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 reconstruction 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 [24, 35] to diffusion models [24, 33, 38], 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, 21, 23, 33, 38, 45]. 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 interpretation 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 reset 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 [46], 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 [31], human faces [40], and natural scenes [12, 24, 34].

The advent of high-resolution image synthesis with Latent Diffusion Models [28] and multi-modal contrastive models like CLIP [25], 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 [38], with efforts focusing on improved mapping through self-supervision [3], masked modeling [5], and contrastive learning [33]. Additional explorations involve advanced diffusion models [21] and conditional control [19, 45]. 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, 20, 36, 44, 48] 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 [32], DMs have demonstrated superior performance in the task of text-to-image generation [9, 26, 30, 42, 47] 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 [46], 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 [18]. Following common practices [12, 21, 23, 33, 38, 45], 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 [33], we use preprocessed fMRI voxels from "NSDGeneral" regions of interest (ROI). Due to the inherent variablity 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 serves 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 relys 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.

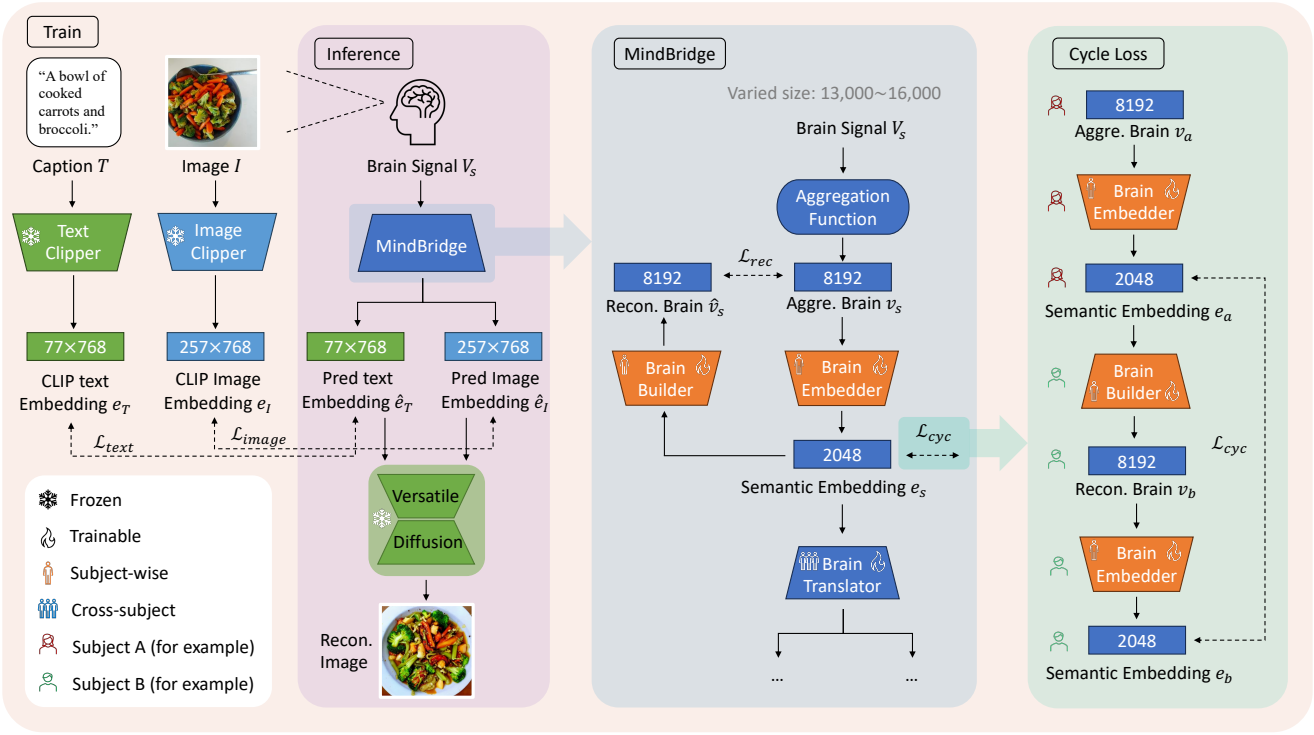

Figure 2. **Overview of MindBridge.** MindBridge is a cross-subject brain decoding framework capable of handling fMRI signals from different subjects. Initially, an aggregation function unifies the size of fMRI signals. Subsequently, subject-wise brain embedders and brain builders are trained to obtain 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, 33, 38] 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) [46] 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 [22, 27, 41] 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, 16]. The activation functions of artificial neural networks such as Sigmoid or ReLU also echo this fundamental principle in neurophysiology [10, 29]. 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 [33], 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.

$$\mathcal{L}_{SoftCLIP}(p,t) = -\sum_{i=1}^{N}\sum_{j=1}^{N}$$

$$\left[\frac{\exp\left(t_i \cdot t_j/\tau\right)}{\sum_{m=1}^{N}\exp\left(\frac{t_i \cdot t_m}{\tau}\right)} \cdot \log\left(\frac{\exp\left(p_i \cdot t_j/\tau\right)}{\sum_{m=1}^{N}\exp\left(\frac{p_i \cdot t_m}{\tau}\right)}\right)\right] \quad (1)$$

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

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

During our exploratory experiments, however, we observed that reconstructed images still exhibited some artifacts when using only the SoftCLIP loss. We hypothesize that this may be due to the SoftCLIP loss's inability to guarantee the authenticity of the learned CLIP embeddings. Therefore, we introduced the second loss, the MSE loss, to ensure a more accurate prediction of CLIP embeddings.

$$\mathcal{L}_{MSE}(p,t) = \frac{1}{N}\sum_{i=1}^{N}(p_i - t_i)^2 \tag{2}$$

Incorporating these two losses ensures a more natural image reconstruction. The complete set of losses for predicting image and text CLIP embeddings includes:

$$\mathcal{L}_{image} = \mathcal{L}_{SoftCLIP}(\hat{e}_I, e_I) + \mathcal{L}_{MSE}(\hat{e}_I, e_I) \tag{3}$$

$$\mathcal{L}_{text} = \mathcal{L}_{SoftCLIP}(\hat{e}_T, e_T) + \mathcal{L}_{MSE}(\hat{e}_T, e_T) \tag{4}$$

Where e_I and e_T are CLIP image and text embeddings of image stimuli I and captions T.

Cyclic fMRI Reconstruction. To facilitate subject-invariant representation learning, the simplest way is to directly minimize the distance between the semantic embeddings e_s from two subjects when they are viewing the same images. Nevertheless, as described in Sec. 3.1, there is no common-viewd image across different subjects in the training set. Therefore, we turn to design a mechanism, wishing to synthesize the fMRI signals even the subject does not actually see the image stimulus. In this way, we can mimic the scenario that two subjects are viewing the same images.

To realize that, we first introduce a brain builder $\mathcal{B}_s$ to reconstruct fMRI signal $\hat{v}_s = \mathcal{B}_s(\mathcal{E}_s(v_s))$ in a AutoEncoder [2] manner. The reconstruction loss is:

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_{i=1}^{N}(\hat{v}_s - v_s)^2 \tag{5}$$

We then randomly select two subjects a and b from all training subjects in every training iteration. Through a cyclic fMRI reconstruction, we can transform subject a's fMRI signal v_a into subject b's $v_b = \mathcal{B}_b(\mathcal{E}_a(v_a))$ like they are viewing the same image. This cycle is tenable only when the involved semantic embeddings $e_a = \mathcal{E}_a(v_a), e_b = \mathcal{E}_b(v_b)$ are really subject-invariant, that is, the same when viewing the same image. So a cycle loss is employed to ensure consistency in this cycle:

$$\mathcal{L}_{cyc} = \frac{1}{N}\sum_{i=1}^{N}(e_b - e_a)^2 \tag{6}$$

MindBridge is trained end-to-end by incorporating all these losses to achieve cross-subject brain decoding.

$$\mathcal{L}_{total} = \mathcal{L}_{image} + \mathcal{L}_{text} + \mathcal{L}_{rec} + \mathcal{L}_{cyc} \tag{7}$$

3.3. New-Subject Adaptation

The scope of "cross-subject" is not limited to previously trained subjects. With MindBridge's ability to handle different subjects within a single model, adapting the model to a new subject is now feasible. This scenario is ubiquitous in real-world applications, such as diagnosing a new patient. However, in practice, acquiring brain signals for a new subject can be extremely costly and time-consuming. For example, the authors of NSD dataset [1] spent an entire year completing all fMRI scan sessions. To address the challenge of limited data for a new subject, we adopt the classic "pretrain-then-finetune" paradigm and propose two techniques: reset-tuning and pseudo data augmentation, to enhance the performance of new subject adaptation.

Reset-Tuning. Contrary to traditional fine-tuning in computer vision, which often freezes shallow layers to leverage generic features [49], MindBridge adopts an inverse manner. The transferable knowledge within MindBridge resides in the deep layers, brain translator $\mathcal{T}$. While the shallow layers, brain embedder $\mathcal{E}_s$ and builder $\mathcal{B}_s$, are subject-specific due to human brain diversity. Hence, we propose reset-tuning strategy: training the brain embedder and builder from reset parameters while freezing the brain translator to retain the pretrained cross-subject knowledge.

Pseudo Data Augmentation. A straightforward approach to mitigating the data scarcity problem is through data augmentation. However, suitable data augmentation method for brain signals is currently unavailable. Nevertheless, the cycle reconstruction mechanism of fMRI can serve as a form of pseudo data augmentation. During the adaptation process, fMRI signals from all previously trained subjects can be utilized to augment new subject's data: converted into the fMRI signals of the new subject through cycle reconstruction, regulated by $\mathcal{L}_{rec}$ and $\mathcal{L}_{cyc}$ too.

4. Experiments and Analysis

Evaluation Metrics. To quantitatively compare with other methods, we adopt eight image quality evaluation metrics following [23]. PixCorr, SSIM [43], AlexNet(2), and AlexNet(5) [17] are used to evaluate low-level properties. Inception [37], CLIP [25], EffNet-B [39], and SwAV [4] are considered for evaluating higher-level properties.

4.1. Cross-Subject Brain Decoding

MindBridge can perform brain decoding for multiple subjects using one single model, whereas other methods require training separate models for each subject. To validate MindBridge's effectiveness, we compare its average image reconstruction performance across all four subjects with that of state-of-the-art methods: Takagi *et al.* [38], Brain-Diffuser [23], and MindEye [33]. We also train a per-subject-per-model MindBridge for fair comparison, which is denoted as "MindBridge (Single)". The quantitative and qualitative results for all methods are presented in Tab. 1 and Fig. 3 respectively. Through subject-invariant representation learning, we achieve comparable performance against

Figure 3. **Brain decoding results with only one model.** Unlike previous methods, which confine one model to a specific subject, our proposed cross-subject brain decoding framework, MindBridge, can reconstruct images from multiple subjects using just one model.

Method	# Models	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Takagi et al. [38]	4	–	–	83.0%	83.0%	76.0%	77.0%	–	–
Brain-Diffuser [23]	4	.254	.356	94.2%	96.2%	87.2%	91.5%	.775	.423
MindEye [33]	4	.309	.323	94.7%	97.8%	93.8%	94.1%	.645	.367
MindBridge (Single)	4	.148	.259	86.9%	95.3%	92.2%	94.3%	.713	.413
MindBridge (Ours)	**1**	.151	.263	87.7%	95.5%	92.4%	**94.7%**	.712	.418

Table 1. **Quantitative comparison of brain decoding between MindBridge and other methods.** Our MindBridge is the first effective cross-subject brain decoding approach that only employs one model to reconstruct images from multiple subjects' fMRI signals. While other methods follows a per-subject-per-model fashion. All metrics are calculated as the average across 4 subjects.

state-of-the-art methods while maintaining just one model, demonstrating our success on cross-subject brain decoding.

4.2. New Subject Adaptation

MindBridge also possesses a capability to transfer its pretrained knowledge for adapting to new subjects, which is valuable in practical applications where collecting brain signals for new subjects is resource-intensive and time-consuming. To simulate scenarios with limited data, we tested our method using subsets of the total 8859 training data – specifically, 500, 1500, and 4000 data points – for new subject adaptation. We selected three subjects for pretraining (source subjects) and one additional subject for adaptation (target subject). We choose to compare our method with the "vanilla" approach, which involves training MindBridge from "scratch" on the same target data in a per-subject-per-model fashion. In Fig. 1, we present a qualitative comparison of our method with vanilla method. The vanilla method struggles to reconstruct reasonable images, which can be attributed to two main factors. Firstly, our pretrained brain translator serves as a robust prior backbone, transferring highly useful knowledge that significantly enhances our method's performance. Secondly, the full-parameter model used in the vanilla approach tends to overfit when data is limited. In contrast, our approach employs reset-tuning, updating only the parameters within the lightweight brain embedder and brain builder, effectively preventing overfitting. The quantitative comparison shown in Tab. 2 demonstrates that our method not only significantly outperforms traditional approaches but also highlights the feasibility of reliable brain decoding with substantially less data. This advancement opens up exciting

Method	# Data	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Vanilla	500	.079	.171	73.5%	83.3%	74.4%	80.1%	.894	.587
MindBridge (Ours)	500	**.112**	**.229**	**79.6%**	**89.0%**	**82.3%**	**86.7%**	**.840**	**.521**
Vanilla	1500	.107	.206	79.4%	90.0%	82.4%	87.2%	.844	.523
MindBridge (Ours)	1500	**.140**	**.250**	**84.6%**	**92.6%**	**85.8%**	**91.0%**	**.796**	**.485**
Vanilla	4000	.114	.232	81.4%	92.2%	85.3%	89.8%	.815	.491
MindBridge (Ours)	4000	**.156**	**.258**	**85.7%**	**94.1%**	**88.9%**	**92.5%**	**.765**	**.458**

Table 2. **Reseults of new subject adaptation in limited data scenario.** Here we report results from models that were trained on subsets of 500, 1500, and 4000 data points, selected from a total of 8859 training data points for subject 7. MindBridge (Ours) is fine-tuned using reset-tuning from the pretrained MindBridge model in subjects 1, 2, and 5. Compared to vanilla methods, which trains per-subject-per-model MindBridge from scratch, our MindBridge achieves superior brain decoding performance, benefiting from its use of pretrained cross-subject knowledge. For adaptation of other subjects, please refer to supplementary materials.

Aggregation Function	Low-Level				High-Level			
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Interpolation	.151	.260	87.1%	95.4%	92.1%	94.4%	.712	.413
AdaAvgPool	.163	.274	87.4%	95.7%	92.8%	94.5%	.707	.405
AdaMaxPool (Ours)	**.165**	**.284**	**88.7%**	**96.2%**	**93.7%**	**95.0%**	**.697**	**.400**

Table 3. **Ablation of different aggregation functions.** Models are trained and evaluated on subject 1.

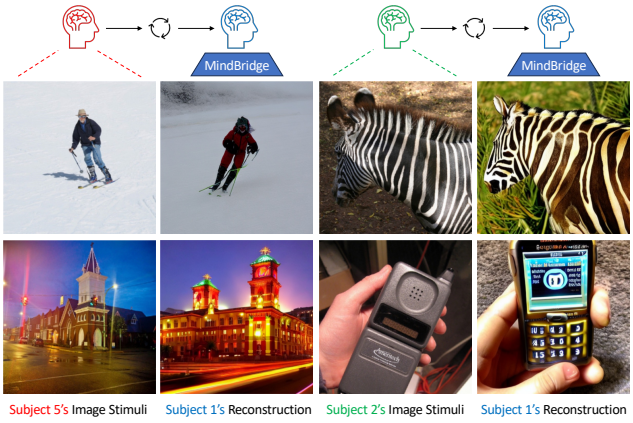

Figure 4. **Novel fMRI synthesis within MindBridge pretrained on subject 1, 2, 5**. The fMRI signals of subjects 5 and 2 are converted into subject 1's fMRI signals through cycle reconstruction, then subject 1's brain embedder are utilized for brain decoding.

prospects for considerably reducing scan times in practical applications, paving the way for more cost-efficient and generalizable brain decoding strategies.

4.3. Novel fMRI Synthesis

Utilizing our cycle reconstruction mechanism, we have enabled a new task: novel fMRI synthesis. This process can transform one's fMRI signal into another's, while preserving the same semantic content as the original stimuli. By employing the pretrained MindBridge model on subject 1,2, and 5, we converted fMRI signals of subject 5 and 2 into those of subjects 1 using their respective brain embedders and brain builders. To validate the quality of these novel fMRI signals, we display the reconstructed images from the synthesized novel fMRI signals in Fig. 4. Notably, the stimuli corresponding to these novel fMRI signals have never been viewed by subject 1. Yet, they can still be reconstructed faithfully, demonstrating the effectiveness of our proposed cycle reconstruction mechanism in synthesize fMRI as well as facilitating subject-invariant representation learning.

4.4. Ablation Study

Ablation on Aggregation Functions. A key component in cross-subject brain decoding is the aggregation function. The ability of this function to retain valuable information while unifying the dimensions of brain signals is crucial. The more effectively it preserves useful information, the more accurate the results will be during the brain decoding process. We show ablation comparison with other functions in Tab. 3. Compared to adaptive average pooling and interpolation functions, our chosen function, adaptive max pooling, not only offers better biological interpretability but also achieves superior performance.

Ablation on Pretraining Losses. We present the re-

Pretrain Loss	Low-Level				High-Level			
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
SoftCLIP loss	.085	**.336**	76.9%	83.1%	79.1%	80.6%	.877	.542
+ MSE loss	.158	.272	88.3%	95.7%	92.2%	94.5%	.720	.418
+ Recon + Cycle Loss (Ours)	**.168**	.277	**88.7%**	**96.1%**	**92.7%**	**94.9%**	**.707**	**.410**

Table 4. **Ablation of different losses at pretraining stage.** Models are trained on subject 1,2 and 5, then evaluated on subject 1.

Finetune Strategy	Low-Level				High-Level			
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Full-tuning + $\mathcal{L}_{image}$ + $\mathcal{L}_{text}$	.110	**.232**	79.5%	88.1%	79.6%	86.4%	.847	.526
Full-tuning + $\mathcal{L}_{total}$	.100	.220	78.6%	88.0%	79.8%	86.0%	.851	.529
Reset-tuning + $\mathcal{L}_{image}$ + $\mathcal{L}_{text}$	**.116**	.227	**79.7%**	88.9%	80.5%	86.1%	.851	.525
Reset-tuning + $\mathcal{L}_{total}$ (Ours)	.112	.229	79.6%	**89.0%**	**82.3%**	**86.7%**	**.840**	**.520**

Table 5. **Ablation of different finetuning strategies.** Models are trained on subject 1,2 and 5, then finetuned and evaluated on subject 7.

sults of involving different losses at the pretraining stage in Tab. 4. When only the SoftCLIP loss is applied, the model struggles to fully learn the reasonable CLIP embeddings and can only achieve a resemblance to the target CLIP embeddings. The inclusion of MSE loss enhances the naturalness of the reconstructed images. Finally, the addition of both reconstruction loss and cycle loss improves the integrity of subject-invariant representation learning, thereby enhancing cross-subject brain decoding performance.

Ablation on Finetuning Strategies. Once we have obtained the pretrained model, aside from our proposed reset-tuning strategy, we have several options for fine-tuning it to adapt to a new subject. Full-tuning involves fine-tuning the brain translator, while reset-tuning entails keeping the brain translator frozen. Both fine-tuning methods involve training a new brain embedder and brain builder, if applicable. We also conducted an ablation study of losses during fine-tuning to assess the benefits of using pseudo data augmentation, which corresponds to the application of losses related to cycle reconstruction. Tab. 5 presents a quantitative comparison among these fine-tuning strategies. The results indicate that the strategy combining reset-tuning with pseudo data augmentation yields the most satisfactory results. This outcome suggests that reset-tuning, which only establishes a projection between brain signals and semantic embeddings, can already sufficiently adapt to a new subject. Moreover, the incorporation of our novel pseudo data augmentation can further imporove performance.

5. Discussion

Currently, due to the limited availability of high-quality fMRI data, one limitation of our paper is the evaluation is only restricted to a small dataset NSD. As a cross-subject framework, the generalizability could be further validated on a more diverse and large-scale dataset in the future. While considering the high cost of acquiring fMRI data, our method also offers a potential solution to reduce scan time. Another limitation is that the fMRI signals are serialized as 1D vectors, which may ruin the original spatial relationship.

Although brain decoding holds promise for assisting visual impaired people, ethical concerns arise regarding misuse for malicious or immoral purposes. Thus, a consented data privacy protocol and a responsible research code of conduct must be established with broader considerations.

6. Conclusion

In this paper, we introduced "MindBridge", a novel cross-subject brain decoding framework that successfully challenges the conventional per-subject-per-model paradigm in brain decoding. By innovatively addressing the critical issues of size variability, diverse neural responses, and data scarcity for new subjects, MindBridge demonstrates significant advancements in cross-subject brain decoding. Our approach, characterized by adaptive signal aggregation, cyclic fMRI reconstruction for subject-invariant representation, and reset-tuning for new subject adaptation, has proven effective in our experiments with the NSD dataset. These achievements not only enhance the decoding accuracy across multiple subjects but also open new avenues for fMRI synthesis and practical applications in neuroscience.

7. Acknowledgement

This project is supported by the Ministry of Education Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006).

References

[1] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience, 25(1):116–126, 2022. 1, 2, 3, 5

[2] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. Machine learning for data science handbook: data mining and knowledge discovery handbook, pages 353–374, 2023. 5

[3] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. Advances in Neural Information Processing Systems, 32, 2019. 3

[4] 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. 5

[5] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. 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, pages 22710–22720, 2023. 3, 4

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34:8780–8794, 2021. 3

[7] Bing Du, Xiaomu Cheng, Yiping Duan, and Huansheng Ning. fmri brain decoding and its applications in brain–computer interface: A survey. Brain Sciences, 12(2):228, 2022. 2

[8] Chengbin Du, Yanxi Li, Zhongwei Qiu, and Chang Xu. Stable diffusion is unstable. Advances in Neural Information Processing Systems, 36, 2023. 3

[9] Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. In Advances in Neural Information Processing Systems, 2023. 3

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016. 4

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020. 3

[12] Zijin Gu, Keith Jamison, Amy Kuceyeski, and Mert Sabuncu. Decoding natural image stimuli from fmri data with a surface-based convolutional network. arXiv preprint arXiv:2212.02409, 2022. 2, 3

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020. 3

[14] Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. The Journal of physiology, 117 (4):500, 1952. 4

[15] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. Nature communications, 8(1):15037, 2017. 3

[16] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. Principles of neural science. McGraw-hill New York, 2000. 4

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 2012. 5

[18] 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. 3

[19] Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In Proceedings of the 31st ACM International Conference on Multimedia, pages 5899–5908, 2023. 3

[20] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In The IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 3

[21] Weijian Mai and Zhijun Zhang. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. arXiv preprint arXiv:2308.07428, 2023. 2, 3

[22] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381(6583):607–609, 1996. 4

[23] Furkan Ozcelik and Rufin VanRullen. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion. arXiv preprint arXiv:2303.05334, 2023. 2, 3, 5, 6

[24] Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In 2022 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE, 2022. 2, 3

[25] 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. 3, 5

[26] 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

[27] 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

[28] 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

[29] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. nature, 323(6088):533–536, 1986. 4

[30] 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

[31] 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

[32] 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

[33] 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

[34] 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

[35] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. PLoS computational biology, 15(1):e1006633, 2019. 2

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020. 3

[37] 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

[38] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. biorxiv. 2022. 2, 3, 4, 5, 6

[39] 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

[40] Rufin VanRullen and Leila Reddy. Reconstructing faces from fmri patterns using deep generative neural networks. Communications biology, 2(1):193, 2019. 3

[41] 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

[42] 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

[43] 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

[44] 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

[45] Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. arXiv preprint arXiv:2310.02265, 2023. 2, 3

[46] 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

[47] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In IEEE/CVF International Conference on Computer Vision, 2023. 3

[48] 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

[49] 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