Towards Interpretable Visual Decoding with Attention to Brain Representations

Columbia University in the City of New York

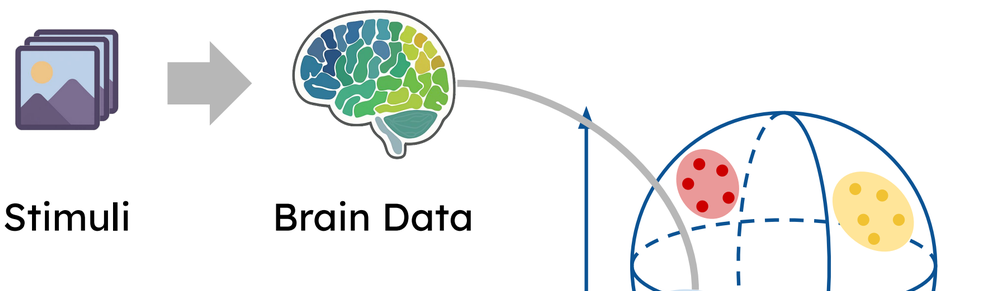
Visual Inference Kriegeskorte Lab

Pinyuan Feng, Hossein Adeli, Wenxuan Guo, Ethan Hwang, Fan L. Cheng, Nikolaus Kriegeskorte

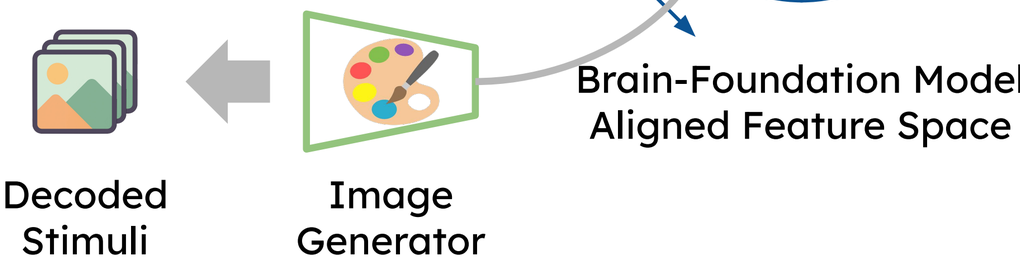
Columbia | Zuckerman Institute

Prior Work

Stage 1

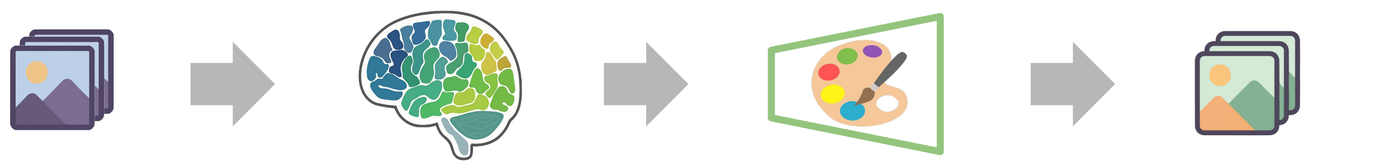


Stage 2

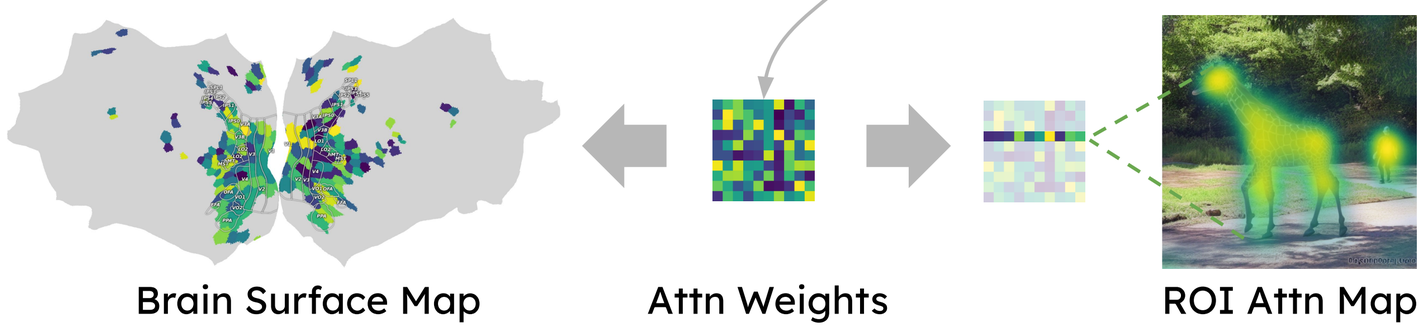


Our Work

End-to-End Decoding



Interpretability Probing



TL;DR: Typical two-stage decoding pipelines first map brain activity to intermediate feature spaces (e.g., CLIP/DINO) and then use those embeddings to guide a generative model. Our end-to-end brain-to-image approach conditions a latent diffusion model directly on brain activity, enabling interpretations of the generative dynamics in both image and brain spaces.

A. NeuroAdapter Training Pipeline.

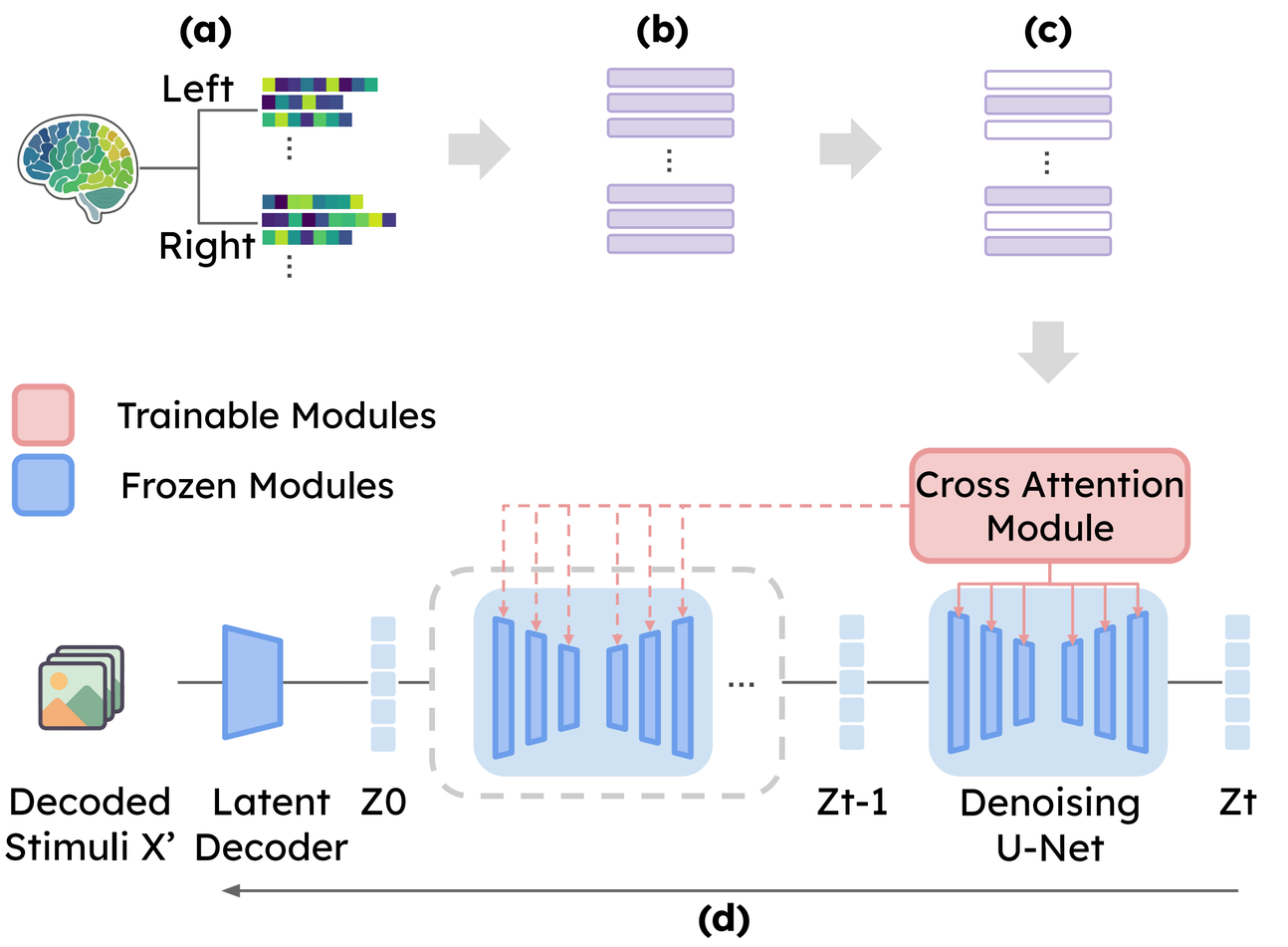


Fig A: (a) cortical parcellation [1], (b) parcel-wise linear mapping from vertices to brain representation tokens, (c) fMRI token dropout, and (d) conditioning a latent diffusion model.

B. Brain Encoder-based Image Selection

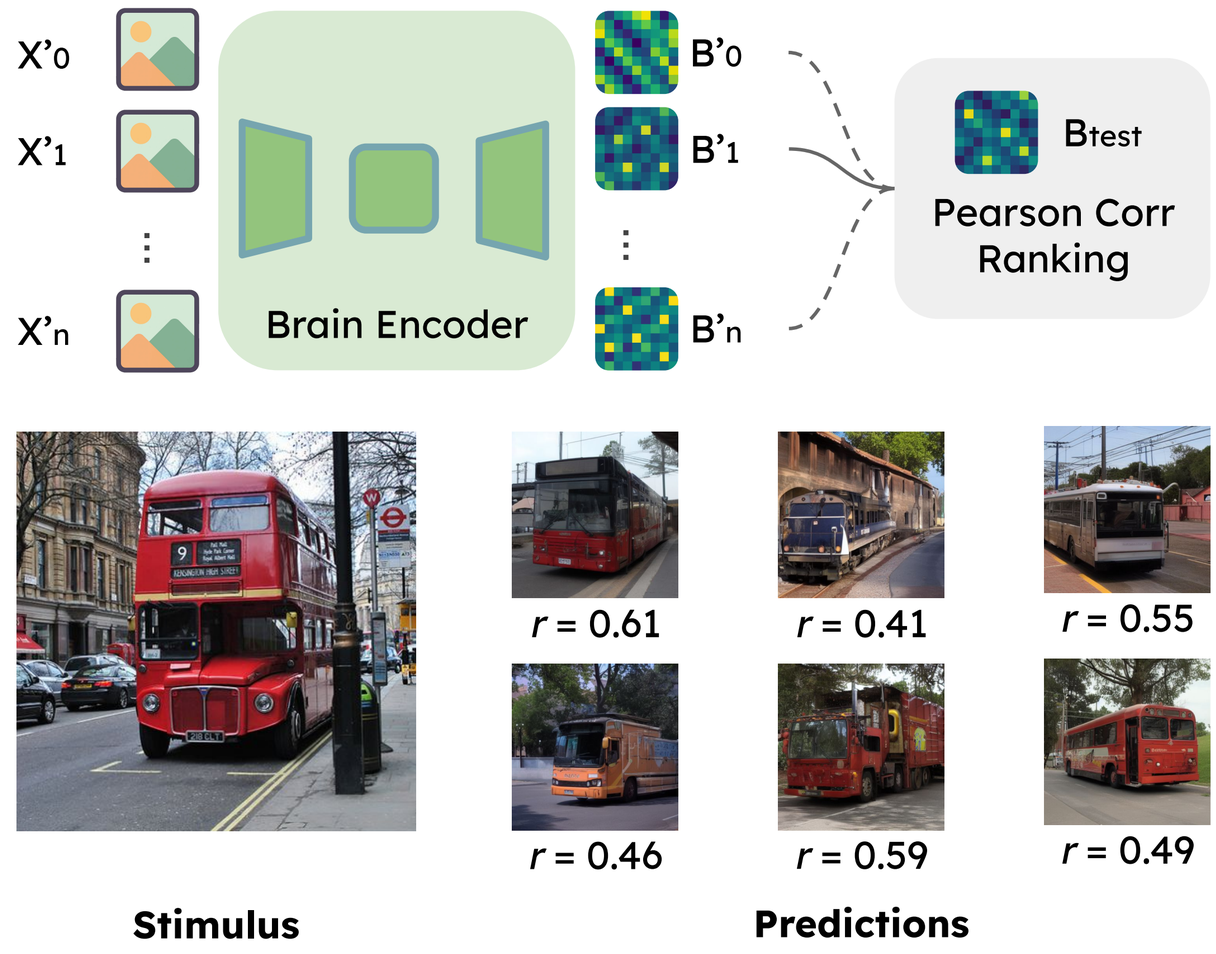
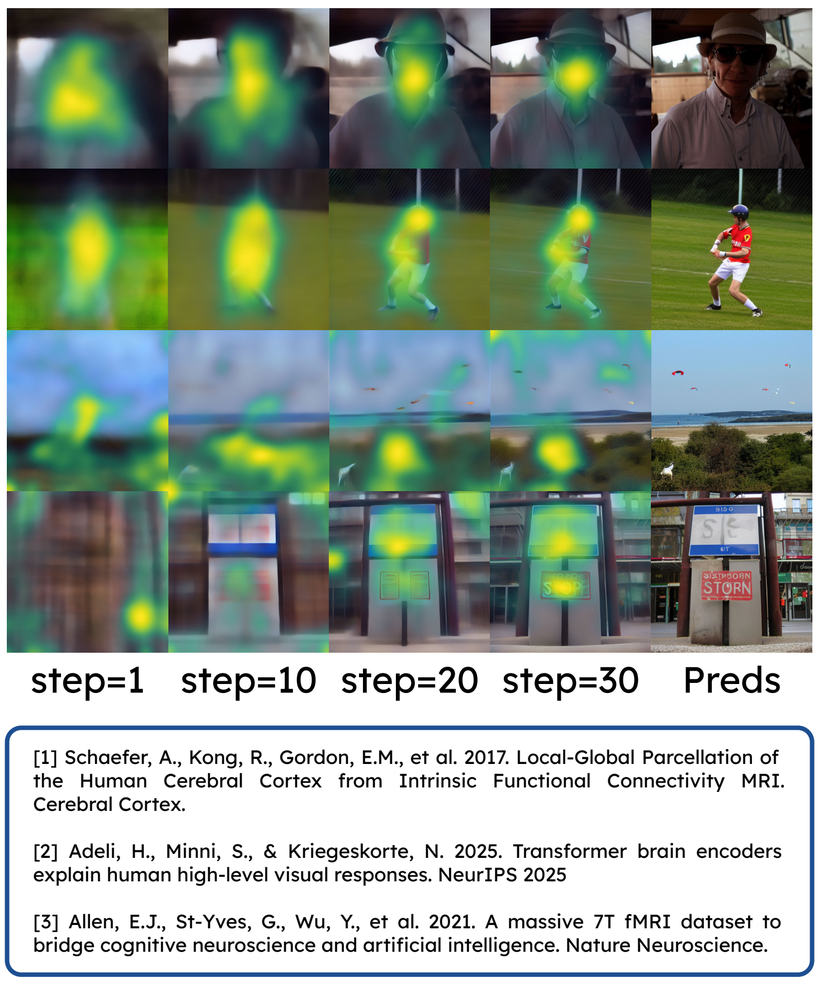


Fig B: Optimal decoding results are selected based on Pearson correlation between actual fMRI activity and predicted fMRI responses from these images using a brain encoder [2].

C. Decoding Examples on NSD [3]

D. ROI Attention Maps



[1] Schaefer, A., Kong, R., Gordon, E.M., et al. 2017. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cerebral Cortex.

[2] Adeli, H., Minni, S., & Kriegeskorte, N. 2025. Transformer brain encoders explain human high-level visual responses. NeurIPS 2025

[3] Allen, E.J., St-Yves, G., Wu, Y., et al. 2021. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nature Neuroscience.

Scan Me