Discovering visual categorical selectivity across the whole brain in silico using transformer-based encoders and large-scale generative models

Ethan Hwang{eh2976@columbia.edu}, Hossein Adeli, Wenxuan Guo, Andrew Luo, Nikolaus Kriegeskorte
Zuckerman Mind Brain Behavior Institute, Columbia University, New York, USA

Abstract

The human visual cortex contains several regions that selectively respond to particular categories (e.g. faces, places, bodies). However, it is unclear whether there are regions responsive to additional (possibly more complex) categories, either inside or outside the visual cortex. Jointly discovering the categories and the corresponding selective regions, without relying on the researchers' biased imagination, remains a methodological challenge. Here, we take an in-silico approach to discovering category-selective regions. We trained a state-of-the-art transformer-based encoding model that predicts neural responses from natural scenes. We then used this model to generate hypotheses about category-selectivity of different regions throughout the human brain by performing in-silico mapping, using large amounts of computation. We use diffusion-based generative models and retrieval from large image datasets to find images that maximally activate different parcels. We found many parcels with complex selectivity, transcending simple categorical concepts: scenes with multiple objects (sport events), specific subcategories (places with vanishing points or parallel lines), and specific interactions (tool use). Our study demonstrates a data-driven paradigm for discovery of visual selectivity for each region with sets of optimal images. The category-selectivity hypotheses generated can be tested in future fMRI experiments.

Keywords: Neural encoding; brain mapping; transformers; generative diffusion model

Introduction

Recent decades have seen great progress in understanding the brain's visual hierarchy: how neurons map low-level features such as orientation to mid-level categorical concepts. Extensive neuroimaging experiments, especially functional magnetic resonance imaging (fMRI), have mapped prominent categories—faces, places, words, bodies, and food—to dedicated brain regions. But visual perception goes beyond simple categories and it remains open what higher-level visual concepts enable humans to make sense of the complex world.

Common mapping methods however are limited to experimenter-curated concepts, and empirically-driven alternatives require more data and expensive fMRI experiments. Our encoding model, leveraging recent advances in AI and large-scale neural datasets, serves as a "digital twin" that is fully observable, upon which we perform extensive experimentation to better hypothesize neuron selectivity beyond the visual cortex. These hypotheses can motivate targeted future fMRI experiments.

Model Design

Parcellation Strategy

We divided up the 327,684 cortical vertices across the whole brain into 1,000 regions using the Schaefer-1000 functional connectivity-based parcellation (Schaefer et al., 2018).

Brain Encoder Architecture

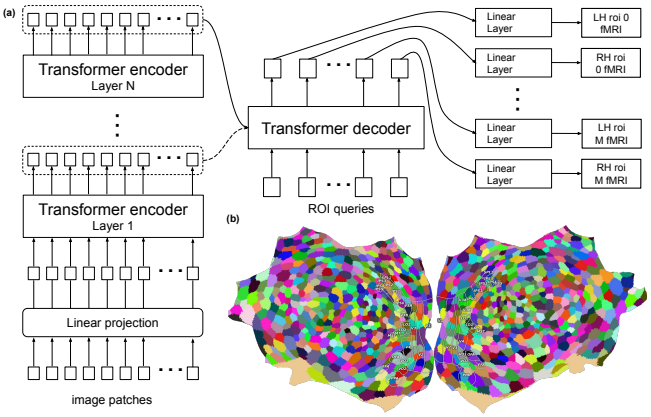

Figure 1: (a) Brain encoder (b) Schaefer-1000 parcellation

Extending the work of Adeli, Minni, and Kriegeskorte (2023), our brain encoder predicts vertex-wise activity for the whole brain from an input image. Patch embeddings are extracted from a DINOv2 (Oquab et al., 2024) backbone. The transformer decoder uses parcel-specific queries to attend to relevant patch embeddings via cross attention. The resulting representations are linearly mapped to predict neural activity.

To improve the accuracy of our predictions, we ensembled several instances of the brain encoder. For each subject, we trained two random seeds with features from four different DINOv2 backbone layers (0, 2, 4, 6 layers from last). To predict a vertex, we take the weighted average across model predictions, scaled by the confidence of each model for that vertex based on validation set accuracy. We train these models on the Natural Scenes Dataset (Allen et al., 2022), the largest fMRI dataset to date, with up to 10,000 images per subject.

Model Results

Prediction Accuracy

Fig. 2 shows the encoding accuracy of our ensemble model for subject 1 projected onto the cortical surface using Pycortex (Gao, Huth, Lescroart, & Gallant, 2015). As expected,

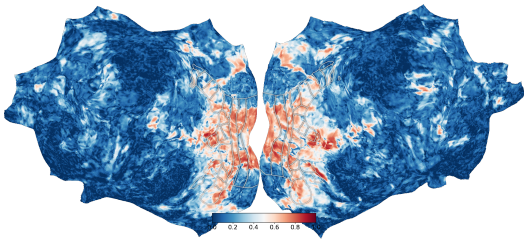

Figure 2: Pearson correlation between model predictions and ground truth data for subject 1 on the held-out test set

the model performs well on predicting the activity in the visual cortex, but also on several regions beyond the typical visual pathways. We first validate our paradigm by replicating the demonstrated selectivity of ventral pathway categorical areas, and then move on to areas beyond the visual cortex. For our exploration, we choose parcels that have high noise ceiling (high visual selectivity) and are well-predicted by our model.

Superstimulus Generation Process

We choose images that maximally activate (mean z-scored betas) a parcel of interest using three different methods: (1) held-out NSD images based on ground truth data. (2) the BrainDIVE model (Luo, Henderson, Wehbe, & Tarr, 2023) (a generative backbone guided using the gradient from a brain encoder, to generate images that can maximally activate certain brain regions). (3) Imagenet (Deng et al., 2009) images that maximally activate the parcel, according to the encoder.

Sanity Check on Known Regions

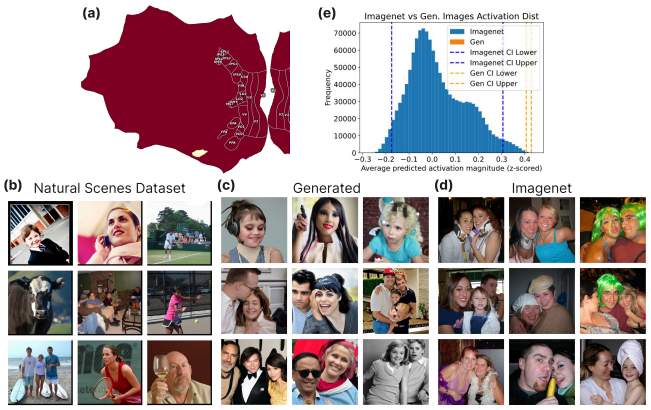

Figure 3: **(a)** The location of the parcel. **(b)** Held-out NSD that maximally activate the parcel (based on ground-truth fMRI). **(c)** BrainDIVE generated and reranked (top 2.25%) images optimized for the parcel. **(d)** Maximally activating images from Imagenet according to the encoder. **(b)** Distribution of parcel activation for all of Imagenet compared to images in (c).

The sample parcel we chose significantly overlaps with aTL-faces (47.6% of the vertices in the parcel overlap with aTL-faces). As shown in fig. 3, all the images prominently feature faces, which agrees with previous work on the selectivity of this area (Sergent, Ohta, & Macdonald, 1992).

Complex Cross-subject Selectivity

Figure 4: **(a)** Parcel location. **(b)** A comparison of the activation magnitude of the parcel from all Imagenet images and the top 9 images, with 95% confidence intervals. **(c)** Selected from the 9 images in Imagenet that maximally activate the parcel of interest, across subjects 1, 2, 5, and 7. **(d)** ChatGPT 4o (04/06/2025) labels for top-25 Imagenet images.

We now explore the selectivity of a parcel outside the visual cortex (fig. 4a) by examining images from Imagenet predicted by the encoder to maximally activate this parcel. The maximally activating images fall substantially outside the distribution of activations of the parcel by the rest of Imagenet (fig. 4b). The selected images appear to depict hands with tools, such as a writing or cooking utensil. When prompted to describe a unifying theme in the top 25 Imagenet images, ChatGPT identifies hands with objects as a prominent theme. Recent work has identified areas that represent tool use in similar areas of the brain (Cortinovis, Peelen, & Bracci, 2025).

Discussion

Leveraging recent advances in AI and the availability of large-scale datasets, we demonstrate a data-driven paradigm to discover the selectivity of parcels beyond the visual cortex, paving the way for more systematic labeling of the whole brain to understand higher-order visual processing.

References

Adeli, H., Minni, S., & Kriegeskorte, N. (2023, August). *Predicting brain activity using Transformers.* Neuroscience. Retrieved 2025-02-28, from http://biorxiv.org/lookup/doi/10.1101/2023.08.02.551743 doi: 10.1101/2023.08.02.551743

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... Kay, K. (2022, January). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. Retrieved 2025-02-28, from https://www.nature.com/articles/s41593-021-00962-x doi: 10.1038/s41593-021-00962-x

Cortinovis, D., Peelen, M. V., & Bracci, S. (2025, March). Tool Representations in Human Visual Cortex. *Journal of Cognitive Neuroscience*, *37*(3), 515–531. Retrieved 2025-04-10, from https://doi.org/10.1162/jocn_a_02281 doi: 10.1162/jocn_a_02281

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009, June). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). Miami, FL: IEEE. Retrieved 2025-02-28, from https://ieeexplore.ieee.org/document/5206848/ doi: 10.1109/CVPR.2009.5206848

Gao, J. S., Huth, A. G., Lescroart, M. D., & Gallant, J. L. (2015, September). Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, *9*. Retrieved 2025-02-28, from http://journal.frontiersin.org/Article/10.3389/fninf.2015.00023/abstract doi: 10.3389/fninf.2015.00023

Luo, A. F., Henderson, M. M., Wehbe, L., & Tarr, M. J. (2023, November). *Brain Diffusion for Visual Exploration: Cortical Discovery using Large Scale Generative Models.* arXiv. Retrieved 2025-02-26, from http://arxiv.org/abs/2306.03089 (arXiv:2306.03089 [cs]) doi: 10.48550/arXiv.2306.03089

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... Bojanowski, P. (2024, February). *DINOv2: Learning Robust Visual Features without Supervision.* arXiv. Retrieved 2025-02-28, from http://arxiv.org/abs/2304.07193 (arXiv:2304.07193 [cs]) doi: 10.48550/arXiv.2304.07193

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., ... Yeo, B. T. T. (2018, September). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex*, *28*(9), 3095–3114. Retrieved 2025-04-05, from https://academic.oup.com/cercor/article/28/9/3095/3978804 doi: 10.1093/cercor/bhx179

Sergent, J., Ohta, S., & Macdonald, B. (1992, February). Functional Neuroanatomy of Face and Object Processing: A Positron Emission Tomography Study. *Brain*, *115*(1), 15–36. Retrieved from https://doi.org/10.1093/brain/115.1.15 (_eprint: https://academic.oup.com/brain/article-pdf/115/1/15/836448/115-1-15.pdf) doi: 10.1093/brain/115.1.15