MindCross: Fast New Subject Adaptation with Limited Data for Cross-subject Video Reconstruction from Brain Signals

Xuan-Hao Liu, Yan-Kai Liu, Tianyi Zhou, Bao-Liang Lu, Wei-Long Zheng*

School of Computer Science, Shanghai Jiao Tong University

{haogram_sjtu, liu-yankai, tianyizhou, bllu, weilong}@sjtu.edu.cn

Abstract

Reconstructing video from brain signals is an important brain decoding task. Existing brain decoding frameworks are primarily built on a subject-dependent paradigm, which requires large amounts of brain data for each subject. However, the expensive cost of collecting brain-video data causes severe data scarcity. Although some cross-subject methods being introduced, they often overfocus with subject-invariant information while neglecting subject-specific information, resulting in slow fine-tune-based adaptation strategy. To achieve fast and data-efficient new subject adaptation, we propose **MindCross**, a novel cross-subject framework. MindCross's N specific encoders and one shared encoder are designed to extract subject-specific and subject-invariant information, respectively. Additionally, a Top-K collaboration module is adopted to enhance new subject decoding with the knowledge learned from previous subjects' encoders. Extensive experiments on fMRI/EEG-to-video benchmarks demonstrate MindCross's efficacy and efficiency of cross-subject decoding and new subject adaptation using only one model.

Code — https://github.com/XuanhaoLiu/MindCross

Introduction

Our daily visual perceptions are composed of a series of seamless scenes, it is of great importance to investigate the neurologic mechanism of how human brain perceives and processes such dynamic visual perceptions (Makeig et al. 2002). One of brain decoding tasks is reconstructing the video from brain signals, which garnered significant interest recently (Yu et al. 2025; Sun, Li, and Moens 2025; Gong et al. 2025). These research deepens our understanding of brain function and offers promising advances for brain-computer interfaces (BCIs).

Spurred by the rapid development of deep multimodal models such as CLIP (Radford et al. 2021) and Stable Diffusion (Rombach et al. 2022), numerous previous works demonstrate the ability to reconstruct high-fidelity visual perception from brain signals (Benchetrit, Banville, and King 2024; Li et al. 2024b; Scotti et al. 2023) such as functional magnetic resonance imaging (fMRI) and electroencephalogram (EEG), which indirectly measures neural activity by detecting changes in blood oxygenation and spontaneous electrical activity from the scalp.

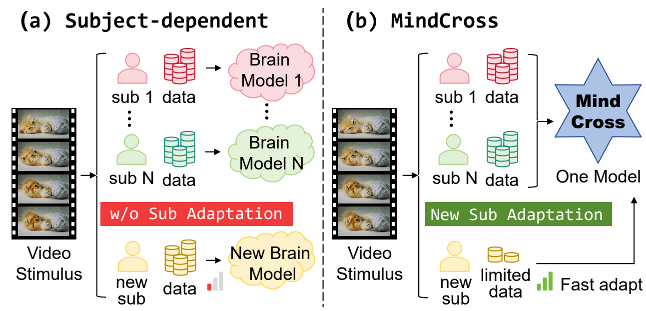

Figure 1: Illustration of cross-subject problem: (a) Subject-dependent paradigm: one single model for each subject. (b) MindCross paradigm: one universal model for all subjects and it can fast adapt to new subjects with limited data.

Unfortunately, there is a significant limitation in nowadays brain decoding frameworks that inherently restricts their wide applications. As depicted in Figure 1 (a), current works predominantly adhere to the subject-dependent paradigm (Chen et al. 2024a,b, 2025). This paradigm necessitates extensive per-subject data acquisition and prolonged model training durations, thereby posing great constraints on its adaptability to new subjects and motivating us to develop more efficient alternatives. Specifically, the cross-subject paradigm emerges as a promising solution, showing two key advantages: (1) decode neural patterns across multiple subjects using a single unified model, and (2) rapid adaptation for new subjects.

However, there are many difficulties in cross-subject brain decoding: **1) Subject Variability** Different subjects have different brain responses to the same stimuli, especially when processing high-level semantic information which requires complex cognition progress. **2) Data Scarcity for New Subject** The high expense of time and resources to conduct BCI experiments limits the amount of new subject data. **3) Utilization of Existing Subjects Data** Since existing subjects' brain data are always larger and more diverse than that of the new subject, it is a waste not to adequately utilize their valuable data.

*Corresponding author.

Several cross-subject frameworks have been designed on an image brain decoding dataset (Allen et al. 2022), whose subjects possess 27750 trials. However, since videos are viewed longer than images, the brain-video benchmark exhibits more significant data scarcity. For instance, each subject of the EEG-video benchmark only has 1400 trials. Furthermore, previous cross-subject methods mostly adopt a fine-tune-based adaptation strategy (Scotti et al. 2024; Wang et al. 2024), which costs a lot of time. These above issues motivate us to design a fast and data-efficient framework.

In this paper, we devise "**MindCross**", a novel cross-subject brain decoding framework designed for rapid adaptation to new subjects. MindCross addresses key challenges through three core innovations: **1) Shared-specific Encoder Architecture** MindCross learns subject-invariant information by a shared encoder and allocates each subject a specific encoder to learn subject-related information. Through this architecture, MindCross is able to decouple learning subject-invariant and subject-related information to solve subject variability. **2) Fast New Subject Calibration** MindCross can rapidly adapt to him/her by only updating the parameter of the new subject while other modules are all frozen. This calibration method provides an efficient and effective solution for new subject adaptation while not harming the decoding performance of the existing subjects, solving data scarcity. **3) Collaboration Decoding** When the new subject has limited training data, MindCross decodes the semantic embeddings not only from the specific encoder of the new subject, but also from the encoders of Top K existing subjects similar to the new subject. This Top-K Collaboration Module largely helps MindCross to decode more precisely by adequately utilizing existing subjects' data.

In conclusion, our contributions are as follows:

- We design MindCross, a shared-specific encoder architecture for decoupling learning subject-invariant and subject-related information.
- We design a novel calibration method for new subject adaptation that only needs updating a small number of parameters for the new subject while keeping other parts frozen, thus is fast and will not affect the decoding performance of existing subjects.
- We design a novel collaboration decoding module that utilizes existing subjects' data to help decoding a new subject with limited training data.
- Extensive experiments on fMRI/EEG-to-video benchmarks demonstrate the efficacy of MindCross.

Related Work

Brain Decoding

Researchers have been trying to decode visual perception (e.g., shape, color, and position) from brain activities for decades (Miyawaki et al. 2008). The rapid development of large generation models (Rombach et al. 2022) enables high-fidelity images/videos generation, bringing a game-changing technique into the brain decoding field. The current brain decoding pipeline (Chen, Qing, and Zhou 2023; Liu et al. 2024b; Liu, Lu, and Zheng 2025a; Sun, Li, and Moens 2025; Lu et al. 2025) can be summarized in two steps: mapping brain signals to CLIP embeddings and then using these embeddings to guide generation models in generating reconstructed images/videos. However, most works fall into the subject-dependent fashion, leaving cross-subject brain decoding largely unexplored.

Cross-subject Algorithm

Cross-subject brain decoding has always been a hot topic because of the practical requirement of quickly adapting a BCI model to new subjects (Liu et al. 2024a; Liu, Lu, and Zheng 2025b). For visual decoding task, MindBridge assigns each subject a particular encoder and adopts a cyclic auto-encoder training strategy to align different subject's brain features (Wang et al. 2024). GLFA (Li et al. 2024a), STTM (Liu et al. 2025) and MindAligner (Dai et al. 2025) applies the functional alignment to align different subject's brain features. MindEye2 (Scotti et al. 2024) and MindTuner (Gong et al. 2025) implicitly aligned all subject brain signals to a shared latent space. Wills Aligner used a mixture of brain experts adapter to achieve few-shot learning (Bao et al. 2025). These methods put excessive attention on the subject-invariant information and fail to capture subject-related information, as their loss function encourages the model to learn similar brain representations from different subjects.

Moreover, for new subject adaptation, GLFA and MindTuner directly fine-tunes the whole model, while MindBridge freezes the deep layers and fine-tunes the shallow layers. Such fine-tune-based strategies not only harm the decoding performance of the previous subjects but also cost a long time as they all update the encoders of previous subjects. In contrast, MindCross achieves fast and data-efficient new subject adaptation with a novel calibration phase, where we update the new subject's specific encoder. For the following comparison, we select the most representative cross-subject framework from image (MindBridge) and video (GLFA) filed.

Video Generation Models

With large-scale image-text pair datasets (Schuhmann et al. 2022), nowadays diffusion models have demonstrated superior performance in the task of text-to-image generation (Rombach et al. 2022). Compared to text-to-image (T2I) models (Mirza and Osindero 2014), text-to-video (T2V) generation models must maintain the temporal consistency between each frame. Tune-A-Video (Wu et al. 2023) design an inflated network technique by adding a temporal attention layer in Unet to augment a T2I model to a T2V model, which was used as the video generation module in many previous brain decoding work (Chen, Qing, and Zhou 2023; Liu et al. 2024b; Lu et al. 2025). However, Tune-A-Video needs to fine-tune on the training set of video clips, consuming lots of time and computation resources. Recently, more wonderful T2V models are proposed, such as AnimateDiff (Guo et al. 2024) and PyramidFlow (Jin et al. 2025). In this paper, we focus on learning cross-subject brain representations, rather than taming SOTA T2V models for brain decoding, thus we simply used these off-the-shelf T2V models without further modification.

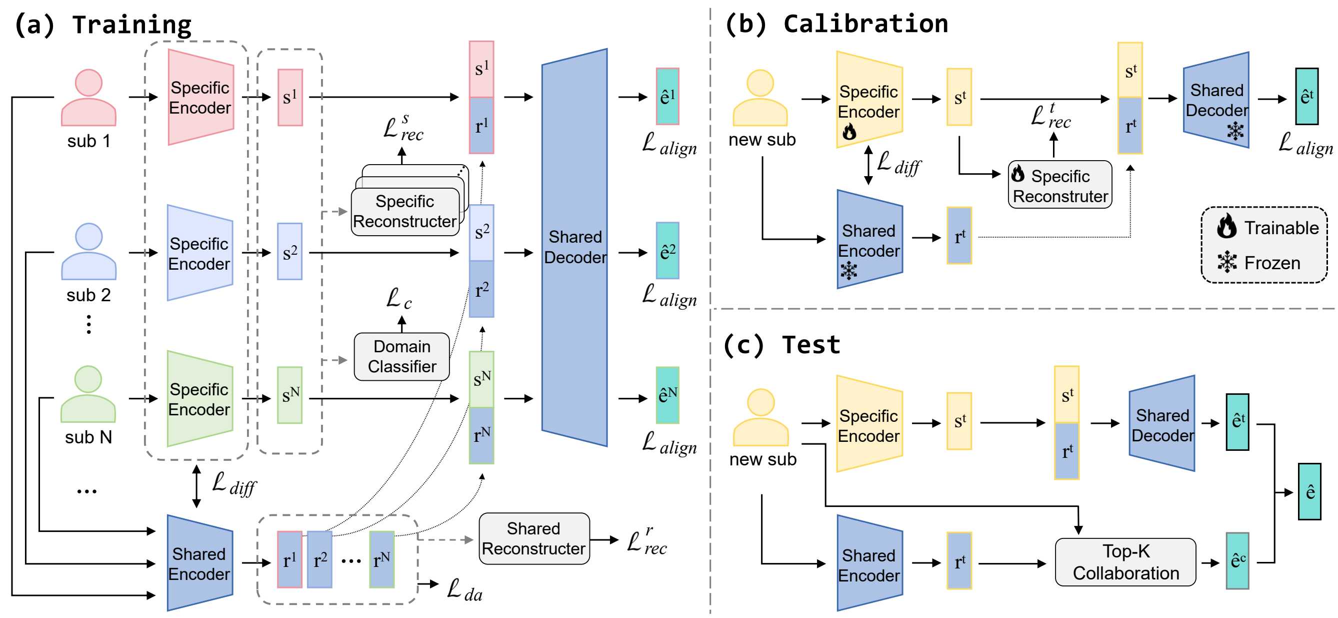

Figure 2: The framework of proposed MindCross consisting of training, calibration, and test phase. **(a)** In the training phase, each specific encoder and shared encoder are optimized by several loss functions. **(b)** In the calibration phase, only the specific encoder of the new subject, marked with a flame icon, will be updated. **(c)** In the test phase, the final predictions are obtained from shared decoder and Top-*K* Collaborate module.

Methodology

Overall Architecture

Current methods of video generation from brain signals all follow a subject-dependent paradigm, requiring a large amount of data from one subject. To overcome inter-subject variability and achieve rapid adaptation on new subjects, we propose MindCross, a shared-specific feature based framework inspired by ShaSpec (Wang et al. 2023). Depicted in Figure 2, the whole framework consists of three phases: training, calibration, and test. In the training phase, N specific encoders and one shared encoder are trained on each subject's data. If a new subject with limited data comes, then in the calibration phase, a new specific encoder will be trained while other modules in MindCross freezing. That is, we only update the new specific encoder and reconstructer. In the test phase, the new subject's final output are obatined by its own specific encoder and a Top-*K* Collaboration module. Let us denote the N subjects' data as $\mathbf{X} = \{\mathbf{x}_j^i, \mathbf{y}_j^i\}_{i=1}^N$, where $\mathbf{x}_j^i$ represents the j^{th} data sample and the superscript ^{i} indexes subject. To simplify the notation, we omit the subscript j when the contextual information is clear enough. The specifc and shared encoder, reconstructer, and decoder are all MLP-like networks.

Training Phase

Semantic Prediction As depicted in Figure 2 (a), N specific encoders and one shared encoder are trained on each subject's data respectively. The intuition behind the design of the specific and shared encoders is clear, we want the specific encoders to focus on the subject-related information and the shared encoders to focus on the subject-invariant information. To achieve this, we introduce three losses here: the domain classification loss, domain alignment loss, and difference loss. The process starts with the shared and specific branches running in parallel, with

$$\mathbf{r}^i = \mathbf{E}_r(\mathbf{x}^i), \text{ and } \mathbf{s}^i = \mathbf{E}_s(\mathbf{x}^i), \quad i \in \{1, \ldots, N\}. \tag{1}$$

Then the specific feature and the shared feature will be fused and fed into the universal decoder to calculate the semantic predictions (text CLIP embeddings):

$$\hat{\mathbf{e}}^i = \mathbf{D}(\text{ResFuse}(\mathbf{s}^i, \mathbf{r}^i)), \tag{2}$$

where ResFuse stands for the residual fusion module: the specific and shared feature are first concatenated as the input of a projection layer f, whose output is added as a residual to the shared features to form the semantically rich brain embedding, as follows:

$$\text{ResFuse}(\mathbf{s}^i, \mathbf{r}^i) = f(\text{concat}(\mathbf{s}^i, \mathbf{r}^i)) + \mathbf{r}^i. \tag{3}$$

Reconstruction Loss To force each encoder to extract the features that contain sufficient information from the original brain data, we introduce an auto-encoder architecture. That is, there are some reconstructers to reconstruct the original brain data from extracted features:

$$\hat{\mathbf{x}}_r^i = \mathbf{R}_r(\mathbf{r}^i), \text{ and } \hat{\mathbf{x}}_s^i = \mathbf{R}_s(\mathbf{s}^i), \tag{4}$$

we adopt the mean squared error (MSE) to calculate the reconstruction loss $\mathcal{L}_{rec}$:

$$\mathcal{L}_{rec} = \mathcal{L}_{rec}^s + \mathcal{L}_{rec}^r = \frac{1}{N}\sum_i^N \left(\frac{1}{m}\|\hat{\mathbf{x}}_s^i - x^i\|_2^2 + \frac{1}{m}\|\hat{\mathbf{x}}_r^i - x^i\|_2^2\right), \tag{5}$$

where m is the length of the original brain data and $\|\|_2$ is the squared L_2-norm.

Domain Classification Loss A domain classifier $\mathbf{C}_{dc}$ is applied to predict which subject the specific feature was extracted from: $\hat{\mathbf{y}}_{dc} = \mathbf{C}_{dc}(\mathbf{s}^i)$, the domain classification loss is calculated using cross-entropy loss:

$$\mathcal{L}_c = -\sum_{i=1}^{N}\sum_{l=1}^{N} \mathbf{y}_{dc}^i \log \hat{\mathbf{y}}_{dc}^i, \tag{6}$$

where $\mathbf{y}_{dc}$ is the domain label, i.e., an one-hot vector indicate the subject index.

Domain Alignment Loss In contrast to the specific encoder, the shared encoder is expected to extract the subject-invariant information. Hence, domain labels are supposed to be removed from shared features $\mathbf{r}^i$, $i \in \{1, \ldots, N\}$. That is, in an ideal situation, it is impossible to distinguish which subject the shared feature was extracted from. One way to achieve this is to confuse the domain classifier by minimizing the cross-entropy loss via a gradient reversal layer (GRL) (Ganin and Lempitsky 2015):

$$\mathcal{L}_{da} = \sum_{i=1}^{N}\sum_{l=1}^{N} \mathbf{y}_{da}^i \log \hat{\mathbf{y}}_{da}^i, \tag{7}$$

where $\mathbf{y}_{da}$ is the domain label, and $\hat{\mathbf{y}}_{da} = \mathbf{C}_{da}(\mathbf{r}^i)$, $\mathbf{C}_{da}$ is another domain classifier for domain alignment. The GRL acts as identity transform for forwardprop, and reverses the gradient for backprop.

Difference Loss Through the shared-specific design, a subject's brain data is processed by two encoders. There will be a waste of computation if both of the specific and the shared encoders extracted very similar information from one brain data. Thus, we adopt the Hadamard product $\odot$ to enhance the orthogonality and simplify the difference loss as follows:

$$\mathcal{L}_{diff} = \frac{1}{N}\sum_{i}^{N} \|\mathbf{s}^i \odot \mathbf{r}^i\|_2^2. \tag{8}$$

Alignment Loss MindCross adopts the contrastive loss and MSE loss to align the semantic predictions with text CLIP embeddings. In order to focus the research on cross-subject brain decoding, we only discuss the simplified situation of only predicting the text CLIP embedding rather than both frame latents and text embeddings. Although previous works have shown that the predicted frame latents help to improve low-level metrics (Liu et al. 2024b; Gong et al. 2024; Lu et al. 2025), overall is not critical in our study and we omit this part for a more focused paper (since this part is just adding an additional decoder to predict frame latents from the same feature used to predict the text CLIP embeddings).

For contrastive loss, MindCross adopt the SoftCLIP loss which was introducted in MindEye (Scotti et al. 2023):

$$\mathcal{L}_{\text{SoftCLIP}}(\mathbf{e}, \hat{\mathbf{e}}) = -\sum\nolimits_{k=1}^{B}\sum\nolimits_{l=1}^{B} \left[\frac{\exp(\mathbf{e}_k \cdot \mathbf{e}_l)/\tau}{\sum_{m=1}^{B}\exp(\mathbf{e}_k \cdot \mathbf{e}_m/\tau)} \cdot \log\left(\frac{\exp(\hat{\mathbf{e}}_k \cdot \mathbf{e}_l)/\tau}{\sum_{m=1}^{B}\exp(\hat{\mathbf{e}}_k \cdot \mathbf{e}_m/\tau)}\right)\right], \tag{9}$$

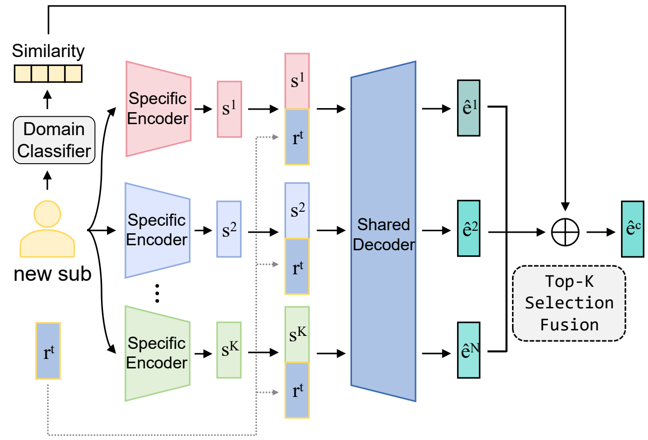

Figure 3: Top-K Collaborate Module: The similarity vector is obtained by feeding the new subject's feature s^t into the domain classifier, then we select the Top-K similar domains for collaborating to calculate the final output.

where $\mathbf{e}$ and $\hat{\mathbf{e}}$ are the ground truth text embedding and the semantic predictions in a batch of size B. τ is a temperature hyperparameter. Besides the contrastive loss for better clustering of different classes, MindCross adds an MSE loss to improve alignment and guarantee the generative quality of the video:

$$\mathcal{L}_{align} = \mathcal{L}_{\text{SoftCLIP}}(\mathbf{e}, \hat{\mathbf{e}}) + \frac{1}{N}\sum_{i}^{N} \|\mathbf{e}^i - \hat{\mathbf{e}}^i\|_2^2. \tag{10}$$

Finally, MindCross is trained end-to-end by incorporating all these losses to achieve cross-subject brain decoding. These Greek letters are all hyperparameters for balancing each loss function:

$$\mathcal{L}_{train} = \mathcal{L}_{align} + \alpha\mathcal{L}_{rec} + \beta\mathcal{L}_c + \gamma\mathcal{L}_{da} + \zeta\mathcal{L}_{diff}. \tag{11}$$

Calibration Phase

If there comes a new subject, the traditional per-subject per-model paradigm needs to train a new brain decoding model from scratch, which is time and computation consuming and requires massive data from the new subject. Our MindCross propose a calibration way to fully utilize the universal model trained on previous subjects' brain data to achieve fast new subject adaptation. Specifically, we freeze the previously trained MindCross and only update the specific encoder and reconstructer of the new subject. By freezing most of the parameters, we save lots of time and computation resources. Most importantly, it is possible to achieve comparative decoding results with limited data, as the shared encoder and decoder are pre-trained on a large amount of subjects' brain data. The calibration loss is written as follows, each loss function is defined in the previous section:

$$\mathcal{L}_{calib} = \mathcal{L}_{align} + \alpha'\mathcal{L}_{rec}^t + \beta'\mathcal{L}_{diff}. \tag{12}$$

Test Phase

In the test phase, MindCross predict the text CLIP embeddings by two branches: one uses the specific encoder of the new subject $\mathbf{E}_t$, and another one utilizes the trained specific encoders $\{\mathbf{E}_s^i\}_{i=1}^N$ through a **Top-*K* Collaboration Module** as depicted in Figure 3. Firstly, the specific feature of the new subject $\mathbf{s}^t$ is fed into the domain classifier $\mathbf{C}_{dc}$ to calculate the similarity p between the brain data of new subject and previous subjects. We select the Top-K subjects' specific encoders and predict the text CLIP embeddings. Afterwards, the K predictions are added together according to the similarity weights. The higher weight indicates that the distribution is more similar to the new subject, so more trust can be given to the corresponding prediction.

$$\hat{\mathbf{e}}^c = \sum_{k \in \mathrm{TopK}(p)} p_k \cdot \hat{\mathbf{e}}^k, \tag{13}$$

where TopK selects the top K index in a logits vector p. Finally, the semantic prediction of the new subject is combining $\hat{\mathbf{e}}^c$ and $\hat{\mathbf{e}}^t$ as demonstrated in Figure 2, $\lambda = 1e-2$:

$$\hat{\mathbf{e}} = \hat{\mathbf{e}}^t + \lambda \hat{\mathbf{e}}^c. \tag{14}$$

Video Generation Module

The video generation model plays a crucial roles in all brain decoding works as pre-training text-to-video (T2V) diffusion models possess a large amount of prior knowledge from the graphics, image, and video domains. However, as demonstrated before, our research focus is not on taming the cut-the-edge T2V models for generating smooth and high-fidelity videos, but on how to learn cross-subject brain representations. Therefore, we adopted a T2V models called PyramidFlow (Jin et al. 2025) for video generation without further modification. Specifically, the videos are generated using the semantic prediction $\hat{\mathbf{e}}$ as the condition.

Experiments

Datasets

In this study, we utilize two publicly available brain-video datasets, which includes paired stimulus videos and their corresponding brain responses. More details of each dataset are written in appendix.

For EEG-to-video reconstruction, we use the SEED-DV dataset (Liu et al. 2024b). This dataset contains 20 subjects' EEG signals while they were watching 7 video blocks containing 1400 two-second video clips of 40 concepts, covering various animals, scenes, and activities. We use the first 6 blocks (1200 trials) as the training set and the last block (200 EEG-video trials) as the test set.

For fMRI-to-video reconstruction, we use the CC2017 dataset (Wen et al. 2018) recording 3 subjects' fMRI data using a 3-T MRI system with two-second temporal resolution. The dataset consists of a training set containing 18 8-minute video clips and a test set containing five 8-minute video clips. Each subject viewed the training and testing video clips 2 and 10 times, respectively, and the test set was averaged across trials. We divide them into two-second small clips. Thus there are 8640 training samples and 1200 testing samples of fMRI-video pairs.

Figure 4: Comparison on EEG-to-video benchmark.

Figure 5: Comparison on fMRI-to-video benchmark.

Evaluation Metrics

Following previous work (Gong et al. 2024), we perform quantitative evaluation using both frame-based and video-based metrics. Frame-based metrics evaluate each frame individually, providing a snapshot evaluation, while video-based metrics evaluate the quality of the video, emphasizing the consistency and smoothness. For semantic evaluation, we use the N-way top-K metric and set K to 1, which means a video is considered successfully reconstructed if the ground truth (GT) class is in the top 1 probabilities using a pretrained classifier. For frame-based metric, the classifier is a CLIP-based classifier (Radford et al. 2021) trained on ImageNet (Deng et al. 2009). For video-based metric, the classifier is a VideoMAE-based (Tong et al. 2022) video classifier trained on Kinetics-400 dataset (Kay et al. 2017). For spatiotemporal-level metrics to measure video consistency, we compute the average cosine similarity between all pairs of adjacent video frames' CLIP image embeddings, CLIP-pcc. The structural similarity index metric (SSIM) and peak signal-to-noise ratio (PSNR) are used as pixel-level metrics. More details of all metrics are written in appendix.

Cross-Subject Video Reconstruction

Our MindCross can achieve cross-subject brain decoding across several subjects with only one model, while most

	Method	Venue	# Models	Video-based			Frame-based			
				Semantic-Level		ST-Level	Semantic-Level		Pixel-Level	
				2-way	40/50-way	CLIP-pcc	2-way	40/50-way	SSIM	PSNR
SEED-DV	MinD-Video	NeurIPS 23	20	0.805±0.02	0.156±0.03	0.411±0.47	0.755±0.03	0.128±0.03	0.176±0.06	8.579±1.42
	NeuroClips	NeurIPS 24	20	0.809±0.03	0.154±0.03	0.756±0.28	0.785±0.03	0.151±0.04	0.238±0.08	8.703±1.37
	Mind-Animator	ICLR 25	20	0.799±0.03	0.158±0.02	0.421±0.56	0.768±0.02	0.142±0.04	0.253±0.12	8.679±1.52
	EEG2Video	NeurIPS 24	20	0.800±0.03	0.161±0.01	0.412±0.45	0.772±0.03	0.146±0.01	0.258±0.08	8.684±1.46
	GLFA	ECCV 24	1	0.778±0.02	0.152±0.02	0.751±0.52	0.743±0.03	**0.136**±0.03	0.192±0.08	8.642±1.48
	MindBridge	CVPR 24	1	0.782±0.03	0.148±0.02	0.753±0.43	0.749±0.03	0.125±0.02	0.185±0.07	8.625±1.39
	MindCross	Ours	1	**0.786**±0.02	**0.154**±0.03	**0.758**±0.32	**0.752**±0.03	0.128±0.02	**0.197**±0.06	**8.658**±1.43
CC2017	MinD-Video	NeurIPS 23	3	0.839±0.03	0.197±0.02	0.408±0.46	0.796±0.03	0.174±0.03	0.171±0.08	8.662±1.52
	NeuroClips	NeurIPS 24	3	0.834±0.03	0.220±0.01	0.738±0.17	0.806±0.03	0.203±0.01	0.390±0.08	9.211±1.46
	Mind-Animator	ICLR 25	3	0.830±0.02	0.185±0.04	0.425±0.52	0.816±0.03	0.182±0.03	0.321±0.12	9.220±1.48
	EEG2Video	NeurIPS 24	3	0.833±0.03	0.209±0.02	0.413±0.37	0.811±0.04	0.191±0.03	0.318±0.14	8.763±1.45
	GLFA	ECCV 24	1	**0.840**±0.03	0.209±0.02	0.742±0.49	0.805±0.04	0.187±0.03	0.134±0.07	8.763±1.45
	MindBridge	CVPR 24	1	0.821±0.03	0.206±0.02	0.754±0.42	**0.813**±0.03	0.195±0.02	0.156±0.07	8.972±1.67
	MindCross	Ours	1	0.830±0.03	**0.210**±0.03	**0.762**±0.34	0.802±0.03	**0.198**±0.03	**0.163**±0.07	**8.983**±1.54

Table 1: **Quantitative comparison of brain decoding between MindCross and other methods.** The best performance of cross-subject frameworks are marked in BOLD. The best performance of per-subject per-model methods are underlined.

previous methods require to train a particular model for each subject. We compare its average video reconstruction performance across all subjects with that of state-of-the-art subject-dependent methods: MinD-Video (Chen, Qing, and Zhou 2023), NeuroClips (Gong et al. 2024), Mind-Animator (Lu et al. 2025), EEG2Video (Liu et al. 2024b), and two cross-subject methods: MindBridge (Wang et al. 2024), and GLFA (Li et al. 2024a). The quantitative and qualitative results for all methods are presented in Table 1, and Figure 4, and Figure 5 respectively. It can be seen that our MindCross achieve comparable performance against per-subject per-model methods while maintaining just one model in semantic level metrics, and outperforms other cross-subject methods, demonstrating our success on cross-subject brain decoding. Figure 4 demonstrates our MindCross achieves more accurate in semantic decoding across different subjects, e.g., MindBridge decoded bird as plane, and GLFA decoded beach as ship.

New-Subject Adaptation

MindCross can effectively transfer its pretrained knowledge to adapt to new subjects, offering significant advantages in real-world applications. To simulate the case when the new subject only has limited data, we select 40, 200, and 600 samples from the new subjects for the EEG dataset and 500, 1500, and 4000 samples from the new subjects for the fMRI dataset. We perform a leave-one-subject-out (LOSO) experiment here. The results are displayed in Figure 6, where MindCross performs comparable on the new subject to the existing subjects. MindCross significantly outperforms the variant 'w/o train' (training MindCross on new subject from scratch), which indicates the efficacy of the training phase. We also evaluate the performance of calibration with only $\mathcal{L}_{align}$. It can be observed that the difference loss and reconstruction loss improve the decoding results. Furthermore, the comparison with baselines is displayed in Table 2, we calibrated on the new subject using 200/500 data on EEG/fMRI datasets, respectively. It can be seen that our MindCross significantly reduces the adapation time and updating parameter (especially when the number of subjects is large), while outperforming or tying the baselines, demonstrating the superiority of MindCross on new subject adaptation.

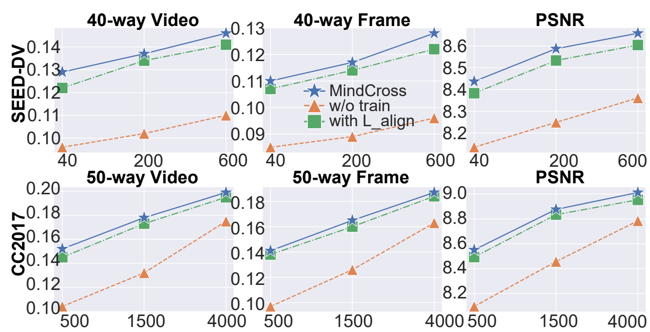

Figure 6: Results of new subject adaptation in limited data scenario. The x-axis is the number of data from new subjects, the y-axis is the metric.

	Training Loss	40-V	40-F	PSNR	Time/sec	# Para
EEG	MindBridge	**0.142**	0.104	8.514	5.104	126.81M
	GLFA	0.135	**0.121**	8.522	10.651	247.27M
	MindCross	0.137	0.117	**8.587**	**1.090**	**9.77M**
fMRI	MindBridge	0.147	**0.145**	8.547	9.267	293.43M
	GLFA	**0.153**	0.137	**8.601**	17.513	504.23M
	MindCross	0.152	0.141	8.551	**2.724**	**109.10M**

Table 2: Comparison with other cross-subject baselines on new subject adaptation with limited data task.

Ablation Study

Ablation on Training Loss Table 3 displays the results of introducing different losses in the training phase. It can be seen that MindCross with $\mathcal{L}_{da} + \mathcal{L}_{dc} + \mathcal{L}_{rec}$ outperforms MindCross with only $\mathcal{L}_{align}$, demonstrating the shared-specific architecture helps model learn cross-subject representations. Adding $\mathcal{L}_{diff}$ does not affect the decoding performance much, but is useful for new subject adaptation in the calibration phase. Thus, we add the difference loss in the training phase to encourage the shared-specific encoder focus on the different aspects of brain data.

	Training Loss	2-way-V	2-way-F	SSIM	PSNR
EEG	$\mathcal{L}_{align}$	0.756	0.693	0.187	8.627
	$+\mathcal{L}_{rec} + \mathcal{L}_{da} + \mathcal{L}_{dc}$	**0.789**	**0.757**	0.195	8.647
	$+\mathcal{L}_{diff}$ (Ours)	0.786	0.752	**0.197**	**8.658**
fMRI	$\mathcal{L}_{align}$	0.812	0.785	0.152	8.576
	$+\mathcal{L}_{rec} + \mathcal{L}_{da} + \mathcal{L}_{dc}$	0.829	**0.810**	**0.168**	8.957
	$+\mathcal{L}_{diff}$ (Ours)	**0.830**	0.802	0.163	**8.983**

Table 3: Ablation of different losses at the training phase. Models are trained and tested across all subjects.

Ablation on Top-K Collaboration Module In this case, we use the $\hat{\mathbf{e}}^t$ to generate videos as the model 'w/o Top-K'. The results are depicted in Figure 7, which demonstrates the efficacy of the Top-K Collaboration Module. There is no significant difference between $K = 1$ and $K = 2$. Thus, we set $K = 1$ in this paper.

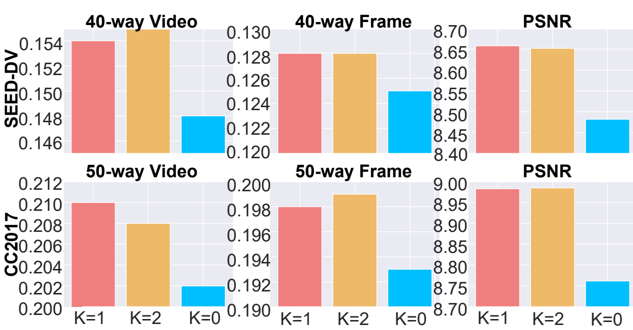

Figure 7: Ablation study of Top-K Collaboration Module.

Visualization

Top-K Subject Selection To demonstrate the effectiveness of the Top-K collaboration module, We enumerate the new subject in the SEED-DV dataset and train a MindCross model on the other 19 subjects. We visualize the subject selection probability of each subject in Figure 8, where $K = 1$ in this case. It can be seen that sub 10 and sub 14 will cooperate with each other with a high probability, which indicates it is useful to introduce a mechanism like memory retrieval process to recall the existing subjects' similar data. Although some subjects like sub 4 and sub 19 have low similarity with all subjects, the Top-K module still improves the performance as shown in Figure 3.

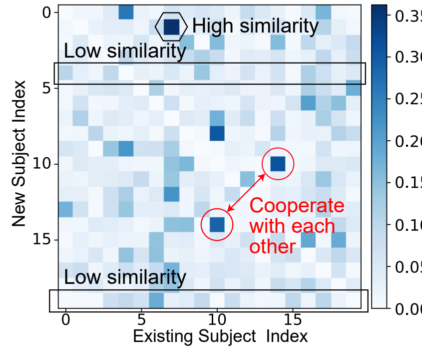

Figure 8: The heat map of the subject selection probability of the Top-K Collaboration Module.

Shared and Specific Features Our MindCross is supposed to separate the subject-related and subject-invariant components out from the original brain signals. Here we trained our model on 19 subjects from the SEED-DV dataset and obtained their original data $\mathbf{x}^i$ along with the specific feature $\mathbf{s}^i$ and shared feature $\mathbf{r}^i$ learned by our MindCross. The 19 subject's original brain data are displayed in Figure 9 (a), where their brain data differ significantly from each other. After the process of MindCross, the subject-related and subject-invariant components are successfully extracted, as shown in Figure 9 (b). Compared to other previous cross-subject methods which only extract the subject-invariant information, our MindCross keeps the subject-related information. Therefore, we can calculate the similarity between new subject and existing subjects to enhance the brain decoding for new subject.

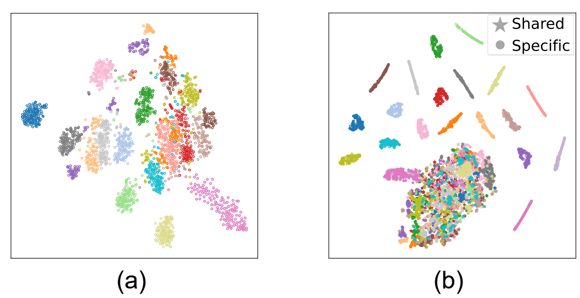

Figure 9: Feature visualization of MindCross using t-SNE. (a) Original brain data. The color represents different subjects. (b) Learned features of MindCross.

Conclusion

In this paper, we propose a cross-subject brain decoding framework **MindCross**, which can effectively extract both subject-invariant and subject-related information. Through a novel calibration phase and collaboration module, MindCross significantly reduces the adaptation time and updating parameter, demonstrating great superiority and efficiency in new subject adaptation with limited data. Our work not only enhances cross-subject brain decoding, but also showcases promising practical applications in the BCI field.

Acknowledgements

We have published an extended version on arxiv, please kindly refer to https://arxiv.org/abs/2511.14196 for more results and analysis. This work was supported in part by grants from Brain Science and Brain-like Intelligence Technology-National Science and Technology Major Project (2025ZD0218900), National Natural Science Foundation of China (62376158), STI 2030-Major Projects+2022ZD0208500, Medical-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University "Jiao Tong Star" Program (YG2023ZD25, YG2024ZD25 and YG2024QNA03), Shanghai Jiao Tong University 2030 Initiative, the Lingang Laboratory (Grant No. LGL-1987), GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine, and Shanghai Jiao Tong University SCS-Shanghai Emotionhelper Technology Co., Ltd Joint Laboratory of Affective Brain-Computer Interfaces.

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.

Bao, G.; Zhang, Q.; Gong, Z.; Zhou, J.; Fan, W.; Yi, K.; Naseem, U.; Hu, L.; and Miao, D. 2025. Wills Aligner: Multi-Subject Collaborative Brain Visual Decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14194–14202.

Benchetrit, Y.; Banville, H.; and King, J.-R. 2024. Brain decoding: toward real-time reconstruction of visual perception. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Chen, J.; Qi, Y.; Wang, Y.; and Pan, G. 2024a. Bridging the semantic latent space between brain and machine: Similarity is all you need. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 11302–11310.

Chen, J.; Qi, Y.; Wang, Y.; and Pan, G. 2024b. Mind artist: Creating artistic snapshots with human thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27207–27217.

Chen, J.; Qi, Y.; Wang, Y.; and Pan, G. 2025. Mindgpt: Interpreting what you see with non-invasive brain recordings. *IEEE Transactions on Image Processing*.

Chen, Z.; Qing, J.; and Zhou, J. H. 2023. Cinematic Mindscapes: High-quality Video Reconstruction from Brain Activity. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dai, Y.; Yao, Z.; Song, C.; Zheng, Q.; Mai, W.; Peng, K.; Lu, S.; Ouyang, W.; Yang, J.; and Wu, J. 2025. MindAligner: Explicit Brain Functional Alignment for Cross-Subject Visual Decoding from Limited fMRI Data. In *International Conference on Machine Learning (ICML)*. PMLR.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, 1180–1189. PMLR.

Gong, Z.; Bao, G.; Zhang, Q.; Wan, Z.; Miao, D.; Wang, S.; Zhu, L.; Wang, C.; Xu, R.; Hu, L.; Liu, K.; and Zhang, Y. 2024. NeuroClips: Towards High-fidelity and Smooth fMRI-to-Video Reconstruction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Gong, Z.; Zhang, Q.; Bao, G.; Zhu, L.; Xu, R.; Liu, K.; Hu, L.; and Miao, D. 2025. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14247–14255.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *The Twelfth International Conference on Learning Representations*.

Jin, Y.; Sun, Z.; Li, N.; Xu, K.; Xu, K.; Jiang, H.; Zhuang, N.; Huang, Q.; Song, Y.; MU, Y.; and Lin, Z. 2025. Pyramidal Flow Matching for Efficient Video Generative Modeling. In *The Thirteenth International Conference on Learning Representations*.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Li, C.; Qian, X.; Wang, Y.; Huo, J.; Xue, X.; Fu, Y.; and Feng, J. 2024a. Enhancing cross-subject fmri-to-video decoding with global-local functional alignment. In *European Conference on Computer Vision*, 353–369. Springer.

Li, D.; Wei, C.; Li, S.; Zou, J.; and Liu, Q. 2024b. Visual Decoding and Reconstruction via EEG Embeddings with Guided Diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Liu, X.-H.; Jiang, W.-B.; Zheng, W.-L.; and Lu, B.-L. 2024a. Moge: Mixture of graph experts for cross-subject emotion recognition via decomposing eeg. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3515–3520. IEEE.

Liu, X.-H.; Liu, Y.-K.; Wang, Y.; Ren, K.; Shi, H.; Wang, Z.; Li, D.; Lu, B.-L.; and Zheng, W.-L. 2024b. EEG2Video: Towards Decoding Dynamic Visual Perception from EEG Signals. In *Advances in Neural Information Processing Systems*, volume 37, 72245–72273.

Liu, X.-H.; Lu, B.-L.; and Zheng, W.-L. 2025a. EEGmirror: leveraging EEG data in the wild via montage-agnostic self-supervision for EEG to video decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Liu, X.-H.; Lu, B.-L.; and Zheng, W.-L. 2025b. mixEEG: Enhancing EEG Federated Learning for Cross-subject EEG Classification with Tailored mixup. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Liu, Y.; Ma, Y.; Zhu, G.; Jing, H.; and Zheng, N. 2025. See Through Their Minds: Learning Transferable Brain Decoding Models from Cross-Subject fMRI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5730–5738.

Lu, Y.; Du, C.; Wang, C.; Zhu, X.; Jiang, L.; Li, X.; and He, H. 2025. Animate Your Thoughts: Reconstruction of Dynamic Natural Vision from Human Brain Activity. In *The Thirteenth International Conference on Learning Representations*.

Makeig, S.; Westerfield, M.; Jung, T.-P.; Enghoff, S.; Townsend, J.; Courchesne, E.; and Sejnowski, T. J. 2002. Dynamic brain sources of visual evoked responses. *Science*, 295(5555): 690–694.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Miyawaki, Y.; Uchida, H.; Yamashita, O.; Sato, M.-a.; Morito, Y.; Tanabe, H. C.; Sadato, N.; and Kamitani, Y. 2008. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5): 915–929.

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 (ICML)*, 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 (CVPR)*, 10684–10695.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C. W.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S. R.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Scotti, P.; Banerjee, A.; Goode, J.; Shabalin, S.; Nguyen, A.; Dempster, A.; Verlinde, N.; Yundler, E.; Weisberg, D.; Norman, K.; et al. 2023. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36: 24705–24728.

Scotti, P. S.; Tripathy, M.; Torrico, C.; Kneeland, R.; Chen, T.; Narang, A.; Santhirasegaran, C.; Xu, J.; Naselaris, T.; Norman, K. A.; and Abraham, T. M. 2024. MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data. In *International Conference on Machine Learning (ICML)*.

Sun, J.; Li, M.; and Moens, M.-F. 2025. Neuralflix: A simple while effective framework for semantic decoding of videos from non-invasive brain recordings. In *Annual AAAI Conference on Artificial Intelligence*.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 10078–10093.

Wang, H.; Chen, Y.; Ma, C.; Avery, J.; Hull, L.; and Carneiro, G. 2023. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15878–15887.

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.

Wen, H.; Shi, J.; Zhang, Y.; Lu, K.-H.; Cao, J.; and Liu, Z. 2018. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, 28(12): 4136–4160.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7623–7633.

Yu, M.; Lin, S.; Yan, H.; and Ma, K. 2025. MindPainter: Efficient Brain-Conditioned Painting of Natural Images via Cross-Modal Self-Supervised Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14468–14476.