NeuroAdapter: Visual Reconstruction with Masked Brain Representation

Hossein Adeli{ha2366@columbia.edu}, Wenxuan Guo, Pinyuan Feng, Ethan Hwang, Fan Cheng, Nikolaus Kriegeskorte
Zuckerman Mind Brain Behavior Institute, Columbia University, New York, USA

Abstract

Recent advances in generative decoding models have shown that complex visual scenes can be reconstructed from brain activity. However, current models rely on an intermediate step that maps the brain data to rich image and text feature spaces, resulting in overly large and computationally intensive models. This intermediate process may also cause loss of information deriving from the selectivity and receptive field location of individual brain units. In this work, we explore the capabilities of visual decoding in the absence of intermediate representations. We propose NeuroAdapter, a simple modular framework that directly encodes the neural data from different brain regions to condition the diffusion process. To avoid overfitting, our model incorporates a random token-masking strategy. We train our model on the 7T-fMRI Natural Scenes Dataset (NSD) and evaluate it on multiple metrics. NeuroAdapter excels at capturing high-level semantic visual content from fMRI signals, outperforming more complex models. Our model demonstrates a promising direction for scaling decoding models up to whole-brain image reconstruction.

Keywords: Brain Decoding; Stable Diffusion; fMRI; Visual Perception; Representation Learning; NeuroAI

Introduction

Current approaches to decoding visual content from the brain (Chen et al., 2022; Lin et al., 2022; Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023; Li et al., 2025) that leverage image-generative models typically implement a two-step process: (1) Brain activity is first mapped to multiple intermediate representations in image or text space (such as Clip; Radford et al., 2021). (2) These intermediate representations are passed to an image generator for reconstruction. Faithful reconstruction of perceived images critically depends on the ability of intermediate network representations to extract image information from neural activity. A recent study found that decoding models use only a small amount of information from the brain for image reconstruction. The intermediate representation may form an unnecessary bottleneck that leads to significant information loss (Mayo et al., 2024).

Moreover, existing approaches mostly utilize early and ventral pathway visual areas for reconstruction. While visual responses and category selectivity are well-characterized in those areas (Stigliani et al., 2015; Benson et al., 2018), several studies suggest that conscious perception of visual location and certain aspects of visual processing can occur outside of the visual cortex, specifically in higher-order brain regions (Liu et al., 2019). A method that can process whole-brain data and automatically learn the most relevant information embedded in neural activity is needed.

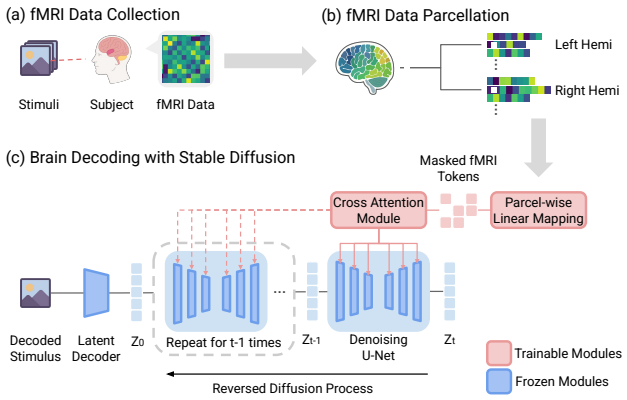

Figure 1: **Pipeline of NeuroAdapter**

Here, we tackle the problem by developing a lightweight, end-to-end mapping framework that can transform whole-brain activity into visual reconstructions without relying on intermediate pre-defined representations.

Approach

Our model, NeuroAdapter, built on the IP-Adapter framework (Ye, Zhang, Liu, Han, & Yang, 2023), conditions a pre-trained Stable Diffusion model (Rombach, Blattmann, Lorenz, Esser, & Ommer, 2022) on sparse fMRI-derived features via a cross-attention mechanism to reconstruct perceived visual stimuli. An overview of our approach is presented in Fig. 1.

Neural Data Processing and Parcellation

We train our model using the surface-based fMRI data in *fsaverage* space from a single subject (subject 1) in the 7T-fMRI NSD. The dataset includes high-resolution fMRI recordings from participants viewing up to 10,000 natural images (Allen et al., 2022). Subject 1's data is split to training, validation, and test sets. We average the vertex responses across image repetitions to obtain a single response pattern per image. To transform the high-dimensional fMRI data into structured inputs for conditioning the diffusion model, we apply the Schaefer parcellation (Schaefer et al., 2017), which clusters cortical vertices into 500 parcels per hemisphere.

Parcel-wise Linear Mapping

We compute vertex-wise Signal-to-Noise Ratio (SNR) and select top 100 parcels per hemisphere with the highest average SNR (Fig. 2 shown on pycortex; Gao, Huth, Lescroart, & Gallant, 2015), yielding a total of $P = 200$ parcels as fMRI inputs to the model. Since the number of vertices varies

Method	Low-Level				High-Level			
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
ImageNet retrieval	.128	.242	81.9%	90.3%	75.1%	80.6%	.869	.522
Ozcelik & VanRullen, 2023	**.254**	**.356**	**94.2%**	**96.2%**	87.2%	91.5%	.775	.423
NeuroAdapter (Ours)	.130	.289	86.1%	93.5%	**91.2%**	**92.4%**	**.697**	**.394**

Table 1: **Performance across different metrics**

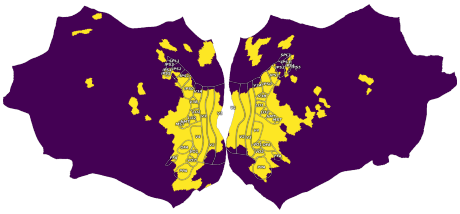

Figure 2: **Top-100-SNR Parcels for each brain hemisphere**

across parcels, we pad each parcel's vertex response vector to match the vertex count of the largest selected parcel, comprising V_{max} vertices. This yields raw brain data $B \in \mathbb{R}^{N \times P \times V_{\max}}$, where N is the batch size. We then apply a linear projection from vertex space to the token space. Each parcel is assigned a unique projection matrix $W_p \in \mathbb{R}^{V_{\max} \times D}$, transforming padded vertex response into parcel embeddings $E \in \mathbb{R}^{N \times P \times D}$, where $D = 768$.

fMRI-Guided Diffusion Process

To enable fMRI-guided image reconstruction, we modify the IP-Adpater framework by replacing its image-based cross-attention module with a mechanism that attends to the fMRI token embeddings described above. The text input to the diffusion model's text encoder is set to an empty string, which removes textual guidance. The fMRI token representations are the only conditioning input during the reverse diffusion process. To prevent overfitting and ensure robustness, we apply a stochastic masking procedure to the fMRI token embeddings. Given the tokens E, we randomly mask a subset of them for each training sample. We sample a masking probability $r \sim \mathcal{U}(0,1)$ and retain each of the P tokens independently with probability r. This results in a binary mask $M \in \{0,1\}^{N \times P \times 1}$, which is applied element-wise to the fMRI token embeddings $E' = E \odot M$. This masking strategy forces the model to perform image reconstruction using subsets of parcels and we find it to be crucial for good performance. During training, only the parcel-wise linear projection and cross-attention modules are updated using mean squared error (MSE) loss, with the Stable Diffusion model remaining frozen.

Brain Encoder

Extending the work of Adeli, Minni, and Kriegeskorte (2023), we use a brain encoder to predict vertex-wise activity across the whole brain from an input image. We apply this encoder in two ways. First, we establish a baseline model by retrieving an image from 1.3 million ImageNet images (Deng et al., 2009)

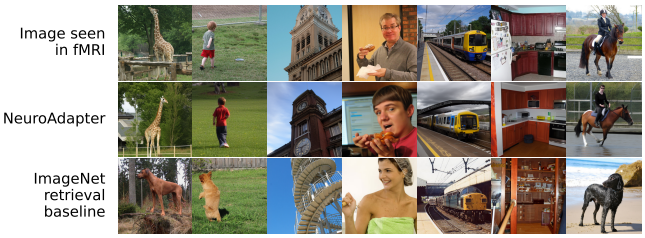

Figure 3: **Example reconstruction images**

whose predicted neural activity from our encoder best correlates with the ground truth fMRI response, a retrieval-based decoding method inspired by Kay, Naselaris, Prenger, and Gallant (2008). Second, we use the encoder to rank eight candidate images generated by our fMRI-guided diffusion model (from eight different seeds) for a given test fMRI sample, selecting the image whose predicted activity best correlates the measured brain response.

Results

We evaluate our approach on eight image reconstruction metrics, comparing it against the ImageNet retrieval baseline and the method proposed by Ozcelik and VanRullen (2023). As shown in Table 1, NeuroAdapter consistently outperforms the retrieval-based baseline across all metrics. Furthermore, our method achieves better reconstruction performance on high-level metrics than Ozcelik and VanRullen's approach, while achieving lower performance on the low-level metrics. This pattern suggests that NeuroAdapter, despite its simplicity, is particularly effective at capturing semantic content encoded in the fMRI signals without an intermediate representation, even if it is less accurate in reproducing low-level visual details, such as color, texture, and pixel-level structure. Future work will explore augmenting the MSE loss to better capture the low-level perceptual features in the representation.

Discussion

We present a simple yet effective brain-decoding framework that directly conditions the diffusion denoising process on brain activity. Our model encodes each parcel with a distinct token embedding, an approach that can easily be extended to using fMRI activity from the whole brain for decoding. Conditioning directly on the neural data from different regions will allow systematic exploration of the representational content in each region.

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., et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*, 116–126. doi: 10.1038/s41593-021-00962-x

Benson, N. C., et al. (2018). The human connectome project 7 tesla retinotopy dataset: Description and population receptive field analysis. *Journal of Vision*, *18*(13), 23. doi: 10.1167/18.13.23

Chen, Z., et al. (2022, November). Seeing beyond the brain: Masked modeling conditioned diffusion model for human vision decoding. In *arxiv.* Retrieved from https://arxiv.org/abs/2211.06956

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

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

Kay, K., Naselaris, T., Prenger, R., & Gallant, J. (2008, March 20). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355. doi: 10.1038/nature06713

Li, H., et al. (2025). Neuraldiffuser: Neuroscience-inspired diffusion guidance for fmri visual reconstruction. *IEEE Transactions on Image Processing*, *34*, 552–565. doi: 10.1109/tip.2025.3526051

Lin, S., et al. (2022). *Mind reader: Reconstructing complex images from brain activities.* Retrieved from https://arxiv.org/abs/2210.01769

Liu, S., et al. (2019). Neural correlates of the conscious perception of visual location lie outside visual cortex. *Current Biology*, *29*(23), 4036-4044.e4. doi: https://doi.org/10.1016/j.cub.2019.10.033

Mayo, D., Wang, C., Harbin, A., Alabdulkareem, A., Shaw, A. E., Katz, B., & Barbu, A. (2024). Brainbits: How much of the brain are generative reconstruction methods using? In *The thirty-eighth annual conference on neural information processing systems.* Retrieved from https://openreview.net/forum?id=KAAUvi4kpb

Ozcelik, F., & VanRullen, R. (2023). *Natural scene reconstruction from fmri signals using generative latent diffusion.* Retrieved from https://arxiv.org/abs/2303.05334

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022, June). High-resolution image synthesis with latent diffusion models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 10684-10695).

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., ... Yeo, B. T. T. (2017, July). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral Cortex*, *28*(9), 3095–3114. doi: 10.1093/cercor/bhx179

Scotti, P. S., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Cohen, E., ... Abraham, T. M. (2023). *Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors.* Retrieved from https://arxiv.org/abs/2305.18274

Stigliani, A., et al. (2015). Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, *35*(36), 12412–12424.

Takagi, Y., & Nishimoto, S. (2023, June). High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 14453-14463).

Ye, H., Zhang, J., Liu, S., Han, X., & Yang, W. (2023). Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models.