This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

MindLink: Subject-agnostic Cross-Subject Brain Decoding Framework

Sungyoon Jung, Donghyun Lee, and Won Hwa Kim

Pohang University of Science and Technology (POSTECH), Pohang, South Korea
{syjung, dongtak, wonhwa}@postech.ac.kr

Abstract. Brain decoding is a pivotal topic in neuroscience, aiming to reconstruct stimuli (e.g., image) from brain activity (e.g., fMRI). However, existing methods rely on subject-specific modules and flatten 3D voxel grids, limiting generalization and discarding spatial information. To address these issues, we propose *MindLink*, a scalable cross-subject brain decoding framework designed to link multiple subjects into a single model by extracting subject-invariant features while preserving the spatial structure of 3D fMRI data. This is achieved by parcellating 3D fMRI into standardized cubic patches processed by a 3D Vision Transformer for informative representations. Domain adversarial training enhances cross-subject generalizability by extracting subject-agnostic features within a single model structure. We also introduce a two-level alignment strategy that effectively bridges fMRI and stimuli image embeddings through instance-level consistency and flexible token-level matching. *MindLink* achieves comparable or even better performance over state-of-the-art methods on the NSD dataset with a constant parameter size across subjects and demonstrates strong adaptability to new subject.

Keywords: Cross-subject Brain Decoding · Domain Adversarial Training · Multi-modal Alignment

1 Introduction

Brain decoding is important for understanding the intricate mechanisms of human cognition and perception. It aims to interpret neural activities given certain conditions, and functional magnetic resonance imaging (fMRI) facilitates such analysis by providing detailed neural activity *in vivo* non-invasively with high spatial and temporal resolution [11]. The decoding component relies on recent advances in generative models [5, 7, 14] for reconstructing sensory and cognitive representations. While recent works have demonstrated a significant potential in brain decoding, high variability in neural activity across subjects highlight the need for more robust and generalizable brain decoding approaches [3].

Conventionally, brain decoding has been limited to subject-specific applications [12, 15, 19], i.e., models trained on the brain activity of a specific subject could not be adopted for others. While efforts have been made for cross-subject brain decoding for generality [22, 24], existing approaches still rely on subject-specific modules, increasing model complexity as the number of subjects grows.

This poses a significant challenge by limiting scalability for larger populations in real-world scenarios. Moreover, in prior studies [12, 15, 19, 22, 24], 3D voxel grids are flattened into a vector during the fMRI preprocessing stage. This simplification discards spatial information and limits the model's ability to capture complex spatial relationships inherent in neural activity patterns.

To address the two major issues above, we propose *MindLink*, a scalable cross-subject brain decoding framework, designed to **link** multiple subjects into a single model while preserving spatial structure. *MindLink* achieves this by parcellating 3D fMRI into standardized cubic patches, which are processed by a 3D Vision Transformer [9] to preserve spatial information. Domain adversarial training is used to extract subject-agnostic features while maintaining a single model structure to avoid subject-specific biases. Afterward, we introduce a two-level alignment strategy to bridge the gap between fMRI and image embeddings. At the instance level, fMRI embeddings are aligned with pretrained image embeddings in scale and direction, ensuring compatibility of image generation without fine-tuning. At the token level, we introduce a cross-attention mechanism dynamically matches fMRI and image tokens, capturing context and preventing misalignment. By jointly leveraging instance-level consistency and token-level flexibility, our approach achieves accurate and coherent image reconstructions.

To this end, our **main contributions** are summarized as follows: **1)** We present *MindLink*, which links multiple subjects with a *single model* to extract subject-invariant features while preserving the spatial structure of 3D fMRI. **2)** We propose a two-level alignment strategy that effectively bridges fMRI and image embeddings through instance-level consistency and flexible token-level matching. **3)** Our method achieves comparable or even better performance on the Natural Scenes Dataset (NSD), maintaining a constant parameter size across subjects and demonstrating strong adaptability to new subjects.

Related Work on Brain Decoding Recent approaches in brain decoding focus on aligning brain activity and stimuli within pretrained embedding spaces using regression [12,19] or contrastive learning [15]. As the previous methods follow per-subject-per-model paradigm, cross-subject brain decoding methods have been proposed, which extract subject-invariant features through cyclic fMRI reconstruction mechanism [22] or ridge regression [24]. Nevertheless, these methods either require subject-specific parameters [22, 24] or face efficiency issues due to fine-tuning large language model (LLM) [16], moreover, all the previous methods discard spatial information in the brain [12, 15, 19, 22, 24]. In contrast, we preserve spatial structure while ensuring efficient cross-subject generalization by parcellating 3D fMRI into standardized cubic patches and using a single model.

2 Method

2.1 Space-preserving 3D fMRI Data Processing

fMRI data captures neural activity by measuring blood-oxygen-level-dependent (BOLD) changes in 3D voxel grids. However, traditional preprocessing methods

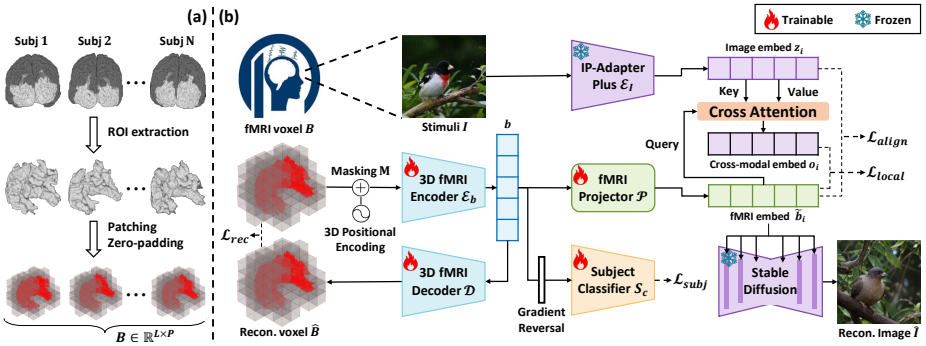

Fig. 1. Overview of MindLink. (a) 3D fMRI preprocessing standardizes brain volumes into cubic patches. (b) 3D fMRI Encoder $\mathcal{E}_b$ is trained to extract subject-invariant fMRI embeddings. The fMRI projector $\mathcal{P}$ then projects fMRI embeddings into image latent space, which is utilized to reconstruct images through stable diffusion model.

typically flatten this 3D structure into 1D vectors [12, 15, 19, 22, 24], leading to significant loss of spatial structural information. To address this challenge, we propose a minimal preprocessing strategy that preserves spatial structure through ROI extraction, zero-padding and patching, as illustrated in Fig. 1a.

Given an original 3D fMRI voxel grid $B^s_{\text{ori}} \in \mathbb{R}^{X_s\times Y_s\times Z_s}$ of a subject s, we first extract the "nsdgeneral" brain regions of interest (ROI), which contain voxels most responsive to visual stimuli. The data is then divided into cubic patches of size $P = r^3$, with zero-padding applied to include all voxels within the ROI. To standardize the representation across subjects, we unify the extracted patches so that each subject shares the same set of patches. This process results in patched data $B \in \mathbb{R}^{L\times P}$, where L denotes the number of patches common to all subjects. By enforcing a standardized patch structure, this approach preserves the spatial organization of BOLD signals, allowing a single model to be trained across subjects without requiring subject-specific modules.

2.2 Model Architecture: *MindLink*

Fig. 1b illustrates the overall model architecture, which is trained in an end-to-end manner to extract subject-invariant fMRI embeddings and align them with image embeddings for visual stimuli reconstruction. The preprocessed fMRI data B, combined with 3D positional encoding, is passed into a 3D fMRI encoder $\mathcal{E}_b$, producing fMRI embeddings $b = \mathcal{E}_b(B) \in \mathbb{R}^{L\times d}$. A masked autoencoder framework is applied, where a decoder $\mathcal{D}$ reconstructs masked portions of B to ensure the embeddings preserve spatial structure while capturing essential brain activity patterns. Simultaneously, a subject classifier $\mathcal{S}_c$ is introduced with domain adversarial training to discard subject-specific variations, making the embeddings more generalizable across subjects. To facilitate alignment with image embeddings, the fMRI embeddings b are projected into a shared latent space using

an fMRI projector $\mathcal{P}$ as $\tilde{b} = \mathcal{P}(b) \in \mathbb{R}^{N_I \times d}$ with the same dimension as image embeddings $z = \mathcal{E}_I(I) \in \mathbb{R}^{N_I \times d}$ from an image encoder such as IP-Adapter Plus [25]. A two-level alignment strategy refines $\tilde{b}$ to establish fine-grained correspondences with z, effectively capturing visual information.

During inference, the aligned fMRI embeddings $\tilde{b}$ replace the image embeddings z as conditions for a pretrained generative model, i.e., Stable Diffusion [14]. reconstructing images that reflect the perceived visual stimuli of individuals.

Masked Voxel Modeling fMRI exhibits spatial redundancy due to the brain's structural organization, where adjacent regions process similar information [3, 17]. To leverage this, we propose Masked Voxel Modeling, a self-supervised framework that reconstructs missing voxels from partially masked fMRI data, learning robust neural representations. Given preprocessed fMRI data B, the model minimizes the MSE reconstruction loss $\mathcal{L}_{\text{rec}}$, formulated as:

$$\mathcal{L}_{\text{rec}} = \|\mathcal{D}\left(\mathcal{E}_b(B \odot M)\right) - B \odot (1 - M)\|_2^2 , \tag{1}$$

where $M \in \{0, 1\}^N$ is a binary mask with random masking, and $\odot$ is element-wise dot product. $B \odot M$ represents the visible patches of the input.

Domain Adversarial training The fMRI embedding b inherently contains subject-specific variations due to individual neural patterns, which degrade the quality of global visual reconstructions. To eliminate such variations while preserving image-relevant features, we employ domain adversarial training. Specifically, we introduce a subject classifier $\mathcal{S}_c$ and apply adversarial training using a gradient reversal layer (GRL) [4], which inverts the gradient sign between $\mathcal{E}_b$ and $\mathcal{S}_c$. The subject adversarial loss $\mathcal{L}_{\text{subj}}$ is defined as:

$$\mathcal{L}_{\text{subj}} = \ell^{ce}\left(s, \mathcal{S}_c(\mathcal{E}_b(B))\right), \tag{2}$$

where ℓ^{ce} indicates the cross-entropy. The GRL ensures that $\mathcal{S}_c$ learns subject identity while $\mathcal{E}_b$ is adversarially trained to discard subject-specific information.

Two-Level fMRI-Image Alignment. *(Instance-level Alignment)* Pretrained Stable Diffusion [14] is adopted to generate images conditioned on image embeddings z from IP-Adapter Plus [25]. To utilize this pretrained model without fine-tuning, we align projected fMRI embeddings $\tilde{b}$ with image embeddings z in both scale and direction to ensure compatibility. To achieve this, we minimize the MSE between $\tilde{b}_i$ and the corresponding z_i for the i-th fMRI-image pair as follows:

$$\mathcal{L}_{\text{align}} = \frac{1}{N} \sum_{i=1}^{N} \left\| \tilde{b}_i - z_i \right\|^2 , \tag{3}$$

where N is the number of data samples. This ensures compatibility with Stable Diffusion, enabling effective image reconstruction without fine-tuning.

(Token-level Alignment.) The image token embeddings z extracted by IP-Adapter Plus has $N_I = 16$ tokens, which exhibit interdependencies. To leverage

this characteristic, we introduce a cross-attention-based token-level alignment mechanism that enables soft matching between fMRI and image token embeddings. Instead of enforcing strict positional correspondences, each fMRI token embedding dynamically attends to all image token embeddings, capturing their contextual dependencies in a more flexible manner. For the i-th image-fMRI pair, the j-th token of fMRI token embedding $\tilde{b}_i^j$, and its corresponding cross-modal representation o_i^j is computed via attention $\alpha_i^{j,k}$ between the j-th fMRI token embedding and k-th image token embedding as:

$$o_i^j = \sum_{k=1}^{N_I} \alpha_i^{j,k}(z_i^k V), \quad \alpha_i^{j,k} = \text{softmax}\left(\frac{(\tilde{b}_i^j Q)(z_i^k K)^T}{\sqrt{d}}\right), \tag{4}$$

where $Q, K, V \in \mathbb{R}^{d\times d}$ are learnable matrices. To further refine the alignment, we apply a local alignment loss $\mathcal{L}_{\text{local}}$, which leverages contrastive learning to maximize the similarity between fMRI token embedding $\tilde{b}_i^j$ and its corresponding cross-modal embedding o_i^j, formulated as:

$$\mathcal{L}_{\text{local}} = -\frac{1}{2NN_I}\sum_{i=1}^{N}\sum_{j=1}^{N_I}\log\frac{\exp(\text{sim}(\tilde{b}_i^j, o_i^j)/\tau)}{\sum_{k=1}^{N_I}\exp(\text{sim}(\tilde{b}_i^j, o_i^k)/\tau)} + \log\frac{\exp(\text{sim}(o_i^j, \tilde{b}_i^j)/\tau)}{\sum_{k=1}^{N_I}\exp(\text{sim}(o_i^j, \tilde{b}_i^k)/\tau)}, \tag{5}$$

where τ is a temperature hyperparameter. This flexibility captures the contextual dependencies among visual tokens, preventing misalignment when semantically related tokens occur at different positions.

Overall objective Our model is trained by combining all the loss terms as:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{rec}} + \lambda_2\mathcal{L}_{\text{subj}} + \lambda_3\mathcal{L}_{\text{align}} + \lambda_4\mathcal{L}_{\text{local}}, \tag{6}$$

where λ_1, λ_2, λ_3, and λ_4 are hyperparameters.

3 Experiments

3.1 Experimental Setup

Dataset and Preprocessing. We conducted experiments on the Natural Scenes Dataset (NSD) [1], containing high resolution 7-Tesla fMRI scans paired with visual stimuli from MS-COCO dataset [10]. Among the eight subjects available, we used four subjects (subj01, subj02, subj05, and subj07) who completed all sessions as in [22]. The 982 images commonly viewed by all subjects were used as the test set, while the remaining 8859 distinct images, unique to each subject, were used for training. To preserve spatial structure, the fMRI data was partitioned into cubic patches of size $10 \times 10 \times 10$.

Implementation Details. Our architecture integrates ViT-L/14 [13] and IP-Adapter Plus [25] with $N_I = 16$ tokens for image feature extraction. fMRI data is processed through a 16-layer Transformer Encoder [21] and a 4-layer Perceiver [6]. The model is trained for 150 epochs with a batch size of 256, starting with

Fig. 2. Qualitative comparison of image reconstructions. *MindLink* reconstructs semantically relevant images with robust cross-subject generalization.

a 50-epoch warmup using only $\mathcal{L}_{\text{rec}}$. *MindLink* is optimized using AdamW with a one-cycle scheduler, with a maximum learning rate of 3e-4. The temperature $\tau = 0.1$ in $\mathcal{L}_{\text{local}}$. The weights of the losses λ_1, λ_2, λ_3 and λ_4 are set to 1, 1, 1, 1e-2, respectively. For visual reconstructions, the SD v1.5 [14] model is used with a DDIM sampler configured for 50 steps and a guidance scale of 7.5.

Evaluation. To quantitatively evaluate image quality, we use eight metrics consistent with the protocol in [22]. Low-level features are measured by PixCorr, SSIM [23], AlexNet(2), and AlexNet(5) [8], while high-level features are evaluated with Inception [18], CLIP [13], EffNet-B [20], and SwAV [2].

3.2 Cross-Subject Brain Decoding

We evaluate *MindLink*'s by comparing its average reconstruction performance across all subjects with state-of-the-art methods, including Takagi et al. [19], Brain-Diffuser [12], MindEye [15], MindBridge [22], and UMBRAE [24].

Qualitative Results. As shown in Fig. 2, *MindLink* demonstrates superior visual quality and semantic accuracy compared to baselines. Specifically, in the second row where the stimuli show food in a bowl with a spoon, *MindLink* reconstructs the spoon for all subjects, whereas baselines fail. This illustrates that *MindLink* properly captures fine-grained details and contextual elements for accurate reconstructions. Furthermore, our model consistently maintains semantic content across all subjects, as seen in the rightmost columns. Despite individual variations in brain activity, the reconstructed images exhibit shared perception, highlighting *MindLink*'s ability to generalize across subjects and achieve accurate image reconstruction without subject-specific modules.

Quantitative Results. As shown in Table 1, *MindLink* achieves notable improvements in high-level metrics, with gains of 0.6% in Inception, 0.036 in EffNet-B, and 0.031 in SwAV over the second-best results in the single model fashion. It also performs comparably in SwAV against subject-specific methods like

Table 1. Quantitative comparison of brain decoding between *MindLink* and other methods. All metrics are averaged over 4 subjects. † indicates models trained in a per-subject-per-model fashion for comparison.

Method	Subject agnostic	Low-Level				High-Level				# Params
		PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	
Per-subject-per-model fashion										
Takagi et al. [19]	✗	-	-	83.0%	83.0%	76.0%	77.0%	-	-	487M
Brain-Diffuser [12]	✗	.254	.356	94.2%	96.2%	87.2%	91.5%	.775	.423	1.45B
MindEye [15]	✗	.309	.323	94.7%	97.8%	93.8%	94.1%	.645	.367	1.21B
MindBridge† [22]	✗	.148	.259	86.9%	95.3%	92.2%	94.3%	.713	.418	561M
MindLink†	✗	.219	.323	92.1%	97.0%	92.7%	92.3%	.684	.374	**159M**
Single model fashion										
MindBridge [22]	✗	.151	.263	87.7%	95.5%	92.4%	94.7%	.712	.418	694M
UMBRAE [24]	✗	.283	.341	95.5%	97.0%	91.7%	93.5%	.700	.393	146M
MindLink (Ours)	✓	.227	.333	92.5%	97.0%	93.0%	92.8%	.664	.362	**159M**

Table 2. Ablation study on $\mathcal{L}_{\text{subj}}$ and cross-attention in token-level alignment.

Method	Low-Level				High-Level			
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
w/o Cross-att.	.215	.330	90.6%	95.6%	90.5%	91.6%	.694	.383
w/o $\mathcal{L}_{\text{subj}}$	.222	.329	91.7%	96.8%	92.8%	92.5%	.672	.367
Ours	**.227**	**.333**	**92.5%**	**97.0%**	**93.0%**	**92.8%**	**.664**	**.362**

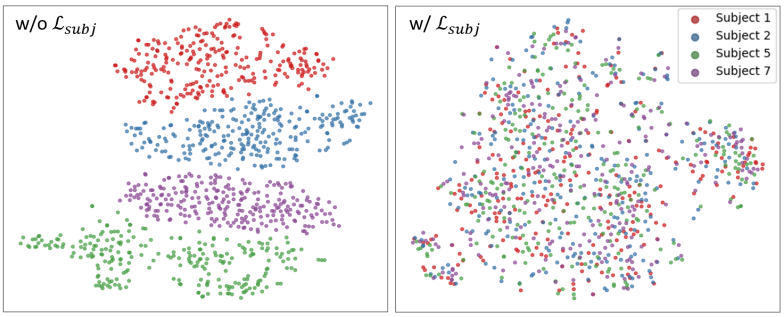

Fig. 3. t-SNE visualization of fMRI embeddings with and without subject loss ($\mathcal{L}_{\text{subj}}$). Embeddings are obtained from the fMRI Encoder $\mathcal{E}_b$ in *MindLink* on the NSD test set.

MindEye, indicating robust semantic alignment and consistent high-level concept representations across subjects. Unlike MindBridge [22], which increases by 133M parameters when transitioning from single-subject to cross-subject settings, *MindLink* maintains a constant size of 159M, ensuring superior efficiency and scalability. While [16] reports better results, it fine-tunes a large language model with >8B parameters, so it was excluded in our evaluation. Nevertheless, *MindLink* achieves competitive results with only 159M parameters, highlighting its balance of efficiency and accuracy.

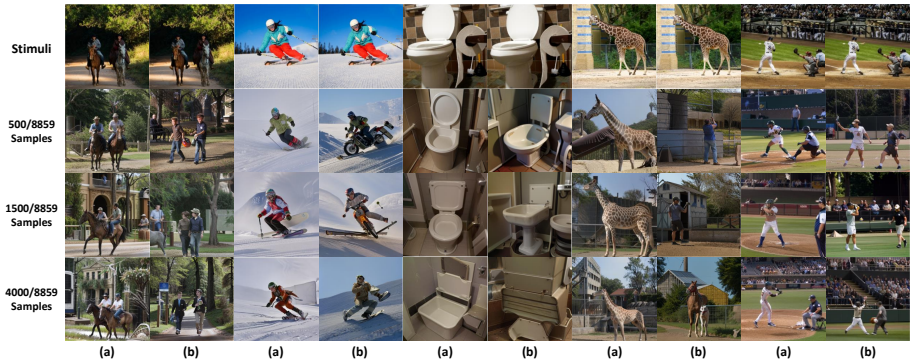

Fig. 4. Qualitative comparison with limited data for a new subject (subj 7): (a) fine-tuning a model pretrained on other subjects (subj 1, 2, 5) and (b) training from scratch.

Ablation study. Ablation study results on domain adversarial training and cross-attention in token-level alignment are reported in Table 2. Utilizing a domain adversarial training with a subject-adversarial loss $\mathcal{L}_{\text{subj}}$ consistently improves performance across all eight metrics, demonstrating that discarding subject-specific information enhances cross-subject generalizability. As illustrated in the t-SNE visualization (Fig. 3), applying $\mathcal{L}_{\text{subj}}$ alleviates subject-wise clustering, resulting in a more subject-invariant embedding space. Moreover, our cross-attention-based token-level alignment outperforms rigid position-based matching by dynamically computing token correspondences, highlighting the importance of leveraging token interdependencies for robust decoding.

3.3 New Subject Adaptation

MindLink demonstrates strong adaptability to new subjects. To evaluate this capability, we employed a cross-subject transfer learning approach under limited data conditions. Specifically, we pretrained our model on fMRI data from three source subjects and adapted it to a new target subject using subsets of 500, 1500, and 4000 samples. We compared two approaches: (a) fine-tuning our pretrained model on each subset of the new target subject's data and (b) training the model from scratch on the same subsets, following a per-subject-per-model paradigm.

Table 3. Quantitative comparison with limited data for a new subject (subj 7).

Training Strategy	# Data	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Scratch	500	.148	.307	84.1%	92.1%	84.9%	85.2%	.791	.448
Fine-tuning	500	**.175**	**.312**	**87.5%**	**94.9%**	**89.7%**	**90.1%**	**.731**	**.406**
Scratch	1500	.159	.309	86.1%	93.6%	86.9%	87.5%	.767	.430
Fine-tuning	1500	**.188**	**.317**	**88.8%**	**94.8%**	**90.1%**	**90.4%**	**.720**	**.399**
Scratch	4000	.171	.317	88.8%	95.1%	89.8%	89.7%	.730	.408
Fine-tuning	4000	**.199**	**.319**	**90.1%**	**95.6%**	**91.0%**	**91.2%**	**.704**	**.389**

As shown in Figure 4, our fine-tuning approach consistently outperforms scratch-trained models across all cases. While scratch-trained models struggle with limited data, *MindLink* leverages cross-subject pretrained knowledge for effective generalization. Table 3 confirms superior reconstruction accuracy across all metrics, demonstrating robust performance even with just 500 samples and highlighting *MindLink*'s efficiency in knowledge transfer and scalability.

4 Conclusion

In this work, we propose *MindLink*, which links multiple subjects into a single model by extracting subject-invariant features while preserving the spatial structure of 3D fMRI. *MindLink* achieves this using standardized cubic patches of fMRI and applying domain adversarial training to enhance cross-subject generalizability. The two-level alignment strategy bridges fMRI and image embeddings through instance-level consistency and flexible token-level matching. Extensive experiments on the NSD dataset confirm that our model achieves comparable performance across multiple subjects and adapts effectively to new subjects.

Acknowledgments. This research was supported by RS-2022-II220290 (90%) and RS-2019-II191906 (AI Graduate Program at POSTECH, 10%).

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., et al.: A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. Nature Neuroscience (2022)
2. Caron, M., Misra, I., Mairal, J., et al.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems (2020)
3. Chen, Z., Qing, J., Xiang, T., et al.: Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In: Conference on Computer Vision and Pattern Recognition (2023)
4. Ganin, Y., Ustinova, E., Ajakan, H., et al.: Domain-adversarial training of neural networks. Journal of Machine Learning Research (2016)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial networks. Communications of the ACM (2020)
6. Jaegle, A., Gimeno, F., Brock, A., et al.: Perceiver: General perception with iterative attention. In: International Conference on Machine Learning. PMLR (2021)
7. Jeong, M., Cho, H., Jung, S., Kim, W.H.: Uncertainty-aware diffusion-based adversarial attack for realistic colonoscopy image synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2024)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems (2012)

9. Lahoud, J., Cao, J., Khan, F.S., et al.: 3D vision with transformers: A survey. arXiv preprint arXiv:2208.04309 (2022)
10. Lin, T.Y., Maire, M., Belongie, S., et al.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision (2014)
11. Logothetis, N.K.: What we can do and what we cannot do with fmri. Nature (2008)
12. Ozcelik, F., VanRullen, R.: Natural scene reconstruction from fmri signals using generative latent diffusion. Scientific Reports (2023)
13. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR (2021)
14. Rombach, R., Blattmann, A., Lorenz, D., et al.: High-resolution image synthesis with latent diffusion models. In: Conference on Computer Vision and Pattern Recognition (2022)
15. Scotti, P., Banerjee, A., Goode, J., et al.: Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. Advances in Neural Information Processing Systems (2024)
16. Shen, G., Zhao, D., He, X., et al.: Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. Advances in Neural Information Processing Systems (2025)
17. Shmuel, A., Yacoub, E., Chaimow, D., et al.: Spatio-temporal point-spread function of fmri signal in human gray matter at 7 tesla. Neuroimage (2007)
18. Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: Rethinking the inception architecture for computer vision. In: Conference on Computer Vision and Pattern Recognition (2016)
19. Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. In: Conference on Computer Vision and Pattern Recognition (2023)
20. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR (2019)
21. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
22. Wang, S., Liu, S., Tan, Z., et al.: Mindbridge: A cross-subject brain decoding framework. In: Conference on Computer Vision and Pattern Recognition (2024)
23. Wang, Z., Bovik, A.C., Sheikh, H.R., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (2004)
24. Xia, W., de Charette, R., Oztireli, C., et al.: Umbrae: Unified multimodal brain decoding. In: European Conference on Computer Vision (2024)
25. Ye, H., Zhang, J., Liu, S., et al.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)