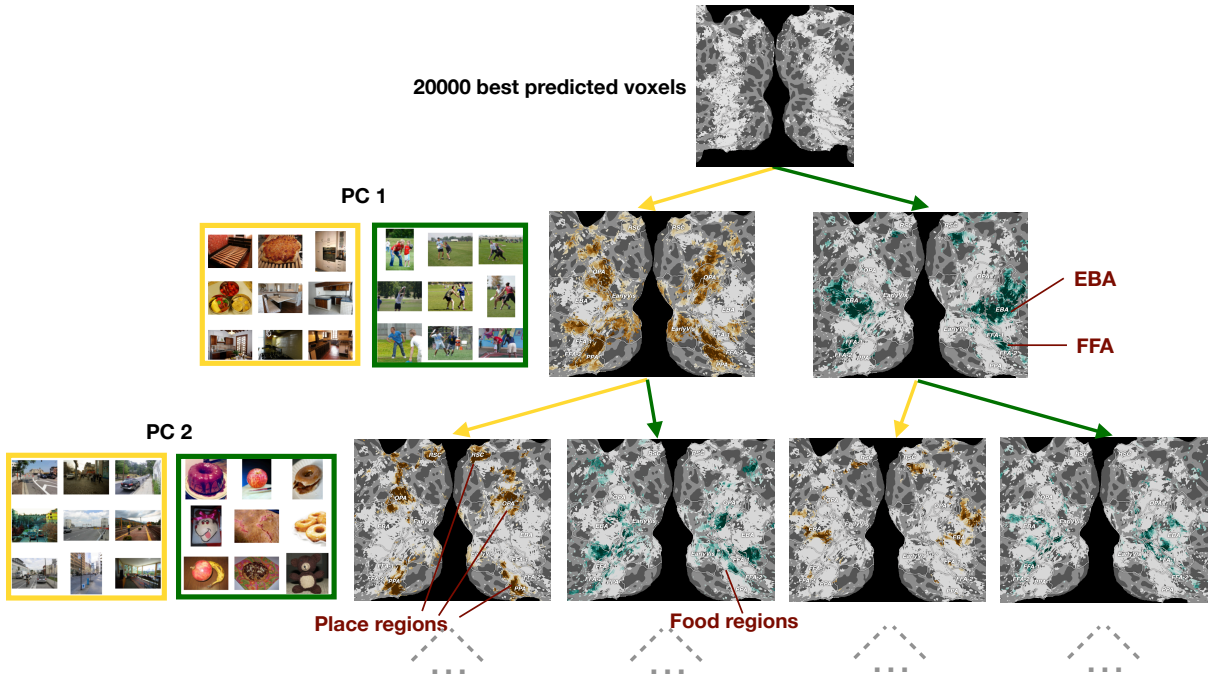



Figure 6.17: **Cortical semantic organization as revealed by the principal components of the CLIP encoding model.** Brain regions well predicted by ResNet$_{CLIP}$ can be hierarchically decomposed using the model PCs. PC1 separates animacy regions (EBA and FFA) from other regions, which are themselves separated by PC2 into place and food regions [Jain et al., 2023]. The rest of the tree is not shown due to space constraints.

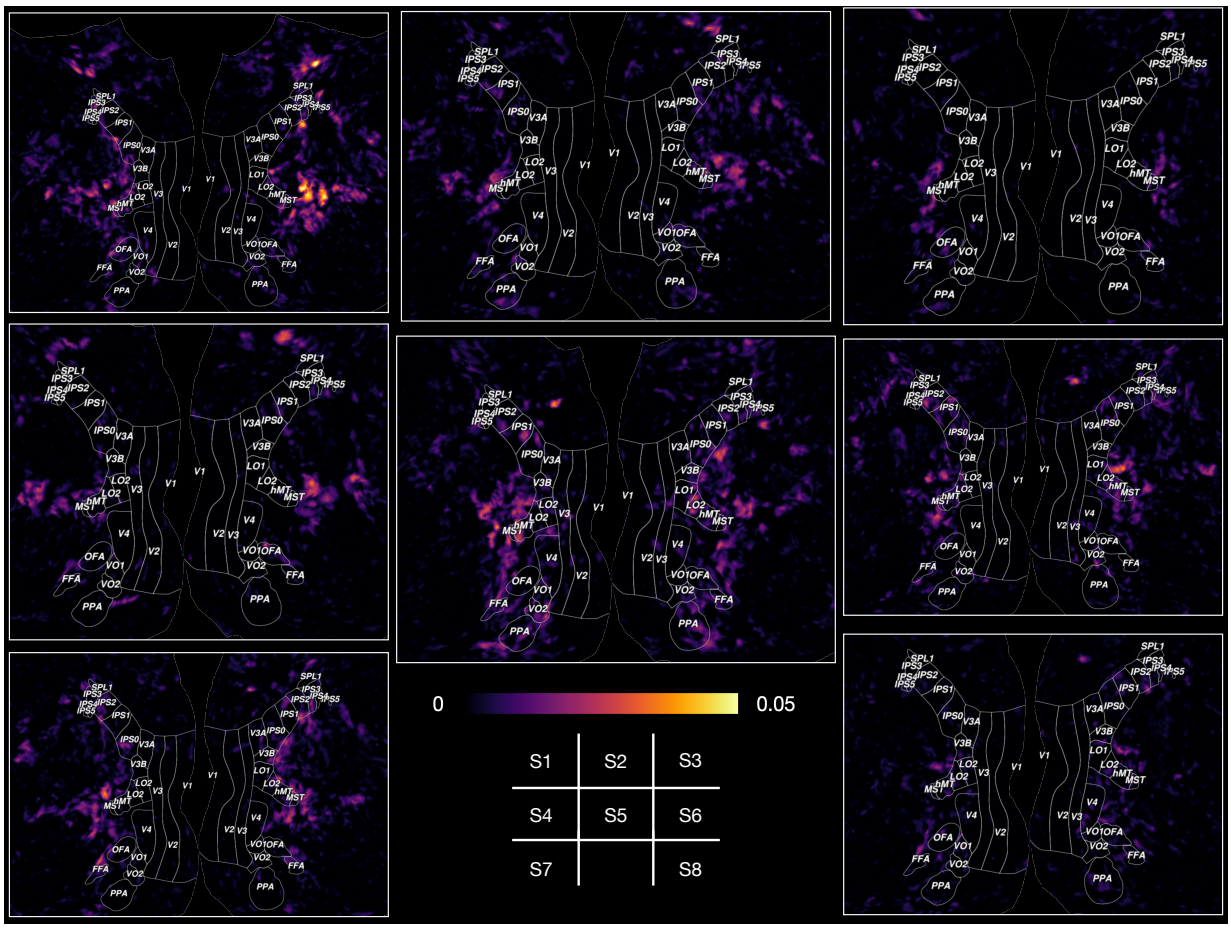



Figure 6.18: **Unique variances by YFCC SLIP compared to YFCC simCLR across all 8 subjects in MNI space.** Unique variances explained by YFCC SLIP from each subjects are projected into MNI space. Only unique variances above significance threshold are projected ($p < 0.05$, FDR-corrected [Benjamini and Hochberg, 1995].

# Chapter 7

# Interplay of language and visual representations

## 7.1 Introduction

In the prior sections, we have used task driven and multimodal models to map out the representation of visual and semantic information in the human brain. Notably, these data are collected using fMRI responses while viewing static images under free viewing (albeit with fixation) conditions. In particular, there is no manipulation of attention with respect to different scene locations or objects in the scene. However, as demonstrated by Çukur et al. [2013], different attention conditions can warp the representation of different semantic categories. For example, they show that instructing subjects to pay attention to people versus cars warps the representational space towards the attended to category so that the attended categories and its adjacent categories are more heavily represented in the brain.

The fact that neural representations change their structure with shifts in attention is relevant to our interactions with dynamic scenes. As we go about our daily life, we constantly switch attention to different elements of the scene and, sometimes, we simultaneously process information from two or more modalities while attending to both. In this chapter, we explore how

humans manage to accomplish this. The specific questions we ask include:

1. When we see and hear at the simultaneously (e.g., watching a movie), does the brain recruit different representations depending on which modality we are attending to?
2. If so, which neural mechanisms mediate switching between modalities and where in the brain do we observe representation shifts when there is a shift in attending to a given modality?

To address these questions, we employed a state-of-the-art visual and language semantics model to model brain responses arising from watching episodes of the "Friends" TV show as collected in the *Friends* movie dataset [Boyle et al., 2021], as shown in Figure 7.1. Critically, subjects were presented with simultaneous visual and audio stimuli, but were free to attend to whichever modality they preferred. One interesting aspect of *Friends* is that incongruency between the semantic content between visual and audio is common during the episodes. One common example of such incongruency is at a dinner table, where the visual inputs switch between scenes of people's faces and food, while the audio inputs carry a conversation on a wide array of topics – for example, discussions of other people, life updates, jobs, etc. The goal of this project is to use encoding models that are based on both the visual and language modalities to map out the processing of both streams of information in the brain, thereby revealing potential neural areas that have flexible semantic representations for the attended modality. We hypothesized that we would be able to observe fluctuations of model prediction between the two modalities across time, as pictured in Figure 7.2. Potential factors that we hypothesize can drive these fluctuations include: 1) presence and absence of conversation; 2) presence and absence of body and faces as the focus of the scene; 3) frequency of switching of visual frames (density of visual information); 4) congruency between linguistics and visual content. In exploring each of these hypothesized drivers, we aim to account for how the human brain is able to process multi-modal information in parallel.

This work is still ongoing at the time this thesis was written. Therefore, this chapter provides

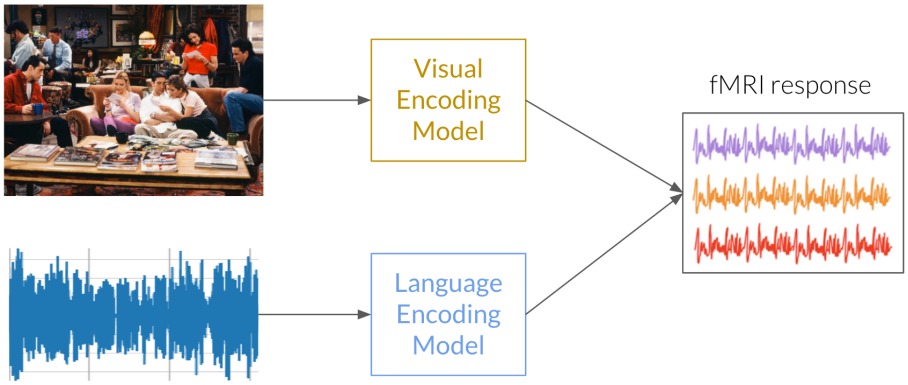



Figure 7.1: Model pipeline.

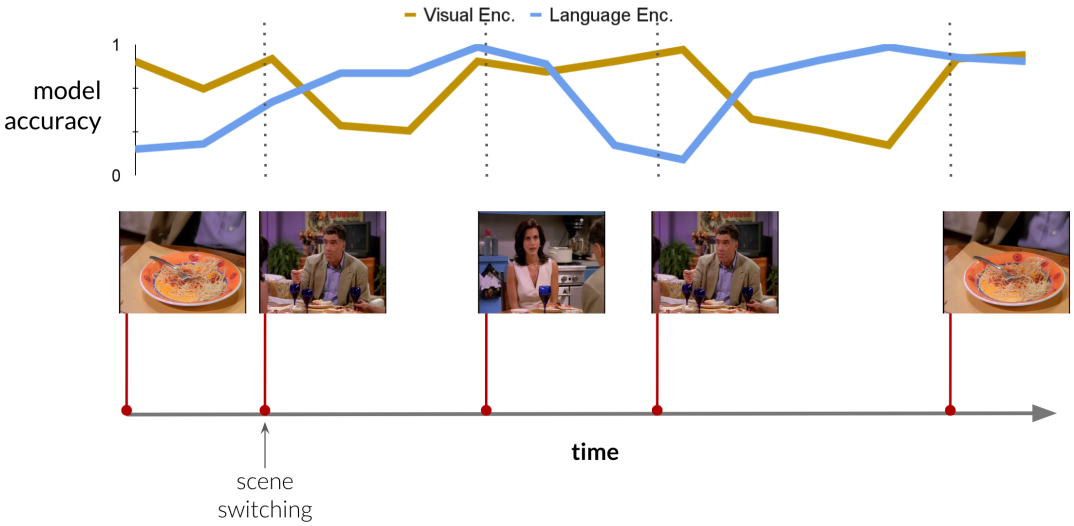



Figure 7.2: Hypothesized fluctuations of model accuracy between visual and language modalities.

an overview of the current results and presents directions for future research.

## 7.2 Methods

### 7.2.1 fMRI data

The Friends Dataset is provided by the Courtois NeuroMod group (data release cneuromod-2022)[Boyle et al., 2021]. This dataset contains functional data acquired while showing 6 subjects many episodes of the Friends TV show with English dialog. It includes seasons 1-6 for all subjects, except sub-04 who only completed seasons 1-4 (and a few segments of season 5). Each episode is cut in two segments (a/b) to allow for more flexibility while scanning and to give subjects opportunities for breaks. There is a small overlap between the segments to allow subjects to catch up with the storyline. The fMRI sampling rate (TR) was 1.49s. The data were prepossessed using fMRIPrep 20.1.0 (ref. 59). These data are available on request here.

Before fitting encoding models, fMRI data are z-scored across runs. 20 TRs are removed from the beginning of the run and 15 TRs are removed from the end of the run to ensure data quality for model fitting. Data across subjects 1-5 are averaged for better signal to noise ratio before model fitting. Data from subject 6 is reserved for future testing.

### 7.2.2 Feature Extraction

To process videos and extract the rough semantic content from each distinct scenes, we first used content based scene detect algorithms ('scenedetect' Python library) to split episodes into distinct scenes. We then input the first frame of each distinct scene into vision model and extracted the network activations. We tested the last layer of the CLIP visual encoder[Radford et al., 2021] and, as a control, the second-to-last layer of a ImageNet[Deng et al., 2009a] trained ResNet50[He et al., 2016]. To align with brain data, when more than one distinct scene (as determined by the 'scenedetect' algorithm) is present within a TR, the scene embeddings for each distinct scene are averaged.

To process the dialog in each episode, we used the transcripts provided by the Friends Dataset

and input them into a BERT model with a sequence length of 7, and extracted the last layer activations from the model for each token in the transcripts. To align the token embedding with brain data, we used a Lanczos filter to interpolate the token embedding for each TR, similar to what has been done with other language datasets [Huth et al., 2016].

### 7.2.3 Encoding Model

Features from both vision and language are delayed 2s, 4s, 6s, 8s and then concatenated together to model the hemodynamic responses in voxels. We built a ridge regression model (implemented in PyTorch; see [Koushik, 2017]) to predict one averaged fMRI response for each scene per voxel. We chose to use a ridge regression model instead of more complex models in order to retain the interpretability of model weights, which may provide insights into the underlying dimensions of the brain responses.

We did not shuffle the data but split season 1 of Friends into training and test sets with roughly a 4-to-1 ratio. For each subject, each voxel's regularization parameter were chosen independently via 7-fold cross-validation across the training set. We swept through 100 regularization parameters spaced evenly on a log scale from $10^{-8}$ to $10^{10}$, that is, np.logspace(-8, 10, 100). Cross-validation was handled by `sklearn.model_selection.KFold`, where data are split into consecutive folds without shuffling. Each fold is used once as validation while the rest of the set is used for training. Model performance was evaluated on the test data using both Pearson's correlation and coefficient of determination ($R^2$).

## 7.3 Results

The encoding models with CLIP and BERT both show reasonable performance in the whole brain across three runs in the test data (s1e23b, s1e24a, s1e24b). Averaged model performance measured in correlation (Pearson's *r*) for CLIP and BERT across these three runs is shown in

Figure 7.3 and Figure 7.4, respectively. We have also tested an encoding model based on ImageNet trained ResNet50 as an alternative visual encoding model. The prediction accuracy is lower across the brain as compared to CLIP based encoding model. More specifically, both CLIP based and BERT based encoding models showed good performance in high-level visual cortex, and language regions of interest (AG, PTL, ATL). Differences in model performance between visual (CLIP) and language (BERT) models measured in $r^2$ averaged across the 3 test runs are shown in Figure 7.5. Compared to the vision model (shown in orange), the language encoding models (shown in purple) have better prediction across the test data, especially in language ROIs. Visual encoding model based on CLIP visual encoder only have better average prediction accuracy in more ventral part of the higher visual cortex (the orange stripe in the figure).

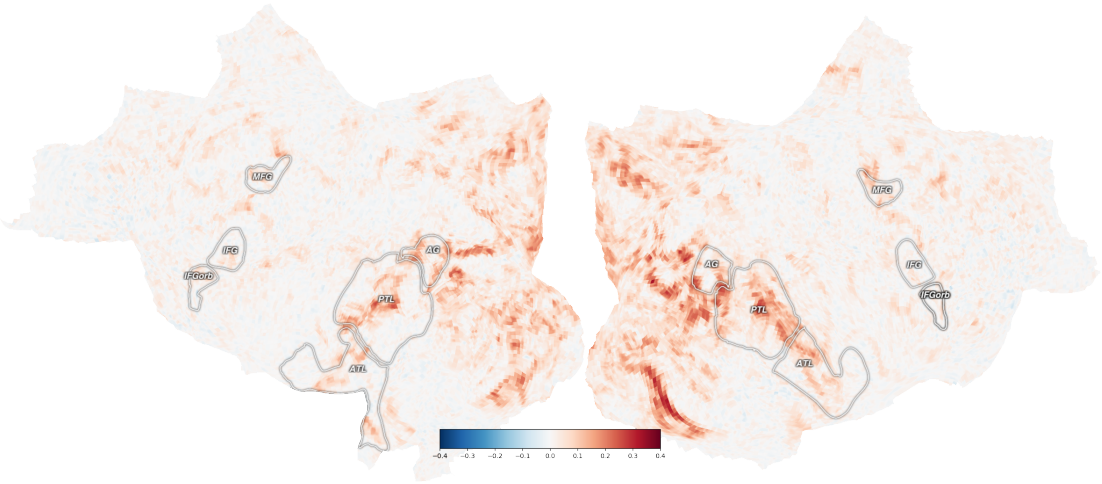



Figure 7.3: Model Performances (measured in correlation) from CLIP visual encoder in test data.

We were interested in prediction accuracy across time during the three test runs and analyzing which factors drive prediction accuracy for each modality. In Figure 7.6, we show four snapshots over s1e23b to demonstrate that visual and language encoding models show fluctuation in accuracy over time. The accuracy is evaluated over a time window of 10 TRs with a padding of 9 TR. Here we measured accuracy with both $R^2$ and negative Mean Square Error (MSE) and generally two metrics show very similar patterns in Pycortex flatmaps while $R^2$ has noisier results. As preliminary results, we found that the absence of conversation and the presence of faces (especially

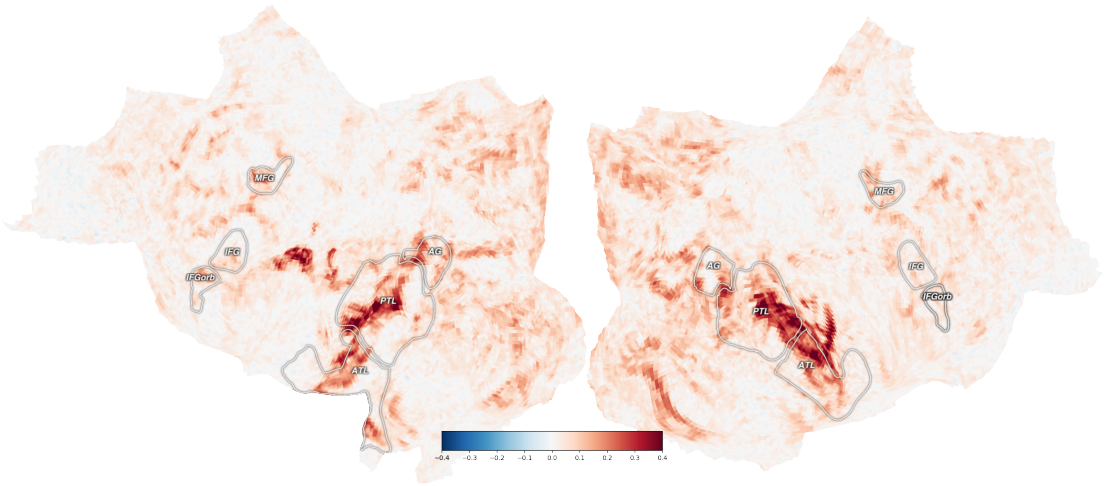



Figure 7.4: Model Performances (measured in correlation) from BERT models in test data.

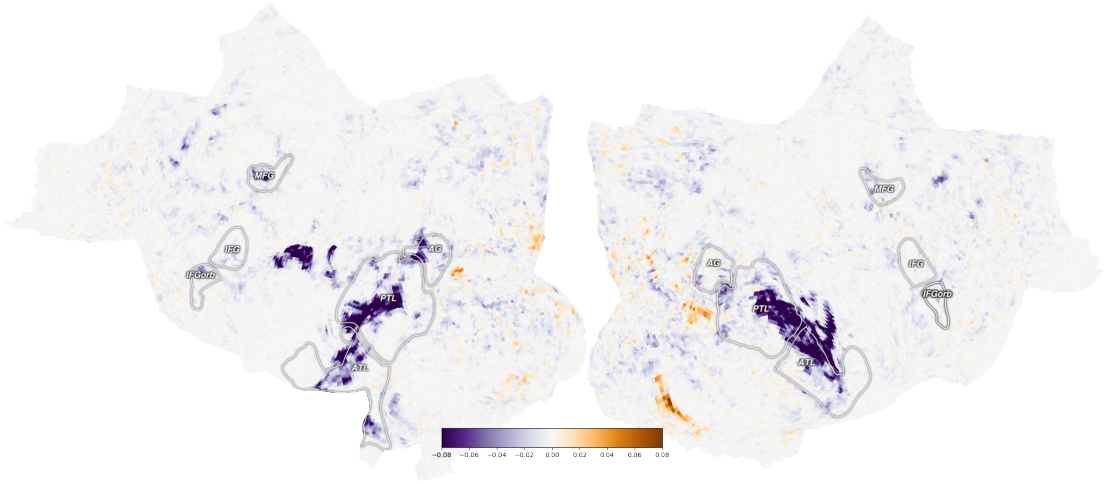



Figure 7.5: Difference in model Performances (measured in $R^2$) between CLIP visual encoder (orange) and BERT (purple) averaged across test data.

baby faces!) are indicative of better performance for the visual encoding model in the ventral cortex.

In future work, we will analyze other potential factors that we hypothesize may drive fluctuations as revealed by differences in predictions across visual and language encoding models. These factors include: 1) frequency of switching of visual frames, which we posit can describe demands in visual processing; and 2) congruency between spoken dialog and visual content. In summary, with these preliminary results, we demonstrate the potential for understanding dy-

namic scenarios such as TV watching using encoding models for multiple modalities based on task driven networks.


Figure 7.6: Four snapshots that show differences in model performances (measured in mean square error) between CLIP visual encoder (orange) and BERT (purple) in sliding time windows (T=10TR, padding=9TR) across run s1e23b. Selective frames from the movie stimuli are shown on the right and the respective differences in model accuracies are shown on the left.

# Chapter 8

# Joint interpretation of representations in neural networks and the brain

## 8.1 Motivation

Given similarity in the task end goals of both artificial and biological systems, it is not surprising that high-performing systems in both domains share representations despite drastically different physical implementations[Yamins and DiCarlo, 2016a]. More broadly, we see a similar convergence in many domains, including vision [Agrawal et al., 2014, Güçlü and van Gerven, 2015, Yamins et al., 2014b, Schrimpf et al., 2018], audition [Kell et al., 2018], language [Wehbe et al., 2014b, Jain and Huth, 2018, Caucheteux and King, 2020, Jain et al., 2020], and both feedforward and recurrent networks [Wehbe et al., 2014b, Nayebi et al., 2021].

The “explanatory arrow" has almost always been unidirectional – what can artificial neural networks and their learned representations tell us about brain representations. Implicit in this directionality is the assumption that neural networks are *good* models for neural systems; that is, that the computations implemented in neural networks help us to better understand the "black box" computations realized in different parts of the brain. Here we take a deeper look into how various choices of network in terms of layers and task the network is trained on could affects how

well these representation can predict the the brain. Our results indicate that the converse is also possible: facts about the brain can help us to better understand computations and representation in artificial neural networks.

Interest in “interpretable AI" and different methods for visualizing representation in artificial neural networks has exploded over the past several years [Bau et al., 2017, Mordvintsev et al., 2015, Olah et al., 2017, 2018]. Yet there are limitations on how much one can learn from visualization of network features - not the least of which is the human tendency to assign a greater semantic meaning and functional relevance to visualizations than otherwise might be warranted. On the other hand, there is a century long history of visual neuroscience on which we can build [Gross, 1994]. For example, Hubel and Wiesel's [Hubel and Wiesel, 1959b] elucidation of the response properties of localized receptive fields – a concept that forms the basis for almost all modern approaches to edge detection [Canny, 1986] – and the well investigated functional regions of interest (ROI) in high level vision of human discovered using fMRI that consistently serve as face and place detectors [Kanwisher et al., 1997a, Epstein et al., 1999]. Outside of the field of vision, [Toneva and Wehbe, 2019] recently demonstrated that the explanatory arrow can be reversed in the domain of language, and that brain activity during reading can be used to facilitate the interpretation of deep neural network language models. In this light, we suggest that our extensive understanding of biological vision will not only enable future advances in artificial vision systems, but that this knowledge will also enable a better understanding of the inner workings of such systems.

## 8.2 Methods

We extract learned representations from different tasks and network architectures and explored how they differ in predicting brain responses to natural images. Layerwise features are extracted from specific networks and then used to build voxelwise encoding models for the cortical area of Participant 1 in NSD. For evaluation, we calculate $R$, the square root of the *coefficient of*

*determination*, as the metric of the goodness of fit for the encoding model. We also show weights learned from the stacking algorithm for each feature.

Stacked regression method [Wolpert, 1992, Breiman, 1996] is adapted such that each encoding model used a different feature space as input. At each voxel, encoding models are trained, then the stacking algorithm learns a convex combination of the predictions of these models for that voxel. The result from stacking is a readily-interpretable combination of individual features that outperforms the performance of the best feature alone. These stacking weights indicate how features are best combined to predict the specific voxel response: generally, the fewer errors a feature makes in its respective encoding model, the higher its corresponding stacked weight; that is, the importance of that feature for prediction.

## 8.3 Results

We extracted learned representations from different tasks and network architectures and explored how they differ in predicting brain responses to natural images. Layerwise features were extracted from specific networks and then used to build voxelwise encoding models for the cortical area of Participant 1 in NSD. For evaluation, we calculated $R$, the square root of the *coefficient of determination*, as the metric of the goodness of fit for the encoding model. We also show weights learned from the stacking algorithm for each feature.

To investigate how tasks influence the representations learned by a network and their ability to predict brain data, we fixed the network architecture and compared representations learned for object and scene classification. We used AlexNet [Krizhevsky et al., 2012] pretrained on ImageNet [Deng et al., 2009a] and Places365 [Zhou et al., 2017] for each task. For each AlexNet model, we extracted features from the following 7 layers in an order consistent with the network architecture: Conv-1, Conv-2, Conv-3, Conv-4, Conv-5, FC-6, FC-7.

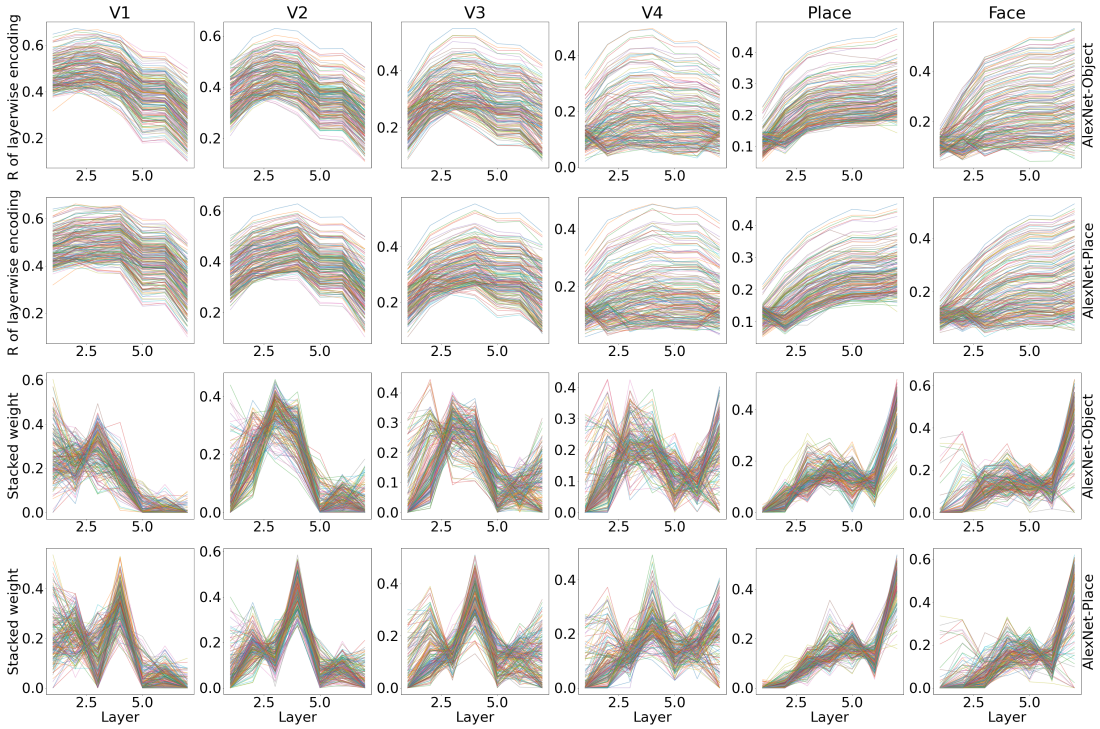
Figure 8.1 shows the result from encoding V1, V2, V3, V4 in the early visual cortex, Place ROIs (OPA, PPA, RSC) and Face ROIs (FFA, OFA, aTLface) using layerwise features from

AlexNet for object and scene classification. Within each subfigure, each individual line is prediction performances and stacking weights across features from different layers for an individual voxel. For each row of the subfigures, we can see a progression of preferred layers across ROIs. Consistent with previous results [Güçlü and van Gerven, 2015, Yamins et al., 2014b], in *AlexNet-Object* features extracted from convolution layers (Conv-2 and Conv-3) encode the early visual areas (especially V1, V2, V3) better while features extracted from fully connected layer (FC-7) outperform those from all other layers in encoding Place and Face ROIs. Comparing the first row with the second, and the third row with the fourth, we can see there is a **peak weight shift** from Conv-3 layer to Conv-4 layer as we change the task from *AlexNet-Object* to *AlexNet-Place*. This indicates that with the same architecture, change of tasks could affect how representations from network predict the brain.

Task differences observed in Alexnet do not replicate when we change the network architecture to ResNet50 [He et al., 2016] while fixing the task and dataset. From ResNet50, we extract features from the following 6 layers in the order consistent with how the network is built: Conv-1, the last layer of Conv-2 to Conv-5 blocks respectively, and the last Avgpool layer before the final layer.

Figure 8.2 shows results from voxelwise encoding models of V1-V4 in the early visual cortex, Place ROIs (OPA, PPA, RSC) and Face ROIs (FFA, OFA, aTLface) using layerwise features from ResNet50 for object and scene classification. Similar to what we see in the AlexNet results in Figure8.1, we observe the same trend that features extracted from early layers represent the early visual cortex better while features extracted from later layers represent Place and Face ROIs better. However, preferred layers by the brain as well as stacking weight are consistent between the two tasks, indicating that the additional depth of networks might lessen the influence of task in terms brain prediction and that these deeper network might just represent more information about the input that are not subject to tasks. One thing to note is that, for Face and Place area prediction, layer 4 in ResNet is assigned the largest stacking weight. Different from what we see

Figure 8.1: AlexNet encoding results. Every line corresponds to one of the best 200 encoded voxels. Each column corresponds to a visual ROI. The first and second rows are R results for *AlexNet-Object* and *AlexNet-Place* layers. The third and fourth row are the stacking weights. For V1-V4 show a reverse pattern when going from *AlexNet-Object* to *AlexNet-Place*: weights of Conv-4 layer surge and weights of Conv-3 layer plunge.




in Alexnet, where the last layer has the largest weight, it indicates that network expressiveness might the key for a good brain prediction instead of the semantic structure in the representations.

Lastly, the commonly observed pattern that early layers in networks predict early visual layers in the brain better while later layers in a network predict higher visual areas better, as shown in Güçlü and van Gerven [2015], Yamins et al. [2014b], does not hold when using representations extracted from edge detection networks. Here we extracted features from Taskonomy [Zamir et al., 2018] encoder trained for 2D and 3D edge detection. The network architecture is similar to ResNet50 but differs in replacing stride 2 convolution with stride 1 convolution in Conv-5 and removing all global average pooling. We extracted features exactly as what we do in ResNet50 but excluded Conv-4 blocks.

Figure 8.2: ResNet50 encoding results. Every line corresponds to one of the best 200 encoded voxels. Each column corresponds to a visual ROI. The first and second rows are R results for *ResNet50-Object* and *ResNet50-Place* layers. The third and fourth row are the stacking weights. Preferred layers for *ResNet50-Object* and *ResNet50-Place* are consistent across ROIs.

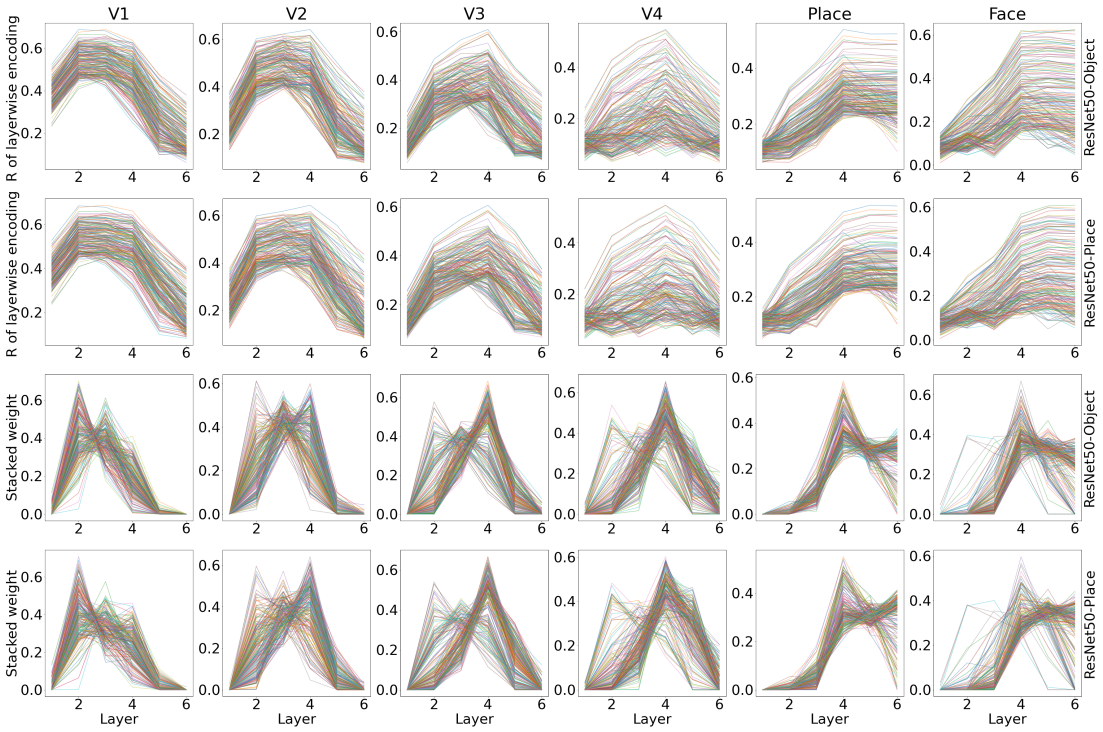
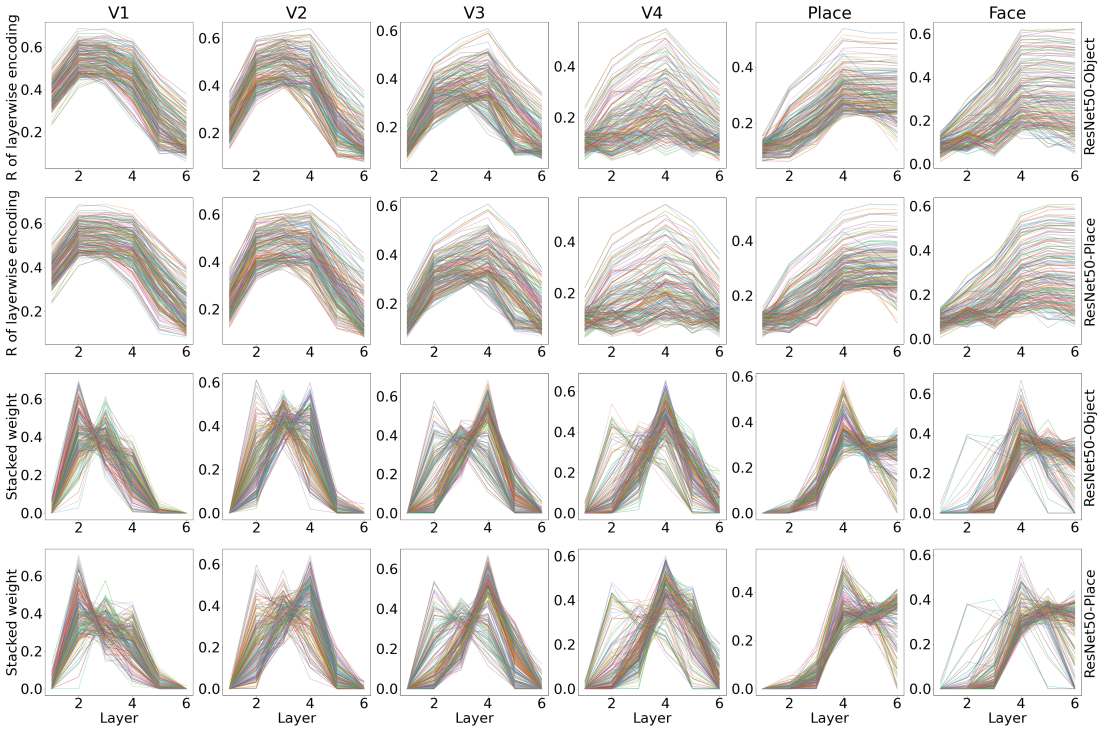



Figure 8.3 shows the encoding results for V1 to V4 in the early visual cortex, Place ROIs (OPA, PPA, RSC) and Face ROIs (FFA, OFA, aTLface) using Taskonomy features for 2D and 3D edge detection. For both *Edge2d* and *Edge3d*, early layers predict consistently better across ROIs. The overall prediction performance is lower, which is not surprising considering how little information edge detection task would normally required compared to more high level semantic tasks. What's surprising here is how early layers of edge detection networks yield higher prediction performances than the later layers even in predicting place and face areas in the brain. From the neuroscience literature, we know fairly well about the consistent responses of place and face images in Place [Epstein et al., 1999, Park and Chun, 2009, Rajimehr et al., 2011] and Face ROIs [Kanwisher et al., 1997a, 2002, Tarr and Gauthier, 2000, Gauthier et al., 2000, Grill-Spector et al., 2004] in the brain respectively. Contrary to the commonly believed view

Figure 8.3: Taskonomy edge detection network encoding results. Content in each subfigure is similar as ones in previous figures. The first and second rows are R results for *Edge2D* and *Edge3D* layers. The third and fourth row are the stacking weights. For both *Edge2d* and *Edge3d*, early layers predict consistently better across ROIs and Conv-3 layer is the most preferred in all ROIs except V1.




that a network trained to do a task should only represent variance relevant to that task [Bruna and Mallat, 2013], our result indicates that a network could possibly represent more information than what is needed in a task among the intermediate layers. Further analysis would be needed to further support this point.

## 8.4 Discussion

We observed that in predicting brain response using representations from a relatively simple neural network (i.e., AlexNet), varying the training task leads to differences in which network layers best predict the brain. Of note, this effect is network dependent and disappears when the

same comparison is done with a much larger network (i.e., ResNet50). As previously shown in multi-task learning, network capacity and expressive power [Bengio and Delalleau, 2011, Raghu et al., 2017] influences the learned task-relevant representations and affects how different tasks may be learned together [Standley et al., 2020]. Thus, our first takeaway is that network structure should be taken into consideration when mapping from network representations to the brain. A second takeaway is that, as exemplified by our results from edge detection networks, one can leverage our extensive understanding of the computations realized in different brain areas to gain a more holistic understanding of learned representations in neural networks - a step beyond visualizing randomly- or hand-picked units. Overall, the methods presented here enable a more comprehensive approach to using neural network representations to model brain function, allowing us to both better understand how choices as to network architecture and task affect predictions for biological systems and, conversely, to further interpret the learned representations realized in artificial systems.

# Chapter 9

# Conclusion

## 9.1 Summary of Contributions

This thesis built upon past works and methods on using task driven neural networks as proxy models for brain mapping, in order to understand how the brain processes visual scenes and semantics. Prior to works in this thesis, only representations from networks trained for object classification is used in prediction of brain responses to images. Works in this thesis expand this approach to multiple tasks and multiple modality to both build a more accurate and ecological model of the visual and semantic processing in the brain as well as to help the interpretation of information landscapes in the brain. Lastly, the thesis also points out the limitation of this method. We detail the contribution as follows:

- We extended the approach of using task driven neural networks as proxy models for brain mapping to a pool of 21 task-trained networks. By building encoding models using representation of each of these 21 task specific networks, we have a map of how low to high level task related visual information is represented in the brain (a.k.a Neural Taskonomy). We show similarity between tasks structure found through transfer learning and brain predictions. More generally, we propose Neural Taskonomy as a tool to help with interpreting prediction by task driven models. This methodology can be widely applied in pools of any

chosen tasks.

- We collected the first affordances (i.e. actions an objects affords to an observer) dataset with human labelers and trained a neural network to predict affordances from images.
- We used a multimodal model with language and vision pre-training (CLIP) to identified the relevant semantic dimension in high level visual cortex. We compared the prediction from this model with vision only and self-supervised models across different data set size and diversity. With the controlled experiment, we found unique variance in the brain explained by using natural language as feedback to vision models.
- We further applied task driven visual and language models in modeling brain response while subjects viewing popular TV sitcom. We showed the potentials of extending this method to modeling attend modality and semantic concepts in more realisitcs and dynamic scenario such as TV watching.
- Lastly, we examined the limitation and potential problems of using task driven methods in modeling visual representations in the brain. We showed that layer correspondence from neural networks representations to layers in human visual cortex is dependent on network architecture and tasks. Our results suggests that as the networks get more and more complex and expressive, hierarchical representation of visual feature as seen in simpler networks such as AlexNet no longer holds and thus making it harder to interpret successful predictions from network representation to brain responses.

# Bibliography

David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959a. 1, 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, 1997a. 1, 2, 3.1, 8.1, 8.3

Daniel L K 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. *Proc Natl Acad Sci U S A*, 111(23):8619–8624, 2014a. doi: 10.1073/pnas.1403112111. 1, 6.1, 6.3

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018. 1, 3.2.1, 4.1, 8.1

Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci*, 19(3):356–365, 2016a. doi: 10.1038/nn.4244. URL `syncii:///Usinggoal-drivendeeplearnin.pdf`. 1, 6.1, 6.3, 8.1

Pulkit Agrawal, Dustin Stansbury, Jitendra Malik, and Jack L Gallant. Pixels to voxels: Modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*, 2014. 1, 2, 3.2.1, 4.1, 5, 8.1

Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27): 10005–10014, 2015. 1, 2, 3.2.1, 4.1, 6.2.1, 6.3, 8.1, 8.3, 8.3

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356, 2016b. 1, 2, 3.2.1, 4.4, 5

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, pages 14928–14938, 2019. 1, 8.1

Elissa M Aminoff and Michael J Tarr. Functional Context Affects Scene Processing. *Journal of Cognitive Neuroscience*, 33(5):933–945, apr 2021. ISSN 0898-929X. doi: 10.1162/jocn_a_01694. URL `https://doi.org/10.1162/jocn_a_01694`. 1

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018. 1, 2, 4.1, 4.2.2, 4.4, 4.3.4, 4.7, 4.4, 8.3

Anna W Roe, Leonardo Chelazzi, Charles E Connor, Bevil R Conway, Ichiro Fujita, Jack L Gallant, Haidong Lu, and Wim Vanduffel. Toward a unified theory of visual area v4. *Neuron*, 74(1):12–29, 2012. 2

Jack L Gallant, Jochen Braun, and David C Van Essen. Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. *Science*, 259(5091):100–103, 1993. 2

Apostolos P Georgopoulos and Adam F Carpenter. Coding of movements in the motor cortex. *Current opinion in neurobiology*, 33:34–39, 2015. 2

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. In *Foundations in social neuroscience*. MIT Press Cambridge, MA, 2002. 2, 8.3

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598, 1998a. 2, 4.1

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, 2001a. 2

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011. 2, 3.2.1, 4.1, 6.1

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008. 2

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011. 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, 2016. 2, 3.1, 3.2.1, 3.2.3, 4.1, 6.2.1, 6.2.3, 7.2.2

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014a. 2

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. 2

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009a. 2, 5.1, 7.2.2, 8.3

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual

models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 6.1, 6.2.4, 7.2.2

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. 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, 2014b. 2, 3.2.1, 4.1, 8.1, 8.3, 8.3

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):49, 2019. 2, 3.3.1, 4.2.3, 4.4

Emily Jean Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Logan T Dowdle, Bradley Caron, Franco Pestilli, Ian Charest, J Benjamin Hutchinson, Thomas Naselaris, et al. A massive 7t fmri dataset to bridge cognitive and computational neuroscience. *bioRxiv*, 2021. 2

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 2, 8.1

Hidenori Tanaka, Aran Nayebi, Niru Maheswaranathan, Lane McIntosh, Stephen Baccus, and Surya Ganguli. From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. *Advances in neural information processing systems*, 32, 2019. 2

Dwight J Kravitz, Cynthia S Peng, and Chris I Baker. Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *Journal of Neuroscience*, 31(20): 7322–7333, 2011. 2, 4.1

Soojin Park, Talia Konkle, and Aude Oliva. Parametric coding of the size and clutter of natural scenes in the human brain. *Cerebral cortex*, 25(7):1792–1805, 2014. 2, 4.1

Assaf Harel, Dwight J Kravitz, and Chris I Baker. Deconstructing visual scenes in cortex: gradients of object and spatial layout information. *Cerebral Cortex*, 23(4):947–957, 2012. 2, 4.1

Mark D Lescroart, Dustin E Stansbury, and Jack L Gallant. Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. *Frontiers in computational neuroscience*, 9:135, 2015. 2, 4.1, 6.2.2

Katrina Ferrara and Soojin Park. Neural representation of scene boundaries. *Neuropsychologia*, 89:180–190, 2016. 2, 4.1

Frederik S Kamps, Joshua B Julian, Jonas Kubilius, Nancy Kanwisher, and Daniel D Dilks. The occipital place area represents the local elements of scenes. *Neuroimage*, 132:417–424, 2016. 2, 4.1

Simon Kornblith, Xueqi Cheng, Shay Ohayon, and Doris Y Tsao. A network for scene processing in the macaque temporal lobe. *Neuron*, 79(4):766–781, 2013. 2, 4.1

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. 2, 4.1

Mark D Lescroart and Jack L Gallant. Human scene-selective areas represent 3d configurations of surfaces. *Neuron*, 101(1):178–192, 2019. 2, 3.2.1, 3.2.3, 4.1, 4.3.1

Dustin E Stansbury, Thomas Naselaris, and Jack L Gallant. Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, 79(5):1025–1034, 2013. 2, 4.1

Russell Epstein, Alison Harris, Damian Stanley, and Nancy Kanwisher. The parahippocampal place area: recognition, navigation, or encoding? *Neuron*, 23(1):115–125, 1999. 3.1, 8.1, 8.3

Rankin Williams McGugin, J Christopher Gatenby, John C Gore, and Isabel Gauthier. High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area

related to perceptual performance. *Proceedings of the National Academy of Sciences*, 109 (42):17063–17068, 2012. 3.1

Kalanit Grill-Spector, Rory Sayres, and David Ress. High-resolution imaging reveals highly selective nonface clusters in the fusiform face area. *Nature neuroscience*, 9(9):1177–1185, 2006. 3.1

MA Woodbury. Inverting modified matrices (statistical research group, memorandum report 42), 1950. 1, 3.2.1

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. 3.2.1, 4.1

Jayanth Koushik. torch-gel. `https://github.com/jayanthkoushik/torch-gel`, 2017. 3.2.1, 4.2.1, 6.4.3, 7.2.3

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. 3.3.1, 4.2.3

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014a. 3.3.1, 4.2.3

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010. 3.3.1, 4.2.3

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience

and artificial intelligence. *Nat. Neurosci.*, 25(1):116–126, 2022. 3.3.2, 6.1, 6.2.1, 6.3, 6.4.1

Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS:740–755, 2014b. 3.3.2, 6.4.1, 6.4.1

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009b. 3.3.2, 6.1, 6.4.1

Anders M. Dale, Bruce Fischl, and Martin I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999. 3.3.2, 6.4.1

Bruce Fischl, Martin I. Sereno, and Anders M. Dale. Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207, 1999. 3.3.2, 6.4.1

James S. Gao, Alexander G. Huth, Mark D. Lescroart, and Jack L. Gallant. Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9, 2015a. 3.3.2, 6.4.1

Julie A. Boyle, Basile Pinsard, Emilie Dessureault, François Lespinasse, Francois Paugam, Pravish Sainath, Valentina Borghesani, Elizabeth DuPre, Eva Alonso Ortiz, Jonathan Armoza, Francois Nadeau, Samie-Jade Allard, Amal Boukhdhir, Agah Karakuzu, Jeni Chen, Arnaud Boré, Andre Cyr, Paul-Henri Mignot, Yann Harel, Sylvie Belleville, Simona Brambati, Julien Cohen-Adad, Adrian Fuente, Martin N. Hebart, Karim Jerbi, Pierre Rainville, and Pierre Bellec. The courtois project on neuronal modelling - 2021 data release. In *Annual Meeting of the Organization for Human Brain Mapping*, 2021. 3.3.3, 7.1, 7.2.1

Kshitij Dwivedi and Gemma Roig. Task-specific vision models explain task-specific areas of visual cortex. *BioRxiv*, page 402735, 2018. 4.1

James S Gao, Alexander G Huth, Mark D Lescroart, and Jack L Gallant. Pycortex: an interactive

surface visualizer for fmri. *Frontiers in neuroinformatics*, 9:23, 2015b. 4.3.2

JJ Gibson. The ecological approach to visual perception (pp. 127-143), 1979. 5

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 5, 8.3

Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. Mining semantic affordances of visual object categories. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 4259–4267. IEEE, 2015. 5.1

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995. 5.1

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5.3.1, 6.1, 7.2.2, 8.3

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5.3.1

Xusheng Ai, Jian Wu, Victor S Sheng, Yufeng Yao, Pengpeng Zhao, and Zhiming Cui. Best first over-sampling for multilabel classification. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1803–1806. ACM, 2015. 5.3.2

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 5.4.2

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015a. 6.1

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature computational science*, 2(11):

745–757, 2022. 6.1

Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 6.1

Elissa M. Aminoff and Michael J. Tarr. Associative processing is inherent in scene perception. *PLOS ONE*, 10(6):1–19, 06 2015. 6.1

Isabel Gauthier, Thomas W James, Kim M Curby, and Michael J Tarr. The influence of conceptual knowledge on visual discrimination. *Cogn Neuropsychol*, 20(3):507–523, 2003. 6.1, 6.3

Jonathan Schaffner, Sherry Dongqi Bao, Philippe N Tobler, Todd A Hare, and Rafael Polania. Sensory perception relies on fitness-maximizing codes. *Nat. Hum. Behav.*, 2023. 6.1

Gary Lupyan, Sharon L. Thompson-Schill, and Daniel Swingley. Conceptual penetration of visual processing. *Psychological Science*, 21(5):682–691, 2010. 6.1

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965, 2022. 6.1

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision, 2021. URL `https://arxiv.org/abs/2111.11432`. 6.1

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representa-

tion learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*. PMLR, 2021. 6.1

Wu dao 2.0. `https://gpt3demo.com/apps/wu-dao-20`. Accessed: 2022-10-20. 6.1

Steven Pinker. *The language instinct: How the mind creates language*. HarperCollins, New York, NY, 2007. 6.1

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (CLIP). In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 6216–6234. PMLR, 2022. 6.1, 6.2.4, 6.3

Barry J Devereux, Alex Clarke, and Lorraine K Tyler. Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Sci. Rep.*, 8(1):10636, 2018. 6.1

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. SLIP: Self-supervision meets language-image pre-training. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, page 529–544, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19808-3. 6.1

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 6.1

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 6.1

Scott O Murray, Huseyin Boyaci, and Daniel Kersten. The representation of perceived angular size in human primary visual cortex. *Nat Neurosci*, 9(3):429–434, 2006. 6.1

Charles D Gilbert and Wu Li. Top-down influences on visual processing. *Nat. Rev. Neurosci.*, 14(5):350–363, 2013. 6.1

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 6.1

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*. JMLR.org, 2020. 6.1

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa 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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc., 2022a. 6.1, 6.2.4

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, jan 2016. 6.1, 6.2.4

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995. 6.2.1, 6.2.4, 6.18

Russell A. Epstein and Chris I. Baker. Scene perception in the human brain. *Annual Review of*

*Vision Science*, 5(1):373–397, 2019. 6.2.1

P E Downing, Y Jiang, M Shuman, and N Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001b. 6.2.1

J Sergent, S Ohta, and B MacDonald. Functional neuroanatomy of face and object processing: A positron emission tomography study. *Brain*, 115:15–36, 1992. 6.2.1

N Kanwisher, J McDermott, and M M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci*, 17(11):4302–4311, 1997b. 6.2.1

Wendy A de Heer, Alexander G Huth, Thomas L Griffiths, Jack L Gallant, and Frédéric E Theunissen. The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, 37(27):6539–6557, 2017. 6.2.2

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. 6.2.2

Tolga Çukur, Shinji Nishimoto, Alexander G Huth, and Jack L Gallant. Attention during natural vision warps semantic representation across the human brain. *Nature neuroscience*, 16(6): 763–770, 2013. 6.2.3, 7.1

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. *Commun. Biol.*, 6(1):175, 2023. 6.2.3, 6.17

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, 2023. 6.2.3

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, 2022. 6.2.3

Colin Conwell, Jacob S Prince, Christopher J Hamblin, and George A Alvarez. Controlled assessment of clip-style language-aligned vision models in prediction of brain & behavioral data. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*. 6.2.4, 6.3

Daniel D. Leeds, Darren A. Seibert, John A. Pyles, and Michael J. Tarr. Comparing visual representations across human fMRI and computational vision. *Journal of Vision*, 13(13):25–25, 11 2013. 6.2.4

Colin Conwell, Jacob S. Prince, George A. Alvarez, and Talia Konkle. Large-scale benchmarking of diverse artificial vision models in prediction of 7T human neuroimaging data. *bioRxiv*, 2022a. 6.3

Colin Conwell, Jacob Prince, George Alvarez, Talia Konkle, and Kendrick Kay. Opportunistic experiments on a large-scale survey of diverse artificial vision models in prediction of 7t human fMRI data. In *Conference on Cognitive Computational Neuroscience*, 2022b. URL `https://2022.ccneuro.org/proceedings/0000070.pdf`. 6.3

Stefania Bracci and Hans P. Op de Beeck. Understanding human object vision: A picture is worth a thousand representations. *Annual Review of Psychology*, 74(1):113–135, 2023. 6.3

Martin Maier and Rasha Abdel Rahman. No matter how: Top-down effects of verbal and semantic category knowledge on early visual perception. *Cognitive, Affective, & Behavioral Neuroscience*, 19(4):859–876, 2019. 6.3

Ian Charest, Emily Allen, Yihan Wu, Thomas Naselaris, and Kendrick Kay. Precise identification of semantic representations in the human brain. *Journal of Vision*, 20(11):539–539, 2020. 6.3

Rebecca Nappa, Allison Wessel, Katherine L. McEldoon, Lila R. Gleitman, and John C. Trueswell. Use of Speaker's Gaze and Syntax in Verb Learning. *Lang. Learn. Dev.*, 5(4): 203–234, sep 2009. ISSN 1547-5441. doi: 10.1080/15475440903167528. URL `https:`

//pubmed.ncbi.nlm.nih.gov/24465183/. 6.3

Sandra R. Waxman and Dana B. Markow. Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cogn. Psychol.*, 29(3):257–302, 1995. ISSN 0010-0285. doi: 10.1006/COGP.1995.1016. URL https://pubmed.ncbi.nlm.nih.gov/8556847/. 6.3

Gary Lupyan, David H. Rakison, and James L. McClelland. Language is not just for talking: redundant labels facilitate learning of novel categories. *Psychol. Sci.*, 18(12):1077–1083, dec 2007. ISSN 0956-7976. doi: 10.1111/J.1467-9280.2007.02028.X. URL https://pubmed.ncbi.nlm.nih.gov/18031415/. 6.3

Anna Shusterman and Elizabeth Spelke. Language and the Development of Spatial Reasoning. In Peter Carruthers, Stephen Laurence, and Stephen Stich, editors, *The Innate Mind: Structure and Contents*, pages 89–106. Oxford University Press, 07 2005a. 6.3

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 2018. doi: 10.1101/407007. 8.1

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014b. 8.1

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In *Advances in neural information processing systems*, pages 6628–6637, 2018. 8.1

Charlotte Caucheteux and Jean-Rémi King. Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv*, 2020. 8.1

Shailee Jain, Vy A Vo, Shivangi Mahto, Amanda LeBel, Javier S Turek, and Alexander G Huth. Interpretable multi-timescale models for predicting fmri responses to continuous natural

speech. *Advances in Neural Information Processing Systems*, 2020. 8.1

Aran Nayebi, Javier Sagastuy-Brena, Daniel M Bear, Kohitij Kar, Jonas Kubilius, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel LK Yamins. Goal-driven recurrent neural network models of the ventral visual stream. *bioRxiv*, 2021. 8.1

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015. URL `https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html`. 8.1

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization. 8.1

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. doi: 10.23915/distill.00010. https://distill.pub/2018/building-blocks. 8.1

C G Gross. How inferior temporal cortex became a visual area. *Cerebral Cortex*, 4(5):455–469, 1994. 8.1

D H Hubel and T N Wiesel. Receptive fields of single neurons in the cat's striate cortex. *J. Physiol.*, 148:574–591, 1959b. 8.1

J Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, June 1986. ISSN 0162-8828. doi: 10.1109/TPAMI.1986.4767851. URL `https://doi.org/10.1109/TPAMI.1986.4767851`. 8.1

David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992. 8.2

Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996. 8.2

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 8.3

Soojin Park and Marvin M Chun. Different roles of the parahippocampal place area (ppa) and

retrosplenial cortex (rsc) in panoramic scene perception. *Neuroimage*, 47(4):1747–1756, 2009. 8.3

Reza Rajimehr, Kathryn J Devaney, Natalia Y Bilenko, Jeremy C Young, and Roger BH Tootell. The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys. *PLoS Biol*, 9(4):e1000608, 2011. 8.3

Michael J Tarr and Isabel Gauthier. Ffa: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature neuroscience*, 3(8):764–769, 2000. 8.3

Isabel Gauthier, Michael J Tarr, Jill Moylan, Pawel Skudlarski, John C Gore, and Adam W Anderson. The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of cognitive neuroscience*, 12(3):495–504, 2000. 8.3

Kalanit Grill-Spector, Nicholas Knouf, and Nancy Kanwisher. The fusiform face area subserves face perception, not generic within-category identification. *Nature neuroscience*, 7(5):555–562, 2004. 8.3

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013. 8.3

Yoshua Bengio and Olivier Delalleau. On the expressive power of deep architectures. In *International conference on algorithmic learning theory*, pages 18–36. Springer, 2011. 8.4

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *international conference on machine learning*, pages 2847–2854. PMLR, 2017. 8.4

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pages 9120–9132. PMLR, 2020. 8.4

Sarah E. Koopman, Bradford Z. Mahon, and Jessica F. Cantlon. Evolutionary Constraints on Human Object Perception. *Cognitive Science*, 41(8):2126–

2148, nov 2017. ISSN 1551-6709. doi: 10.1111/COGS.12470. URL https://onlinelibrary.wiley.com/doi/full/10.1111/cogs.12470https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12470https://onlinelibrary.wiley.com/doi/10.1111/cogs.12470.

Soojin Park, Timothy F Brady, Michelle R Greene, and Aude Oliva. Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *Journal of Neuroscience*, 31(4):1333–1340, 2011.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

L G Ungerleider and M Mishkin. Two cortical visual systems. *Analysis of visual behavior*, page 549–586, 1982.

D Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, 1982.

Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015b. doi: 10.1038/nature14539. NULL.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

Anna Shusterman and ES Spelke. Language and the development of spatial reasoning. *The innate mind: Structure and contents*, pages 89–106, 2005b.

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.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Richard Antonello, Javier S Turek, Vy Vo, and Alexander Huth. Low-dimensional structure in the space of language representations is reflected in brain responses. *Advances in Neural Information Processing Systems*, 34:8332–8344, 2021.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, pages 1–12, 2023.

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.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training, 2021. URL `https://arxiv.org/abs/2112.03857`.

Ian Morgan Leo Pennock, Chris Racey, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, Anna Franklin, and Jenny Bosten. Color-biased regions in the ventral visual pathway are food-selective. *bioRxiv*, 2022.

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

P E Downing, A W Chan, M V Peelen, C M Dodds, and N Kanwisher. Domain specificity in visual cortex. *Cerebral cortex (New York, N.Y. : 1991)*, 16:1453–1461, 2006.

R Epstein and N Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998b.

Emilie L. Josephs and Talia Konkle. Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex. *Proceedings of the National Academy of Sciences*, 117(47):29354–29362, 2020.

Talia Konkle and Alfonso Caramazza. Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, 33(25):10235–10242, 2013.

Elissa M Aminoff, Kestutis Kveraga, and Moshe Bar. The role of the parahippocampal cortex in cognition. *Trends Cogn Sci*, 17(8):379–390, 2013.

Grace W. Lindsay. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 33(10):2017–2031, 2021.

Mariya Toneva, Tom M. Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals new aspects of meaning composition. *bioRxiv*, 2020.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008a.

Tiwalayo Eisape, Roger Levy, Joshua B. Tenenbaum, and Noga Zaslavsky. Toward human-like object naming in artificial neural systems. In *International Conference on Learning Representations (ICLR 2020), Bridging AI and Cognitive Science workshop*, Virtual conference (due to Covid-19), 04/2020 2020.

Kevin S Weiner and Kalanit Grill-Spector. Not one extrastriate body area: using anatomical landmarks, hmt+, and visual field maps to parcellate limb-selective activations in human lateral occipitotemporal cortex. *Neuroimage*, 56(4):2183–2199, 2011.

Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

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

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*, 2022b.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. SLIP: Self-supervision meets language-image pre-training, 2021.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

William E Vinje and Jack L Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Emin Celik, Umit Keles, İbrahim Kiremitçi, Jack L Gallant, and Tolga Cukur. Cortical networks of dynamic scene category representation in the human brain. *cortex*, 143:127–147, 2021.

Mo Shahdloo, Emin Çelik, Burcu A Urgen, Jack L Gallant, and Tolga Çukur. Task-dependent warping of semantic representations during search for visual action categories. *Journal of Neuroscience*, 42(35):6782–6799, 2022.