ReMindAR: Reconstruct Mind Autoregressively

Jiani Wang
2023010748

Chenxiao Yang
2023010573

Abstract

Reconstructing visual experiences from brain activity remains a central challenge in neuroscience and brain-computer interface research. While recent advances have predominantly leveraged diffusion-based generative models for fMRI-to-image reconstruction, these approaches often struggle with capturing the sequential and hierarchical nature of human visual perception. In this work, we propose the first autoregressive (AR)-based framework for visual reconstruction from fMRI signals, built upon the Visual Autoregressive Model (VAR). Our method decodes fMRI voxel patterns into multi-scale latent representations, which are then used to guide a coarse-to-fine autoregressive generation process that reconstructs realistic and semantically meaningful images. By aligning with the brain's hierarchical visual processing pathways, our framework achieves perceptually coherent and semantically aligned reconstructions. We further analyze the strengths and limitations of VAR in the context of neural decoding, offering new insights into its applicability for brain-to-image translation tasks. Our results suggest that AR-based models represent a promising alternative to existing diffusion-based approaches and open new avenues for future research in brain-aligned generative modeling. Code and models are publicly available at https://github.com/99ninew/ReMindAR

1. Introduction

The task of brain decoding—reconstructing visual stimuli or cognitive states from brain activity—remains a fundamental challenge in neuroscience. If we succeed in unraveling the secrets of human perception and cognition, there will be an advancement in brain-computer interfaces and artificial intelligence.

Recent advances have been driven by two complementary trends: advancements in neuroscience methodologies, which aim to acquire more precise and informative brain activity data (particularly via functional magnetic resonance imaging, fMRI), and the rise of deep learning, which offers powerful tools for modeling complex brain-image relationships and decoding perceptual experiences from neural patterns.

Building on these developments, many model architectures have been proposed to reconstruct visual stimuli from preprocessed fMRI data.

Inspired by the improvement in the field of visual generation, we explore an alternative decoding framework by shifting from the commonly used GAN or stable diffusion model to an autoregressive pattern, the visual autoregressive model (VAR) [16]. VAR is a new visual generative framework that first achieves a performance comparable to that of language-model-based AR models with strong diffusion models in terms of image quality, diversity, data efficiency, and inference speed.

This replacement is intended to better capture the sequential dependencies in visual data and improve the overall quality of image reconstruction. We aim to leverage the power of VAR, as its autoregressive learning and multi-scale prediction process aligns with the natural coarse-to-fine progression observed in human visual perception, which is highly correlated with our fMRI-to-image reconstruction framework.

Similar to the motivation behind most state-of-the-art (SOTA) models, our framework is grounded in the structure of human visual perception, which is generally understood as a two-stage process: (1) A high-level semantic pathway, responsible for extracting the content conveyed in the image by integrating prior knowledge and contextual cues. (2) A low-level perceptual pathway, which encodes basic visual elements such as the global structure of the image, its background, color, orientation, and texture. We define our main VAR pipeline as a balanced integration of both the low-level and high-level pathways. By effectively constructing both pathways, we aim to generate reconstructions that are both semantically accurate and perceptually coherent.

Our work makes the following contributions:

1. We propose the first AR-based pipeline for fMRI-to-image reconstruction.
2. We conduct a comprehensive investigation into the unique characteristics of VAR, highlighting its strengths in integrating novel perspectives from fMRI data, as well as examining its limitations.

2. Related Work

fMRI-to-Image Reconstruction. The task of brain decoding—reconstructing visual stimuli from functional magnetic resonance imaging (fMRI) data—has garnered increasing attention due to its significant implications for cognitive neuroscience and brain-computer interfaces. Since the release of the Natural Scenes Dataset (NSD), methods for this task have evolved iteratively. These approaches can be broadly categorized into GAN-based models, stable-diffusion-based generative frameworks, and optimized perceptual-semantic dual-pathway models. Early studies employed Generative Adversarial Networks (GANs)[3], Variational Autoencoders (VAE)[6], and self-supervised learning techniques to extract latent features for pixel-level reconstruction of visual stimuli. These methods, while effective in their respective domains, typically relied on relatively simple feature extraction strategies.

With the rise of deep neural networks, more sophisticated and complex paradigms have emerged for decoding brain signals. Notably, approaches using IC-GAN [4] or Stable diffusion [13],[8] have shown promise by mapping fMRI data to generative models capable of producing high-quality visual stimuli. These models leverage advanced architectures to enhance both image fidelity and perceptual relevance.

In addition, recent works such as Mindeye2[14], MindBridge[17], and NeuralDiffuser[8], have introduced cross-subject brain decoding frameworks. These studies address the challenge of generalizing over subject-specific fMRI data by aligning subject-independent voxels into a shared space, enabling more robust and transferable models for brain decoding across different individuals.

Visual feature guidance by neuroscience theory. Several approaches explore multi-pathway or multi-modal representations to align fMRI data with various levels of visual processing. For instance, Brain-Streams [5] employs cascaded pipelines involving high-level textual semantic encoders using BERT, mid-level visual semantic encoders using CLIP ViT-L/14, and low-level pixel-based representations via SD encoders. In most SOTA models, regardless of their specific architecture, a common approach is to use a high-level pipeline for extracting semantic information (textual and visual) and a low-level pipeline to capture perceptual features. These dual-stream frameworks effectively capture the hierarchical nature of human visual processing, though challenges remain in ensuring semantic accuracy with perceptual detailed reconstructions.

Autoregressive Image Generation. AutoRegressive (AR) models have recently emerged as competitive alternatives to diffusion models in image generation tasks. Recent advancements in visual AR modeling [16] have closed the performance gap with diffusion models, yielding higher ImageNet scores and paving the way for their application in neuroscientific contexts.

Position of Our Work. In contrast to previous efforts primarily based on stable diffusion models, our work investigates the potential of using a Visual autoregressive model (VAR) for brain decoding. While AR models have not been extensively explored in this domain, their structured generation process and multi-scale representation offer promising avenues for improved performance. We compare our approach with SOTA diffusion-based pipelines and evaluate its effectiveness through both perceptual and semantic metrics.

3. Method

3.1. Overview

Given an fMRI volume $\mathbf{v} \in \mathbb{R}^{N_v}$ containing N_v voxels, the goal is to recover the corresponding visual image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ perceived by the subject. This is achieved by mapping the high-dimensional, noisy voxel data into a compact and structured latent space. The reconstruction is then completed by a powerful generative model that decodes the latent codes back into images.

For the generative model, we are the first to try the autoregressive image generation model, instead of the diffusion model, to reconstruct the image from the latent representation. This will lead to different latent representations and different decoding processes compared to former methods. The overall pipeline is illustrated in Figure 1.

3.2. Voxel Decoding and Latent Space Supervision

Image encoding. For training, given an input image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, we use the vector quantized variational autoencoder trained in VAR model to map it to a latent representation $\{\mathbf{z}_k\}_{k=1}^{K}$ ($\mathbf{z}_k \in \mathbb{R}^{h_k\times w_k}$), where h_k and w_k are the height and width of the latent representation at scale k, respectively. The latent representation is perceived by the quantization process that selects the nearest discrete latent code from a predefined codebook $\mathcal{Z}$:

$$\mathbf{z}_k(i,j) = \underset{\mathbf{z}_{r_k}\in\mathcal{Z}}{\operatorname{argmin}} \|\mathcal{I}(\mathcal{E}(\mathbf{I}), h_k, w_k) - \mathbf{z}_{r_k}\|_2,$$

where $\mathcal{I}$ is the interpolation function and r_k is "image tokens" used in VAR. This representation provides a structured latent space that serves as the target domain for our neural decoding process, offering improved stability compared to discrete token spaces, particularly when working with the noisy, high-dimensional nature of fMRI data.

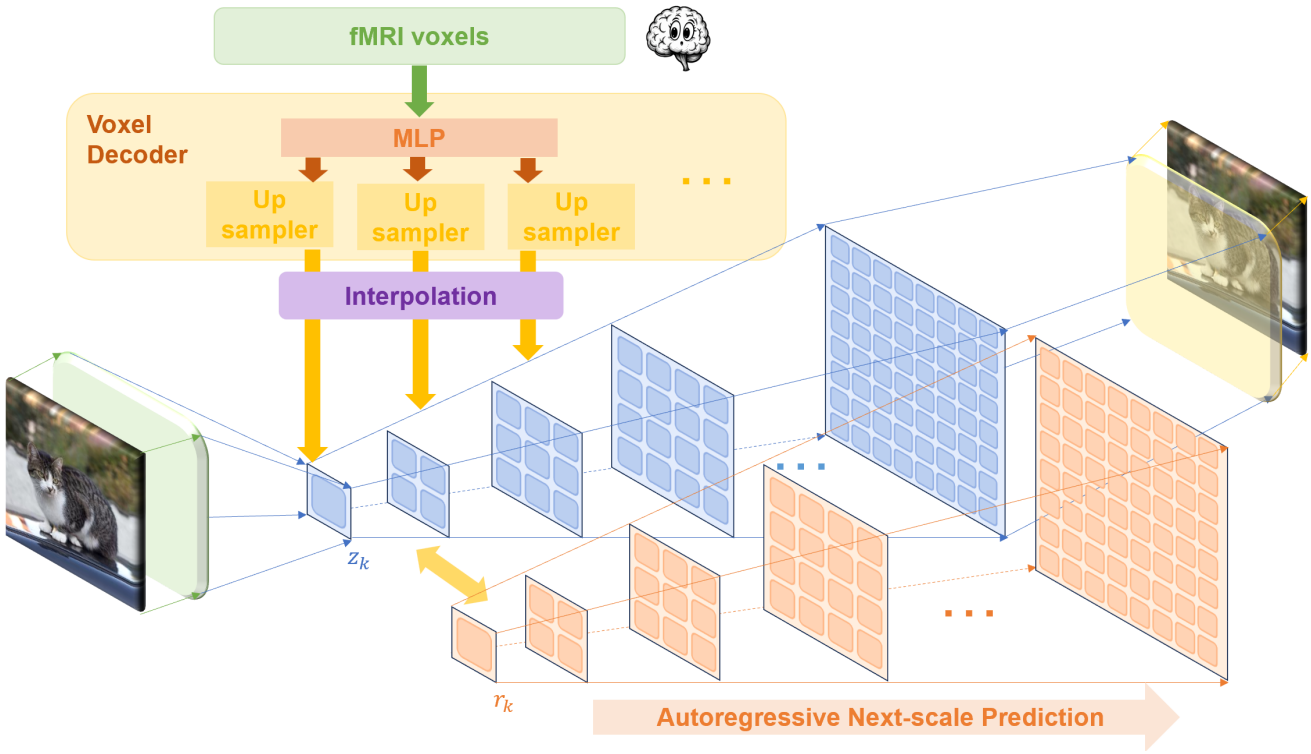

Figure 1. Overview of the proposed AR-based fMRI-to-image reconstruction pipeline. The model first decodes fMRI voxel data into multi-scale latent features using an MLP and upsampling modules. These features are then used to guide a visual autoregressive (VAR) model, which progressively predicts finer-scale representations and reconstructs the final image.

Voxel decoding. The core decoding model initially employs a Multi-Layer Perceptron (MLP), followed by K individual VAE decoders to upsample a hidden vector of size $(B, 256, 8, 8)$ into K tensors of size $(B, 32, 32, 32)$. Bilinear interpolation is then applied to obtain the targeted multi-scale embeddings, which is the final step in mapping voxel activity patterns to the corresponding VQVAE latent representations. Formally, we define a mapping function $f_\psi : \mathbb{R}^{N_v} \to \mathbb{R}^{K' \times H \times W}$ parameterized by ψ:

$$\hat{\mathbf{z}} = f_\psi(\mathbf{v})$$

where $0 \leq K' \leq K$ is a hyperparameter representing the number of scales of latent representations used in the decoding process.

Training loss. The model is trained to minimize the mean squared error (MSE) between the predicted latent representations $\hat{\mathbf{z}}$ and the ground truth latent representations $\mathbf{z}$ from the VQVAE encoder:

$$\mathcal{L}_{\text{MSE}} = \sum_{k=1}^{K'} \frac{1}{h_k w_k} \sum_{i=1}^{h_k} \sum_{j=1}^{w_k} \|\mathbf{z}_k(i,j) - \hat{\mathbf{z}}_k(i,j)\|_2^2$$

This training objective ensures that the MLP learns to accurately predict latent codes that encode the essential visual features of the input stimuli.

3.3. Image Reconstruction

To reconstruct the image from the predicted latent representation, we generate images autoregressively using the VAR architecture in a next-scale generation manner. We replace the first K' scales of latent representations $\{\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{K'}\}$ with the predicted latent codes $\hat{\mathbf{z}}_0, \hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_{K'}$ from the voxel decoding step as the guidance for the beginning of the autoregressive generation process. Then the model generates the remaining latent representations and finally reconstructs the image with prior knowledge learned from the VAR model. We argue that only a few scales of latent representations are sufficient to learn the global structure of the image and reconstruct the most important visual features.

Our method can also be combined with the diffusion-based pipeline, such as the high level semantic pipeline mentioned in MindEye[13] to achieve better semantic representation capabilities and image quality, incorporating the strengths of the diffusion model.

4. Experiments

All experiments were conducted on a single A100 GPU.

Figure 2. The reconstruction results generated in the VAR pipeline. For each pair, the ground truth stimulus appears on the left, and the corresponding reconstruction is shown on the right.

4.1. Dataset

We utilize the **Natural Scenes Dataset (NSD)** [1], a state-of-the-art, large-scale fMRI dataset for studying brain activity in response to naturalistic visual stimuli selected from part of the MS COCO dataset [10]. Each subject in the dataset underwent approximately 30 to 40 scanning sessions while viewing thousands of richly detailed natural scene images.

Among the four individuals (subjects 1, 2, 5, and 7) who completed all 40 sessions, we used fMRI data from subject 1 (subj01) to train a subject-specific model. In order to fairly compare with the performance across fMRI-to-Image NSD papers, we use the same standardized train/test splits as other NSD reconstruction papers [13] [5]. The training dataset consists of 24,980 fMRI-image pairs, while the test dataset comprises 982 images that are disjoint from the training data. Those images were viewed by all four subjects. For more data selection and preprocessing details, see the supplementary material at 6.

4.2. fMRI-to-Image Reconstruction results

4.2.1. Qualitative results

Figure 2 presents the reconstructions generated by our model from fMRI data, alongside the corresponding ground truth stimuli. The generation results of joint pipeline leveraging diffusion-based semantic pipeline are presented in Figure 3.

4.2.2. Quatitative results

Following the standard NSD metrics to assess reconstruction quality at different levels, **ReMindAR** performs on par with the SOTA methods, [15] [4] [5] [7][8] [10] [9] [11] [12] [13] [18]. As shown in Table1, we emphasize the **best**, the second, and the third.

Among all the evaluation metrics, AlexNet(2), AlexNet(5), CLIP and InceptionV3 are two-way identification metrics to evaluate feature similarity; The structural similarity index metric (SSIM) and pixel-wise correlation (PixCorr) measure pixel-level similarity between reconstructed images and original stimuli; EfficientNet-B1 (Eff) and SwAV-ResNet50 (SwAV) refer to average correlation distance.

Figure 3. The reconstruction results generated in the joint pipeline.

4.2.3. Ablation studies

Reconstruction strategies. To separate the effect of each pipeline, we conducted ablation experiments on them to evaluate their performance. A quantitative comparison across different pipelines is provided in Table 2.

It shows that the VAR pipeline performs better on pixel-to-pixel similarity (SSIM, PixCorr) than the CLIP pipeline, which confirms that by using a multi-scale, coarse-to-fine progression learning architecture, VAR is relatively capable of balancedly capturing the global structure and local details.

While certain artifacts, such as irregular lines, incomplete items, and patterns composed of repeating elements sometimes occur in the reconstructed images when only applying the VAR pipeline, our analysis suggests that it could be attributed to training biases inherent in the VAR architecture. These biases likely stem from the use of a pre-trained codebook, originally trained on ImageNet [2], to obtain embeddings.

Method		Low-Level (perceptual)				High-Level (semantic)			
		SSIM↑	PixCorr↑	AlexNet(2)↑	AlexNet(5)↑	IncepV3↑	CLIP↑	Eff↓	SwAV↓
GANs	MindReader (NeurIPS'22)[9]	–	–	–	–	78.2%	–	–	–
	Gu et al. (MIDL'23) [4]	0.325	0.150	–	–	–	–	0.862	0.465
Stable-Diffusion	Takagi et al. (CVPR'23) [15]	–	–	83.0%	83.0%	76.0%	77.0%	–	–
	BrainDiffuser (Sci-Rep'23) [12]	0.356	0.254	94.2%	96.2%	87.2%	91.5%	0.775	0.423
	MindEye (NeurIPS'23) [13]	0.337	**0.390**	97.4%	98.7%	94.5%	94.6%	**0.630**	0.358
Optimized	MindDiffuser (ACM MM'23) [11]	0.354	0.278	–	97.7%	93.9%	93.9%	0.645	0.367
	MindEye+BOI (ArXiv'23) [7]	0.329	0.259	93.9%	96.7%	93.7%	94.1%	0.645	0.418
	DREAM (WACV'24) [18]	0.338	0.288	93.9%	96.7%	93.7%	94.1%	–	–
	Brain-Streams (24) [5]	**0.365**	0.342	94.7%	97.0%	94.0%	95.2%	0.651	**0.357**
	NeuralDiffuser (25)[8]	0.330	0.378	**98.3%**	**99.4%**	**95.3%**	**95.8%**	0.642	0.380
VAR	**ReMindAR (Ours)**	0.322	0.288	94.9%	98.1%	94.4%	94.9%	0.637	0.360

Table 1. Quantitative comparison of ReMindAR's reconstruction performance on perceptual and semantic evaluation metrics against other models. (If other methods have evaluation for the subject-specific model on subject 1 in their original paper, we will use them; otherwise, we use the average value among 4 subjects, which is also acceptable since the variances are small.) **Best**, second and third are emphasized.

Method	Low-Level (perceptual)				High-Level (semantic)			
	SSIM↑	PixCorr↑	AlexNet(2)↑	AlexNet(5)↑	IncepV3↑	CLIP↑	Eff↓	SwAV↓
ReMindAR - only VAR pipeline	**0.392**	0.280	91.4%	92.1%	75.5%	74.8%	0.873	0.555
ReMindAR - only CLIP pipeline	0.318	0.209	92.8%	98.0%	**94.5%**	94.8%	**0.635**	0.361
ReMindAR	0.322	**0.288**	**94.9%**	**98.1%**	94.4%	**94.9%**	0.637	**0.360**

Table 2. Comparing reconstruction results via the usage of each pipeline. The **best** is in bold.

Method	Low-Level (perceptual)				High-Level (semantic)			
	SSIM↑	PixCorr↑	AlexNet(2)↑	AlexNet(5)↑	IncepV3↑	CLIP↑	Eff↓	SwAV↓
ReMindAR (K'=3, w/ lookup)	**0.390**	0.266	89.4%	91.4%	74.4%	75.1%	0.887	0.536
ReMindAR (K'=2, w/o lookup)	0.389	0.261	89.4%	91.5%	75.0%	75.2%	0.885	0.532
ReMindAR (K'=3, w/o lookup)	0.322	**0.288**	**94.9%**	**98.1%**	**94.4%**	**94.9%**	**0.637**	**0.360**

Table 3. Comparative Analysis of Different Learning Objectives in Training the VAR Pipeline. st represents the number of scales input to the Transformer for guidance. The terms w/ and w/o refer to whether or not a vector quantization step is applied to the predicted latent embeddings during the inference process. The **best** is in bold.

Investigation in the VAR architecture. Although VAR has achieved near-optimal ImageNet FID scores, its performance on more challenging image generation tasks has not yet been proven when compared to other state-of-the-art (SOTA) generative models. In our pipeline, a learnable embedding model maps voxels from fMRI data into multi-scale latent vectors, aligning them with the corresponding ground truth encoded in the VAR architecture. This introduces a novel setting with several aspects requiring further investigation:

- The optimal number of scales used as latent representation K' remains unclear. Therefore, the optimal number of scales for the predicted latent embeddings from the fMRI voxels, which guide the generation process, needs to be verified.
- After getting the latent representations $\{\mathbf{z}_k\}$, we explored whether conducting an additional nearest neighbor looking-up in the codebook $\mathcal{Z}$ for every $\mathbf{z}_k$ would help.

To explore the above aspects of VAR, we conducted ablation experiments, as shown in Table 3. For each of the aspect, we discover that:

- Utilizing four scales of predicted embeddings from the voxels as input to the Transformer for guidance results in better performance than using only three scales, which are both much better than any other number of scales. Although previous experiments [1] indicate that if the first two scales are generated well, the remaining tokens have limited impact on the final image quality, incorporating more guidance at different resolutions leads to higher reconstruction accuracy, especially when dealing with challenging and unfamiliar voxel inputs. While the voxel decoder has limited prediction ability, too much number of scales would increase error and lead to poor image generation quality.
- We found that directly using the latent representations predicted by the voxel decoders without conducting the re-lookup yielded better reconstruction quality. This could be because of the domain gap between the codebook learned in VAR and the actual images. The decoders would possibly have learned features in the latent space and done generalization in the face of images out of the domain. The re-lookup operation would ruin the robustness of the model.

5. Discussion

Conclusion. In this work, we present the first autoregressive (AR)-based pipeline for reconstructing visual images from fMRI data, leveraging the recently proposed Visual Autoregressive Model (VAR). Our approach departs from the prevalent use of diffusion models and instead explores the potential of AR models to more effectively capture the sequential dependencies inherent in visual perception. By aligning the reconstruction process with the coarse-to-fine dynamics of human visual cognition, our framework integrates low-level perceptual and high-level semantic pathways to produce reconstructions that are both structurally coherent and semantically meaningful.

Future work. While significant progress has been made in designing the model architecture and constructing the VAR pipeline in conjunction with the CLIP pipeline, there remain several opportunities for further refinement. Specifically, the integration of auxiliary losses to better define the learning objective could enhance the model's performance. In future work, we aim to optimize our model to achieve superior reconstruction quality.

In addition, as highlighted in the ablation studies and supported by our empirical results, the current VAR model still exhibits domain bias due to the use of vector quantization (VQ) and the unproven effects of multi-scale guidance. Further research into the key factors influencing the effectiveness of the VAR architecture is crucial, not only for developing a superior generative model but also for guiding the focus of future iterations of fMRI-to-image reconstruction models. Such research will be essential for advancing conditional image generation across a wide range of applications.

[1] https://zhouyifan.net/blog-en/2024/12/21/20241218-VAR/

References

[1] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022. 4, 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 2

[4] Zijin Gu, Keith Jamison, Amy Kuceyeski, and Mert Sabuncu. Decoding natural image stimuli from fmri data with a surface-based convolutional network, 2023. 2, 4, 5

[5] Jaehoon Joo, Taejin Jeong, and Seongjae Hwang. Brainstreams: fmri-to-image reconstruction with multi-modal guidance, 2024. 2, 4, 5, 1

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 2

[7] Reese Kneeland, Jordyn Ojeda, Ghislain St-Yves, and Thomas Naselaris. Brain-optimized inference improves reconstructions of fmri brain activity, 2023. 4, 5

[8] Haoyu Li, Hao Wu, and Badong Chen. Neuraldiffuser: Neuroscience-inspired diffusion guidance for fmri visual reconstruction. *IEEE Transactions on Image Processing*, 34: 552–565, 2025. 2, 4, 5

[9] Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities, 2022. 4, 5

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 4, 1

[11] Yizhuo Lu, Changde Du, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion, 2023. 4, 5

[12] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion, 2023. 4, 5

[13] Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Demp-

ster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors, 2023. 2, 3, 4, 5, 1

[14] Paul S. Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data, 2024. 2

[15] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14453–14463, 2023. 4, 5

[16] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction, 2024. 1, 2

[17] Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework, 2024. 2

[18] Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system, 2024. 4, 5

ReMindAR: Reconstruct Mind Autoregressively

Supplementary Material

6. Additional Dataset Information

Natural Scenes Dataset (NSD) [1] provides high-resolution 7-Tesla fMRI data paired with visual stimuli from part of the MS COCO dataset [10]. Each subject in the dataset underwent approximately 30 to 40 scanning sessions while viewing thousands of richly detailed natural scene images.

Among the four individuals (subjects 1, 2, 5, and 7) who completed all 40 sessions, we used the fMRI data of subject 1 (subj01) to train a subject-specific model. To fairly compare the performance across fMRI-to-Image NSD papers, we use the same standardized train/test splits as other NSD reconstruction papers, [13], [5]. The training dataset consists of 24,980 fMRI-image pairs, while the test dataset comprises 982 images that are disjoint from the training data. All four subjects viewed those images.

Following the procedure used in other reconstruction papers [13], [5], we use the single-trial beta estimates generated from a generalized linear model (GLM) with fitted hemodynamic response functions, further refined by the GLMDenoise and ridge regression procedures (`betas_fithrf_GLMdenoise_RR`). These outputs provide 1.8 mm isotropic voxel-wise resolution, offering more precise and analyzable brain activation patterns. In addition, to narrow the focus to regions of the brain most relevant to visual processing, the model uses the NSDGeneral ROI (Region of Interest) mask provided, restricting our analysis to 15,724 voxels for the subject (subj01). Additionally, we leverage the corresponding MS COCO captions as auxiliary semantic inputs in our decoding framework.