

MindBridge: A Cross-Subject Brain Decoding Framework

Shizun Wang Songhua Liu Zhenxiong Tan Xinchao Wang[†]

National University of Singapore

{shizun.wang, songhua.liu, zhenxiong}@u.nus.edu, xinchao@nus.edu.sg



Figure 1. **Image stimuli and images reconstructed from captured brain signals.** Given the limited training data from a new subject (subj07 from the NSD dataset [1]), our proposed **MindBridge** can faithfully reconstruct natural images using less data, benefiting from pretrained **cross-subject** knowledge. In contrast, the **Vanilla** method, which represents current methods following a **per-subject-per-model** paradigm, fails to learn effectively from limited data.

Abstract

*Brain decoding, a pivotal field in neuroscience, aims to reconstruct stimuli from acquired brain signals, primarily utilizing functional magnetic resonance imaging (fMRI). Currently, brain decoding is confined to a **per-subject-per-model** paradigm, limiting its applicability to the same individual for whom the decoding model is trained. This constraint stems from three key challenges: 1) the inherent variability in input dimensions across subjects due to differences in brain size; 2) the unique intrinsic neural patterns, influencing how different individuals perceive and process sensory information; 3) limited data availability for new subjects in real-world scenarios hampers the performance of decoding models.*

*In this paper, we present a novel approach, **MindBridge**, that achieves **cross-subject brain decoding** by employing only one model. Our proposed framework establishes a generic paradigm capable of addressing these challenges by introducing biological-inspired aggregation function and novel cyclic fMRI reconstruction mechanism for subject-invariant representation learning. Notably, by cycle re-*

construction of fMRI, MindBridge can enable novel fMRI synthesis, which also can serve as pseudo data augmentation. Within the framework, we also devise a novel reset-tuning method for adapting a pretrained model to a new subject. Experimental results demonstrate MindBridge's ability to reconstruct images for multiple subjects, which is competitive with dedicated subject-specific models. Furthermore, with limited data for a new subject, we achieve a high level of decoding accuracy, surpassing that of subject-specific models. This advancement in cross-subject brain decoding suggests promising directions for wider applications in neuroscience and indicates potential for more efficient utilization of limited fMRI data in real-world scenarios. Project page: <https://littlepure2333.github.io/MindBridge>

1. Introduction

The human brain, an intricate web of neurons, possesses the remarkable ability to encode the sensory stimuli that we encounter every day, making sense of our perceptual world. While the reverse process, known as brain decoding, aims to reconstructs image stimulus from brain signals, which are primarily captured by functional magnetic reso-

[†]Corresponding author.

nance imaging (fMRI). Brain decoding has been a subject of intense interest, as it offers the tantalizing prospect of unraveling the secrets of cognition and perception, and present a potential advancement for brain-computer interface (BCI) [7] and beyond. From GANs [25, 37] to diffusion models [25, 35, 40], the use of increasingly powerful generative models has enabled brain decoding to reconstruct more realistic and faithful images. However, nowadays brain decoding is confronted with significant challenges that hinder its application on a broader scale.

Specifically, the current practice of brain decoding is confined to subject-specific applications [12, 22, 24, 35, 40, 47]. In other words, a decoding model trained on a one subject’s brain can only be effectively applied to that same subject, and can not be applied to other subjects, which results in high expense of model storage and training. This limitation motivates us to move towards **cross-subject brain decoding**, which is capable of using one model to decode brain signals from multiple subjects and adapting to new subjects. Such a paradigm thereby can expand the utility of brain decoding in a more general way and bring more applicability in real scenarios.

However, it requires adequately addressing various substantial challenges in pursuit of this beautiful vision: **1) Size Variability:** fMRI signals exhibit substantial size differences across subjects , largely due to the inherent variability in brain size and structure, which necessitates a flexible approach to handle this variability. **2) Diverse Neural Responses:** The intricacies of the brain extend beyond structural variations. The way each subject’s brain processes stimuli is uniquely shaped by their experiences, biases, and cognitive patterns, posing a challenge in unifying the interpretation of brain signals. **3) Data Scarcity for New Subjects:** In real-world applications, it is highly cumbersome to acquire extensive fMRI data for new subjects if not infeasible at all. The high costs in both resources and time significantly hinder the training and adaptation of brain-decoding models for new subjects.

To address these challenges, we devise “**MindBridge**”, a novel framework designed to achieve cross-subject brain decoding. MindBridge employs innovative strategies to tackle each of the identified obstacles: **1) Adaptive Signal Aggregation:** Inspired by neural-science findings that the brain activation is sparse and only neurons exceeding a certain threshold activate, we propose to use an aggregation function based on adaptive max pooling to extract most useful information , and unify the input dimension of fMRI signals across different subjects. **2) Subject-Invariant Representation:** We extract subject-invariant semantic embeddings from disparate subjects’ fMRI signals by utilizing a novel cycle reconstruction mechanism. These embeddings are then translated and aligned within a consistent CLIP embedding space, facilitating a standardized interpre-

tation across varying neural responses. **3) Efficient Adaptation Strategy:** To mitigate the data scarcity issue for new subjects, we introduce a novel finetuning method, reset-tuning. Because transferable knowledge from cross-subject pretraining is held in the deep layers, while the shallow layers are responsible for projecting diverse subjects’ fMRI signals into subject-invariant embeddings. Reset-tuning re-set the shallow layers but reuse the deep layers.

Furthermore, MindBridge incorporates additional enhancements to improve semantic accuracy and expand its application. Utilizing a multi-modal versatile diffusion (VD) model [48], we can incorporate not only image stimuli but also the text caption as training data, then predict corresponding image and text embeddings to reconstruct more semantically faithful images. Moreover, MindBridge opens new possibility to synthesize new brain signals using data from other subjects while preserving same semantic meaning by cyclic fMRI reconstruction. Therefore, MindBridge not only bridges the gap among different brains but also potentially augments the volume of available data.

To verify our approach, we conducted experiments on the publicly available NSD dataset [1]. Notably, the absence of common images across different subjects in the training set poses an additional challenge to cross-subject brain decoding. Surprisingly, MindBridge, employing only one model, achieves performance comparable to subject-specific methods, which require multiple models. Additionally, experiments on new subject adaptation validate that our method surpasses methods trained from scratch, showcasing the benefits of transferable knowledge from cross-subject pretraining and our proposed reset-tuning.

In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to effectively addresses the challenge of cross-subject brain decoding. We design a novel framework, MindBridge, equipped with an adaptive signal aggregation function and novel cyclic fMRI reconstruction mechanism for subject-invariant representations learning.
- We introduce a novel “reset-tuning” strategy, which efficiently adapts the MindBridge model to new subjects, and effectively overcoming the limitations posed by data scarcity for new subjects.
- MindBridge enables new capability for the synthesis of brain signals, leveraging data across various subjects while maintaining consistent semantic interpretation.
- Extensive experiments demonstrate MindBridge’s efficacy and adaptability, showcasing its potential to significantly advance the field of brain decoding.

2. Related Work

2.1. Brain Decoding

The evolution of brain decoding has been marked by the integration of advanced modeling approaches. Earlier work

[15] applies sparse linear regression on fMRI data to predict features from early convolutional layers of pretrained CNNs. With the introduction of generative adversarial networks (GANs) [11], there has been a shift towards visual decoding techniques that map brain signals to the latent spaces of GANs, facilitating the reconstruction of handwritten digits [33], human faces [42], and natural scenes [12, 25, 36].

The advent of high-resolution image synthesis with Latent Diffusion Models [30] and multi-modal contrastive models like CLIP [27], along with extensive fMRI datasets [1], has propelled researchers to map fMRI signals into the CLIP embedding space. This mapping guides latent diffusion models for image reconstruction [40], with efforts focusing on improved mapping through self-supervision [3], masked modeling [5], and contrastive learning [35]. Additional explorations involve advanced diffusion models [22] and conditional control [20, 47]. Unlike almost all brain decoding research that requires training multiple models for different subjects, MindBridge stands out by aiming to achieve cross-subject brain decoding with a single model.

2.2. Diffusion Models

Diffusion models (DMs) [6, 8, 13, 21, 38, 46, 50] have recently emerged as a focal point in deep generative model research, known for their ability to generate high-quality images. DMs utilize iterative denoising to recover a sampled variable from Gaussian noise and transform it into a sample conforming to the learned data distribution. With large-scale image-text pair datasets [34], DMs have demonstrated superior performance in the task of text-to-image generation [9, 28, 32, 44, 49] and achieves unprecedented image quality. Building on this, latent diffusion models (LDMs), also known as Stable Diffusion (SD), have furthered DMs by reducing computational demands through denoising in a latent space produced by autoencoders. An advanced form of LDMs, the Versatile Diffusion (VD) model [48], demonstrates the capability to produce high-quality images, guided by both image and text inputs. Consequently, we have adopted the VD model for its dual-input capacity, leveraging its enhanced image generation potential.

3. MindBridge

3.1. Data Elaboration

To better understand the task at hand, we illustrate the data we used in ahead. We have chosen the widely-used Natural Scenes Dataset (NSD) [1] for our brain decoding research. This dataset consists of high-resolution 7-Tesla fMRI scans collected from 8 healthy adult subjects, who were instructed to view thousands of natural images from MS-COCO dataset [19]. Following common practices [12, 22, 24, 35, 40, 47], our research mainly use data from

4 subjects (subj01, 02, 05, 07), who completed all the scan sessions. Notably, only a subset of data, 982 images, were **commonly viewed** by all four subjects. Those data were used as the **test set**. While the remaining data, each of 8,859 **distinct** images viewed by each subject were used as the **training set**. Following prior work [35], we use preprocessed fMRI voxels from “NSDGeneral” regions of interest (ROI). Due to the inherent variability in brain size and structure, the fMRI signals within the ROI exhibit different sizes (about 13,000 to 16,000 voxels per subject), which is the first challenge we need to tackle in cross-subject brain decoding. The original acquired fMRI data is 4D (3D+t), which is firstly averaged among time dimension, and then is flattened from 3D to 1D. ROIs serve as masks on the 1D vector. So the dimensionality of input fMRI voxels is 1D.

3.2. Cross-Subject Brain Decoding

Current brain decoding pipeline can be summarized in two steps: mapping fMRI voxels to CLIP embeddings and then using these embeddings to guide diffusion models in generating reconstructed images. While previous brain decoding methods all fall in a per-subject-per-model fashion. Here we argue the key insight for achieving cross-subject brain decoding lies in establishing a shared common representation space that is subject-invariant. However, there are two main barriers to realizing this objective. The first is the variation in the size of fMRI signals among different subjects, as explained in Sec. 3.1. The second challenge is how to effectively model subject-invariant representation learning.

To address these challenges, we propose MindBridge, a novel framework for cross-subject brain decoding. Formally, we denote the 1D fMRI voxels from subject s as $V_s \in \mathbb{R}^F$, where F represents fMRI voxels’ size. The corresponding image stimulus I and image caption T can be extracted by a pretrained CLIP model as image embedding e_I and text embedding e_T .

Pipeline. MindBridge first adaptively unifies the fMRI voxels V_s to a unified size $v_s = f(V_s)$ using a biologically-inspired aggregation function f . Unlike previous methods that directly learn the projection between fMRI voxels and corresponding CLIP embeddings, MindBridge projects different subjects’ aggregated fMRI voxels v_s to an intermediate semantic embedding $e_s = \mathcal{E}_s(v_s)$ using a subject-wise brain embedder \mathcal{E}_s . To ensure that semantic embeddings from different subjects reside in a common shared space, we propose a novel cyclic fMRI reconstruction mechanism. This mechanism relies on an additional subject-wise brain builder \mathcal{B}_s to reconstruct the unified fMRI voxels $\hat{v}_s = \mathcal{B}_s(e_s)$. Once the semantic embeddings are obtained, a brain translator \mathcal{T} translates them into two embeddings, $(\hat{e}_I, \hat{e}_T) = \mathcal{T}(e_s)$, representing the predicted CLIP image and text embeddings. The brain embedder, brain builder and brain translator are all MLP-like networks.

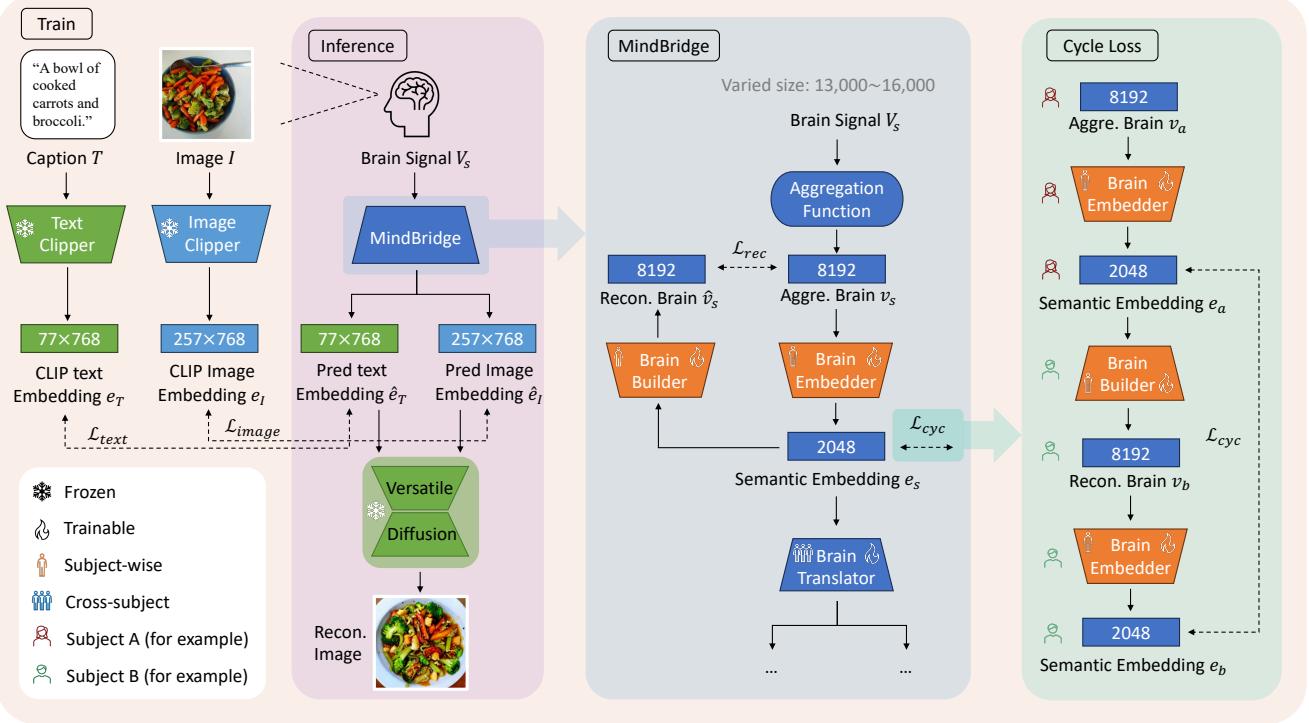


Figure 2. **Overview of MindBridge.** MindBridge is a cross-subject brain decoding framework capable of handling fMRI signals from different subjects. Initially, an aggregation function unifies the size of fMRI signals. Subsequently, subject-wise brain embedders and brain builders are trained to obtain subject-invariant semantic embeddings. The Brain Translator then generates text and image embeddings, which are utilized to reconstruct images through versatile diffusion model. The dimension of data is denoted within the box.

Diffusion Model. Due to the limited volume of brain data, a generative model trained on a large-scale dataset is necessary to aid the image reconstruction process. Previous methods [5, 35, 40] have demonstrated the superiority of using diffusion models as an interface for image generation. In this work, we have chosen to employ the versatile diffusion (VD) [48] model, a multimodal latent diffusion model that is guided by image and text CLIP embeddings and achieves state-of-the-art performance in image generation. Its exceptional capabilities give us an opportunity to utilize both visual and semantic information, as represented by CLIP image and text embeddings predicted by MindBridge, to reconstruct images at inference time.

Adaptive fMRI Aggregation. Modern neural-science research [23, 29, 43] reveals that visual stimuli are encoded sparsely in the primary visual cortex, activating only a few neurons for most natural images. Also, neurons require a certain level of threshold, to become active and fire an action potential [14, 17]. The activation functions of artificial neural networks such as Sigmoid or ReLU also echo this fundamental principle in neurophysiology [10, 31]. Inspired by these findings, we posit that brain signals can be aggregated sparsely, with higher values tending to be more valuable. Consequently, we propose employing ‘‘Adaptive

Max Pooling’’¹ as the aggregation function. This function unifies the size of input fMRI signals by dynamically adjusting its pooling size to produce a fixed output size.

Learning Objectives. MindBridge learns CLIP image and text embeddings through two types of losses. One is the SoftCLIP loss, introduced in [35], which has proven effective in aligning the fMRI modality with the embedding space of the pretrained CLIP model. This loss facilitates contrastive learning by maximizing the similarity of positive pairs while minimizing the similarity of negative pairs. Positive pairs are defined as the soft labels produced by the dot product of embeddings within a batch, and the loss considers both CLIP-CLIP and Brain-CLIP scenarios.

$$\mathcal{L}_{SoftCLIP}(p, t) = - \sum_{i=1}^N \sum_{j=1}^N \left[\frac{\exp(t_i \cdot t_j / \tau)}{\sum_{m=1}^N \exp(\frac{t_i \cdot t_m}{\tau})} \cdot \log \left(\frac{\exp(p_i \cdot t_j / \tau)}{\sum_{m=1}^N \exp(\frac{p_i \cdot t_m}{\tau})} \right) \right] \quad (1)$$

Where p, t are the predicted CLIP embedding and target CLIP embedding in a batch of size N , respectively. τ is a temperature hyperparameter.

¹PyTorch implementation: <https://pytorch.org/docs/stable/generated/torch.nn.AdaptiveMaxPool1d.html>

- [28] 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. 3
- [29] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999. 4
- [30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3
- [31] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986. 4
- [32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3
- [33] Sanne Schoenmakers, Markus Barth, Tom Heskes, and Marcel Van Gerven. Linear reconstruction of perceived images from human brain activity. *NeuroImage*, 83:951–961, 2013. 3
- [34] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3
- [35] Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, et al. Reconstructing the mind’s eye: fmri-to-image with contrastive learning and diffusion priors. *arXiv preprint arXiv:2305.18274*, 2023. 2, 3, 4, 5, 6, 1
- [36] Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel AJ van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, 2018. 3
- [37] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019. 2
- [38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3
- [39] 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. 5
- [40] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. *biorxiv*. 2022. 2, 3, 4, 5, 6, 1
- [41] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5
- [42] Rufin VanRullen and Leila Reddy. Reconstructing faces from fmri patterns using deep generative neural networks. *Communications biology*, 2(1):193, 2019. 3
- [43] William E Vinje and Jack L Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000. 4
- [44] Yunhe Wang, Hanting Chen, Yehui Tang, Tianyu Guo, Kai Han, et al. Pangu-π: Enhancing language model architectures via nonlinearity compensation. In *arXiv:2312.17276*, 2023. 3
- [45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5
- [46] Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, and Mingyuan Zhou. Patch diffusion: Faster and more data-efficient training of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [47] Weihao Xia, Raoul de Charette, Cengiz Özireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. *arXiv preprint arXiv:2310.02265*, 2023. 2, 3, 1
- [48] 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. 2, 3, 4, 1
- [49] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *IEEE/CVF International Conference on Computer Vision*, 2023. 3
- [50] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3
- [51] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 5
- [52] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *arXiv preprint arXiv:2302.04867*, 2023. 1

MindBridge: A Cross-Subject Brain Decoding Framework

Supplementary Material

A. More Dataset Information

We used the Natural Scenes Dataset (NSD)[1] for all experiments. This dataset consists of high-resolution 7-Tesla fMRI scans collected from 8 healthy adult subjects. Every subject was instructed to view thousands of natural images from MS-COCO dataset[19] over the course of 30–40 scan sessions. Each participant was exposed to 9,000-10,000 distinct images, each displayed for three seconds and repeated three times, resulting in a total of 22,000-30,000 fMRI response trials per participant. The fMRI responses were derived from GLMSingle, with outputs being session-wise z-scored single-trial betas [26]. Following common practices [12, 22, 24, 35, 40, 47], our research mainly use data from 4 subjects (subj01, subj02, subj05, subj07), who completed all the scan sessions, as our experimental data. Following prior work [35], we use preprocessed fMRI voxels from “NSDGeneral” regions of interest (ROI). This ROI is defined by NSD authors as a subset of voxels in the posterior cortex that respond most strongly to the visual stimuli presented.

Each subject was instructed to view up to three repeated trials per image. For the training set, we randomly choose one single-trial fMRI signal per image for training. For the test set, we average fMRI signals across the three same-image repetitions for the test set, similar to [35, 40].

B. Implementation Details

Our method’s image reconstruction part relies on versatile diffusion model[48]. In the diffusion process, we employ the UniPCMultistep scheduler[52] to execute 20 diffusion steps with a guidance scale of 5, setting the text-image ratio at 0.5 to integrate both visual and semantic information. For each test sample, we reconstruct 8 images and select the one with the highest CLIP similarity to the stimuli as the final result.

We implemented the aggregation function as Adaptive-MaxPool1D in our PyTorch code, setting the output size to 8192, which is the largest power-of-two integer smaller than all the fMRI signal sizes. In practice, we incorporated AdaptiveMaxPool1D into the data preprocessing pipeline of the dataloader to ensure uniform data dimensions within each batch. We use the last hidden layer of CLIP ViT-L/14 for the CLIP image embedding and CLIP text embedding, results a shape of 257×768 and 77×768 respectively.

The total loss is balanced by setting the weights of \mathcal{L}_{image} , \mathcal{L}_{text} , \mathcal{L}_{rec} , and \mathcal{L}_{cyc} to 1, $1e4$, 1, and 1, respectively. The MindBridge models for cross-subject experiments are trained for 600 epochs, and those for new-subject

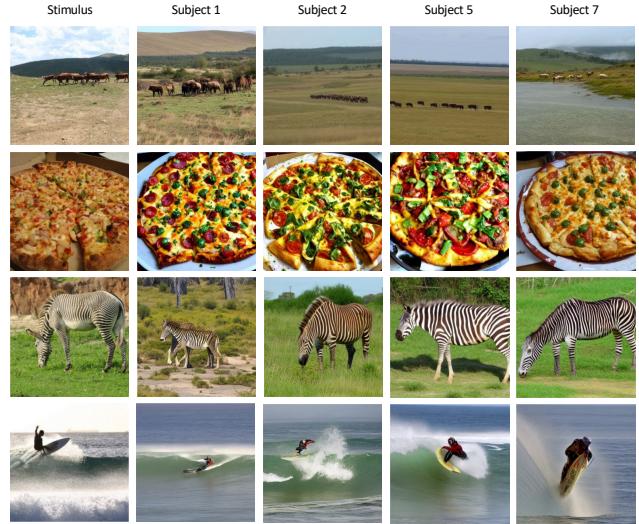


Figure 5. More cross-subject reconstructions of MindBridge on subject 1, 2, 5 and 7.

adaptation are trained for 200 epochs. Across all experiments, the batch size is set to 50 per GPU. Since \mathcal{L}_{rec} and \mathcal{L}_{cyc} are designed for subject-invariant representation learning, we do not employ them in single-subject model training. Therefore, we only use \mathcal{L}_{image} and \mathcal{L}_{text} for single-subject model training.

Data augmentation are applied to target images in all experiments to enhance model robustness. These include random cropping, resizing, and random adjustments of brightness, contrast, gamma, saturation, hue, sharpness, and gray scale. In the pseudo data augmentation for new subject adaptation, we use the same number of data points from a previously trained subject as that of the new subject. And all previously trained subjects are involved in the training process.

C. More Reconstruction Results

Here we show more visual results of cross-subject reconstruction from brain signals using one MindBridge model trained on subject 1, 2, 5 and 7 in Figure 5.

D. Architecture of MindBridge

MindBridge comprises an aggregation function, a brain embedder, a brain builder, and a brain translator. The aggregation function is adaptive max pool. The brain embedder consists of a LoRA[16]-like adapter and one linear layer. The brain builder is basically the reverse of brain embed-

der. The brain translator is a 4-layer residual MLP equipped with two linear heads, and the hidden layer size is 2048.

The PyTorch-like architectural code for the brain embedder, brain builder, and brain translator is illustrated in Fig 6.

E. More Results for New Subject Adaptation

Here, we present additional results of new subject adaptation for subjects 1, 2, and 5 in Table 6, Table 7, and Table 8, respectively. The results show that our method consistently yields superior performance compared to the vanilla method.

F. Effect of Text-Image Ratio

Because versatile diffusion[48] model can accept both image and text inputs, we demonstrate how varying the image-text ratio can influence the trade-off between image consistency and diversity in Figure 7. A greater proportion of text input ensures more semantic correctness while introducing more diversity. To balance semantic accuracy and visual consistency, we have chosen to set the image-text ratio to 0.5 in this work.

G. Ethic and Social Impact

As brain decoding technology advances, it brings critical ethical considerations to the forefront. While it promises enhanced communication for those with speech or motor impairments, its potential for involuntary mind reading necessitates stringent ethical frameworks. Key elements include informed consent, robust data privacy, and a thorough consideration of societal implications. It is crucial for the scientific community and society to address these ethical challenges to prevent misuse and ensure the responsible development of brain decoding technology.

```

1 class MindBridge(nn.Module):
2     def __init__(self, in_dim=8196, out_dim_image=257*768, out_dim_text=77*768, h=2048,
3      ↪ subj_list=[1,2,5,7])
4         self.embedder = nn.ModuleDict(subj: nn.Sequential(
5             Adapter(in_dim, 128),
6             nn.Linear(in_dim, h),
7             nn.LayerNorm(h),
8             nn.GELU(),
9             nn.Dropout(0.5)
10            ) for subj in subj_list))
11
12         self.builder = nn.ModuleDict({subj: nn.Sequential(
13             nn.Linear(h, in_dim),
14             nn.LayerNorm(in_dim),
15             nn.GELU(),
16             Adapter(in_dim, 128)
17             ) for subj in subj_list})
18
19         self.mlp = nn.ModuleList([
20             nn.Sequential(
21                 nn.Linear(h, h),
22                 nn.LayerNorm(h),
23                 nn.GELU(),
24                 nn.Dropout(0.15)
25             ) for _ in range(4)])
26
27         self.head_image = nn.Linear(h, out_dim_image)
28         self.head_text = nn.Linear(h, out_dim_text)

```

Figure 6. PyTorch-like pseudo code of MindBridge architecture.

Method	# Data	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
MindBridge (Scratch)	500	.097	.199	76.9%	87.3%	75.7%	82.8%	.885	.560
MindBridge (Ours)	500	.142	.263	84.4%	91.4%	82.3%	89.4%	.831	.500
MindBridge (Scratch)	1500	.122	.235	81.6%	91.5%	82.8%	87.7%	.838	.506
MindBridge (Ours)	1500	.161	.275	87.2%	94.2%	88.5%	92.5%	.784	.460
MindBridge (Scratch)	4000	.138	.266	85.8%	94.3%	87.8%	91.5%	.800	.465
MindBridge (Ours)	4000	.157	.275	88.1%	95.5%	90.0%	93.9%	.747	.436

Table 6. **Results of new subject adaptation in limited data scenario.** MindBridge(Ours) is finetuned on subject 1 from model pretrained on subject 2, 5 and 7.

