ConnecToMind: Connectome-Aware fMRI Decoding for Visual Image Reconstruction

Gunwoo Bae[1,†], Yeonwoo Kim[2,†], and Mansu Kim[1,3,*]

[1] Department of AI convergence, Gwangju Institute of Science and Technology, Korea
[2] Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Korea
[3] GIST InnoCORE AI-Nano Convergence Initiative for Early Detection of Neurodegenerative Diseases, Gwanjgu Institute of Science and Technology, Korea
mansu.kim@gist.ac.kr

Abstract. Recent deep-learning approaches have achieved significant improvements in reconstructing visual images from human brain activity. However, existing methods typically represent brain activity as flattened voxel-wise signals, overlooking the detailed anatomical and functional organization of visual cortical regions. Here, we propose ConnecToMind, a novel decoding framework that employs a region-level fMRI embedding module to preserve distinct functional representations across visual cortical sub-regions, while leveraging functional connectivity (FC) derived from resting-state fMRI. Experiments on the Natural Scenes Dataset (NSD) demonstrate that ConnecToMind outperforms the MindEye in both the semantic and perceptual fidelity of reconstructed images, validating the effectiveness of preserving distinct functional representations with FC prior. Moreover, ConnecToMind shows competitive performance in image retrieval tasks. Ablation analyses further reveal that low-level (e.g., V1–V3) and high-level (e.g., Lateral Occipital, Fusiform) visual regions distinctly contribute to the reconstruction quality, highlighting the importance of region-specific embeddings in visual reconstruction. All codes for this study are publicly available at GitHub (`https://github.com/aimed-gist/ConneToMind`).

Keywords: fMRI-to-image · Diffusion model · functional connectivity · Natural Scene Dataset

1 Introduction

Reconstructing visual images from human brain activity has recently seen remarkable progress with deep learning-based decoding models. For example, Takagi & Nishimoto pioneered latent-diffusion-based reconstruction, mapping fMRI signals directly into the Stable Diffusion latent space to generate high-resolution outputs [1]; MindEye decomposed diffusion-prior submodules to effectively reduce the modality gap between CLIP image embeddings and fMRI-derived latent

[†]G. Bae and Y. Kim contributed equally to this work.

representations [2]; and MindEye2 aligned data in a shared-subject latent space for pretraining and then fine-tuned on new-subject data, thereby achieving robust cross-subject generalizability [3]. These approaches typically map voxel activity patterns from the visual cortex measured by fMRI into the latent feature space of generative models [1–3].

Emerging neuroscience studies emphasize that brain anatomical structure and inter-regional circuits are crucial for visual processing. The primate visual cortex comprises distinct regions (e.g., V1–V4, IT), each with specialized functional roles interconnected by a hierarchical network [4,5]. In particular, feedback connections (signals sent back from higher to earlier visual areas) from mid-level areas like V4 to the primary visual cortex refine and contextualize visual representations, enhancing figure-ground segmentation [6]. Moreover, recognizing even well-defined objects requires recurrent loops between the visual cortex and frontal cortex, underscoring the essential role of inter-regional connectivity [7].

Despite their success, most recent models represent brain activity as unstructured voxel-wise signals. For instance, MindEye directly embeds visual cortical fMRI signals into the CLIP image embedding space, enabling high-fidelity reconstructions via diffusion models [2]. However, by flattening the visual cortex into simple vector representations, these approaches overlook its fine-grained anatomical and functional organization [2].

To address these limitations, we propose a novel decoding framework, termed ConnecToMind, which integrates region-level fMRI embeddings with functional connectivity (FC) prior derived from resting-state fMRI data. Our primary scientific contributions are as follows: (1) we introduce a region-level fMRI embedding module that explicitly preserves anatomical and functional distinctions between visual cortical regions; (2) we apply a Connectome-Transformer that utilizes FC within its self-attention mechanism to explicitly model intrinsic inter-regional relationships; (3) we demonstrate that combining FC prior with positional embeddings significantly improves the semantic coherence and perceptual fidelity of reconstructed images. Overall, ConnecToMind leverages hierarchical brain connectivity to achieve more accurate and interpretable visual reconstruction from fMRI signals.

2 Methodology

In this study, we propose ConnecToMind, a novel framework that extends the MindEye [2] by integrating region-level fMRI embeddings and a Connectome-Transformer guided by resting-state FC. A diffusion prior subsequently aligns the embeddings generated by the Connectome-Transformer into the CLIP image embedding space for visual reconstruction. The overall framework is illustrated in Figure 1.

2.1 Region-level fMRI embedding module

We develop a region-level fMRI embedding module to represent brain activity in the visual cortex, as presented in Figure 1-(a). Specifically, we parcellate the

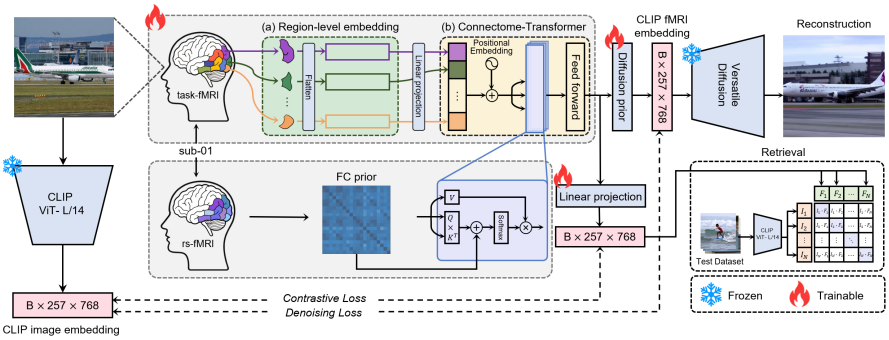

Figure 1. Overview of ConnecToMind. ConnecToMind is consisted of (a) region-level fMRI embeddings module and (b) Connectome-Transformer guided by FC prior. The output of Connectome-Transformer is subsequently fed in parallel to diffusion models for image reconstruction and an image retrieval head.

NSDgeneral mask [8], which includes visual areas from early visual cortex to higher-order visual regions, into 20 cortical sub-regions based on the Desikan-Killiany atlas [9]. These sub-regions include the Cuneus, Fusiform, Inferior parietal, Inferior temporal, Lateral occipital, Lingual, Middle temporal, Parahippocampal, Pericalcarine, and Superior parietal areas in both hemispheres. We then linearly project the fMRI response values from each region into a shared embedding space.

Given the flattened voxel responses in the i-th region as $x_i \in \mathbb{R}^{1\times d_i}$, the projected embedding for the i-th region $z_i \in \mathbb{R}^{1\times 768}$ is computed as follows:

$$z_i = x_i w_i + e_i, \tag{1}$$

where $w_i \in \mathbb{R}^{d_i\times 768}$ represents a trainable weight matrix, $e_i \in \mathbb{R}^{1\times 768}$ is a bias term, and d_i denotes the number of voxels in the i-th brain region. The embeddings from all regions are concatenated to form the region-level fMRI embeddings, $Z = [z_1, z_2, \ldots, z_{20}] \in \mathbb{R}^{20\times 768}$, which is then used as input tokens to the Connectome-Transformer.

2.2 Connectome-Transformer with FC-attention mechanism

In this section, we extend the original connectome transformer [10] as shown in Figure 1-(b) by adding FC prior into its self-attention mechanism and incorporating an inter-regional positional embedding.

Given a region-level fMRI embedding $Z \in \mathbb{R}^{20\times 768}$, we first add a learnable positional embedding $P \in \mathbb{R}^{20\times 768}$ to obtain the region-aware fMRI embedding $\hat{Z} = Z + P$. Multi-head self-attention (MSA) is then computed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + FC\right)V, \tag{2}$$

where $FC \in \mathbb{R}^{20\times20}$ is the resting-state FC prior, and $Q = \hat{Z}W_q$, $K = \hat{Z}W_k$, and $V = \hat{Z}W_v$ are computed via linear projections with learnable projection matrices $W_q, W_k, W_v \in \mathbb{R}^{768\times768}$, and per-head dimensionality $d_k = \frac{768}{n_h}$ is determined by the number of heads n_h for scaling the dot product. Splitting these into n_h heads, each computes attention independently, and then their outputs are concatenated and linearly projected by a matrix W^O to produce the multi-head self-attention:

$$\mathrm{MSA}(\hat{Z}) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_{n_h})W^O, \tag{3}$$

where $\mathrm{head}_i = \mathrm{Attention}(Q_i, K_i, V_i)$ and $W^O \in \mathbb{R}^{768\times768}$ is a learnable linear projection.

2.3 Diffusion model

Motivated by recent successes in diffusion-based models for fMRI-based image reconstruction [1–3], we employ a diffusion prior [11] to map CLIP fMRI embeddings into the corresponding CLIP image embedding space [2]. The diffusion process comprises two phases: forward diffusion and reverse diffusion.

During the forward diffusion process, Gaussian noise is gradually added to the CLIP fMRI embeddings y_t over multiple diffusion timesteps $t \in \{0, 1, \ldots, T\}$:

$$y_t = \sqrt{\alpha_t}\, y_{t-1} + \sqrt{1-\alpha_t}\, \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I), \tag{4}$$

where α_t denotes the noise scaling factor at timestep t, and ϵ_t represents the Gaussian noise sampled from a standard normal distribution $\mathcal{N}(0, I)$.

In the reverse diffusion process, our denoising network learns to reconstruct the target CLIP image embedding y from the noisy input y_t, obtained by adding noise to the CLIP fMRI embedding. Following [11], we train the prior by minimizing the expected denoising loss:

$$\mathcal{L}_{prior} = \mathbb{E}_{t\sim[1,T], y_t\sim q_t}\left[\|f_\theta(y_t, t, Z) - y\|_2^2\right], \tag{5}$$

where $f_\theta(y_t, t, Z)$ denotes a transformer-based diffusion model that takes the noisy embedding y_t, the diffusion timestep t, and the CLIP fMRI embedding Z as inputs to predict the target CLIP image embedding y. By aligning the CLIP fMRI embedding with the CLIP image embedding, this formulation effectively reduces the modality gap between brain signals and visual representations during the reconstruction process [2]. Finally, the denoised CLIP fMRI embedding $\hat{y}$ is fed into a pretrained Versatile Diffusion model to generate the final reconstructed image [12].

2.4 Loss functions

The loss function of our model consists of two main components: a denoising loss for the diffusion process and a contrastive loss. Specifically, the denoising loss (Equation 5) directly estimates the denoised embedding $\hat{y}$ and employs a Mean Squared Error (MSE) loss [11]. For the contrastive loss, we follow a two-stage

training approach inspired by the BiMixCo and SoftCLIP losses [2]. During the initial one-third of training epochs, we utilize the BiMixCo loss, which incorporates bidirectional MixCo augmentation and hard labels to explicitly maximize the cosine similarity between positive embedding pairs while minimizing it between negative pairs. For the subsequent two-thirds of training epochs, we transition to the SoftCLIP loss, employing soft labels derived from CLIP image embeddings without augmentation. The total loss is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{prior}} + \alpha_1 \, \mathcal{L}_{\text{BiMixCo|SoftCLIP}}, \tag{6}$$

where α_1 balances the contributions of the diffusion prior loss and the contrastive loss [2].

2.5 Implementation details

We train models for 220 epochs with a batch size of 160, using AdamW [13] and the OneCycleLR scheduler [14] on NVIDIA H100 GPUs. Region-level fMRI embeddings (Sec. 2.1) are computed as $z_i = x_i w_i + e_i$, using independently learned weights $w_i \in \mathbb{R}^{d_i \times 768}$ and biases e_i using PyTorch's einsum function [15]. The Connectome-Transformer (Sec. 2.2) comprises one attention layer (8 heads), a feedforward dimension of 2048, and dropout of 0.1. Its output is linearly projected from $\mathbb{R}^{20\times768}$ to $\mathbb{R}^{257\times768}$. Diffusion models (Sec. 2.3) follow hyperparameters from MindEye [2].

3 Experimental results

3.1 Experimental setup

Data description and preparation We utilize multi-modal neuroimaging data, including T1-weighted MRI, rs-fMRI, and task-fMRI, collected from the Natural Scenes Dataset (NSD) [8]. NSD provides high resolution neuroimaging data acquired from eight participants over 30–40 sessions using 7-Tesla MRI. Details are available on the NSD website. For each session, participants view images sampled from the COCO dataset [16] while performing a continuous recognition test.

Following the experimental setup of a previous study [1], we select subject 1, who completed 28,557 task-fMRI trials by viewing 9,519 unique images, each repeated three times. Among these, 25,611 trials corresponding to 8,537 images are used as the training set. The remaining 2,946 trials are averaged across the three repetitions per image, resulting in a final test set of 982 samples.

We preprocess the fMRI scans following the procedure used in the NSD dataset. Specifically, we perform temporal resampling to correct for slice-timing differences, and spatial resampling to correct the head motion within and across scan sessions, EPI distortion, and gradient nonlinearities [17]. After preprocessing, we estimate single-trial beta responses using a generalized linear model (GLM) implemented in GLMsingle [17]. We then apply session-wise z-scoring to the beta responses and construct region-level representations for each of the 20 cortical sub-regions defined in Section 2.1.

Evaluation We conduct two comprehensive evaluation strategies: image reconstruction and retrieval. These two evaluations quantify our model's capability to reconstruct accurate visual stimuli from fMRI signals and to retrieve corresponding visual stimuli based on learning representations from ConnecToMind.

For image reconstruction, we compare reconstructed image against its ground-truth image using both low-level and high-level metrics. Low-level perceptual fidelity is assessed by pixel correlation (PixCorr), structural similarity index (SSIM), and feature-space correlations derived from the 2nd and 5th convolutional layers of a pretrained AlexNet [18]. High-level semantic fidelity is computed via inception scores [19], CLIP similarity scores [20], embedding distances computed from EfficientNet-B [21] and SwAV networks [22] which quantify semantic similarity between reconstructions and visual stimulus images in pretrained deep visual representations. These metrics are computed using the same settings as in previous studies [2, 3, 23, 24] to ensure consistent evaluation.

For image retrieval, we quantify the alignment between neural representations derived from fMRI signals and their corresponding visual stimulus embeddings. For example, we compute cosine similarities between the fMRI-derived query embedding $F_{(i)}$ and the stimulus-derived target embeddings $I_{(j)}$. Retrieval accuracy is calculated over randomly sampled batches for computational efficiency. To ensure robustness, we repeat this procedure 30 times and report the average as the retrieval accuracy.

3.2 fMRI-to-Image Reconstruction

We evaluate the quality of fMRI-to-image reconstruction by comparing our proposed ConnecToMind model against the baseline MindEye, which utilizes flattened voxel-wise signals. We further assess ablated models: ConnecToMind-FC (without FC prior) and ConnecToMind-FCP (without FC prior and positional embedding). Reconstruction fidelity is evaluated using eight metrics capturing low-level perceptual and high-level semantic quality.

Quantitative and qualitative results are shown in Table 1 and Figure 2. ConnecToMind achieves the best overall performance across both low-level and high-level metrics. As shown in Figure 2, reconstructions from ConnecToMind preserve higher perceptual fidelity and semantic accuracy compared to other models. In the evaluation of perceptual fidelity, ConnecToMind outperforms other models in low-level metrics (PixCorr: 0.373, SSIM: 0.357, AlexNet(2): 95.8%),

Method	Low-Level				High-Level				Retrieval	
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Image↑	Brain↑
MindEye	**.373**	.356	95.6%	97.5%	93.8%	93.7%	.657	.388	**95.6%**	**92.2%**
ConnecToMind-FCP	.368	.353	95.5%	97.6%	94.4%	94.3%	.649	.378	88.9%	85.0%
ConnecToMind-FC	.370	.355	95.4%	**97.7%**	**94.7%**	94.4%	.645	.378	93.8%	90.8%
ConnecToMind(Ours)	**.373**	**.357**	**95.8%**	97.6%	94.3%	**94.7%**	**.644**	**.376**	94.3%	90.1%

Table 1. Model performance across low-level, high-level, and retrieval metrics.

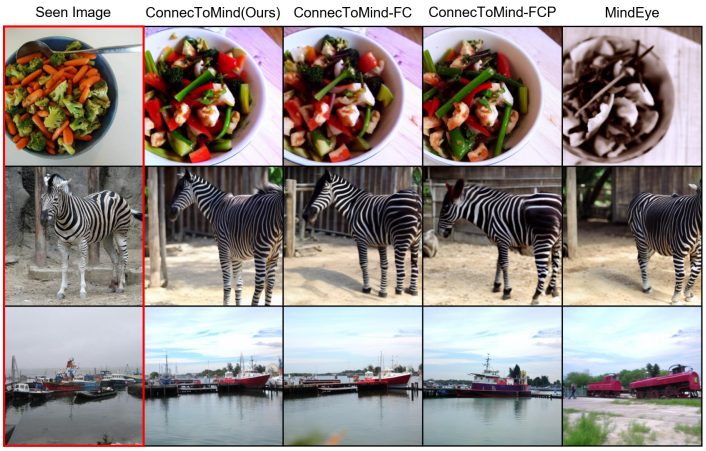

Figure 2. Qualitative comparison of reconstructed images across models.

indicating superior preservation of perceptual information. For semantic accuracy, it achieves the highest scores in high-level metrics (CLIP: 94.7%, EfficientNet: 0.644, SwAV: 0.376). ConnecToMind-FC shows the strongest performance in the remaining metrics (AlexNet(5): 97.7% and Inception: 94.7%). These results underscore the effectiveness of region-level modeling compared to flattened voxel-based methods.

3.3 Image/Brain Retrieval

We evaluate retrieval performance to measure how accurately neural embeddings from fMRI signals can be matched to their corresponding visual stimulus embeddings. Specifically, we adopt the retrieval evaluation protocol proposed by Mind Reader [25], which consists of two complementary retrieval tasks: forward retrieval (Image-based) and backward retrieval (Brain-based). Forward retrieval assesses the ability to identify the target embedding derived from the stimulus given an fMRI-derived query embedding, while backward retrieval measures the ability to identify the target embedding derived from fMRI given a stimulus-derived query embedding. Retrieval success for each sample is defined as the case where the matched query-target pair yields the highest cosine similarity among all pairs in the batch. We perform this evaluation with a batch size of 300, randomly sampling batches and repeating the retrieval process over 30 iterations across the entire test set to obtain robust and reliable performance estimates.

Table 1 presents retrieval results across evaluated models. MindEye achieves the highest retrieval accuracy (Image-based: 95.6%, Brain-based: 92.2%), while our proposed ConnecToMind yields slightly lower but still comparable performance (Image-based: 94.3%, Brain-based: 90.1%). The ablated models, ConnecTo Mind-FC and ConnecToMind-FCP, show lower retrieval accuracies (93.8%, 88.9% Image-based; 90.8%, 85.0% Brain-based, respectively).

Method	Low-Level				High-Level			
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
ConnecToMind	0.373	0.357	95.8%	97.6%	94.3%	94.7%	0.644	0.376
ConnecToMind-L	0.367	0.350	95.0%	97.4%	94.1%	94.2%	0.656	0.384
ConnecToMind-H	0.372	0.353	95.6%	97.1%	92.7%	93.4%	0.669	0.392

Table 2. Impact of sub-regional visual cortex on image reconstruction.

3.4 Ablation Study on Impact of Visual Cortical Sub-regions

Cortical regions in the visual cortex process information at distinct hierarchical levels. Early visual regions, including V1–V3, primarily encode low-level perceptual details, while higher-order regions such as the Lateral Occipital and Fusiform specialize in semantic processing [4]. To examine the contributions of these cortical regions, we trained two additional model variants: ConnecToMind-L (excluding early visual regions), and ConnecToMind-H (excluding higher-order visual regions).

Quantitative results are shown in Table 2, and qualitative comparisons are illustrated in Appendix Figure A1. The full ConnecToMind model achieves the best overall performance across both low-level and high-level reconstruction metrics. Notably, the ConnecToMind-L variant exhibits significant reductions in low-level metrics (PixCorr: 0.367, SSIM: 0.350, AlexNet(2): 95.0%), with relatively minor decreases in high-level semantic metrics. Conversely, the ConnecToMind-H variant shows substantial drops in high-level semantic metrics (Inception: 92.7%, CLIP similarity: 93.4%, EfficientNet: 0.669, SwAV: 0.392), while maintaining comparatively stable low-level perceptual performance. These results confirm that each cortical region subset makes distinct and meaningful contributions to visual image reconstruction quality, thereby supporting our ConnecToMind approach, which separately processes fMRI signals according to distinct cortical sub-regions.

4 Conclusion

In this study, we propose ConnecToMind, a novel decoding framework for reconstructing visual images from fMRI signals. ConnecToMind introduces a region-level fMRI embedding module that preserves anatomical and functional distinctions across visual cortical regions, addressing limitations of previous voxel-based approaches. We further present a Connectome-Transformer that leverages FC prior as a guiding bias within its self-attention mechanism, effectively capturing intrinsic inter-regional relationships. Our results demonstrate that ConnecToMind significantly enhances both the perceptual and semantic fidelity of reconstructed images. By explicitly modeling hierarchical cortical organization and incorporating FC, ConnecToMind offers a more accurate and interpretable approach for fMRI-to-image reconstruction.

Acknowledgments This work was partly supported by the NRF grant funded by the Korean government (No. RS-2025-00521250), by the IITP grants funded by the Korean government (No. RS-2021-II212068, Artificial Intelligence Innovation Hub; No. 2019-0-01842, Artificial Intelligence Graduate School Program [GIST], No.2022–0–00448/RS-2022-II220448, Deep Total Recall: Continual Learning for Human-Like Recall of Artificial Neural Networks), by the MOTIE under the "Infrastructure Program for Industrial Innovation" supervised by the Korea Institute for Advancement of Technology (KIAT) (No. RS-2024-00434342), and by InnoCORE-GIST program of the Ministry of Science and ICT.

References

1. Yu Takagi and Shinji Nishimoto. 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*, pages 14453–14463, 2023.
2. Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.
3. Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.
4. Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.
5. Nikola T Markov, Julien Vezoli, Pascal Chameau, Arnaud Falchier, René Quilodran, Cyril Huissoud, Camille Lamy, Pierre Misery, Pascale Giroud, Shimon Ullman, et al. Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *Journal of comparative neurology*, 522(1):225–259, 2014.
6. Charles D Gilbert and Wu Li. Top-down influences on visual processing. *Nature reviews neuroscience*, 14(5):350–363, 2013.
7. Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B Issa, and James J DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience*, 22(6):974–983, 2019.
8. 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, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.
9. Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.
10. Diego Machado-Reyes, Mansu Kim, Hanqing Chao, Li Shen, and Pingkun Yan. Connectome transformer with anatomically inspired attention for parkinson's diagnosis. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–4, 2022.

11. Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
12. Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023.
13. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
14. Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.
15. A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.
16. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.
17. Jacob S Prince, Ian Charest, Jan W Kurzawski, John A Pyles, Michael J Tarr, and Kendrick N Kay. Improving the accuracy of single-trial fmri response estimates using glmsingle. *Elife*, 11:e77599, 2022.
18. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
19. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
20. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
21. Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
22. Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.
23. Furkan Ozcelik and Rufin VanRullen. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion. arxiv 2023. *arXiv preprint arXiv:2303.05334*, 2023.
24. Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11333–11342, 2024.
25. Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*, 35:29624–29636, 2022.