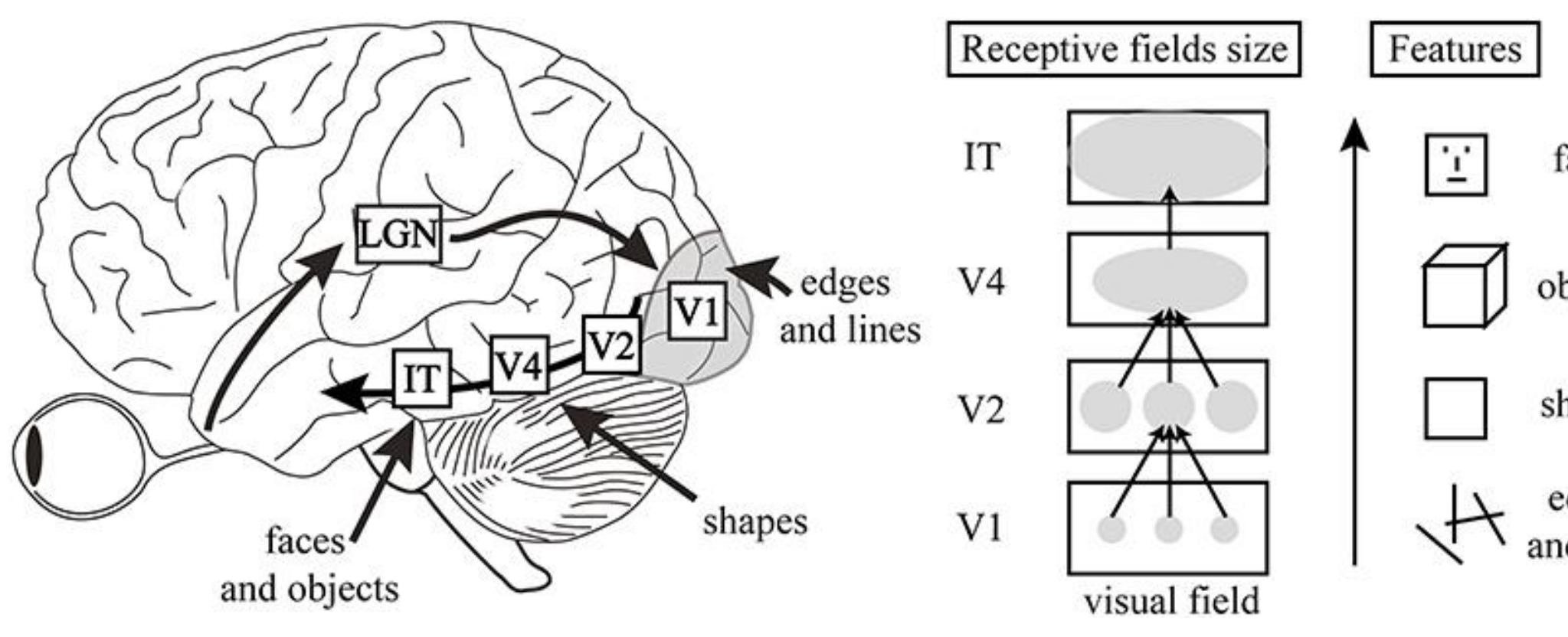
Mapping visual categorical selectivity across the whole brain using transformer-based encoders and large-scale generative models

Columbia University
In the City of New York

Columbia | Zuckerman Institute

Visual Inference Lab
Kriegeskorte

Ethan Hwang[1], Hossein Adeli[1], Wenxuan Guo[1], Andrew Luo[2], Nikolaus Kriegeskorte[1]
[1]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, USA
[2]The University of Hong Kong, Hong Kong, China



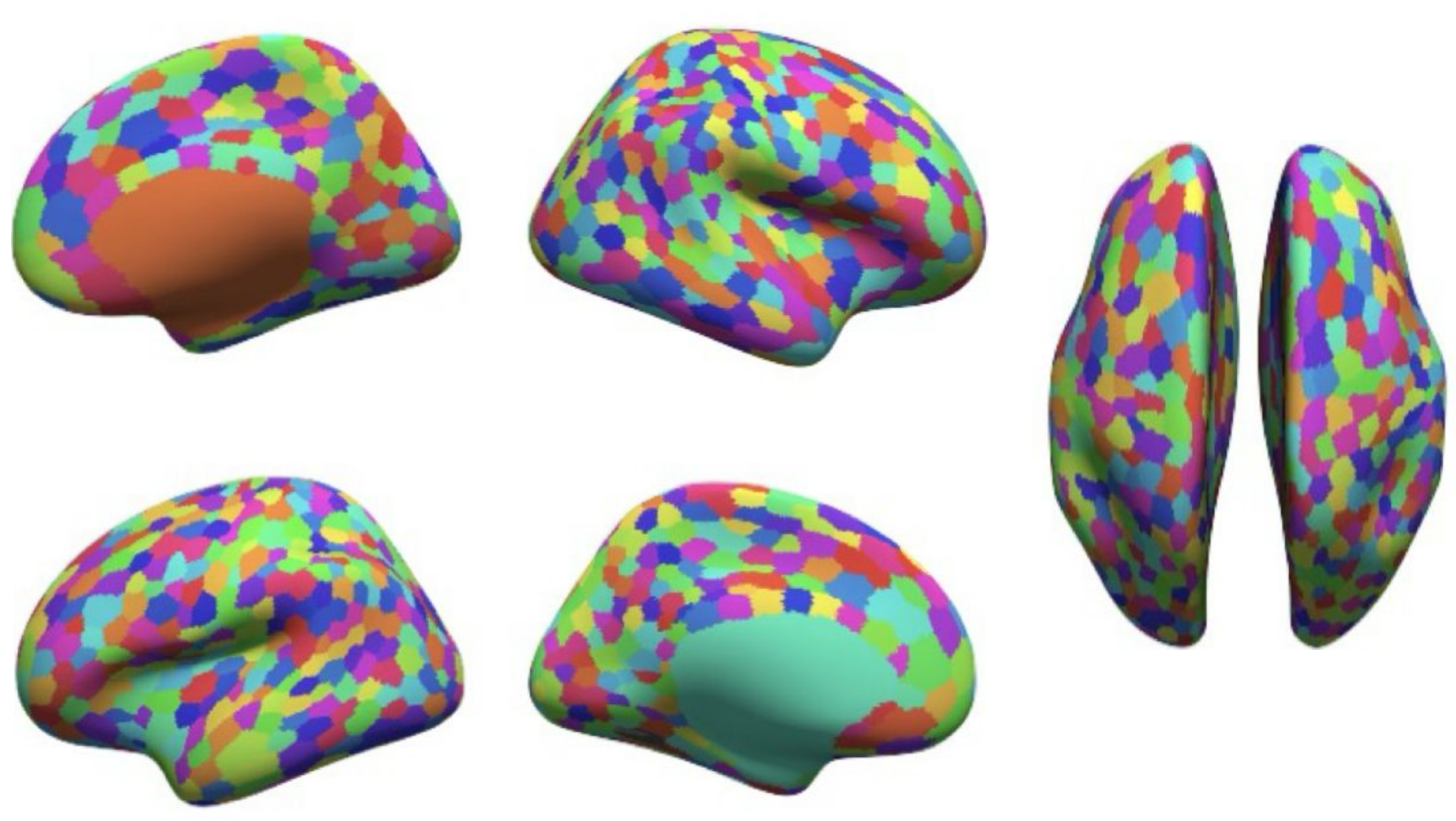
Reproduced from Figure 1 in [8]

Extensive neuroimaging experiments, especially functional magnetic resonance imaging (fMRI), have mapped a few prominent categories—faces, places, words, bodies, and food—to dedicated brain regions. But visual perception goes beyond simple categories and it remains an open question what higher-level visual concepts enable humans to make sense of the complex world.

Here, we leverage recent advances in artificial intelligence and a large-scale fMRI dataset to explore the visual selectivity of brain regions beyond the well-studied visual cortex. We found many parcels with more complex selectivity transcending simple categorical concepts. Finally, we verify our labels by testing how well CLIP representations of our "superstimuli" can predict ground-truth activation on a held-out set.

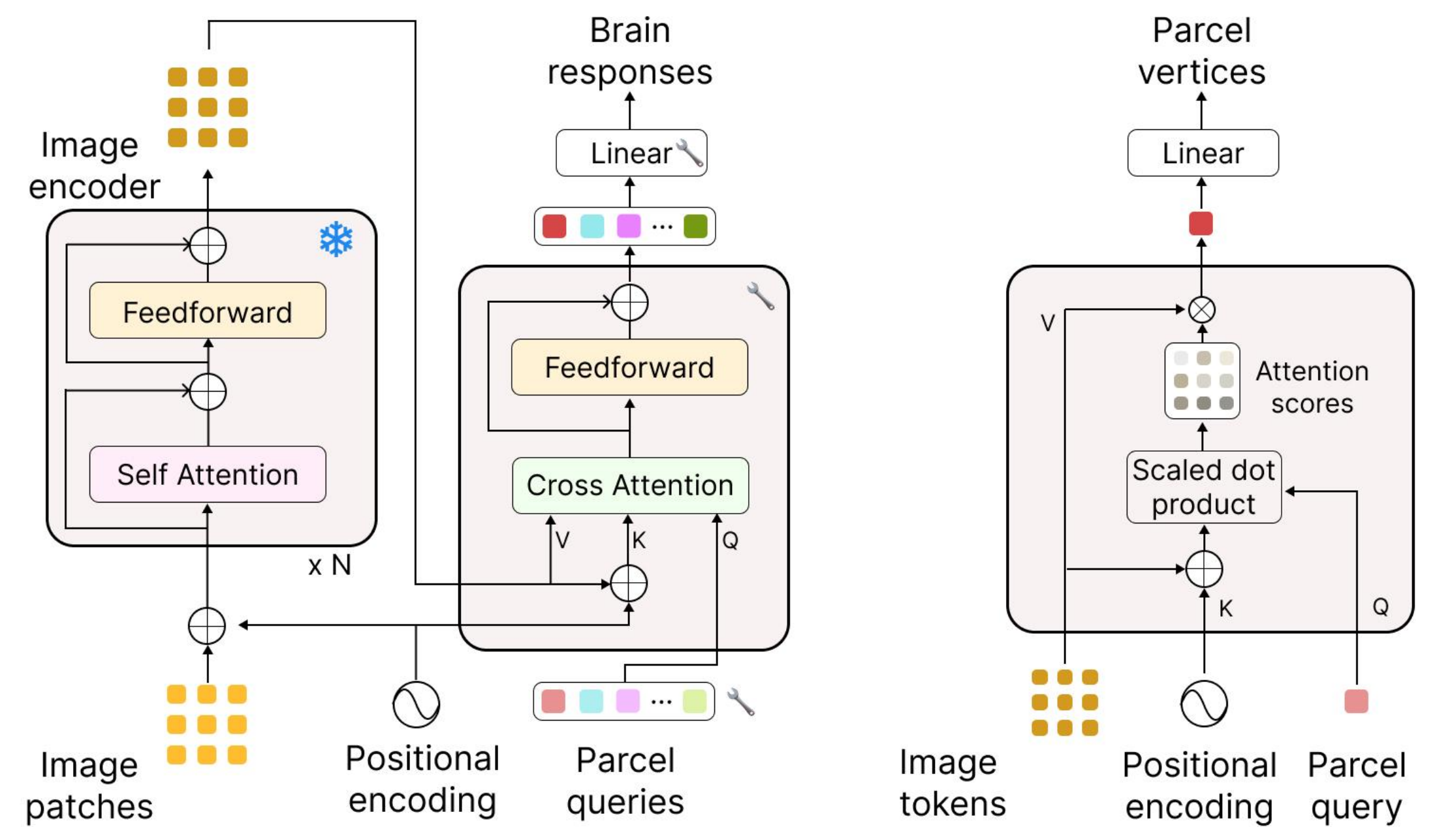
Data-driven functional parcellation

We partitioned the 327,684 cortical vertices across the whole brain into 1,000 functionally homogeneous parcels using the Schaefer resting-state functional connectivity parcellation [9].



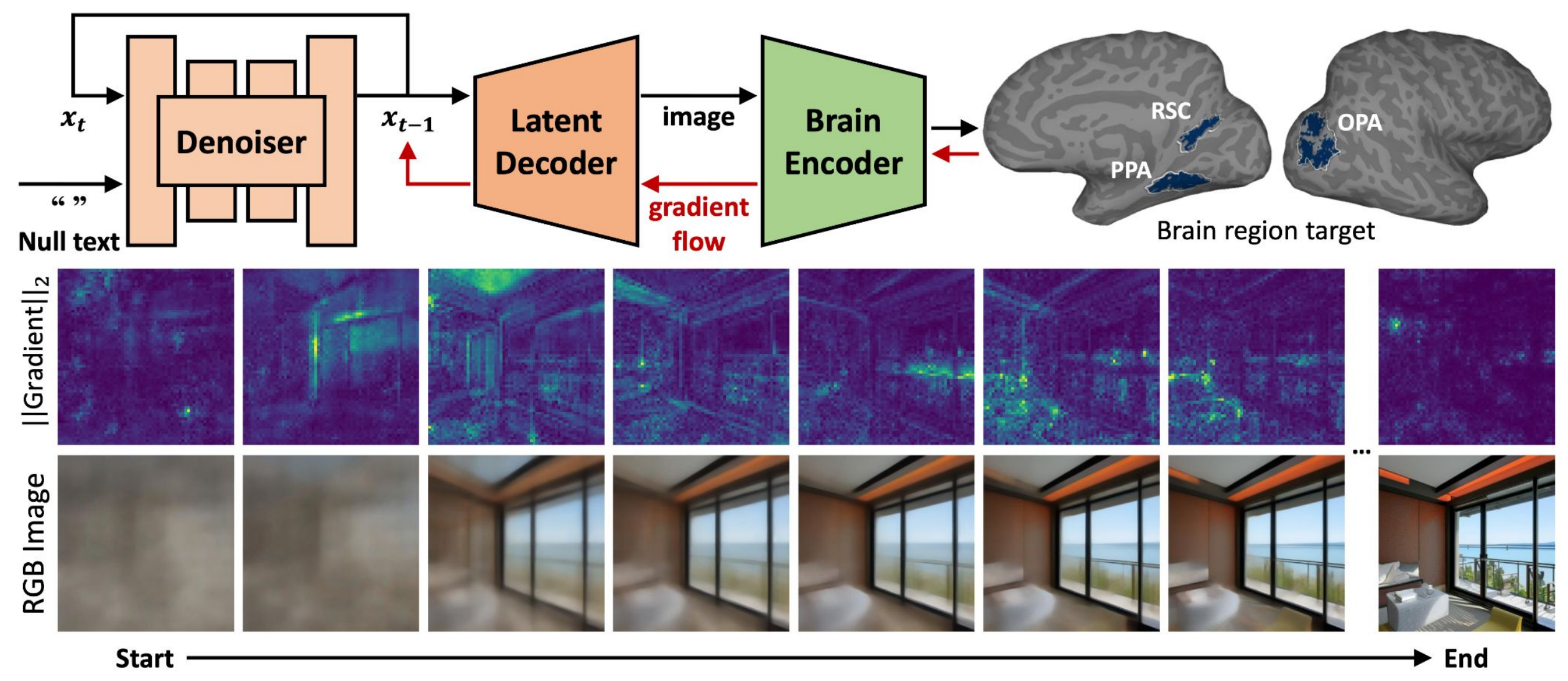
Brain encoder

We built a transformer-based brain encoder to predict brain activity from visual features [3] [5]. We trained the model on the NSD dataset [1], which includes up to 30,000 image-fMRI pairs from 8 subjects.



Generating images with diffusion

We used a diffusion model, BrainDIVE [2], to generate images that would maximally activate a given parcel. The encoder model serves as a "digital twin" that is fully observable, upon which we perform extensive experimentation to better understand selectivity beyond the visual cortex.

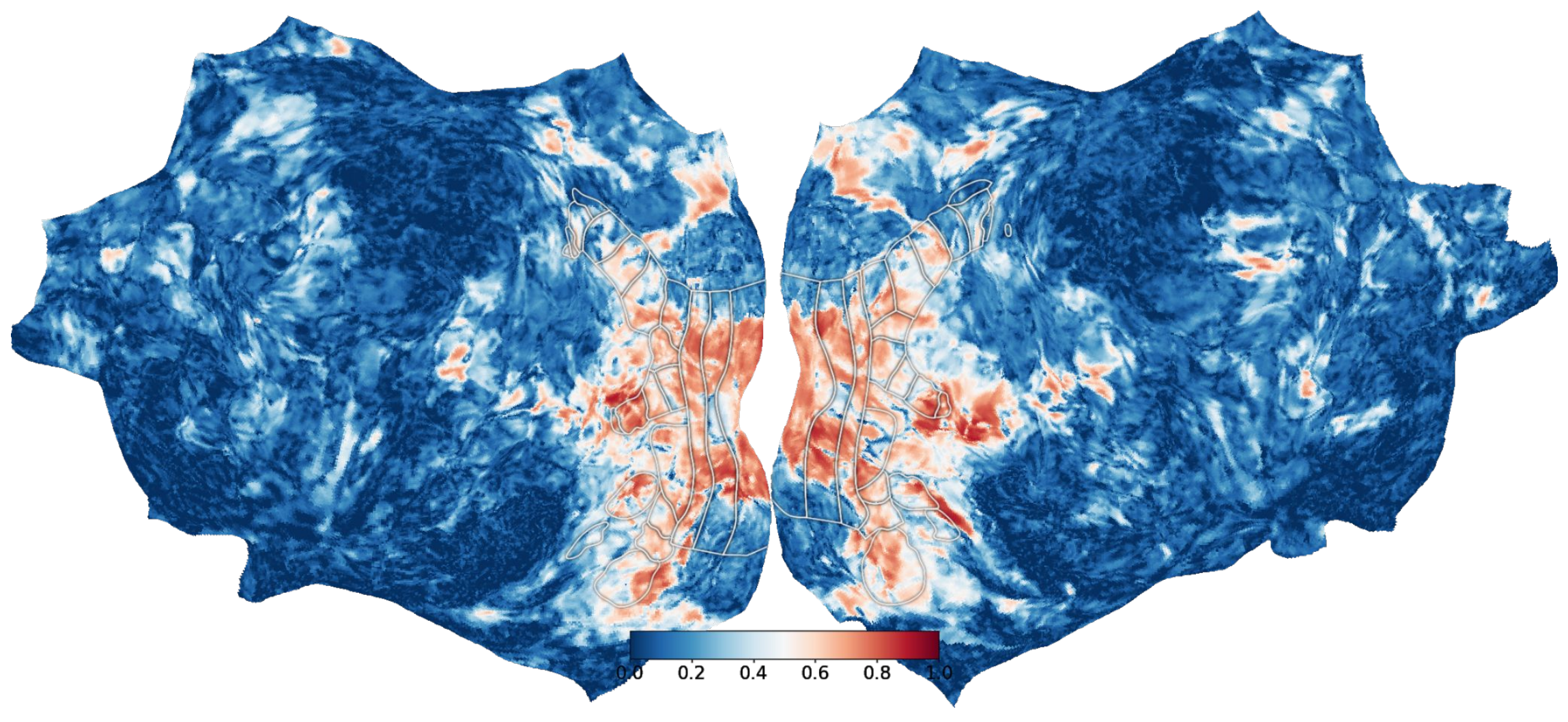


References

[1] E. J. Allen et al., "A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence," Nat Neurosci, vol. 25, no. 1, pp. 116–126, Jan. 2022, doi: 10.1038/s41593-021-00962-x.
[2] A. F. Luo, M. M. Henderson, L. Wehbe, and M. J. Tarr, "Brain Diffusion for Visual Exploration: Cortical Discovery using Large Scale Generative Models," Nov. 28, 2023, arXiv: arXiv:2306.03089. doi: 10.48550/arXiv.2306.03089.
[3] M. Oquab et al., "DINOv2: Learning Robust Visual Features without Supervision," Feb. 02, 2024, arXiv: arXiv:2304.07193. doi: 10.48550/arXiv.2304.07193.
[4] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL: IEEE, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
[5] H. Adeli, S. Minni, and N. Kriegeskorte, "Transformer brain encoders explain human high-level visual responses," May 22, 2025, arXiv: arXiv:2505.17329. Doi: 10.48550/arXiv.2505.17329
[6] J. S. Gao, A. G. Huth, M. D. Lescroart, and J. L. Gallant, "Pycortex: an interactive surface visualizer for fMRI," Front. Neuroinform., vol. 9, Sep. 2015, doi: 10.3389/fninf.2015.00023.
[7] J. Sergent, S. Ohta, and B. Macdonald, "Functional neuroanatomy of face and object processing: A positron emission tomography study," Brain, vol. 115, no. 1, pp. 15–36, Feb. 1992, doi: 10.1093/brain/115.1.15.
[8] M. H. Herzog and A. M. Clarke, "Why vision is not both hierarchical and feedforward," Front. Comput. Neurosci., vol. 8, Oct. 2014, doi: 10.3389/fncom.2014.00135.
[9] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. T. Yeo, "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI," Cereb. Cortex, vol. 28, no. 9, pp. 3095–3114, Sep. 2018, doi: 10.1093/cercor/bhx179.

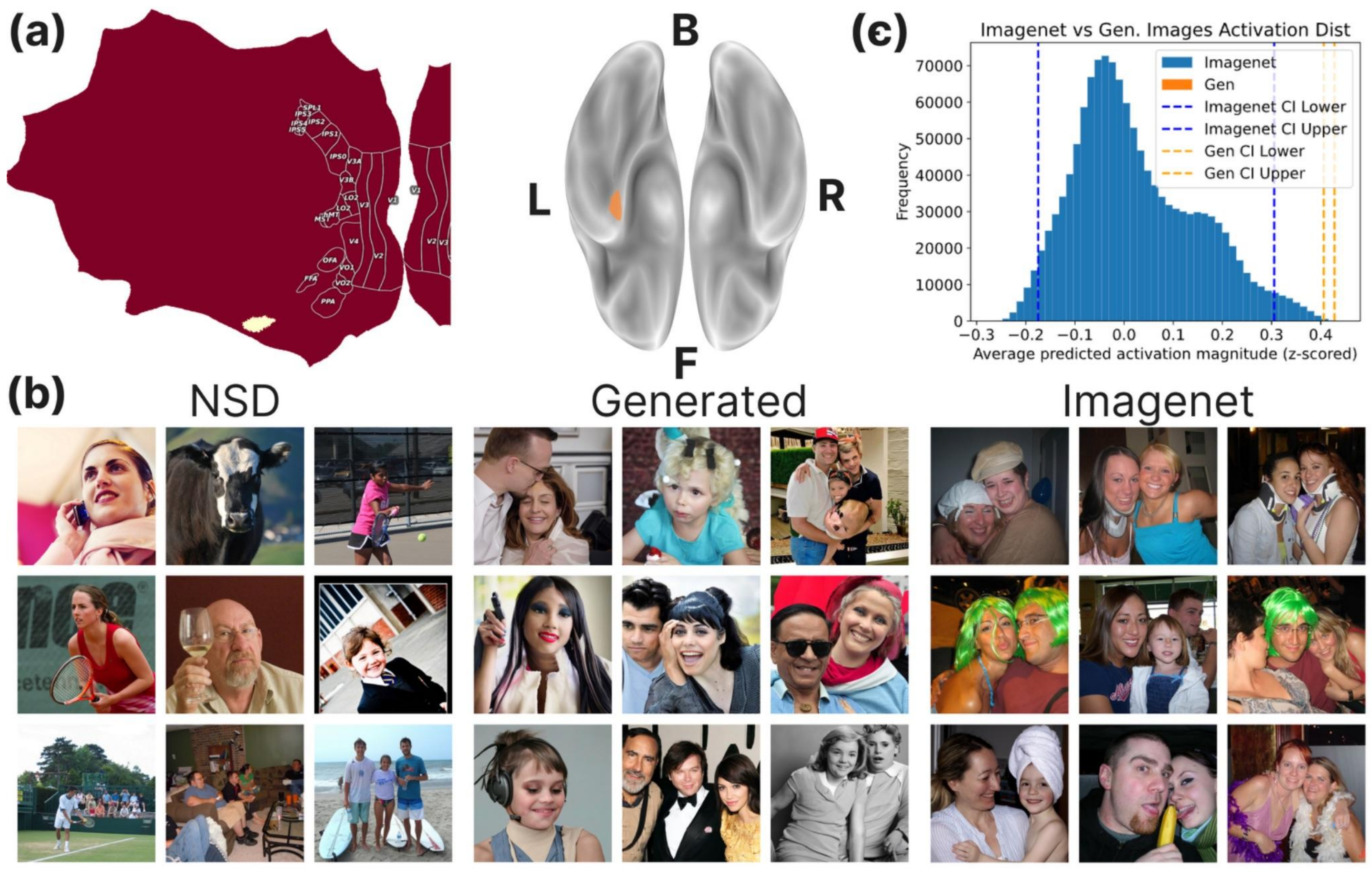
Model results

Prediction accuracy (pearson correlation) on held-out set:



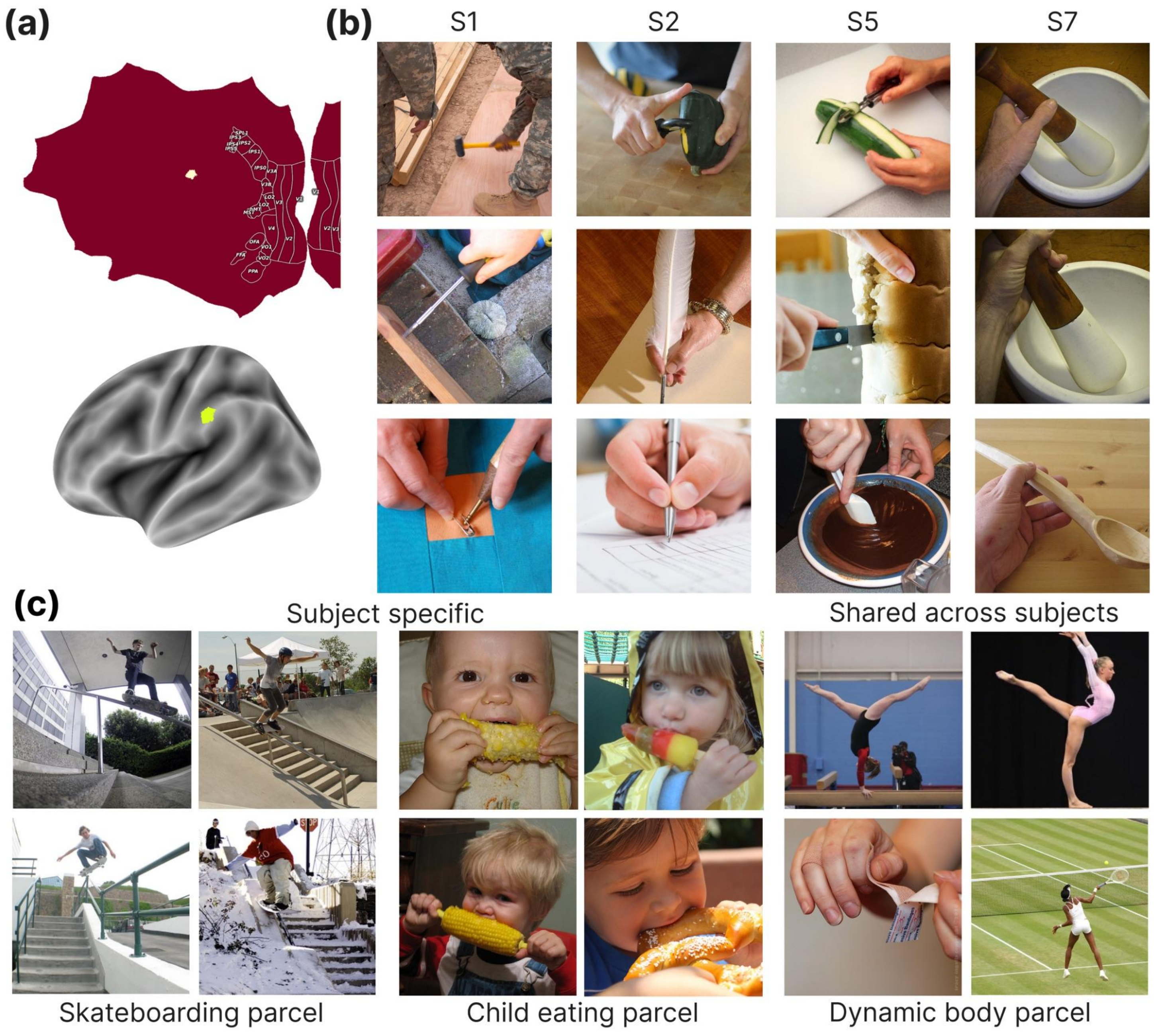
As expected, our model performs well on predicting activity in the visual cortex. Several areas beyond the visual cortex are also well-predicted, which may offer insight into downstream visual processing.

Sanity check: selectivity of known regions (aTL-faces)



We start by demonstrating the selectivity of a parcel that overlaps significantly with a known region (a), aTL-faces, which is known to respond to faces [7]. We chose images from the NSD, BrainDIVE, and Imagenet [4] that maximally activate the parcel (b). The images consistently include faces, as expected. The model-predicted activation from generated images is at the very tail end of ImageNet activations (c).

Main result: demonstrating complex selectivity



We next explore the selectivity of parcels outside the canonical visual area, chosen because they are visually responsive and well-predicted by our model. The images that maximally activate parcel (a) seem to depict tool use (b). We uncover several other parcels, subject-specific and shared, that exhibit consistent and interesting selectivity (c). The model trained on NSD (a relatively small dataset) reveals complex selectivity that can only be discovered using generative models or large-scale image retrieval.