Towards Interpretable Visual Decoding with Attention to Brain Representations

Pinyuan Feng[1]* **Hossein Adeli**[1] **Wenxuan Guo**[1]

Fan Cheng[1] **Ethan Hwang**[1] **Nikolaus Kriegeskorte**[1]

[1]**Zuckerman Mind Brain Behavior Institute, Columbia University, USA**

Abstract

Recent work has demonstrated that complex visual stimuli can be decoded from human brain activity using deep generative models. However, most current approaches rely on mapping brain data into intermediate image or text feature spaces before guiding the generative process, masking the effect of responses from different brain areas on the final reconstruction output. In this work, we propose *NeuroAdapter*, a framework that directly conditions a latent diffusion model on brain representations, bypassing the need for intermediate feature spaces. Our method demonstrates competitive visual reconstruction quality on the Natural Scenes Dataset (NSD). To reveal how different cortical areas influence the unfolding generative trajectory, we contribute an Image–Brain BI-directional interpretability framework (*IBBI*), investigating cross attention mechanisms across diffusion steps. Our work enables new approaches for interpreting latent diffusion models through the lens of visual neuroscience.

1 Introduction

Recent years have witnessed remarkable success of deep generative models, which drives the NeuroAI frontier of decoding complex natural images from brain activities, bringing the prospect of “mind reading” closer to reality.

Current approaches to reconstructing visual stimuli from the brain (Lin et al., 2022; Cheng et al., 2023; Takagi and Nishimoto, 2023; Ozcelik and VanRullen, 2023; Scotti et al., 2023; Li et al., 2025; Kamitani et al., 2025; Ferrante et al., 2025) typically implement a two-stage pipeline: (1) brain activity is first mapped to intermediate representations in latent space derived from large foundation models, such as CLIP (Radford et al., 2021) and DINO (Caron et al., 2021; Oquab et al., 2023); (2) these intermediate representations are then used to condition a visual generative model for stimulus reconstruction.

Mapping brain data into an intermediate representation space leverages rich priors in embedding spaces to improve reconstruction quality and has proved highly effective for reconstruction. However, the use of this intermediate representation can introduce an information bottleneck (Mayo et al., 2024) with successful reconstruction of perceived stimuli depending on the alignment between neural representations and the embedding space. This intermediate step can also mask the effect of different brain areas and their content on the final reconstruction, limiting the interpretability of the approach.

Many state-of-the-art decoding frameworks (Scotti et al., 2023, 2024; Xia et al., 2024; Gong et al., 2025) introduce multiple specialized components to enhance reconstruction fidelity, such as low-level

[*] pf2477@columbia.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

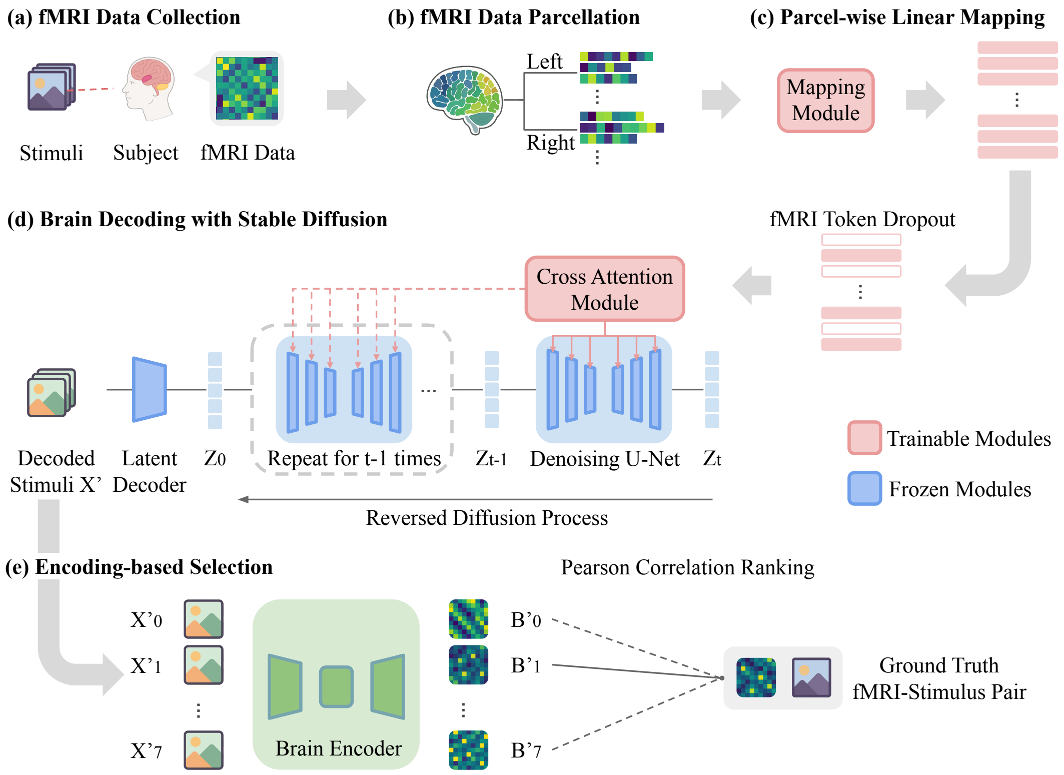

Figure 1: An overview of our brain decoding pipeline. NSD fMRI data are parcellated into cortical regions, then linearly mapped into parcel-wise tokens, followed by random token dropout. These brain-derived tokens condition a latent diffusion model during training and inference. A brain encoding model is used to search for the best decoded stimulus.

pathways for detailed texture and color recovery, and high-level pathways for semantic consistency. This practice, although it favors the decoding performance, has significantly increased the complexity of decoding. The recent streamlined approach Careil et al. (2025) proposed a single-stage solution with LoRA finetuning (Hu et al., 2022) for dynamic visual decoding from time-resolved fMRI signals. In our work, we explore an interpretable alternative by conditioning latent diffusion models directly on the brain activity.

Contributions of our paper Our contributions are as follows: (1) we design *NeuroAdapter* as a flexible and extensible module that can be integrated into generative backbones; (2) we demonstrate that high-quality visual decoding can be achieved directly from brain representations, without reliance on external embedding spaces; and (3) we provide a bi-directional interpretability framework, namely *IBBI*, to reveal both the relative contribution of brain parcels and their spatial influence in the reconstructed images.

2 Model Training and Evaluation

Our brain decoding framework, *NeuroAdapter*, as shown in Fig.1 (more examples are shown in Appendix A), is built on the IP-Adapter framework (Ye et al., 2023). We conditioned a pre-trained Stable Diffusion model (SD, Rombach et al. (2022)) on fMRI-derived features via cross attention to reconstruct perceived visual stimuli.

2.1 Neural Data Processing and Parcellation

We trained our model using the surface-based fMRI data in *fsaverage* space on Natural Scene Dataset (NSD, Appendix B, Allen et al. (2022)). We first averaged the vertex responses across image

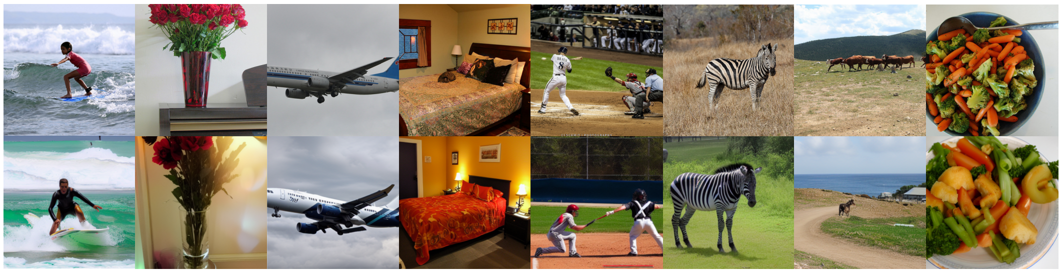

Figure 2: Ground Truth (first row) vs. Decoded Stimuli (second row)

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 applied the Schaefer parcellation (Schaefer et al. (2017), Appendix C). This clusters cortical vertices into 500 parcels per hemisphere and has been shown to be an effective practice for brain tokenization (Bosch et al., 2025).

2.2 Parcel-wise Linear Mapping

We computed vertex-wise Signal-to-Noise Ratio (SNR) and selected top 100 parcels per hemisphere with the highest average SNR, yielding a total of $p = 200$ parcels as fMRI inputs to the model. Since the number of vertices varies across parcels, we padded each parcel's vertex response vector to match the largest vertex count across parcels v_{max}. This yields processed neural data $\mathrm{D}_{\text{fMRI}} \in \mathbb{R}^{n \times p \times v_{\text{max}}}$, where n is the batch size. Then, each parcel was assigned a unique projection matrix $w \in \mathbb{R}^{v_{\text{max}} \times f}$, transforming padded vertex response into fMRI embeddings $\mathrm{E} \in \mathbb{R}^{n \times p \times f}$, where $f = 768$.

2.3 Latent Diffusion Process with fMRI Conditioning

We introduce a cross-attention module that enables U-Net in SD to attend to the fMRI embeddings directly. To ensure that fMRI embeddings were the only conditioning input, the text encoder in SD received an empty input during both training and inference. During training, only the parcel-wise linear projection and cross-attention modules were updated, with the SD parameters kept frozen.

fMRI Token Dropout. We applied a stochastic token dropout strategy during training to the fMRI embeddings E to ensure robustness of visual decoding. We randomly dropped out parcel-wise tokens for each training sample. A dropout probability $r \sim \mathcal{U}(0, 1)$ was drawn, and each of the p tokens was independently retained with probability r. This produced in a binary mask $\mathrm{M} \in \{0, 1\}^{n \times p \times 1}$, which was applied parcel-wise to the fMRI embeddings $\mathrm{E}' = \mathrm{E} \odot \mathrm{M}$. We found this regularization to be crucial for strong decoding performance, as supported by the ablation results in Appendix D.

Min-SNR Loss Weighting. To stabilize training and improve sample quality, we adopted the min-SNR weighting strategy (Hang et al., 2023) recently introduced in diffusion models. This approach down-weights the contribution of easy high-SNR steps, where reconstructions are clean, while preserving the weight of noisy low-SNR steps, yielding a more balanced training signal across the diffusion process (please view Appendix F for details).

2.4 Decoded Image Selection with Brain Encoding Model

During evaluation, we used a brain encoder (Adeli et al., 2023, 2025) trained on the NSD dataset to identify the best decoded stimuli. For each fMRI sample in the test set, we generated eight candidate images with different random seeds. The brain encoder predicted vertex-wise fMRI activity for each candidate image, which was correlated with the measured fMRI response. The candidate image with the highest Pearson correlation was selected as the decoded image. An ablation study assessing the impact of the brain encoder to decoding performance is reported in Appendix E.

3 Image-Brain Bi-directional Interpretability Analysis

Beyond decoding performance, we also investigated the interpretability of generative process in our model. Since the conditioning input to SD was parcel-wise embeddings, this can be represented as a token matrix $\mathrm{E} \in \mathbb{R}^{p\times f}$ (batch size $n = 1$ for simplicity), where each row $\mathrm{e}_i \in \mathbb{R}^f$ corresponds to the embedding of parcel P_i. Since, anatomical or functional labels are available for parcels, this formulation enables ROI-level probing of the cross-attention mechanism.

Following this idea, we propose the **I**mage-**B**rain **BI**-directional interpretability analysis framework (*IBBI*), which links brain activity and image features during decoding. In *NeuroAdapter*, each cross-attention layer computes attention scores $\mathtt{Attn}(\mathrm{Q}, \mathrm{K}, \mathrm{V})$, where queries $\mathrm{Q} \in \mathbb{R}^{q\times d}$ come from spatial tokens in the U-Net of SD, and keys and values $(\mathrm{K}, \mathrm{V}) \in \mathbb{R}^{p\times d}$ are derived from the fMRI embeddings E. At each denoising timestep t, the attention weight matrix $\mathrm{A}^{(\ell,h,t)} \in \mathbb{R}^{q\times p}$ for head h in layer ℓ quantifies the influence of each parcel token on each spatial query. Intuitively, each entry of this matrix reflects the degree of attention from a particular spatial query in the image to a specific parcel. Our bidirectional interpretability analysis further exploits this matrix from two complementary views (please refer to Appendix G for more details).

We summarize the attention weight matrix $\mathrm{A}^{(\ell,h,t)}$ over parcels at each timestep into a vector $\mathrm{B}^{(t)}$ (Parcel Contribution Vector, *PCV*), normalized to unit mass. The vector represents the relative contribution of different parcels at timestep t. We then project this vector onto the cortical surface using `pycortex` (Gao et al., 2015), visualizing how strongly each parcel influences denoising at current timestep. With spatial structure in $\mathrm{A}^{(\ell,h,t)}$, we first pool attention across heads and ROI-related parcel tokens to obtain a query-wise matrix. We then reshape the matrix into a 2D map, which matches the layer's 2D grid, and upsample to full image resolution. We normalize each map to unit mass and average across layers, yielding $\mathrm{I}^{(t)}$ (ROI Attention Map, *RAM*) that shows where the ROI attends to at timestep t.

4 Results

4.1 Decoding Performance

We evaluate our approach on 8 image quality metrics that are commonly used in prior literature, comparing it against *Cortex2Image* (Gu et al., 2024), Takagi's approach (Takagi and Nishimoto, 2023), *Brain Diffuser* (Ozcelik and VanRullen, 2023). As shown in Table 1, our method substantially outperforms *Cortex2Image* across all reported metrics and achieves performance comparable to approaches that rely on external embeddings from foundation models. In particular, our method achieves the best scores on several high-level semantic metrics, such as Inception and EfficientNet distance, while maintaining competitive CLIP similarity. 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 (Fig. 2). Future work will explore techniques to better capture the low-level perceptual features in the representation.

Table 1: **Performance across Different Image Quality Metrics**

Method	**Low-Level**				**High-Level**			
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Cortex2Image	.150	.325	–	–	–	–	.862	.465
Takagi et al.	–	–	83.0%	83.0%	76.0%	77.0%	–	–
Brain Diffuser	**.254**	**.356**	**94.2%**	**96.2%**	87.2%	**91.5%**	.775	.423
NeuroAdapter (ours)	.124	.306	84.8%	93.7%	**90.9%**	91.1%	**.715**	**.405**

4.2 Interpretability Analysis

In this section, we visualize and analyze how brain representations influence the generative process with cross attention in *NeuroAdapter*. As mentioned in Section 3, our proposed *IBBI* framework provides two complementary perspectives, showing how different brain regions contribute to visual reconstruction and where those ROIs direct their attention in the pixel-level stimulus space.

Brain-directed View. Based on Parcel Contribution Vector (*PCV*), we average $\mathrm{B}^{(t)}$ across timesteps, to obtain a global view $\overline{\mathrm{B}^{(t)}}$ throughout the generative process. 200 parcels with corresponding contribution weights are visualized on the brain surface.

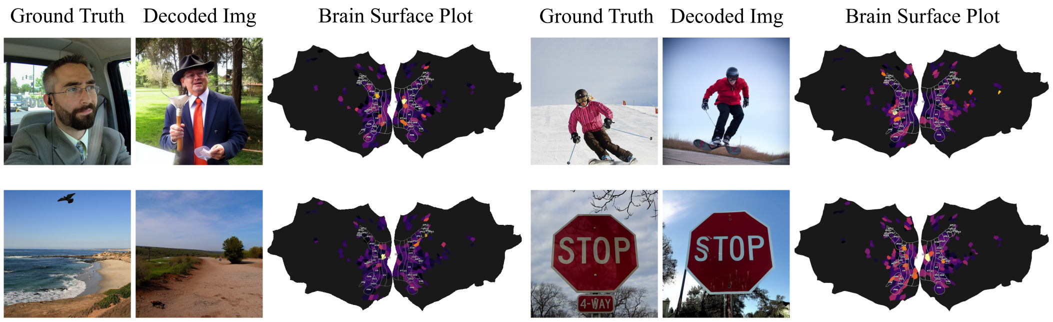

Figure 3: Examples of Parcel Contribution Vector (*PCV*) Projection on Cortex Surface Plot.

Image-directed View. To better interpret the ROI Attention Map (*RAM*), we connect them to well-established functional regions. Specifically, we map the anatomical/functional labels provided in NSD onto our parcellation scheme. A parcel is assigned to a label if more than 50% of its vertices overlap with that region. Here, we visualize the ROI attention maps across generative timesteps for representative category-selective regions, including *Face*, *Body*, *Scene*, and *Word*. Fig. 4 reveals how different cortical ROIs guide attention toward distinct spatial locations in the image during the denoising process, thereby linking regional neural signals to specific pixel-level features.

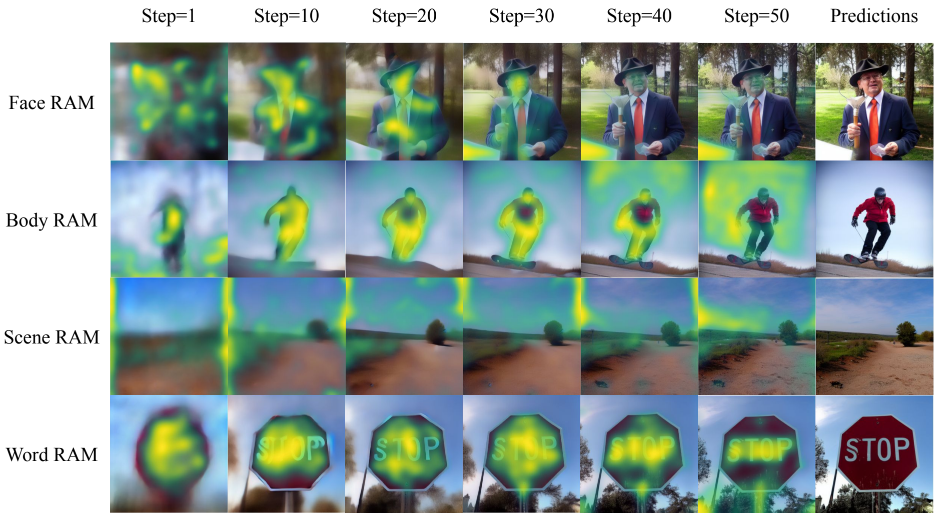

Figure 4: Examples of ROI Attention Maps (*RAM*) Overlaid on Decoded Images.

5 Discussion

We present a simple yet effective brain-decoding framework that directly conditions the diffusion denoising process on brain activity, bypassing intermediate feature spaces and enabling both effective decoding and mechanistic interpretability. Our results show that this approach achieves competitive reconstruction quality, particularly on high-level semantic metrics. Through our *IBBI* framework, we further reveal how different cortical parcels contribute to and shape the unfolding generative process, linking brain activity with image features in a bi-directional manner.

6 Acknowledgment

Research reported in this publication was supported in part by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award numbers 1RF1NS128897 and 4R01NS128897. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Adeli, H., Minni, S., and Kriegeskorte, N. (2023). Predicting brain activity using transformers. *bioRxiv*, pages 2023–08.

Adeli, H., Sun, M., and Kriegeskorte, N. (2025). Transformer brain encoders explain human high-level visual responses. *arXiv preprint arXiv:2505.17329*.

Allen, E. J. et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25:116–126.

Bosch, V., Anthes, D., Doerig, A., Thorat, S., König, P., and Kietzmann, T. C. (2025). Brain-language fusion enables interactive neural readout and in-silico experimentation.

Careil, M., Benchetrit, Y., and King, J.-R. (2025). Dynadiff: Single-stage decoding of images from continuously evolving fmri.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Cheng, F. L., Horikawa, T., Majima, K., Tanaka, M., Abdelhack, M., Aoki, S. C., Hirano, J., and Kamitani, Y. (2023). Reconstructing visual illusory experiences from human brain activity. *Science Advances*, 9(46):eadj3906.

Ferrante, M., Boccato, T., Rashkov, G., and Toschi, N. (2025). Towards neural foundation models for vision: Aligning eeg, meg, and fmri representations for decoding, encoding, and modality conversion. *Information Fusion*, page 103650.

Gao, J. S., Huth, A. G., Lescroart, M. D., and Gallant, J. L. (2015). Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9.

Gong, Z., Zhang, Q., Bao, G., Zhu, L., Xu, R., Liu, K., Hu, L., and Miao, D. (2025). Mindtuner: cross-subject visual decoding with visual fingerprint and semantic correction. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Gu, Z., Jamison, K., Kuceyeski, A., and Sabuncu, M. R. (2024). Decoding natural image stimuli from fmri data with a surface-based convolutional network. In Oguz, I., Noble, J., Li, X., Styner, M., Baumgartner, C., Rusu, M., Heinmann, T., Kontos, D., Landman, B., and Dawant, B., editors, *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pages 107–118. PMLR.

Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., and Guo, B. (2023). Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7441–7451.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Kamitani, Y., Tanaka, M., and Shirakawa, K. (2025). Visual image reconstruction from brain activity via latent representation. *Annual Review of Vision Science*, 11.

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

Lin, S. et al. (2022). Mind reader: Reconstructing complex images from brain activities.

Mayo, D., Wang, C., Harbin, A., Alabdulkareem, A., Shaw, A. E., Katz, B., and 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*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision.

Ozcelik, F. and VanRullen, R. (2023). Natural scene reconstruction from fmri signals using generative latent diffusion.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

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

Scotti, P. S., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Cohen, E., Dempster, A. J., Verlinde, N., Yundler, E., Weisberg, D., Norman, K. A., and Abraham, T. M. (2023). Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors.

Scotti, P. S., Tripathy, M., Villanueva, C. K. T., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K. A., and Abraham, T. M. (2024). Mindeye2: shared-subject models enable fmri-to-image with 1 hour of data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Takagi, Y. and Nishimoto, S. (2023). 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)*, pages 14453–14463.

Xia, W., de Charette, R., Öztireli, C., and Xue, J.-H. (2024). Dream: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

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

Appendix

A Examples of Decoded Stimuli

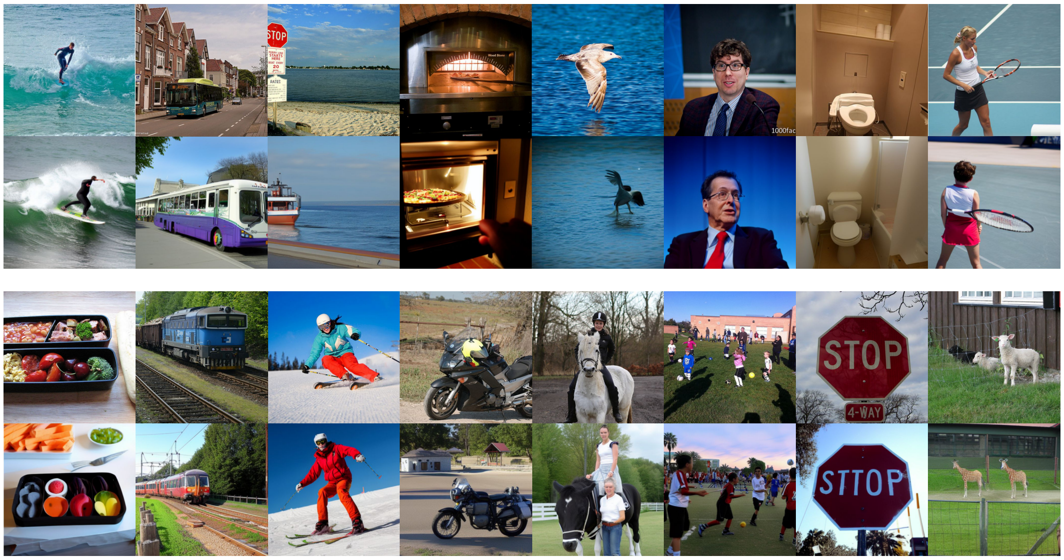

Figure 5: Ground Truth vs. Decoded Stimuli

B Natural Scene Dataset

We use the Natural Scenes Dataset (NSD), a large-scale 7T-fMRI dataset designed for studying visual representations in the human brain. This contains high-resolution brain responses from 8 subjects, each viewing up to 10,000 distinct natural images sampled from the MSCOCO dataset. For our experiments, we train on surface-based fMRI data in fsaverage space following standard preprocessing provided by NSD. Here, we report comparisons with prior work using the averaged results from subjects 1, 2, 5, and 7. For the remaining ablation studies, we restrict our analysis to subject 1 and evaluate models under different experimental conditions on this single-subject dataset.

C Schaefer Parcellation

To represent brain activity at the regional level, we adopt the Schaefer cortical parcellation (Fig. 6). This provides a functional subdivision of the cortex derived from large-scale resting-state fMRI. In our experiments, we compute vertex-wise Signal-to-Noise Ratio (SNR) and select top 100 parcels per hemisphere with the highest average SNR.

D Ablation Study: fMRI Token Dropout

We conduct an ablation study to evaluate the effect of the proposed fMRI token dropout (TD) strategy in training on decoding performance. As shown in Table 2, removing token dropout substantially compromised performance across almost all metrics.

Table 2: **Model Performance with / without Token Dropout (TD)**

Conditions	**Low-Level**				**High-Level**			
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
without TD	.038	**.307**	67.4%	75.5%	61.2%	64.8%	.974	.666
with TD	**.123**	.291	**85.4%**	**93.8%**	**91.8%**	**92.0%**	**.710**	**.407**

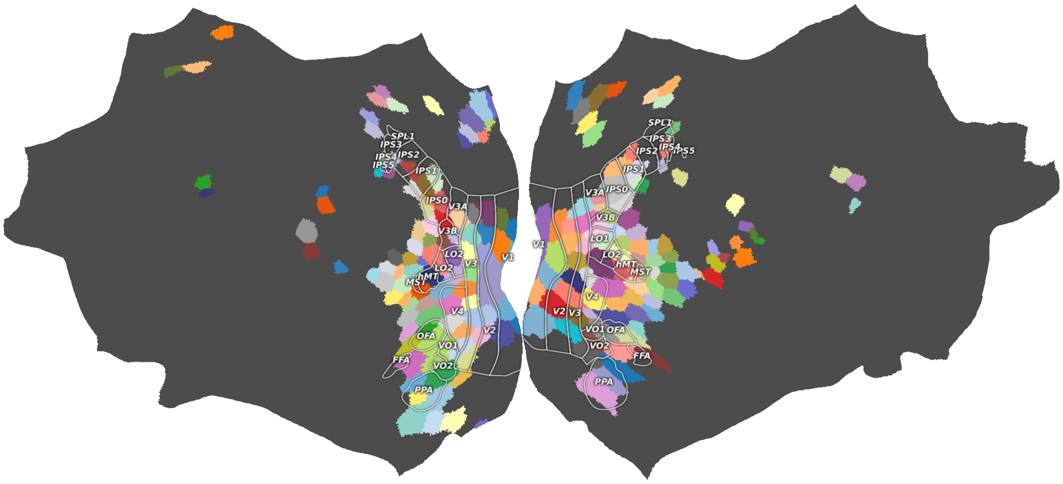

Figure 6: Top-100-SNR Parcels for each brain hemisphere displayed on the cortical surface.

E Ablation Study: Brain Encoder as a Ranking Tool

We further evaluate the role of the brain encoder as a selection mechanism for decoded stimuli. Instead of relying on a single generated stimulus for each fMRI input, we sample multiple candidates and use the brain encoder to rank them against measured neural responses. Table 3 shows that increasing the number of candidate predictions consistently improves decoding performance.

Table 3: **Performance across Different Image Quality Metrics**

Conditions	**Low-Level**				**High-Level**			
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
num of preds = 1	.104	.292	79.0%	90.1%	89.5%	90.8%	.729	.417
num of preds = 2	.105	.292	82.6%	91.7%	89.8%	88.7%	.733	.416
num of preds = 4	.120	**.293**	84.0%	93.5%	90.1%	91.5%	.725	.408
num of preds = 8	**.123**	.291	**85.4%**	**93.8%**	**91.8%**	**92.0%**	**.710**	**.407**

F Explanations of Min-SNR Loss Weighting

At each diffusion timestep t, the effective signal-to-noise ratio is defined as

$$\mathrm{SNR}_t = \frac{\bar{\alpha}_t}{1 - \bar{\alpha}_t},$$

where $\bar{\alpha}_t$ denotes the cumulative product of noise scheduling coefficients.

Without reweighting, high-SNR steps (early timesteps) tend to dominate the mean squared error (MSE) loss, while low-SNR steps (late timesteps) provide weaker gradients despite being more challenging and important for generation.

Ideally, the model should learn more from low-SNR noisy samples rather than overfitting to the easier, cleaner ones. Min-SNR weighting balances this trade-off by rescaling the per-timestep loss with

$$w_t = \frac{\min(\mathrm{SNR}_t, \gamma)}{\mathrm{SNR}_t},$$

where γ is a threshold hyperparameter (we set it to 5.0 in training).

G Details of IBBI Interpretability Analysis Framework

G.1 Preparations and Setups

In *NeuroAdapter*, each cross-attention layer computes attention scores $\mathrm{Attn}(\mathrm{Q}, \mathrm{K}, \mathrm{V})$, where queries $\mathrm{Q} \in \mathbb{R}^{q\times d}$ come from spatial tokens in the U-Net of SD, and keys and values $(\mathrm{K}, \mathrm{V}) \in \mathbb{R}^{p\times d}$ are derived from the fMRI embeddings E. At each denoising timestep t, the attention weight matrix $\mathrm{A}^{(\ell,h,t)} \in \mathbb{R}^{q\times p}$ for head h in layer ℓ encodes the influence of each parcel token on each spatial query. Each entry of the attention weight matrix can be expressed as:

$$\mathrm{A}_{i,j}^{(\ell,h,t)} \;=\; \frac{\exp\Big(\langle \mathrm{Q}_i^{(\ell,h,t)}, \mathrm{K}_j^{(\ell,h,t)}\rangle \,/\, \sqrt{d}\Big)}{\sum_{j'=1}^{P} \exp\Big(\langle \mathrm{Q}_i^{(\ell,h,t)}, \mathrm{K}_{j'}^{(\ell,h,t)}\rangle \,/\, \sqrt{d}\Big)}$$

Where query index $i \in \{1, \ldots, q\}$, and parcel index $j \in \{1, \ldots, p\}$. Specifically, the entry $\mathrm{A}_{i,j}^{(\ell,h,t)}$ refers to the attention from the i-th query vector $\mathrm{Q}_i^{(\ell,h,t)}$ to the j-th parcel token, represented by its key vector $\mathrm{K}_j^{(\ell,h,t)}$. Our bidirectional interpretability analysis further exploits this matrix in two complementary views.

G.2 Brain-directed View

Let q^ℓ be the number of spatial queries in layer ℓ. At denoising step t, each cross-attention map satisfies $\sum_{j=1}^{p} A_{i,j}^{(\ell,h,t)} = 1$ for every (ℓ, h, i). To aggregate the total attention mass assigned to each parcel across layers with different spatial resolutions, we weight every query equally and normalize by the total number of queries, $\sum_{\ell=1}^{L} q^\ell$. For each parcel $j \in \{1, \ldots, p\}$, we define

$$\mathrm{B}_j^{(t)} \;=\; \frac{1}{H\,\sum_{\ell=1}^{L} q^\ell} \sum_{\ell=1}^{L} \sum_{h=1}^{H} \sum_{i=1}^{q^\ell} A_{i,j}^{(\ell,h,t)}$$

Here, $\sum_{j=1}^{p} \mathrm{B}_j^{(t)} = 1$, so $\mathrm{B}^{(t)} \in \mathbb{R}^p$ can be interpreted as a query-weighted share of attention mass over parcels at timestep t. We project $\mathrm{B}^{(t)}$ onto the cortical surface to visualize *relative strength* across parcels.

G.3 Image-directed View

Conversely, for a given ROI group from parcels, denoted as $\mathcal{R} \subseteq \{1, \ldots, p\}$, we average attentions across heads and ROI tokens to form a query-wise attention profile for each layer:

$$m_{\mathcal{R}}^{(\ell,t)}(i) \;=\; \frac{1}{H}\frac{1}{|\mathcal{R}|} \sum_{h=1}^{H} \sum_{j\in\mathcal{R}} A_{i,j}^{(\ell,h,t)}$$

$m_{\mathcal{R}}^{(\ell,t)} \in \mathbb{R}^{q^\ell}$ is then reshaped to the spatial grid of layer ℓ and upsampled to image resolution, yielding $U_{\mathcal{R}}^{(\ell,t)} \in \mathbb{R}^{H_{\text{img}}\times W_{\text{img}}}$. To produce overlays that are comparable for spatial location across ROIs, we normalize each upsampled map to unit L_1 mass and then average uniformly across layers:

$$\mathrm{I}_{\mathcal{R}}^{(t)} \;=\; \frac{1}{L} \sum_{\ell=1}^{L} \frac{U_{\mathcal{R}}^{(\ell,t)}}{\sum_{x,y} U_{\mathcal{R}}^{(\ell,t)}(x,y)}$$

Where $\mathrm{I}_{\mathcal{R}}^{(t)}$ is a unit-mass heatmap that emphasizes *where* the ROI attends at timestep t.

G.4 Complementary Perspectives

The two views in our *IBBI* framework provide complementary insights into the role of brain representations during image reconstruction.

- **Brain-directed view** emphasizes the *relative strength* of influence across cortical parcels. It highlights which brain regions contribute most strongly at different stages of the denoising trajectory, providing a global map of neural contributions.
- **Image-directed view** emphasizes the *spatial footprint* of selected ROIs on the image space. It reveals how different cortical regions guide attention toward specific pixel-level locations in the image during reconstruction.

Together, this dual analysis allows us to interpret not only *which* brain regions matter, but also *how* their information is expressed in the unfolding visual reconstruction.