ConnecToMind2: Inter-Subject fMRI Decoding via Whole-Brain Connectome-Guided Alignment

Gunwoo Bae[1,†], Yeonwoo Kim[2,†], Junha Bae[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, Gwangju Institute of Science and Technology, Korea
mansu.kim@gist.ac.kr

Abstract. Recent advances in diffusion-based generative models have substantially improved image reconstruction from fMRI signals. However, most existing approaches rely on visual cortical activity, flatten voxel responses into unstructured representations, and require subject-specific adaptation. Here we propose ConnecToMind2, a whole-brain neural decoding framework that models fMRI activity at the region level and incorporates structural connectivity (SC) priors. Whole-brain signals are first aligned to a shared anatomical space and projected into regional embeddings that preserve cortical organization. These embeddings are then aligned to the CLIP image embedding space via a Connectome-guided Q-Former, where SC priors guide cross-attention to enforce inter-regional interactions. We evaluate our model on the Natural Scenes Dataset (NSD) under intra-subject and inter-subject settings. ConnecToMind2 achieves competitive performance across perceptual and semantic metrics. Ablation analysis confirms that essential information for reconstruction is distributed throughout the whole-brain rather than confined to the visual network. All codes for this study are publicly available at GitHub (`https://github.com/aimed-gist/ConnecToMind2`).

Keywords: fMRI-to-image · Diffusion model · Structural connectivity · Natural Scenes Dataset.

1 Introduction

Neural decoding reconstructs perceptual and semantic content from brain activity, providing insights into human cognition and enabling brain-computer interfaces [15]. With the advent of diffusion-based generative models, image reconstruction from functional Magnetic Resonance Imaging (fMRI) has substantially improved in quality [21]. Despite these advances, robust fMRI-to-image alignment remains difficult due to pronounced inter-subject variability in neural

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

responses [4,7]. This variability highlights the need for decoding frameworks that learn transferable representations across subject.

Most recent approaches follow a common paradigm: they extract voxel responses primarily from the visual cortex, flatten them into 1-dimensional vectors, and align them to CLIP image embeddings [21,22]. A diffusion model then uses these aligned embeddings to reconstruct images [21,22]. However, this strategy treats complex brain activity as an unstructured vector, despite the fact that visual perception is a hierarchical process that involves the whole-brain, not just the visual cortex [3]. By flattening the voxel responses, existing methods discard the anatomical and functional organization of large-scale brain networks [2,12].

Neuroscience studies indicate that visual perception is not confined to the visual cortex [6,8]. It relies on the coordinated activity of whole-brain functional networks [3]. Beyond early visual areas, visual processing engages distributed sub-networks to interpret stimuli. For instance, the dorsal and ventral attention networks direct spatial focus and detect salient features during visual tasks [29]. Furthermore, higher-order systems such as the frontoparietal control and default mode networks provide top-down modulation, integrating external visual inputs with internal cognitive states [25]. This extensive inter-regional connectivity indicates that accurate visual decoding requires modeling broad cortical networks rather than isolating the visual cortex.

In this study, we propose *ConnecToMind2*, a whole-brain neural decoding framework designed for anatomically informed image reconstruction. Our main contributions are as follows: **1)** We leverage region-level whole brain activities within a shared anatomical space, preserving the intrinsic anatomical organization rather than flattening voxels into unstructured vectors. **2)** We introduce a *Connectome-Guided Q-Former* that aligns regional fMRI embeddings to the CLIP embeddings by explicitly incorporating SC as an anatomical prior within its cross-attention mechanism. **3)** Our method promotes subject-invariant brain embeddings through structurally consistent representations, enabling robust inter-subject image reconstruction without the need for subject-specific adapter.

2 Method

We propose ConnecToMind2, a framework that decodes visual stimuli from whole-brain fMRI signals without subject-specific adapter layers. As illustrated in Fig. 1, the framework consists of three main components. First, the whole-brain regional embedding module divides the input fMRI signals into distinct brain regions and extracts regional embeddings. Second the Connectome-Guided Q-Former aligns these regional fMRI embeddings to the latent space of a pre-trained diffusion model by utilizing SC as an anatomical prior. Finally, the pre-trained diffusion model generates the reconstructed images conditioned on the aligned embeddings.

2.1 Whole-Brain Regional Embedding Module

We develop a whole-brain regional embedding module to capture broad neural representations that may contribute to visual decoding. We incorporate regions

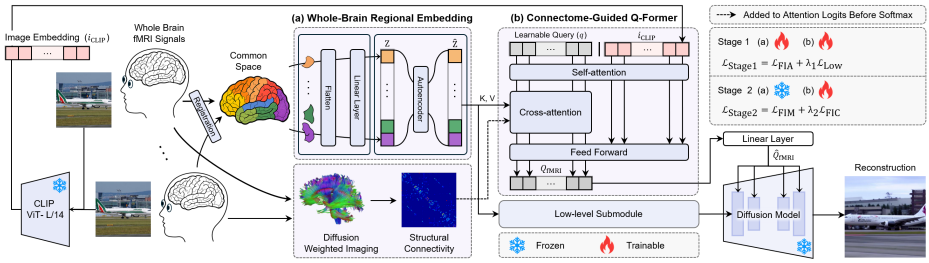

Fig. 1: Overview of ConnecToMind2. ConnecToMind2 consists of (a) whole-brain regional embedding module and (b) Connectome-Guided Q-Former. The output of the Connectome-Guided Q-Former, along with the features from a low-level submodule, is subsequently fed to a pre-trained diffusion model for final image reconstruction.

across multiple functional networks [29], including the visual, dorsal attention, ventral attention, default mode, frontoparietal control, somatomotor, and limbic networks. Specifically, we parcellate the cortex into $N = 100$ regions using the Schaefer atlas [20].

Given the flattened voxel responses in the i-th region as $\mathbf{x}_i \in \mathbb{R}^{1\times d_i}$, where d_i denotes the number of voxels in the i-th region, the region-specific embedding $\mathbf{z}_i \in \mathbb{R}^{1\times 768}$ is computed as follows:

$$\mathbf{z}_i = \sigma(\mathbf{x}_i \mathbf{w}_i + \mathbf{e}_i), \tag{1}$$

where $\mathbf{w}_i \in \mathbb{R}^{d_i\times 768}$ is a trainable weight matrix, $\mathbf{e}_i \in \mathbb{R}^{1\times 768}$ is a bias term, and σ is an activation function. The embeddings from all N regions are then concatenated to form the regional embedding $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N] \in \mathbb{R}^{N\times 768}$.

Since each region is processed by an independent projection, the resulting representations lie in different subspaces. To project these regional embeddings into a common representational space, we process the concatenated embedding $\mathbf{Z}$ through an autoencoder. The projected regional embedding $\hat{\mathbf{Z}} \in \mathbb{R}^{N\times 768}$ is computed as:

$$\hat{\mathbf{Z}} = f_{\mathrm{AE}}(\mathbf{Z}), \tag{2}$$

where $f_{\mathrm{AE}}(\cdot)$ denotes the autoencoder bottleneck network with a hidden dimension of 128.

2.2 Connectome-Guided Q-Former

In this section, we introduce a Connectome-Guided Q-Former to align the projected regional embedding $\hat{\mathbf{Z}}$ into the CLIP representations necessary for diffusion-based generation. This architecture consists of stacked blocks containing self-attention, cross-attention, and feed-forward layers. We define a set of N learnable query embeddings $\mathbf{q} \in \mathbb{R}^{N\times 768}$, where N equals the number of parcellated brain regions.

In the self-attention layers, we concatenate the query tokens $\mathbf{q}$ with the CLIP image embeddings $\mathbf{i}_{\text{CLIP}} \in \mathbb{R}^{257\times768}$ to form $[\mathbf{q}; \mathbf{i}_{\text{CLIP}}] \in \mathbb{R}^{(N+257)\times768}$. This allows the query tokens to interact directly with the image representations. Subsequently, in the cross-attention layers, the query tokens interact with the fMRI embeddings $\hat{\mathbf{Z}}$. To guide the learnable queries to capture region-specific representations, we integrate the subject's $\mathbf{SC} \in \mathbb{R}^{N\times N}$ as an anatomical prior. By adding this $\mathbf{SC}$ prior directly to the attention logits, we bias the network to assign attention weights based on anatomical connections. This enables the queries to extract region-level fMRI information. The cross-attention mechanism is formulated as follows:

$$\text{CrossAttn}(\mathbf{q}, \hat{\mathbf{Z}}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{SC}\right)\mathbf{V}, \tag{3}$$

where $\mathbf{Q} = \mathbf{q}\mathbf{W}_Q$, $\mathbf{K} = \hat{\mathbf{Z}}\mathbf{W}_K$, and $\mathbf{V} = \hat{\mathbf{Z}}\mathbf{W}_V$. The matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{768\times768}$ are trainable linear projections, and $d_k = 768/h$ represents the dimensionality of each of the h attention heads.

Following the cross-attention layers, the tokens pass through feed-forward layers to complete the block operations. After processing through the stacked blocks, we extract the updated representations corresponding only to the query positions, discarding the CLIP image portion. The resulting fMRI-aligned queries, denoted as $\mathbf{Q}_{\text{fMRI}} \in \mathbb{R}^{N\times768}$, are linearly projected along the token dimension to obtain CLIP-aligned fMRI embeddings $\hat{\mathbf{Q}}_{\text{fMRI}} \in \mathbb{R}^{257\times768}$, matching the input shape required by the pre-trained diffusion model.

2.3 Diffusion-Based Image Generation

We employ a pre-trained Versatile Diffusion (VD) [28] for the final image reconstruction. To capture both high-level semantic content and low-level perceptual details, this generation process integrates two distinct inputs: a semantic condition and an initial latent representation [4].

For the semantic condition, we utilize the CLIP-aligned fMRI embeddings $\hat{\mathbf{Q}}_{\text{fMRI}} \in \mathbb{R}^{257\times768}$ obtained from the Connectome-Guided Q-Former. These embeddings replace the standard CLIP image representations and are fed into the cross-attention layers of the diffusion model. To preserve low-level spatial structures, we construct an initial latent representation $\hat{\mathbf{L}} \in \mathbb{R}^{4\times64\times64}$ within VD latent space. Conceptually, a low-level submodule composed of an MLP projector and a CNN upsampler transforms the projected regional embedding $\hat{\mathbf{Z}}$ to formulate this initial latent.

During the image generation process, the initial latent $\hat{\mathbf{L}}$ is noised up to a predefined timestep. VD then iteratively denoises this latent while being continuously conditioned on the CLIP-aligned fMRI embeddings $\hat{\mathbf{Q}}_{\text{fMRI}}$, producing the final reconstructed image.

2.4 Two-Stage Training Strategy

Stage 1: Coarse Alignment and Low-Level Reconstruction In the first stage, we train the whole-brain regional embedding module, the Connectome-

Guided Q-Former, and the low-level submodule. The goal is to establish stable regional embeddings, achieve a coarse alignment with the CLIP image embedding space, and reconstruct the initial latent.

For the high-level semantic objective, we apply the fMRI-Image-Alignment (FIA) loss, which minimizes the L2 distance between the L2-normalized fMRI queries and the CLIP image embeddings:

$$\mathcal{L}_{\text{FIA}} = \|\hat{\mathbf{Q}}_{\text{fMRI}} - \mathbf{i}_{\text{CLIP}}\|_2^2. \tag{4}$$

Simultaneously, the low-level loss supervises the initial latent prediction by minimizing the L1 distance between the predicted latent $\hat{\mathbf{L}}$ and the target latent $\mathbf{L}$ encoded by the pre-trained VAE $\mathcal{L}_{\text{Low}} = \|\hat{\mathbf{L}} - \mathbf{L}\|_1$. The total loss for Stage 1 is defined as:

$$\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{\text{FIA}} + \lambda_1 \mathcal{L}_{\text{Low}}, \tag{5}$$

where λ_1 is scaling hyperparameter.

Stage 2: Fine-Grained Semantic Alignment In the second stage, we focus on training the Connectome-Guided Q-Former to achieve fine-grained semantic alignment. During this phase, the whole-brain regional embedding module is frozen to provide stable input representations. The objective function relies on fMRI-Image-Contrastive (FIC) and fMRI-Image-Matching (FIM) losses [10].

The FIC loss applies symmetric cross-entropy for contrastive learning. Here, self-attention is masked to prevent interactions between query tokens and CLIP image tokens:

$$\mathcal{L}_{\text{FIC}} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ij}/\tau)} + \log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ji}/\tau)} \right], \tag{6}$$

where $s_{ij} = \max_{n=1}^{N} \cos(\mathbf{q}_i^{(n)}, \mathbf{i}_{\text{CLIP}}^{(j)})$ is the maximum cosine similarity between the i-th fMRI and j-th image representations, B is the batch size, and τ is a learnable temperature.

The FIM loss performs binary classification to predict if an fMRI-image pair matches ($y = 1$) or not ($y = 0$). For this task, unmasking self-attention allows bidirectional token interactions. A linear classifier f_{cls} maps each token to a logit, and the predictions are averaged:

$$\mathcal{L}_{\text{FIM}} = \text{CE}\left(\frac{1}{M} \sum_{m=1}^{M} f_{\text{cls}}(\hat{\mathbf{q}}_m),\ y \right), \tag{7}$$

where M is the number of query tokens ($M = 257$), $\hat{\mathbf{q}}_m$ is the m-th token of $\hat{\mathbf{Q}}_{\text{fMRI}}$, and CE is the cross-entropy loss. Following [11], we mine in-batch hard negatives. The total loss for Stage 2 is:

$$\mathcal{L}_{\text{Stage2}} = \mathcal{L}_{\text{FIM}} + \lambda_2 \mathcal{L}_{\text{FIC}}, \tag{8}$$

where λ_2 is a scaling hyperparameter.

2.5 Implementation Details

We train the proposed framework for 150 epochs on five NVIDIA L40 GPUs with a total batch size of 100 (20 per GPU), utilizing DDP, DeepSpeed ZeRO-2 [19], and FP16 mixed precision. The network is optimized using AdamW [14] (weight decay 10^{-2}) and a OneCycleLR scheduler [23]. The learning rate peaks at 3×10^{-4} for the first 40 epochs, transitioning to a constant 1.2×10^{-8} thereafter. As outlined in Sec. 2.4, the two-stage training evenly splits at epoch 75. Loss weights are empirically set to $\lambda_1 = 0.01$, $\lambda_2 = 0.1$.

For the Connectome-Guided Q-Former, we comprises 12 transformer with cross-attention applied at alternating layers. Its self-attention and feed-forward weights are initialized from a pre-trained CLIP ViT-B/16, while target representations $\mathbf{i}_{\text{CLIP}}$ are extracted via a frozen CLIP ViT-L/14 [18] image encoder.

3 Experiments

3.1 Experimental Setup

Dataset preparation We use the Natural Scenes Dataset (NSD) [1], comprising high-resolution fMRI data acquired while subjects viewed images sampled from the COCO dataset [13]. We include four subjects (sub-01, sub-02, sub-05, and sub-07), each completing 40 sessions with 10,000 unique images presented three times, yielding 30,000 single-trial beta responses per subject. We train on 9,000 subject-specific images and test on 1,000 shared images, averaging beta responses across repetitions for evaluation.

Single-trial betas are derived using GLMSingle [17] after several preprocessing steps. The preprocessing steps include fieldmap correction using FUGUE [9], registration to MNI152 space [5], and voxel-wise intensity normalization using training set statistics. Region-level representations are constructed by parcellating beta responses with Schaefer-100 atlas [20]. For diffusion Magnetic Resonance Imaging data, we perform several preprocessing steps, including distortion and motion correction, denoising, Gibbs ringing artifact correction, and registration [9]. Following preprocessing, we compute the fiber orientation distribution and perform whole-brain tractography [24] to calculate the SC matrix based on the Schaefer-100 atlas.

Evaluation We evaluate reconstruction quality using both low-level and high-level metrics under the Brain-Diffuser protocol [16]. We use PCorr and SSIM for low-level evaluation, alongside the features from the 2nd and 5th convolutional layers of a pretrained AlexNet. For high-level semantic evaluation, we employ Inception, CLIP similarity, EfficientNet-B1, and SwAV.

Model compared We compare our method with four state-of-the-art fMRI-to-image reconstruction models, categorized into intra-subject and inter-subject baselines: MindEye1 [21] and ConnecToMind [2], and two inter-subject baselines: MindBridge [26] and Umbrae [27]. To ensure a fair comparison, all baseline

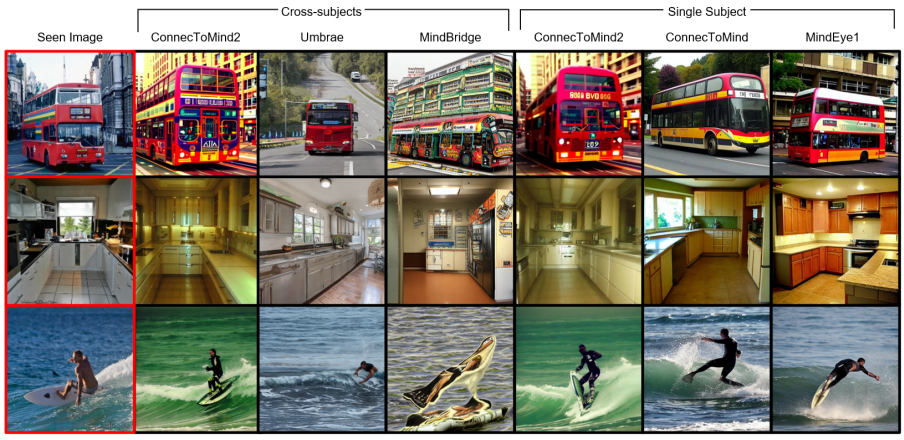

Fig. 2: Qualitative comparison of image reconstructions.

models are trained from scratch utilizing their official implementations and hyperparameter settings. Note that the inputs for these benchmarks are strictly restricted to the visual network regions. In contrast, ConnecToMind2 uses whole-brain regional embeddings in both intra- and inter-subject evaluations.

3.2 Experimental results

fMRI-to-Image Reconstruction We evaluate the fMRI-to-image reconstruction quality under two experimental paradigms: intra- and inter-subject decoding. Quantitative and qualitative results are summarized in Table 1 and Fig. 2, respectively. For intra-subject evaluation, separate models are trained and tested for each subject. Our proposed framework achieves the highest PCorr and SSIM scores, indicating improved preservation of low-level structural details. Furthermore, it yields competitive performance across high-level semantic metrics, indicating that semantic alignment within the CLIP space is well-preserved despite the broader anatomical input.

Table 1: Quantitative comparison of image reconstruction performance under intra-subject and inter-subject settings. Results are averaged across subjects.

	Regions	Model	Low-Level				High-Level			
			PCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Intra	Visual	MindEye1	0.167	0.290	88.5%	95.4%	92.0%	91.0%	0.686	0.402
	Visual	ConnecToMind	0.170	0.297	88.5%	95.9%	93.5%	92.3%	0.666	0.388
	Whole	**ConnecToMind2**	0.278	0.387	87.9%	93.4%	88.1%	89.7%	0.769	0.487
Inter	Visual	MindBridge	0.091	0.214	75.3%	86.7%	85.2%	80.9%	0.853	0.553
	Visual	Umbrae	0.054	0.233	70.7%	85.2%	86.7%	**88.2%**	**0.748**	**0.482**
	Whole	**ConnecToMind2**	**0.263**	**0.407**	**86.5%**	**92.6%**	**86.8%**	87.1%	0.771	0.487

Fig. 3: Comparison of image reconstructions without low- and high-level regions.

For inter-subject evaluation, a single unified model is trained across all subjects. ConnecToMind2 outperforms state-of-the-art models on five out of eight metrics, particularly showing substantial gains in PCorr and SSIM, while preserving strong semantic consistency across Inception, CLIP, and SwAV metrics. Qualitative results further demonstrate that the reconstructed images exhibit similar object structures and semantic content across both intra- and inter-subject settings (Fig. 2), underscoring the inter-subject generalizability of the proposed framework.

Ablation on Early Visual and Higher-Order Regions To validate the distributed characteristics of visual processing, we identify cortical parcels corresponding to *early visual regions* (L) and *higher-order regions* (H) within the whole-brain Schaefer atlas [20]. To evaluate their specific contributions, we introduce two ablated variants: *ConnecToMind2 (w/o L)*, which excludes the early visual areas, and *ConnecToMind2 (w/o H)*, which excludes the higher-order regions. Quantitative and qualitative results are presented in Table 2 and Fig. 3 Our results demonstrate a clear functional dissociation. *ConnecToMind2 (w/o L)* suffers a significant loss in all low-level metrics, suggesting that whole-brain regions associated with low-level visual processing play a critical role in preserving fine-grained spatial details. Conversely, *ConnecToMind2 (w/o H)* exhibits a substantial drop in high-level metrics, such as Inception and CLIP similarity, indicating that distributed higher-order brain regions contribute to semantic and contextual representations in the reconstructed images.

Table 2: Impact of Early Visual and Higher-Order Regions

Method	Low-Level				High-Level			
	PCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
ConnecToMind2	0.263	0.407	86.5%	92.6%	86.8%	87.1%	0.771	0.487
ConnecToMind2 (w/o L)	0.175	0.367	79.5%	88.9%	85.7%	86.2%	0.811	0.517
ConnecToMind2 (w/o H)	0.257	0.418	83.5%	90.0%	84.4%	85.8%	0.799	0.511

4 Conclusion

In this study, we propose ConnecToMind2, a whole-brain neural decoding framework for image reconstruction. ConnecToMind2 introduces a regional embedding module that preserves the intrinsic anatomical and functional organization of the entire cortex. Connectome-Guided Q-Former leverages SC as an anatomical prior to facilitate efficient alignment with the CLIP embedding space. Our results on the NSD dataset demonstrate that ConnecToMind2 achieves robust performance in both intra- and inter-subject reconstruction without subject-specific adaptation. Furthermore, our ablation analysis validates that regions beyond the visual network provide essential information for reconstruction. However, independent validation across different scanning environments and populations remains a challenge for ensuring the model's real-world reliability.

Acknowledgments. This research was supported by the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT (MSIT) (RS-2025-00521250). This work was also partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) funded by MSIT (No.2019-0-01842, Artificial Intelligence Graduate School Program, GIST), and by the InnoCORE program of MSIT (GIST InnoCORE, KH0860).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al.: A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience **25**(1), 116–126 (2022)
2. Bae, G., Kim, Y., Kim, M.: Connectomind: Connectome-aware fmri decoding for visual image reconstruction. In: International Workshop on Machine Learning in Medical Imaging. pp. 400–410. Springer (2025)
3. Bressler, S.L., Menon, V.: Large-scale brain networks in cognition: emerging methods and principles. Trends in cognitive sciences **14**(6), 277–290 (2010)
4. Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22710–22720 (2023)
5. Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., Group, B.D.C., et al.: Unbiased average age-appropriate atlases for pediatric studies. Neuroimage **54**(1), 313–327 (2011)
6. Goodale, M.A., Milner, A.D.: Separate visual pathways for perception and action. Trends in neurosciences **15**(1), 20–25 (1992)
7. Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., Ramadge, P.J.: A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron **72**(2), 404–416 (2011)

8. Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L.: A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron **76**(6), 1210–1224 (2012)
9. Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M.: Fsl. NeuroImage **62**(2), 782–790 (2012), 20 YEARS OF fMRI
10. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
11. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
12. Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L.H., Ventola, P., Duncan, J.S.: Braingnn: Interpretable brain graph neural network for fmri analysis. Medical image analysis **74**, 102233 (2021)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
15. Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.a., Morito, Y., Tanabe, H.C., Sadato, N., Kamitani, Y.: Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron **60**(5), 915–929 (2008)
16. Ozcelik, F., VanRullen, R.: Natural scene reconstruction from fmri signals using generative latent diffusion. Scientific Reports **13**(1), 15666 (2023)
17. Prince, J.S., Charest, I., Kurzawski, J.W., Pyles, J.A., Tarr, M.J., Kay, K.N.: Improving the accuracy of single-trial fmri response estimates using glmsingle. Elife **11**, e77599 (2022)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
19. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: Memory optimizations toward training trillion parameter models. In: SC20: international conference for high performance computing, networking, storage and analysis. pp. 1–16. IEEE (2020)
20. Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.: Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. Cerebral cortex **28**(9), 3095–3114 (2018)
21. Scotti, P., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Dempster, A., Verlinde, N., Yundler, E., Weisberg, D., Norman, K., 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)
22. Scotti, P.S., Tripathy, M., Villanueva, C.K.T., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K.A., et al.: Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. arXiv preprint arXiv:2403.11207 (2024)
23. Smith, L.N.: Cyclical learning rates for training neural networks. In: 2017 IEEE winter conference on applications of computer vision. pp. 464–472. IEEE (2017)

24. Tournier, J.D., Calamante, F., Connelly, A.: Mrtrix: Diffusion tractography in crossing fiber regions. International Journal of Imaging Systems and Technology **22**(1), 53–66 (2012)
25. Vincent, J.L., Kahn, I., Snyder, A.Z., Raichle, M.E., Buckner, R.L.: Evidence for a frontoparietal control system revealed by intrinsic functional connectivity. Journal of neurophysiology (2008)
26. Wang, S., Liu, S., Tan, Z., Wang, X.: Mindbridge: A cross-subject brain decoding framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11333–11342 (2024)
27. Xia, W., de Charette, R., Oztireli, C., Xue, J.H.: Umbrae: Unified multimodal brain decoding. In: European Conference on Computer Vision. pp. 242–259. Springer (2024)
28. Xu, X., Wang, Z., Zhang, G., Wang, K., Shi, H.: Versatile diffusion: Text, images and variations all in one diffusion model. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7754–7765 (2023)
29. Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., et al.: The organization of the human cerebral cortex estimated by intrinsic functional connectivity. Journal of neurophysiology (2011)