Probability Distribution Alignment and Low-Rank Weight Decomposition for Source-Free Domain Adaptive Brain Decoding

Ganxi Xu[1], **Jinyi Long**[1*], **Jia Zhang**[1]
[1]College of Information Science and Technology, Jinan University
xuganxi@outlook.com, jinyil@jnu.edu.cn, jiazhang@jnu.edu.cn

Abstract

Brain decoding currently faces significant challenges in individual differences, modality alignment, and high-dimensional embeddings. To address individual differences, researchers often use source subject data, which leads to issues such as privacy leakage and heavy data storage burdens. In modality alignment, current works focus on aligning the softmax probability distribution but neglect the alignment of marginal probability distributions, resulting in modality misalignment. Additionally, images and text are aligned separately with fMRI without considering the complex interplay between images and text, leading to poor image reconstruction. Finally, the enormous dimensionality of CLIP embeddings causes significant computational costs. Although the dimensionality of CLIP embeddings can be reduced by ignoring the number of patches obtained from images and the number of tokens acquired from text, this comes at the cost of a significant drop in model performance, creating a dilemma. To overcome these limitations, we propose a source-free domain adaptation (SFDA)-based brain decoding framework. Firstly, we apply SFDA, which only acquires the source model without accessing source data during target model adaptation, to brain decoding to address cross-subject variations, privacy concerns, and the heavy burden of data storage. Secondly, we employ maximum mean discrepancy (MMD) to align the marginal probability distributions between embeddings of different modalities. Moreover, to accommodate the complex interplay between image and text, we concatenate the embeddings of image and text and then use singular value decomposition (SVD) to obtain a new embedding. What's more, to achieve better image generation quality, we employ the Wasserstein distance (WD) to align the probability distributions of new embeddings. Finally, in the target model adaptation phase of SFDA, we employ low-rank adaptation (LoRA) to reduce the high expense of tuning the target model. Sufficient experiments demonstrate our work outperforms state-of-the-art methods for brain decoding tasks.

Extended version — https://arxiv.org/abs/2504.09109

Introduction

Brain decoding aims to reconstruct visual stimuli from brain signals, primarily acquired through functional magnetic resonance imaging (fMRI) (Logothetis 2008). This approach not only advances our understanding of neural representation but also holds promise for applications in brain-computer interfaces, neurorehabilitation, and personalized cognitive assessment (Wolpaw et al. 2002; Sitaram et al. 2017; Haynes and Rees 2006). Over the past few years, brain decoding has been able to reconstruct more realistic images with the help of increasingly powerful generative models (Takagi and Nishimoto 2023; Ozcelik and VanRullen 2023; Wang et al. 2024). However, it still faces some significant challenges.

First, the brain activity exhibits high variability across subjects (Chen et al. 2023). This implies that a model trained on one subject (source subject) has limited generalization when applied to another (target subject). Furthermore, source subject data contains subjects' personal information with open access posing privacy leakage risks, while brain decoding tasks require massive amounts of data leading to significant data storage burdens. Although Wang *et al.* (Wang et al. 2024) propose a cross-subject brain decoding framework, during the adaptation phase, this framework requires the use of source subjects' data, thereby introducing not only privacy leakage risks but also increasing data storage burdens.

Additionally, recent studies leverage deep generative models for brain decoding by aligning brain signals with vision-language models (Radford et al. 2021). Current approaches for modal alignment primarily employ the SoftCLIP loss (Scotti et al. 2024). This loss employs softmax probability distribution generated by a robust teacher model to provide a more effective teaching signal to the student model compared to hard labels. However, these approaches overlook directly aligning the marginal probability distributions between embeddings of different modalities, resulting in misalignment between the distributions of fMRI and image/text data, which ultimately degrades image reconstruction performance.

Moreover, existing methods align fMRI with images and text separately without accounting for the complex interplay between image and text modalities (Yang et al. 2024). For example, Wang *et al.* (Wang et al. 2024) align predicted image embeddings with CLIP (Radford et al. 2021) image embeddings and predicted text embeddings with CLIP text embeddings. This approach fails to account for the complex interplay between images and text, leading to misaligned prob-

---

[*] Corresponding authors: Jinyi Long.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ability distributions across different modalities.

Finally, to ensure the quality of image reconstruction, CLIP image embeddings incorporate both category-related and patch embeddings, while CLIP text embeddings account for the number of tokens (Ozcelik and VanRullen 2023). However, the high dimensionality of these embeddings results in significant computational costs.

To address these challenges, we propose a brain decoding framework based on source-free domain adaptation (SFDA), which not only enables cross-subject and cross-modal brain decoding but also mitigates privacy concerns, reduces computational costs, and alleviates data storage burdens. The framework is divided into two phases: source model training, target model adaptation.

In the source model training phase, we train a robust source model prepared for the target model adaptation. The source model comprises four components: an embedder, a translator, an image head, and a text head The image head generates predicted image embeddings to align with CLIP image embeddings, while the text head produces predicted text embeddings to align with CLIP text embeddings. To achieve direct alignment of marginal probability distributions across modalities, we incorporate Maximum Mean Discrepancy (MMD) to perform modality alignment.

During the target model adaptation phase, we access the parameters of the source model without accessing any source subject data, thereby achieving privacy preservation and alleviating data storage burdens. Following the source model training phase, we also utilize MMD to perform marginal distribution alignment.

What's more, building upon the complex interplay between images and text, we concatenate the predicted image and text embeddings with the corresponding CLIP image and text embeddings, forming both the predicted and CLIP unified embeddings. To extract meaningful features from these unified embeddings, we apply Singular Value Decomposition (SVD), where the resulting singular values are used as the new unified embeddings. Inspired by WGAN (Arjovsky, Chintala, and Bottou 2017), we employ Wasserstein distance (WD) to measure the probability distributions discrepancy between the new unified embeddings, thereby achieving superior image generation performance.

Finally, we apply low-rank adaptation (LoRA) to both the image head and text head of the target model to mitigate the excessive computational demands caused by the high dimensionality of image and text embeddings, while ensuring the quality of image reconstruction.

Our main contributions are summarized as follows.

- We introduce source-free domain adaptation (SFDA) in brain decoding to alleviate issues of cross-subject variations, privacy concerns, and the high burden of data storage.
- We employ maximum mean discrepancy (MMD) to align the marginal probability distributions of fMRI with images and text, thereby addressing the incomplete alignment caused by relying solely on the SoftCLIP loss.
- We consider the complex interplay between image and text modalities, concatenate their embeddings, and apply singular value decomposition (SVD) to obtain unified embeddings.
- We leverage Wasserstein distance (WD) to align the probability distributions of unified embeddings, thereby significantly improving image generation quality.
- We apply low-rank adaptation (LoRA) to our model, reducing computational costs while maintaining the model's strong learning capacity.
- We perform extensive experiments and have shown that our method has achieved state-of-the-art performance among the existing brain decoding methods.

Related Works

Brain Decoding

Brain decoding has attracted increasing attention from researchers due to the outstanding performance of Latent Diffusion Models in high-resolution image synthesis (Rombach et al. 2022), the advent of multimodal models like CLIP (Radford et al. 2021), and the availability of new fMRI datasets (Allen et al. 2022). However, most approaches (Takagi and Nishimoto 2023; Ozcelik and VanRullen 2023; Scotti et al. 2024) adopt a per-subject-per-model fashion, which demonstrates significantly degraded cross-subject generalization performance when applied to a new subject. To address this limitation, Wang *et al.* (Wang et al. 2024) develop a cross-subject framework for multiple subjects using only one model. However, their approach does not actually use only one model; instead, it sets up independent embedders for each subject. Additionally, when adapting the model to the target subject, data from source subjects is used, which raises privacy concerns. Our work mitigates above issues through a SFDA (Liang, Hu, and Feng 2020) mechanism.

Cross-Modal Learing with Brain Signals

Recent developments in Vision-Language Models (VLMs) (Radford et al. 2021; Jia et al. 2021; Yao et al. 2021) have greatly influenced the field of computer vision, especially in tasks that integrate language and images. The advent of VLMs has demonstrated their potential as bridges across modalities, enhancing interactions with brain signals (Takagi and Nishimoto 2023; Ozcelik and VanRullen 2023). To achieve better alignment between fMRI and images, Scotti *et al.* (Scotti et al. 2024) design the SoftCLIP loss, where the softmax probability distribution generated by a powerful teacher model serves as a superior teaching signal for a student model compared to hard labels. However, this loss function ignores the direct alignment of the marginal probability distributions between fMRI and other modality. To overcome this problem, our work employs Maximum Mean Discrepancy (MMD) to align the marginal probability distributions.

Source-free Domain Adaptation (SFDA)

Due to the excellent performance of SFDA in protecting data privacy, an increasing number of researchers have applied SFDA to multimodal models. Yin *et al.* (Yin et al. 2023) apply SFDA to multimodal semantic segmentation. Tang *et al.* (Tang et al. 2024) utilize multimodal foundation models in SFDA. Huang *et al.* (Huang et al. 2022) use SFDA to build a multimodal video classification model. Inspired by prior works, we apply SFDA to brain decoding to mitigate cross-subject variations, privacy concerns, and data storage burdens.

Low-Rank Adaptation (LoRA)

LoRA is notable for using low-rank matrices to approximate weight changes during fine-tuning and allowing for seamless integration with pre-trained weights before inference, all without introducing additional computational overhead. Due to LoRA's superior performance, it has been applied in many areas. For example, Hyeon *et al.* (Hyeon-Woo, Ye-Bin, and Oh 2021) concentrate on the low-rank Hadamard product for federated learning. Qiu *et al.* (Qiu et al. 2023) introduce orthogonal fine-tuning to adapt text-to-image diffusion models for downstream tasks. Yeh *et al.* (Yeh et al. 2023) introduce an open-source library that offers a wide selection of fine-tuning methodologies for Stable Diffusion. In brain decoding, the high-dimensional embeddings lead to prohibitive computational overhead. To address this challenge, we employ DoRA (Weight-Decomposed Low-Rank Adaptation) (Liu et al. 2024) to significantly reduce computational costs while maintaining image reconstruction performance.

Methodology

Problem Definition

In this section, we detail the brain decoding task. Specifically, the 1D fMRI voxels from the source subject are represented as $v_s \in \mathbb{R}^F$, where F denotes fMRI voxel's size. The image stimulus I_s and image caption C_s corresponding to fMRI voxels v_s are encoded into image embedding $e_{s,I}$ and text embedding $e_{s,C}$ via a pretrained CLIP model. By replacing s with t, the corresponding representations for the target subject can be obtained.

We clarify our framework in two stages: source model training, target model adaptation. In the source model training phase, we project the source fMRI voxels v_s into an embedding $e_s = \mathcal{E}_s(v_s)$ through an embedder $\mathcal{E}_s$. Subsequently, the embedding e_s is translated through a translator $\mathcal{T}_s$ into two distinct embeddings, $(\hat{e}_{s,I}, \hat{e}_{s,C}) = \mathcal{T}_s(e_s)$. To achieve modality alignment with the CLIP-generated embeddings $e_{s,I}$ and $e_{s,C}$, the embeddings $\hat{e}_{s,I}$ and $\hat{e}_{s,C}$ are projected through an image head $\mathcal{H}_{s,I}$ and a text head $\mathcal{H}_{s,C}$, generating embeddings $\hat{e}'_{s,I} = \mathcal{H}_{s,I}(\hat{e}_{s,I})$ and $\hat{e}'_{s,C} = \mathcal{H}_{s,C}(\hat{e}_{s,C})$, which share the same embedding spaces as the CLIP embeddings $e_{s,I}$ and $e_{s,C}$, respectively. Replacing the subscript s with t can give the process of the target model adaptation phase.

In the following content, we will discuss the source model training and target model adaptation phases separately. Therefore, to simplify the expression, we will omit the subscripts s and t that distinguish between the source model training phase and the target model adaptation phase.

Source Model Training

During the source model training phase, modality alignment is achieved using two types of loss functions. The first is the SoftCLIP loss (Scotti et al. 2024), which is inspired by the concept of knowledge distillation (Hinton, Vinyals, and Dean 2015). This loss utilizes the softmax probability distribution produced by a robust teacher model to deliver a more effective teaching signal to the student model than hard labels.

$$\begin{aligned}\mathcal{L}_{SoftCLIP}(\hat{e}', e) = -\sum_{i=1}^{N}\sum_{j=1}^{N}\Bigg[\frac{exp(\frac{e_i \cdot e_j}{\tau})}{\sum_{m=1}^{N} exp(\frac{e_i \cdot e_m}{\tau})} \cdot \\ log\left(\frac{exp(\frac{\hat{e}'_i \cdot e_j}{\tau})}{\sum_{m=1}^{N} exp(\frac{\hat{e}'_i \cdot e_m}{\tau})}\right)\Bigg],\end{aligned} \tag{1}$$

where $\hat{e}'$ and e represent the predicted CLIP embedding and CLIP embedding, respectively; N denotes the batch size; and τ is the temperature hyperparameter.

While the SoftCLIP loss aligns predicted CLIP embeddings and CLIP embeddings in terms of their softmax probability distributions, it overlooks the direct alignment of their marginal probability distributions. Therefore, we employ maximum mean discrepancy (MMD) (Gretton et al. 2012) to align the marginal probability distributions of the predicted CLIP embedding and CLIP embedding.

$$\mathcal{L}_{MMD}(\hat{e}', e) = MMD(\hat{e}', e). \tag{2}$$

Thus, we can derive the loss functions for the predicted image embeddings and predicted text embeddings.

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

$$\mathcal{L}_{text} = \mathcal{L}_{SoftCLIP}(\hat{e}'_C, e_C) + \mathcal{L}_{MMD}(\hat{e}'_C, e_C). \tag{4}$$

The overall loss function for source model training is formulated as

$$\mathcal{L}_{SRC} = \mathcal{L}_{image} + \mathcal{L}_{text}. \tag{5}$$

Target Model Adaptation

During the target model adaptation phase, we utilize only the source model without accessing the source subject's data, thereby achieving privacy protection and reducing the data storage burden.

Similar to the source model training phase, we employ the SoftCLIP loss and MMD to perform cross-modal alignment.

$$\mathcal{L}_{image} = \mathcal{L}_{SoftCLIP}(\hat{e}'_I, e_I) + \mathcal{L}_{MMD}(\hat{e}'_I, e_I), \tag{6}$$

$$\mathcal{L}_{text} = \mathcal{L}_{SoftCLIP}(\hat{e}'_C, e_C) + \mathcal{L}_{MMD}(\hat{e}'_C, e_C). \tag{7}$$

Current approaches primarily focus on aligning predicted image embeddings with CLIP image embeddings and predicted text embeddings with CLIP text embeddings separately (Wang et al. 2024; Scotti et al. 2024), overlooking the complex interplay between images and text (Yang et al. 2024). To address this limitation, we concatenate the predicted image embedding and predicted text embedding to form a unified predicted embedding, denoted as $\hat{e}'_u$, as

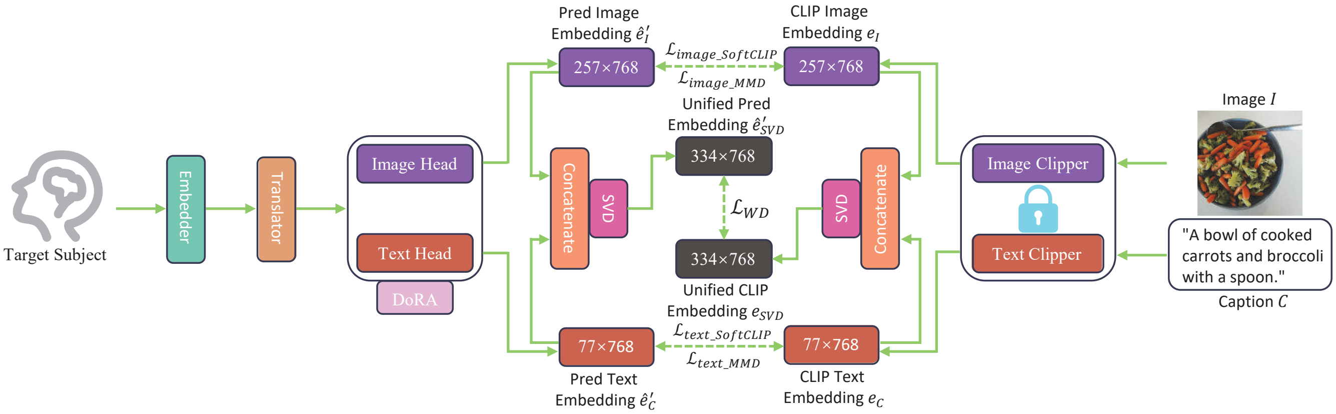

Figure 1: The framework of the target model adaptation.

shown in Figure 1. Similarly, we concatenate the CLIP image embedding and CLIP text embedding to create a unified CLIP embedding, denoted as e_u.

$$\hat{e}'_u = Concatenate(\hat{e}'_I, \hat{e}'_C), \quad (8)$$

$$e_u = Concatenate(e_I, e_C). \quad (9)$$

To extract meaningful features from unified embeddings, we employ SVD (Stewart 1993) to process the obtained unified embeddings, which requires the use of fMRI embeddings for image generation.

$$\hat{U}'_u, \hat{\Sigma}'_u, \hat{V}'_u = SVD(\hat{e}'_u), \quad (10)$$

$$U_u, \Sigma_u, V_u = SVD(e_u), \quad (11)$$

$$\hat{e}'_{SVD} = \hat{\Sigma}'_u, \quad (12)$$

$$e_{SVD} = \Sigma_u, \quad (13)$$

where $\hat{e}'_u$ and e_u have dimension [batch size, 334 × 768], $\hat{e}'_{SVD}$ and e_{SVD} are both vectors of dimension 334 × 768.

The objective of brain decoding is to reconstruct visual stimuli from brain signals, thus necessitating the use of fMRI embeddings for image generation. The application of WD (Rüschendorf 1985) in Generative Adversarial Networks (GANs) (Goodfellow et al. 2020) for image generation offers significant inspiration for this task. The WD is more sensitive to local variations in distributions, enabling it to capture fine-grained differences (Arjovsky, Chintala, and Bottou 2017). Therefore, we employ the Wasserstein-1 distance (Shui et al. 2020) to measure the probability distribution discrepancy between the two unified embeddings.

$$\mathcal{L}_{WD} = WD(\hat{e}'_{SVD}, e_{SVD}). \quad (14)$$

Finally, we need to address the significant computational overhead caused by the high dimensionality of image and text embeddings. We demonstrate this with the NSD dataset utilized in our experiments. When performing modal alignment, we need to generate 257×768-dimensional predicted CLIP image embeddings and 77×768-dimensional predicted CLIP text embeddings from fMRI signals, and align them with CLIP image embeddings (257×768-dim extracted from the corresponding images where the first 768-dimensional vector indicates the category-related embedding and the remaining 256 embeddings represent the patches obtained from the images) and CLIP text embeddings (77×768-dim generated from the COCO captions associated with the corresponding images, where the 77 embeddings correspond to the number of tokens provided as input to the model), respectively (Ozcelik and VanRullen 2023). As a result, the embedding dimensions become extremely large, leading to significant computational cost, as shown in Table 1.

Module	**Input Dim.**	**Output Dim.**
Input	$B \times 8192$	-
Embedder	$B \times 8192$	$B \times 2048$
Translator	$B \times 2048$	$B \times 2048$
Image head	$B \times 2048$	$B \times 197376$ $(= 257 \times 768)$
Text head	$B \times 2048$	$B \times 59136$ $(= 77 \times 768)$

B : Batch Size.

Table 1: The input and output dimensions of the modules

Fortunately, with the rise of large language models (LLMs) (Chang et al. 2024), parameter-efficient fine-tuning (PEFT) (Houlsby et al. 2019) has emerged, allowing fine-tuning with a small number of parameters. Among these methods, LoRA (Hu et al. 2022) has gained significant popularity because it does not require changing the model architecture. Therefore, we apply DoRA (Weight-Decomposed Low-Rank Adaptation) (Liu et al. 2024) to the image head $\mathcal{H}_I$ and text head $\mathcal{H}_C$, which can effectively reduce computational costs while ensuring high-quality image reconstruction. DoRA breaks down the pre-trained weights into two components: magnitude and direction. For fine-tuning, it specifically uses LoRA to make directional updates.

$$W' = \underline{m}\frac{V + \Delta V}{\|V + \Delta V\|_c} = \underline{m}\frac{W_0 + \underline{BA}}{\|W_0 + \underline{BA}\|_c}, \quad (15)$$

where $m \in \mathbb{R}^{1 \times k}$ represents the magnitude vector, ΔV is the incremental directional update obtained by multiplying two low-rank matrices B and A, and the underlined parameters indicate the trainable parameters. Furthermore, $\|\cdot\|_c$ denotes the vector-wise norm of a matrix across each column.

Based on the above, we can obtain the complete loss function for the target model adaptation phase.

$$\mathcal{L}_{ADP} = \mathcal{L}_{image} + \mathcal{L}_{text} + \mathcal{L}_{WD}. \tag{16}$$

Experiment

Datasets

Our work utilizes the Natural Scenes Dataset (NSD) (Allen et al. 2022), a widely-used benchmark in neuroimaging studies. This dataset contains high-resolution 7-Tesla fMRI recordings obtained from eight healthy adult participants during visual perception tasks involving thousands of natural images from the Common Objects in Context (COCO) dataset (Lin et al. 2014).

Consistent with established methodologies in the field, our work focuses on four participants (subj01, subj02, subj05, and subj07) who completed all scanning sessions. The training set for each participant comprises 8859 image stimuli and 24980 fMRI trials (with a maximum of 3 trials per image), while the test set contains 982 image stimuli and 2770 fMRI trials. We employ preprocessed voxel data from the NSDGeneral region of interest (ROI), which varies in dimensionality across participants, containing 15724, 14278, 13039, and 12682 voxels, respectively.

Experimental Settings

Our experimental platform consists of three Tesla V100S PCIe 32GB GPUs. The source model training is executed for 600 epochs, while the target model adaptation phase is executed for 200 epochs. The batch size for both the source model training and target model adaptation phases is set to 50, the optimizer is set to AdamW, where the maximum learning rate is set to 1.5×10^{-4}.

We combine the data from three subjects as the source data, while the data from the remaining subject is used as the target data. For example, if subj07 is selected as the target subject, the data from subj01, subj02, and subj05 are combined to form the source data.

Results

The experimental results on several metrics (See the Metrics section in the Appendix) are shown in Table 2, with the best results highlighted in bold, while we also present the qualitative results in Figure 2. Although three methods (Wang et al. 2024; Xia et al. 2024b; Shen et al. 2025) claim to employ only a single model in their work, they still employ separate modules for each subject during implementation. For instance, MindBridge establishes an independent embedder for each subject. In contrast, our approach does not introduce subject-specific modules. We utilize only one embedder, one translator, one image head, and one text head throughout both the source model training phase and the target model adaptation phase.

The results indicate that using MMD effectively aligns fMRI with both images and text. We concatenate the image embeddings and text embeddings and process them with SVD, which adequately considers the complex interplay between images and text. By using WD, we can achieve a good probabilistic alignment of the unified embeddings, thereby enhancing the quality of image generation. In this regard, the superior performance demonstrates the effectiveness of these approaches. Our work introduces SFDA into brain decoding, effectively safeguarding data privacy and alleviating the data storage burden.

Multi-Embedder VS. Single-Embedder

MindBridge (Wang et al. 2024), UMBRAE (Xia et al. 2024b), and Shen *et al.* (Shen et al. 2025), all claim to use only one model. However, each of these works sets up an independent module for subjects. For example, in MindBridge, an independent embedder is set up for each source subject during the training phase. In the target model adaptation phase, not only is the embedder for each source subject retained, but an independent embedder is also set up for the target subject. In contrast, our method utilizes only one embedder throughout both the source model training phase and the target model adaptation phase. Furthermore, we do not access data from the source subjects during the target model adaptation phase.

We conduct experiments with our method following the experimental setup of MindBridge and compare it with MindBridge, as shown in Table 3. The table shows that our method achieves the best results across all metrics. Moreover, our approach introduces SFDA into brain decoding, effectively overcoming cross-subject variations, protecting data privacy, and reducing data storage burdens.

Ablation Study on MMD and Aggregation Method

In this section, we study the effectiveness of MMD and the aggregation method. We first concatenate image and text embeddings to capture their complex interplay. Then, we apply SVD to the resulting unified embeddings and use the singular values as new unified embeddings. Finally, we utilize WD to probabilistically align the predicted embeddings with CLIP embeddings to achieve higher-quality image generation. We call this multi-step method the aggregation method.

The experimental results are shown in Table 4. From the table, it can be seen that when we use MMD, the performance improves significantly. Although the SoftCLIP loss can provide an effective teaching signal through the softmax probability distribution, it still cannot guarantee the authenticity of the learned predicted embeddings (Wang et al. 2024). Therefore, we use MMD to ensure more accurate predicted embeddings by aligning marginal probability distributions, resulting in notable performance improvements. Table 4 shows that the aggregation method further improves performance. We aim to account for the complex interplay between image and text embeddings without simply concatenating them. Therefore, we use SVD to obtain new unified embeddings. To enhance the quality of image generation, we use WD to align the probability distributions between the unified predicted embeddings and the unified CLIP embeddings. The performance improvement demonstrates the effectiveness of the aggregation method.

(A) Stimulus **(B) Subj01** **(C) Subj02** **(D) Subj05** **(E) Subj07**

Figure 2: Qualitative results.

Method	Models	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Takagi *et al.* (Takagi and Nishimoto 2023)	4	—	—	83.0%	83.0%	76.0%	77.0%	—	—
Brain-Diffuser (Ozcelik and VanRullen 2023)	4	0.254	0.356	94.2%	96.2%	87.2%	91.5%	0.775	0.423
MindEye (Scotti et al. 2024)	4	**0.309**	0.323	94.7%	**97.8%**	93.8%	94.1%	0.645	0.367
DREAM (Xia et al. 2024a)	4	0.288	0.338	**95.0%**	97.5%	94.8%	95.2%	0.638	0.413
MindBridge (Wang et al. 2024)	1	0.151	0.263	87.7%	95.5%	92.4%	94.7%	0.712	0.418
UMBRAE (Xia et al. 2024b)	1	0.283	0.328	93.9%	96.7%	93.4%	94.1%	0.700	0.393
Shen *et al.* (Shen et al. 2025)	1	0.265	0.357	93.1%	97.1%	96.8%	**97.5%**	0.633	0.321
Our work(w/o DoRA)	1	0.286	**0.362**	93.6%	97.5%	**96.9%**	96.0%	0.617	**0.311**
Our work	1	0.291	0.355	93.4%	97.0%	96.8%	96.3%	**0.615**	0.312

Table 2: Quantitative evaluation on brain decoding. All metrics are averaged over the results from the four subjects.

	Low-Level				High-Level			
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Multi-Embedder	0.277	0.335	92.6%	95.8%	96.0%	95.7%	0.635	0.325
Single-Embedder	**0.291**	**0.355**	**93.4%**	**97.0%**	**96.8%**	**96.3%**	**0.615**	**0.312**

Table 3: Quantitative evaluation between multi-embedder and single-embedder. The single-embedder corresponds to our work. For the multi-embedder, we replicate the experimental setup of MindBridge.

Ablation Study on DoRA and LoRA

In this section, we explore the impact of different rank r on DoRA and LoRA. We assess the performance of our method when the rank r is set to 4, 8, 16, 32, and 64. The results are shown in Table 5. As we can see, DoRA outperforms LoRA across all values of r. Table 5 also provides the number of trainable parameters in our model for different values of r. The application of DoRA and LoRA significantly reduces the number of trainable parameters. DoRA decomposes the pre-trained weights into two components: magnitude and di-

MMD	Aggregation Method	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
✗	✗	0.122	0.242	85.1%	90.0%	86.6%	88.5%	0.749	0.443
✗	✓	0.142	0.257	87.9%	92.2%	88.3%	89.9%	0.717	0.392
✓	✗	0.262	0.354	92.3%	95.9%	92.5%	93.0%	0.632	0.346
✓	✓	**0.286**	**0.362**	**93.6%**	**97.5%**	**96.9%**	**96.0%**	**0.617**	**0.311**

Table 4: Ablation study on MMD and the aggregation method.

PEFT	Params(%)	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Our work(Full Fine-tuning)	100.00	0.286	0.362	93.6%	97.5%	96.9%	96.0%	0.617	0.311
DoRA(r=4)	6.18	0.258	0.315	92.0%	95.0%	93.1%	93.8%	0.724	0.403
DoRA(r=8)	6.34	0.291	0.355	93.4%	97.0%	96.8%	96.3%	0.615	0.312
DoRA(r=16)	6.67	0.275	0.335	93.0%	96.5%	95.9%	96.0%	0.628	0.330
DoRA(r=32)	7.31	0.280	0.336	93.2%	96.8%	96.0%	96.3%	0.620	0.321
DoRA(r=64)	8.57	0.266	0.321	92.6%	95.7%	93.7%	94.0%	0.715	0.390
LoRA(r=4)	6.14	0.175	0.248	87.4%	89.6%	86.8%	86.3%	0.812	0.498
LoRA(r=8)	6.30	0.198	0.274	89.4%	91.2%	87.6%	88.9%	0.796	0.472
LoRA(r=16)	6.63	0.246	0.310	92.8%	94.6%	92.3%	92.6%	0.747	0.415
LoRA(r=32)	7.27	0.251	0.313	92.0%	94.7%	92.5%	93.1%	0.741	0.410
LoRA(r=64)	8.53	0.227	0.292	91.2%	92.6%	89.8%	90.2%	0.772	0.445

Table 5: Ablation study on DoRA and LoRA.

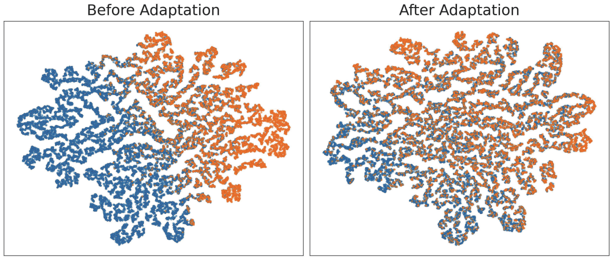

Figure 3: t-SNE Visualization (Van der Maaten and Hinton 2008). Let's take the case where the target subject is subj01 as an example. Blue points represent source data, and orange points represent target data.

rection. During fine-tuning, it employs LoRA for directional updates, effectively reducing the number of trainable parameters. In addition, DoRA enhances the model's learning capacity, further validating the effectiveness of our approach.

Cross-subject Variation Overcoming

Due to the strong variability in brain signals across different subjects, we leverage SFDA to overcome cross-subject variation, as shown in Figure 3. It can be observed that before the target model adaptation phase, source features and target features are not mixed and can be easily distinguished. After the target model adaptation phase, source features and target features are mixed together and cannot be easily differentiated. This demonstrates that we have successfully extracted subject-invariant representations, effectively overcoming cross-subject variation. Furthermore, when training the target model, we do not need to use source subject data, which not only protects the data privacy of the source subject but also reduces the burden of data storage.

Conclusion

In this paper, we introduce SFDA to brain decoding to address the problem of high variability in brain signals. Additionally, since SFDA does not rely on source subject data during the target model adaptation phase, it alleviates concerns about privacy leakage and the heavy burden of data storage. Specifically, in the source model training phase, we use MMD to align the marginal probability distributions of fMRI signals with images and text, achieving modality alignment between fMRI signals and images, as well as between fMRI signals and text. In the target model adaptation phase, we also use MMD to achieve modality alignment, similar to the source model training phase. To account for the complex interplay between images and text, our method concatenates images and text and uses SVD to extract more meaningful embeddings. To achieve higher-quality image generation, we use the WD to align the probability distribution of the unified embeddings. Finally, to reduce the computational burden caused by the high dimensionality of the embeddings while maintaining image reconstruction performance, we leverage DoRA to significantly reduce the number of trainable parameters. We also demonstrate the effectiveness of our method through multiple experiments.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62276115, in part by the Guangzhou Basic and Applied Basic Research Foundation under Grant 2024A04J5486.

References

Allen, E. J.; St-Yves, G.; Wu, Y.; Breedlove, J. L.; Prince, J. S.; Dowdle, L. T.; Nau, M.; Caron, B.; Pestilli, F.; Charest, I.; et al. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1): 116–126.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223. PMLR.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3): 1–45.

Chen, Z.; Qing, J.; Xiang, T.; Yue, W. L.; and Zhou, J. H. 2023. 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*, 22710–22720.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1): 723–773.

Haynes, J.-D.; and Rees, G. 2006. Decoding mental states from brain activity in humans. *Nature reviews neuroscience*, 7(7): 523–534.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, Y.; Yang, X.; Zhang, J.; and Xu, C. 2022. Relative alignment network for source-free multimodal video domain adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1652–1660.

Hyeon-Woo, N.; Ye-Bin, M.; and Oh, T.-H. 2021. Fedpara: Low-rank hadamard product for communication-efficient federated learning. *arXiv preprint arXiv:2108.06098*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Liang, J.; Hu, D.; and Feng, J. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, 6028–6039. PMLR.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, S.-Y.; Wang, C.-Y.; Yin, H.; Molchanov, P.; Wang, Y.-C. F.; Cheng, K.-T.; and Chen, M.-H. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Logothetis, N. K. 2008. What we can do and what we cannot do with fMRI. *Nature*, 453(7197): 869–878.

Ozcelik, F.; and VanRullen, R. 2023. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1): 15666.

Qiu, Z.; Liu, W.; Feng, H.; Xue, Y.; Feng, Y.; Liu, Z.; Zhang, D.; Weller, A.; and Schölkopf, B. 2023. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36: 79320–79362.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rüschendorf, L. 1985. The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1): 117–129.

Scotti, P.; Banerjee, A.; Goode, J.; Shabalin, S.; Nguyen, A.; Dempster, A.; Verlinde, N.; Yundler, E.; Weisberg, D.; Norman, K.; et al. 2024. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36.

Shen, G.; Zhao, D.; He, X.; Feng, L.; Dong, Y.; Wang, J.; Zhang, Q.; and Zeng, Y. 2025. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. *Advances in Neural Information Processing Systems*, 37: 98083–98110.

Shui, C.; Zhou, F.; Gagné, C.; and Wang, B. 2020. Deep active learning: Unified and principled method for query and training. In *International conference on artificial intelligence and statistics*, 1308–1318. PMLR.

Sitaram, R.; Ros, T.; Stoeckel, L.; Haller, S.; Scharnowski, F.; Lewis-Peacock, J.; Weiskopf, N.; Blefari, M. L.; Rana, M.; Oblak, E.; et al. 2017. Closed-loop brain training: the science of neurofeedback. *Nature Reviews Neuroscience*, 18(2): 86–100.

Stewart, G. W. 1993. On the early history of the singular value decomposition. *SIAM review*, 35(4): 551–566.

Takagi, Y.; and Nishimoto, S. 2023. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14453–14463.

Tang, S.; Su, W.; Ye, M.; and Zhu, X. 2024. Source-Free Domain Adaptation with Frozen Multimodal Foundation Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23711–23720.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, S.; Liu, S.; Tan, Z.; and Wang, X. 2024. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11333–11342.

Wolpaw, J. R.; Birbaumer, N.; McFarland, D. J.; Pfurtscheller, G.; and Vaughan, T. M. 2002. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6): 767–791.

Xia, W.; De Charette, R.; Oztireli, C.; and Xue, J.-H. 2024a. Dream: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 8226–8235.

Xia, W.; de Charette, R.; Oztireli, C.; and Xue, J.-H. 2024b. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision*, 242–259. Springer.

Yang, L.; Zhang, R.-Y.; Wang, Y.; and Xie, X. 2024. MMA: Multi-Modal Adapter for Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23826–23837.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Yeh, S.-Y.; Hsieh, Y.-G.; Gao, Z.; Yang, B. B.; Oh, G.; and Gong, Y. 2023. Navigating text-to-image customization: From lycoris fine-tuning to model evaluation. In *The Twelfth International Conference on Learning Representations*.

Yin, Y.; Hu, W.; Liu, Z.; Wang, G.; Xiang, S.; and Zimmermann, R. 2023. CrossMatch: Source-Free Domain Adaptive Semantic Segmentation via Cross-Modal Consistency Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21786–21796.