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[4] 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 [7]. Following common practices [2, 5, 6, 8–10], 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 [8], we use preprocessed fMRI voxels from "NSD-General" 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 [8, 9].

B. Implementation Details

Our method's image reconstruction part relies on versatile diffusion model[11]. In the diffusion process, we employ the UniPCMultistep scheduler[12] 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 AdaptiveMaxPool1D 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 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.

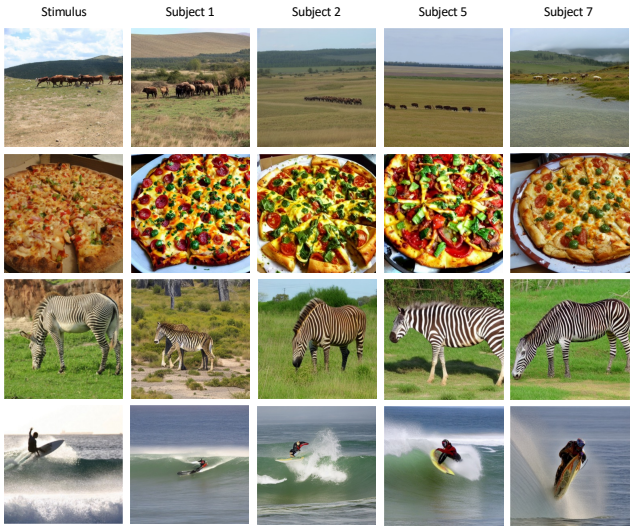

Figure 1. More cross-subject reconstructions of MindBridge on subject 1, 2, 5 and 7.

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 1.

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[3]-like adapter and one linear layer. The brain builder is basically the reverse of brain embedder. 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 2.

E. More Results for New Subject Adaptation

Here, we present additional results of new subject adaptation for subjects 1, 2, and 5 in Table 1, Table 2, and Table 3, 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[11] 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 3. 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.

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] 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. 1

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021. 1

[4] 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. 1

[5] 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. 1

[6] Furkan Ozcelik and Rufin VanRullen. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion. arXiv preprint arXiv:2303.05334, 2023. 1

[7] Jacob S Prince, Ian Charest, Jan W Kurzawski, John A Pyles, Michael J Tarr, and Kendrick N Kay. Improving the accuracy of single-trial fmri response estimates using glmsingle. Elife, 11:e77599, 2022. 1

[8] 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. 1

[9] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. biorxiv. 2022. 1

[10] 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. 1

[11] 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. 1, 2

[12] 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

```
class MindBridge(nn.Module):
    def __init__(self, in_dim=8196, out_dim_image=257*768, out_dim_text=77*768, h=2048,
        ↪ subj_list=[1,2,5,7])
        self.embedder = nn.ModuleDict(subj: nn.Sequential(
            Adapter(in_dim, 128),
            nn.Linear(in_dim, h),
            nn.LayerNorm(h),
            nn.GELU(),
            nn.Dropout(0.5)
        ) for subj in subj_list})

        self.builder = nn.ModuleDict({subj: nn.Sequential(
            nn.Linear(h, in_dim),
            nn.LayerNorm(in_dim),
            nn.GELU(),
            Adapter(in_dim, 128)
        ) for subj in subj_list})

        self.mlp = nn.ModuleList([
            nn.Sequential(
                nn.Linear(h, h),
                nn.LayerNorm(h),
                nn.GELU(),
                nn.Dropout(0.15)
        ) for _ in range(4)])

        self.head_image = nn.Linear(h, out_dim_image)
        self.head_text  = nn.Linear(h, out_dim_text)
```

Figure 2. 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 1. **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.

Method	# Data	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
MindBridge (Scratch)	500	.081	.200	76.0%	87.8%	77.3%	81.6%	.884	.546
MindBridge (Ours)	500	**.122**	**.265**	**83.2%**	**91.0%**	**83.0%**	**87.2%**	**.833**	**.501**
MindBridge (Scratch)	1500	.107	.234	82.0%	93.2%	82.7%	87.1%	.835	.497
MindBridge (Ours)	1500	**.143**	**.273**	**86.6%**	**94.2%**	**88.2%**	**91.7%**	**.780**	**.459**
MindBridge (Scratch)	4000	.128	.257	85.5%	94.2%	87.1%	90.1%	.801	.469
MindBridge (Ours)	4000	**.150**	**.272**	**88.4%**	**95.5%**	**89.8%**	**93.0%**	**.745**	**.436**

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

Method	# Data	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
MindBridge (Scratch)	500	.078	.194	75.0%	88.7%	81.3%	86.7%	.846	.517
MindBridge (Ours)	500	**.124**	**.255**	**83.1%**	**91.9%**	**86.2%**	**91.2%**	**.794**	**.477**
MindBridge (Scratch)	1500	.107	.228	80.9%	92.4%	85.8%	90.8%	.804	.481
MindBridge (Ours)	1500	**.135**	**.262**	**86.1%**	**93.8%**	**89.8%**	**93.5%**	**.746**	**.440**
MindBridge (Scratch)	4000	.123	.239	84.2%	94.6%	90.1%	93.6%	.760	.442
MindBridge (Ours)	4000	**.150**	**.267**	**87.9%**	**95.6%**	**91.9%**	**95.0%**	**.712**	**.418**

Table 3. **Results of new subject adaptation in limited data scenario.** MindBridge(Ours) is finetuned on subject 5 from model pretrained on subject 1, 2 and 5.

Figure 3. Influence of different text-image ratio. Model is trained and tested on subject 1. We reconstruct 8 images using 8 random seeds for illustration.