Enhancing Cross-Subject fMRI-to-Video Decoding with Global-Local Functional Alignment

Chong Li[⋆†], Xuelin Qian[⋆§], Yun Wang[⋆], Jingyang Huo, Xiangyang Xue[†], Yanwei Fu[†], and Jianfeng Feng

Fudan University, Shanghai, China
{lichong23,jyhuo22}@m.fudan.edu.cn,
{xlqian,19110850009,xyxue,yanweifu,jffeng}@fudan.edu.cn

Abstract. Advancements in brain imaging enable the decoding of thoughts and intentions from neural activities. However, the fMRI-to-video decoding of brain signals across multiple subjects encounters challenges arising from structural and coding disparities among individual brains, further compounded by the scarcity of paired fMRI-stimulus data. Addressing this issue, this paper introduces the fMRI Global-Local Functional Alignment (GLFA) projection, a novel approach that aligns fMRI frames from diverse subjects into a unified brain space, thereby enhancing cross-subject decoding. Additionally, we present a meticulously curated fMRI-video paired dataset comprising a total of 75k fMRI-stimulus paired samples from 8 individuals. This dataset is approximately 4.5 times larger than the previous benchmark dataset. Building on this, we augment a transformer-based fMRI encoder with a diffusion video generator, delving into the realm of cross-subject fMRI-based video reconstruction. This innovative methodology faithfully captures semantic information from diverse brain signals, resulting in the generation of vivid videos and achieving an impressive average accuracy of 84.7% in cross-subject semantic classification tasks.

1 Introduction

Deciphering thoughts requires a universal method for decoding brain activity across individuals, with a key challenge being aligning brain signals to preserve shared information [2, 17, 29]. Neuroimaging and deep learning advances focus on understanding cognitive processes, where functional MRI (fMRI) plays a central role in non-invasive whole-brain activity recording and decoding [33]. Recent efforts have decoded fMRI data into text [22], images [5,21,22,24,25] and 3D objects [10]. However, this paper focuses on decoding dynamic videos from scanned fMRI, reflecting continuous visual experiences of individuals, a task pioneered

⋆: Equal contributions.
§: Dr. Xuelin Qian is now with Northwestern Polytechnical University.
†: Corresponding authors.

in [6]. This task poses significant challenges yet holds promise for practical applications.

However, numerous challenges exist in this task. Firstly, newly obtained fMRI data from unknown subjects may significantly differ from existing datasets due to variations in anatomical structure and functional topography among individuals, complicating practical use [7, 11, 39]. This limitation means that fMRI-to-video decoding is often limited to individual-specific approaches, as in [6], which is less than ideal. Secondly, the high costs associated with acquiring fMRI scans limit the availability of data. This poses a significant challenge, especially in fMRI-to-video decoding, where capturing dynamic changes in human brain activity requires frequent snapshots over time.

One potential way to addressing the first challenge involves aligning both anatomical and functional aspects of subjects' neural activities [9, 17]. Anatomical alignment transforms fMRI data into a standard brain template based on structural MRI images [4]. Functional alignment [2, 7] aims to connect neural activities across subjects. These methods reduce variability in fMRI data and enhance cross-subject brain decoding accuracy. However, recent computer vision works focus on within-subject brain decoding, translating brain activity into various formats [5, 6, 10, 21, 22, 22, 24, 25]. Despite promising results, reliance on subject-specific voxel subsets in regions of interest (ROI) still limits cross-subject decoding due to voxel variability. Integrating alignment methods is crucial to improve cross-subject brain decoding capabilities. For example, Qian et al. [26] aligned fMRI to 32k_fs_LR surface space [13], revealing suboptimal cross-subject performance compared to a previous within-subject decoding method [5]. MindTuner [14] integrated alignment methods as preprocessing to achieve cross-subject decoding. Additionally, MindBridge [37] employed deep learning for functional alignment method.

Unfortunately, achieving precise inter-subject functional correspondence based solely on anatomical features remains limited [7]. To address this limitation and enhance cross-subject decoding, we introduce Global-Local Functional Alignment (GLFA). GLFA addresses functional alignment at both local and global levels by learning an fMRI transformation to a unified functional brain space through global and local alignment. This integration is facilitated into the training process using an alignment loss function, enabling easy adaptation to any fMRI decoding model based on gradient descent learning. Further, we integrate GLFA into the fMRI Pretrained Transformer Encoder (fMRI-PTE) [26] to enhance its cross-subject decoding capabilities. Additionally, we augment the model with a spatiotemporal attention module and video stable diffusion pipeline, resulting in the development of a high-quality cross-subject fMRI-to-video decoding pipeline for the first time.

Furthermore, we provide a significant contribution to the fMRI-to-video decoding task, complementing previous data in [6]. We curated an fMRI-video paired dataset from 8 subjects, comprising 75k fMRI-stimulus samples. This dataset is approximately 4.5 times larger than the previous benchmark dataset [40], mitigating the scarcity of fMRI-stimulus paired data. With increased sub-

jects and samples, our dataset serves as a crucial foundation for cross-subject decoding from fMRI to video.

In our experiment, we evaluated the impact of functional alignment on enhancing cross-subject decoding. Applying functional alignment to an unknown subject resulted in an average 35.8% improvement in fMRI spatial correlation with other subjects, and a 17.1% increase in inter-subject correlation. This underscores GLFA's effectiveness in extracting cross-subject correlations. Additionally, we used GLFA for cross-subject fMRI-to-video decoding, achieving high-quality, high-frame-rate video reconstruction from fMRI data. Comparing with state-of-the-art baselines, GLFA demonstrated an average accuracy of 84.7% in semantic classification tasks for cross-subject video reconstruction. Moreover, within-subject decoding showed comparable performance with state-of-the-art methods.

Finally, we would like to highlight some interesting and special points of this paper here: (1) **Focusing on Dynamic Video Decoding**: The paper concentrates on decoding dynamic videos from scanned fMRI, known as MinD-Video [6]. It aligns both anatomical and functional aspects of subjects' neural activities to enhance cross-subject brain decoding accuracy by reducing variability. (2) **Proposed GLFA**: To improve cross-subject decoding, the paper introduces GLFA, which integrates functional alignment into the training process to enhance fMRI decoding models using gradient descent learning, mitigating cross-subject variance. (3) **Integration into fMRI-PTE**: GLFA is added to fMRI-PTE [26] to enhance cross-subject decoding, along with a spatiotemporal attention module and video stable diffusion pipeline. This achieves high-quality cross-subject fMRI-to-video generation, with an average accuracy of 84.7% in cross-subject semantic classification tasks. (4) **Contributions to Data and Experimentation**: The paper provides a large dataset for fMRI-to-video decoding, around 4.5 times larger than previous benchmarks [40], alleviating the shortage of fMRI-stimulus paired data. GLFA experiments demonstrated notable enhancements in fMRI spatial and inter-subject correlation, as well as high-quality video reconstruction and semantic classification tasks for cross-subject video reconstruction.

2 Related Work

2.1 Functional Alignment

Two primary types of functional alignment have been extensively studied, with each focusing on alignment either in anatomical brain space [2, 7, 20, 32] or a common space with high dimensions [11, 16]. For instance, Bazeille et al. [2, 3] divided brain into parcels and applied optimal transport theory to construct a whole-brain template. Conroy et al. [7] proposed pairwise cortical alignment to maximize the functional similarity of inter-subject patterns. Haxby et al. [16] introduced a hyperalignment method to align response-patterns across subjects into a high-dimensional model space. Gao et al. [11] proposed the Deep Cross-subject Adaptation Decoding (DCAD) framework for unsupervised deciphering

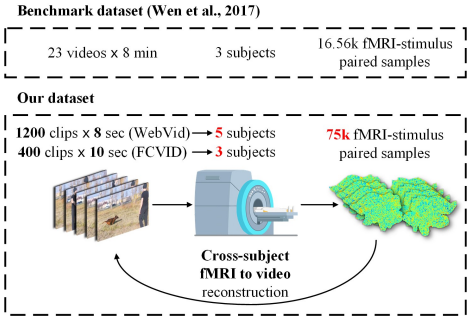

Fig. 1: Comparison between the benchmark dataset [40] and our dataset. Our dataset undergoes fMRI collection with multiple subjects under two distinct video datasets. This facilitates not only cross-subject decoding but also cross-dataset decoding. Additionally, in comparison to the previous benchmark dataset [40], ours encompasses a larger number of subjects and approximately 4.5 times more fMRI stimulus samples. This enhancement better caters to the task of cross-subject fMRI video decoding.

the brain states. In this paper, we amalgamate the principles of these functional alignments and propose the global-local functional alignment, capable of simultaneously aligning fMRI in anatomical and high-dimensional spaces.

2.2 fMRI-to-Video Reconstruction

Humans continually receive visual stimulation from the external environment, and the brain consistently encodes these inputs. Various studies have devoted to reconstruct visual experience from brain activity [6, 15, 40], but their generalization are limited. For example, Wen et al. [40] curated an fMRI-video paired dataset and utilized a CNN for video reconstruction. However, their approach could only capture the position and outline of objects from the original video without extracting semantic information. Furthermore, while evaluating cross-subject reconstruction, their selection of activated voxels restricted its generalization to individuals beyond the dataset. Han et al. [15] employed a pretrained VAE for video frame reconstruction by aligning fMRI to latent space. In contrast, Wang et al. [35] proposed conditional GAN to directly generate video from fMRI. Further, Chen et al. [6] introduced CLIP to train fMRI latents and translated them using a video stable diffusion pipeline. Despite achieving high performance, the reconstructions remain subject-specific. Practical "mind reading" applications demand further exploration of cross-subject methods.

3 Method

In our fMRI-to-video pipeline, all fMRI frames are transformed to a standard anatomical brain template [13] to enable cross-subject decoding. Consequently,

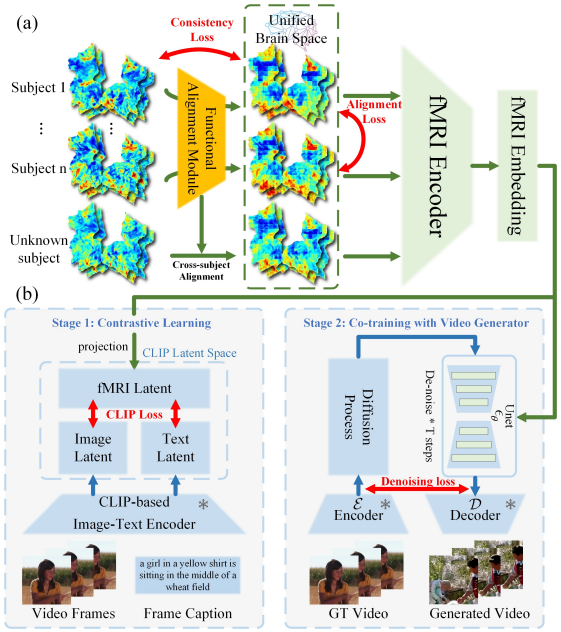

Fig. 2: Diagram of cross-subject fMRI-to-video pipeline. (a) The fMRI frames from different subjects are mapped to a unified brain space through Global-Local Functional Alignment (GLFA) Module. For unknown subjects, the learned GLFA Module can similarly map them to the unified brain space, reducing disparity between individuals. (b) Our method employs a two-stage approach to train an fMRI-to-video pipeline. In the first stage, we project fMRI embedding onto the CLIP latent space and calculate the CLIP loss by comparing them with paired image and text embedding. This allows us to capture rich semantic information in the CLIP space. In the second stage, we integrate a pretrained text-to-video pipeline and train it using fMRI embedding as a replacement for text embedding.

we opt for a large-scale fMRI Pretrained Transformer Encoder (fMRI-PTE) [26], pretrained on anatomically aligned fMRI. Building upon this, we introduce Global-Local Functional Alignment (GLFA), a shared-among-subjects fMRI projection aiming to capture the functional correspondence between voxels from different subjects, enhancing the model's cross-subject decoding capability. Additionally, we augment the fMRI encoder with a video diffusion model, and adopt a progressive learning procedure similar to the MinD-Video [6], comprising fMRI encoder pre-training, multi-modal contrastive learning, and co-training with video generator.

3.1 fMRI data acquisition

As shown in Figure 1, in order to address the challenge of limited data availability in fMRI-video reconstruction, we have meticulously curated an fMRI-

video paired dataset, incorporating data from 8 subjects and comprising 75k fMRI-stimulus samples. This dataset is approximately 4.5 times larger than the previous benchmark dataset [40].

In detail, stimuli videos of dimensions 256×256 and 596×336 are sourced from the FCVID video dataset [18] and WebVid [1] dataset respectively (Figure 1). Within the FCVID dataset, a selection of 100 video categories, encompassing diverse events, scenes, and objects (e.g., accordion performances, amusement parks, and elephants), is chosen. For each category, 4 10-second clips are meticulously selected, with 3 assigned to the training set and 1 to the test set. Random combinations of these clips result in 6 500-second videos for both training and test sets, ensuring each training clip appeared once and each test clip appeared three times.

For the widely used WebVid dataset in text-to-video synthesis, we curated a selection of 1200 8-second clips. Subsequently, 5 subjects are tasked with viewing these videos in random order, with 1000 clips designated for the training set and the remaining 200 clips allocated to the test set. Videos from test set are viewed twice to reduce noise, exclusively originate from the WebVid test set. Additionally, in between consecutive clips, a 2.4-second blank interval is introduced to allow the participants to clear their minds, thereby reducing the overlap of neural activity encoding information from different videos. In the experiment, 8 subjects (6 male and 2 female, aged 23-27, 3 for FCVID and 5 for WebVid) participated, and fMRI data are acquired using a 3T scanner and a 32-channel RF head coil, with the fMRI sampled at 1 frame per 0.8 seconds. Each participant in our fMRI scanning experiment provided informed written consent, and the experimental protocol received approval from the ethical review board.

3.2 The fMRI Pre-processing

In order to enhance the practical applicability of the model, we align fMRI data to the 32k_fs_LR brain surface space through anatomical structure [13]. In contrast to traditional fMRI decoding methods that flatten each frame of fMRI into a 1D signal and select subject-specific activated voxels, our approach transforms fMRI into a unified representation across subjects. We further apply voxel-wise z-transformation to the fMRI and unfold the surface, creating a 2D image that preserves spatial relationships between adjacent voxels. Additionally, given that only a portion of the brain area is activated in visual stimulation tasks [17], we select early and higher visual cortical Regions of Interest (ROIs). These ROIs, totaling 8,405 vertices, are defined by the HCP-MMP atlas [12] in the 32k_fs_LR space.

Finally, each fMRI is transformed to a 256×256 single-channel image. For the dataset with multiple runs for same video, the fMRI frames are averaged across aligned recordings. Due to the time lag between stimulus input and the BOLD signal reaching its peak caused by the natrue of hemodynamic response, a shift of around 6 seconds is applied to the fMRI data.

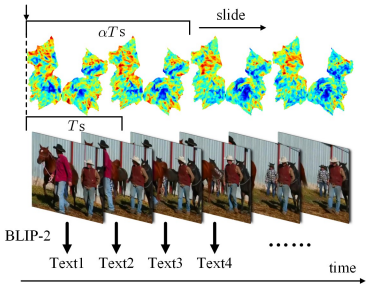

Fig. 3: Presentation of fMRI-video-text paired data sample.

3.3 Paired fMRI-Video-Text Sample

As depicted in Figure 3, we utilize a sliding window in the temporal dimension for aligned fMRI-video data, covering αT seconds of fMRI data and T seconds of video data. Consequently, the fMRI-to-video reconstruction task is reformulated as reconstructing T seconds of video from αT seconds of fMRI signals. Assuming the window comprises n frames of fMRI and m frames of video, for the i-th data sample, the fMRI data $\mathbf{F}_i \in \mathbb{R}^{n \times H_f \times W_f}$, with H_f and W_f denoting the height and width of the unfolded fMRI image, and the video data $\mathbf{V}_i \in \mathbb{R}^{m \times 3 \times H_v \times W_v}$, where H_v and W_v represent the resolution of the video. To facilitate multi-modal contrastive learning, we employ BLIP-2 [19] to describe each frame of the video, yielding m text representations.

3.4 fMRI-to-Video Pipeline

fMRI-PTE We apply the fMRI-PTE [26], featuring a ViT-based autoencoder structure to compress high-dimensional fMRI signals into a low-dimensional feature space, as our fMRI encoder. The model consists of two main parts: compression encoding and decoding reconstruction. During the encoding phase, the 2D fMRI image is partitioned into p square patches, with each patch treated as a token capturing spatial relationships among adjacent voxels. The patchified fMRI is then converted into fMRI token embedding $\mathbf{F}_i^{emb} \in \mathbb{R}^{(p+1) \times D}$ by cascaded spatial attention blocks, where D is embedding size. In the decoding phase, the patch tokens are discarded, retaining only the [CLS] token embedding $\mathbf{F}_i^{CLS} \in \mathbb{R}^{1 \times D}$. After passing through multiple cascaded spatial attention blocks, the [CLS] token embedding is reconstructed back into an fMRI image.

Additionally, the fMRI-PTE is pretrained on resting-state fMRI data sourced from the UK Biobank dataset [23]. To reduce the disparity in datasets, the model undergoes further fine-tuning on the fMRI dataset we used, ensuring the acquisition of general knowledge.

Spatiotemporal Attention for Sequential fMRI Frames To enhance the model's capability to handle sequential fMRI inputs and effectively decode videos

by learning temporal correlations, we augment fMRI-PTE with spatiotemporal attention. Initially, for the input fMRI data $\mathbf{F}_i$, we obtain the fMRI embedding $\mathbf{F}_i^{emb} \in \mathbb{R}^{n\times(p+1)\times D}$ by solely passing each frame to fMRI-PTE and then concatenating the results. To implement spatiotemporal attention, we employ the network inflation trick [41]. Specifically, using the attention calculation as $\text{attn} = \text{softmax}\big(\frac{QK^T}{\sqrt{d}}\big)$, the query and key of spatial attention are derived as $Q = \mathbf{F}_i^{emb} \cdot W_{spat}^Q, K = \mathbf{F}_i^{emb} \cdot W_{spat}^K$. Let $\mathbf{F}_i'^{emb} \in \mathbb{R}^{(p+1)\times n\times D}$ represent the result of exchanging the first two dimensions of $\mathbf{F}_i^{emb}$, the query and key of temporal attention are similarly derived as $Q = \mathbf{F}_i'^{emb} \cdot W_{temp}^Q, K = \mathbf{F}_i'^{emb} \cdot W_{temp}^K$.

Moreover, in contrast to MinD-Video [6], which substitutes spatial attention with spatiotemporal attention within the pretrained encoder, we adopt a simpler approach by appending the encoder with spatiotemporal attention to avoid disturbing the pretrained encoder.

Finally, following spatiotemporal aggregation, two projectors convert the attention results into fMRI embedding $emb_f \in \mathbb{R}^{K\times C}$ and fMRI CLIP latent $emb_f^{CLIP} \in \mathbb{R}^{1\times C}$, where K is the token number of text embedding, and C is the embedding size of text embedding and CLIP latent space.

Video Generator To achieve high-quality video reconstruction, we employ ModelScopeT2V [36], a text-to-video synthesis model including VQGAN [8] and Denoising UNet [31], as our video generator. Building upon the text-to-image Stable Diffusion [30], it incorporates spatiotemporal attention and undergoes pretraining on large-scale video-text paired datasets. This model is able to capture temporal correspondence from text input and generate high-quality continuous videos of varying frame numbers. As illustrated in Figure 2(b), during video reconstruction from fMRI, the fMRI embedding emb_f takes the place of the text embedding as the conditional input for the UNet of video generator.

3.5 Global-Local Functional Alignment

In the fMRI preprocessing, we achieve anatomical alignment in the 32k_fs_LR brain surface space. To further capture the functional correspondence between voxels across subjects, we introduce Global-Local Functional Alignment (GLFA) to our model.

For an fMRI sample $\mathbf{F}_i \in \mathbb{R}^{n\times H_f\times W_f}$, a widely used functional alignment method, such as Pairwise Cortical Alignment [7], employs a voxel-to-voxel mapping, requiring the learning of a large projection matrix $\mathbf{W} \in \mathbb{R}^{H_fW_f\times H_fW_f}$. To reduce the resource consumption, similar to [2], we partition the fMRI image into parcels of $H_p \times W_p$ and only learn a $\mathbf{W}^{parcel} \in \mathbb{R}^{\frac{H_fW_f}{H_pW_p}\times\frac{H_fW_f}{H_pW_p}}$ matrix. Specifically, the fMRI image $\mathbf{F}_i$ is patchified to be $\mathbf{F}_i^{parcel} \in \mathbb{R}^{n\times H_pW_p\times\frac{H_fW_f}{H_pW_p}}$, and the intermediate result of functional alignment is derived as

$$\hat{\mathbf{F}}_i = \mathbf{F}_i^{parcel} \cdot \mathbf{W}^{parcel} \tag{1}$$

As parcel-wise connections are established at the whole-brain level, Eq.(1) plays the role of global functional alignment. Further, $\hat{\mathbf{F}}_i$ is reshaped back to $\mathbb{R}^{n\times H_f\times W_f}$ and a locally connected 2D layer, functioning similarly to the Conv2D layer but with unshared weights, is applied to $\hat{\mathbf{F}}_i$ to capture the local disparities among neighboring voxels. Consequently, the final aligned fMRI is denoted as

$$\mathbf{F}_i^{aligned} = \text{LocallyConnected2D}(\hat{\mathbf{F}}_i, ks) \tag{2}$$

where ks denotes the convolutional kernel size and the stride is set to 1. Through local functional alignment (Eq.(2)), each voxel has its unique kernel to learn correspondence with neiboring voxels. Combining Eq.(1) and Eq.(2), our GLFA achieves functional alignment simultaneously at both the global and local levels.

In order to train the GLFA, we introduce two loss function: alignment loss $\mathcal{L}_{align}$ and consistency loss $\mathcal{L}_{const}$. On the one hand, the aligned fMRI of different individuals with same stimulation should be close to each other. On the other hand, the information in original fMRI should be preserved after the mapping. Thus the loss function can be derived as $\mathcal{L}_{align}(s_a, s_b) = \| \mathbf{F}_{i,s_a}^{aligned} - \mathbf{F}_{i,s_b}^{aligned} \|_F^2$ and $\mathcal{L}_{const}(s_a) = \| \mathbf{F}_{i,s_a}^{aligned} - \mathbf{F}_{i,s_a} \|_F^2$, where $\| \cdot \|_F$ represents Frobenius norm, and s_a, s_b denote different subjects. The loss for GLFA is further derived as

$$\mathcal{L}_{GLFA} = \frac{1}{N_{align}} \sum_{a<b} \mathcal{L}_{align}(s_a, s_b) + \lambda\big(\mathcal{L}_{const}(s_a) + \mathcal{L}_{const}(s_b)\big) \tag{3}$$

3.6 Training Strategy

Pretraining Due to disparities between datasets, we conduct fine-tuning on the fMRI-PTE using the fMRI data of the benchmark dataset [40] and our curated dataset, respectively. We only apply reconstruction loss [26] during fine-tuning, allowing the model to acquire general knowledge.

Contrastive Learning During the stage-1 contrastive learning phase, a random frame image along with its corresponding caption is chosen for each data sample. As shown in Figure 2(b), these are then transformed into image latent and text latent $emb_i^{CLIP}, emb_t^{CLIP} \in \mathbb{R}^{B\times C}$ using a frozen pre-trained CLIP ViT-H/14 Encoder [27]. CLIP loss [28] is subsequently calculated with the fMRI latent. The stage-1 loss is defined as follows.

$$\mathcal{L}_{stage1} = (L_{CLIP}(emb_f^{CLIP}, emb_{img}^{CLIP}) + \mathcal{L}_{CLIP}(emb_f^{CLIP}, emb_{txt}^{CLIP}))/2 \tag{4}$$

where $\mathcal{L}_{CLIP}(a, b) = \text{CrossEntropy}(\epsilon a \cdot b^{\mathrm{T}}, [1, 2, \cdots, B])$, and ϵ is a scaling factor.

Co-training with Video Generator After the contrastive learning phase, we utilize the fMRI embedding as input to the cross-attention module of the UNet for conditioning the video on fMRI.

Each frame of the video sample $\mathbf{V}_i = [v_{i,1}, v_{i,2}, ..., v_{i,m}]$ is individually encoded using VQGAN encoder $\mathcal{E}$, resulting in the ground-truth latent variable

Table 1: Results of cross-subject alignment metrics. All the fMRI spatial correlation and ISC increase significantly after GLFA processing. The metrics related to testing on unknown subjects are marked in bold.

		benchmark dataset [40]						our dataset					
	Task	fSC↑			ISC↑			fSC↑			ISC↑		
		1/2	1/3	2/3	1	2	3	1/2	1/3	2/3	1	2	3
fMRI-PTE [26]	-	.461	.358	.357	.569	.545	.496	.015	.005	.021	.012	.030	.019
GLFA	1,2→3	.692	**.517**	**.538**	.700	.715	**.597**	.040	**.015**	**.045**	.036	.055	**.036**
	1,3→2	**.542**	.559	**.503**	.681	**.630**	.652	**.027**	.012	**.046**	.033	**.054**	.040
	2,3→1	**.552**	**.506**	.581	**.656**	.685	.666	**.033**	**.014**	.055	**.031**	.061	.040

$Z_0^{gt} = [\mathcal{E}(v_{i,1}), \cdots, \mathcal{E}(v_{i,m})] \in \mathbb{R}^{m \times \frac{H_v}{8} \times \frac{W_v}{8} \times 4}$. During training, the diffusion process add Gaussian noise $[\epsilon_1^{gt}, \cdots, \epsilon_T^{gt}]$ for T steps to Z_0^{gt}, obtaining $[Z_0^{gt}, \cdots, Z_T^{gt}]$, with signal-to-noise ratio gradually decreasing. Subsequently, the UNet ϵ_θ operates on the latent space to predict the noise of each step:$\epsilon_t^{pr} = \epsilon_\theta(Z_t^{gt}, emb_f, t)$. A denoising loss, minimizing the difference between the predicted noise ϵ_t^{pr} and ground-truth noise ϵ_t^{gt}, is formulated as Eq. (5):

$$\mathcal{L}_{denoise} = \mathbb{E}_{Z_t^{gt}, \epsilon_t^{gt} \sim \mathcal{N}(0,1), t}\left[\parallel \epsilon_t^{gt} - \epsilon_t^{pr} \parallel_2^2 \right] \tag{5}$$

Given that different subjects share the same ground-truth video, $\mathcal{L}_{denoise}$ can serve a similar function to embedding alignment [16]. Thus, by combining $\mathcal{L}_{GLFA}$ and $\mathcal{L}_{denoise}$, the stage-2 loss enables the model to simultaneously align fMRI in voxel-level and embedding-level spaces.

$$\mathcal{L}_{stage2} = \mathcal{L}_{denoise} + \mathcal{L}_{GLFA} \tag{6}$$

Moreover, to retain the original knowledge in the text-to-video generator, we exclusively fine-tune self attention, cross attention and temporal convolution networks in UNet to achieve fMRI-conditioned video generation.

During the inference phase, the average of all training fMRI is used as the negative condition. The UNet generates $Z_0^{pr} = [z_1^{pr}, \cdots, z_m^{pr}]$ from random noise Z_T^{pr}. Then Z_1^{pr} is decoded by the VQGAN decoder $\mathcal{D}$ to obtain the reconstructed video $\mathbf{V}_i^{pr} = [\mathcal{D}(z_1^{pr}), \cdots, \mathcal{D}(z_m^{pr})]$.

4 Experiments

4.1 Dataset

Pre-training dataset The UK Biobank (UKB) [23] serves as an extensive biomedical database and research repository, encompassing comprehensive genetic and health-related information from approximately half a million participants in the UK. It is utilized for pretraining, specifically using a partial dataset that includes resting-state fMRI data from around 39,630 subjects. Each subject contributes one session, comprising 490 volumes.

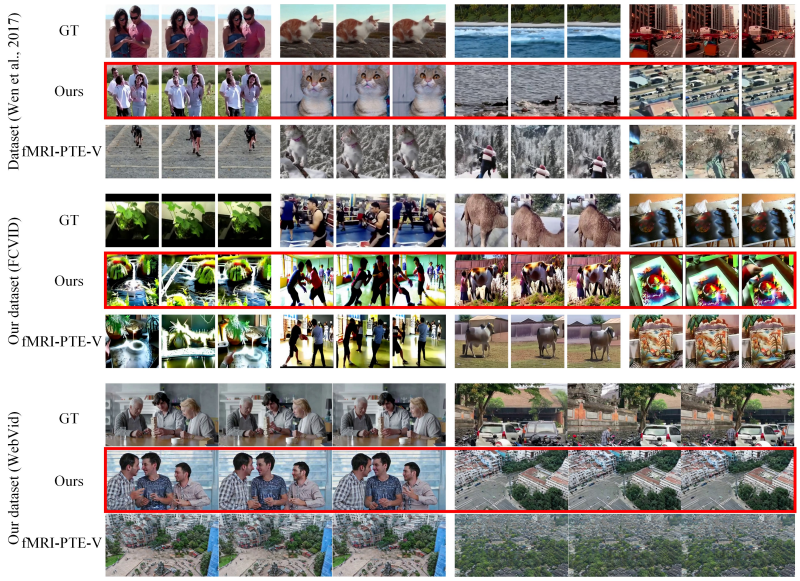

Fig. 4: Results of cross-subject fMRI-to-video Reconstruction. Our model (GLFA) demonstrates the ability to generate videos with greater semantic accuracy.

Paired fMRI-Video Dataset We conducted experiments on both benchmark dataset [40] and our curated dataset. The benchmark fMRI-video dataset [40] involves three subjects, and fMRI frames are acquired using a 3T MRI scanner with a 2-second TR. This dataset consists of totaling ∼3 video hours, providing ∼5.5k paired fMRI-stimulus examples for each individual. Our dataset incorporates data from 8 subjects and comprises 75k paired fMRI-stimulus samples, which is approximately 4.5 times larger than the benchmark dataset [40]. Our dataset is availiable in `https://huggingface.co/datasets/Fudan-fMRI/fMRI-Video`.

4.2 Implementation Details

For both datasets, we employed an fMRI window of $\alpha T = 4s$ to generate videos with a duration of $T = 2s/4s$. The videos from the benchmark dataset were downsampled to 8 FPS, while the videos from our dataset were downsampled to 5 FPS (FCVID) and 3 FPS (WebVid) respectively. All training and inference procedures were carried out on a single NVIDIA A100 GPU. For specific hyperparameter configurations and more detail during each training stage, please refer to the Supplementary. Code and models are available at `https://github.com/chongjg/GLFA-fmri-video`.

Competitors. (1) **MinD-Video** [6]: State-of-the-art within-subject fMRI-to-video pipeline on the benchmark dataset [40]. (2) **LEA** [25]: Baseline method that does not require complex training. It directly employs ridge regression to convert fMRI latents into text embeddings and subsequently reconstructs

Table 2: Results of fMRI-to-video generation performance on cross-subject decoding. Evaluations are provided for diverse subjects on the benchmark dataset [40] and our datasest. All metrics that surpass the baseline (LEA [25]) are indicated in bold.

	Task	Methods	Video-based		Frame-based		
			Semantic-level		Semantic-level		Structure-level
			2-way	50-way	2-way	50-way	SSIM
benchmark dataset [40]	1,2→3	LEA [25]	0.811	0.136	0.750	0.108	0.139
		fMRI-PTE-V	**0.834**	**0.182**	**0.765**	0.107	**0.161**
		GLFA	**0.847**	**0.193**	**0.777**	**0.116**	**0.172**
	1,3→2	LEA	0.809	0.128	0.765	0.101	0.138
		fMRI-PTE-V	**0.844**	**0.174**	0.762	**0.104**	0.130
		GLFA	**0.838**	**0.175**	**0.768**	**0.105**	**0.167**
	2,3→1	LEA	0.811	0.125	0.762	0.120	0.137
		fMRI-PTE-V	**0.846**	**0.179**	**0.771**	**0.123**	**0.151**
		GLFA	**0.837**	**0.179**	**0.781**	**0.126**	**0.181**
our dataset	1,2→3	LEA	0.794	0.105	0.803	0.165	0.169
		GLFA	**0.819**	**0.117**	**0.806**	**0.176**	**0.187**
	1,3→2	LEA	0.806	0.113	0.810	0.171	0.169
		GLFA	**0.808**	**0.117**	**0.815**	**0.174**	**0.178**
	2,3→1	LEA	0.779	0.110	0.792	0.161	0.167
		GLFA	**0.790**	**0.120**	**0.806**	**0.168**	**0.199**

the video through the pretrained text-to-video pipeline. (3) **fMRI-PTE-V**: Enhanced fMRI-PTE [26] by integrating a spatiotemporal attention module and a video diffusion pipeline, resulting in an advanced fMRI-to-video pipeline. (4) **GLFA**: Enhanced fMRI-PTE-V by incorporating GLFA to improve cross-subject performance.

4.3 Metrics

Cross-subject Alignment Metrics To evaluate the effectiveness of functional alignment, we employed commonly used metrics such as fMRI spatial correlation (fSC) and inter-subject correlation (ISC) [7]. For temporally aligned fMRI from k subjects $\mathbf{s} = [s_1, \cdots, s_k]$, we flattened the fMRI signals into 1D signals, represented as $f^{s_1}, \cdots, f^{s_k} \in \mathbb{R}^{T \times N_v}$, where N_v is number of voxel. Then the fMRI spatial correlation r between subject s_i and s_j is calculated as $r(s_i, s_j) = \frac{1}{T}\sum_{t=1}^{T} \text{corr}(f^{s_i}_{t,:}, f^{s_j}_{t,:})$, where $\text{corr}(\cdot)$ is Pearson correlation. Moreover, the ISC represents the average correlation between the functional time series of each subject and the mean time series of the remaining subjects [7], which can be formulated as $\text{ISC} = \frac{1}{k}\sum_{i=1}^{k} \text{ISC}_{s_i}$, where $\text{ISC}_{s_i} = \frac{1}{N_v}\sum_{v=1}^{N_v} \text{corr}(f^{s_i}_{:,v}, \sum_{j \neq i} f^{s_j}_{:,v})$.

Video Reconstruction Metrics Following MinD-Video [6], we employed both frame-based metrics and video-based metrics to evaluate the reconstructed videos.

Table 3: Results of fMRI-to-video generation performance on within-subject decoding. fMRI-PTE-V demonstrates comparable or even superior performance compared to the state-of-the-art model MinD-Video [6], outperforming LEA [25] in most indicators. All optimal metrics are indicated in bold.

	Task	Methods	Video-based		Frame-based		
			Semantic-level		Semantic-level		Structure-level
			2-way↑	50-way↑	2-way↑	50-way↑	SSIM↑
benchmark dataset [40]	subject 1	LEA [25]	0.825	0.149	0.792	0.144	0.137
		MinD-Video [6]	**0.853**	0.202	0.792	**0.172**	0.171
		fMRI-PTE-V	0.851	**0.214**	**0.793**	0.169	**0.193**
	subject 2	LEA	0.826	0.148	**0.785**	0.158	0.145
		MinD-Video	**0.841**	0.173	0.784	0.158	0.171
		fMRI-PTE-V	0.834	**0.192**	0.780	**0.159**	**0.182**
	subject 3	LEA	0.834	0.160	0.803	0.161	0.137
		MinD-Video	0.846	0.216	**0.812**	**0.193**	**0.187**
		fMRI-PTE-V	**0.851**	**0.225**	0.799	0.173	0.176
our dataset	subject 1	LEA	**0.803**	0.099	0.807	0.174	**0.157**
		fMRI-PTE-V	**0.803**	**0.142**	**0.817**	**0.211**	0.144
	subject 2	LEA	0.802	0.106	**0.831**	0.209	**0.160**
		fMRI-PTE-V	**0.814**	**0.138**	0.824	**0.236**	0.137
	subject 3	LEA	**0.808**	0.108	0.811	0.175	**0.168**
		fMRI-PTE-V	0.802	**0.119**	**0.814**	**0.218**	0.149

The frame-based metrics include the structural similarity index measure (SSIM) [38] and the N-way top-K accuracy classification test, serving as indicators for structural-level and semantic-level assessment, respectively. Specifically, SSIM and N-way top-K accuracy are computed for each frame of the ground truth and reconstructed videos, followed by averaging. An ImageNet classifier is utilized for image classification. For the video-based metrics, similar classification tests are applied, and VideoMAE [34] is used for video classification. Additionally, we employ the top-3 classification results as the ground truth (GT) class, following the approach of MinD-Video [6]. For N-way top-K accuracy, N candidates represent ground truth and N-1 randomly sampled classes from all classifier classes.

4.4 Functional Alignment Evaluation

To examine whether GLFA can reduce inter-subject differences, we applied GLFA well trained on the cross-subject decoding task to 3 subjects in the benchmark dataset [40] and our dataset. In Table 1, we presented fSC and ISC metrics for evaluation, where the bolded part is the cross-subject evaluation results. Task "1,2→3" indicates that the alignment learned from subject 1 and 2 is applied to subject 3 and fMRI Spatial Correlation "s_i/s_j" represents $r(s_i, s_j)$. For the ISC metrics, we showed ISC_{s_i} for each subjects.

It can be observed that, following GLFA, there is a significant improvement in fMRI spatial correlation and ISC. For instance, in task "1,2→3" on benchmark dataset, not only does $r(s_1, s_2)$ increase (.461 → .692), but the correlations $r(s_1, s_3)$ (.358 → .517) and $r(s_2, s_3)$ (.357 → .538) on an unknown subject 3 also exhibit notable enhancements, indicating a significant increment in the similarity of functional associations across different subjects after GLFA processing. Moreover, ISC_{s_3} also demonstrates a substantial increase (.496 → .597). These results highlight that GLFA can effectively reduce inter-subject differences.

4.5 Cross-subject fMRI-Video Reconstruction

In order to evaluate the effectiveness of functional alignment in cross-subject decoding, we focused on evaluating the performance of video reconstruction. We conducted experiments on both the benchmark dataset [40] and our fMRI-video dataset, comparing our results with a cross-subject method LEA [25] as baseline.

The results of leave-one-out cross-subject decoding on both datasets are reported in Table 2. After applying GLFA, we observed improvements all metrics. Moreover, in Figure 4, we provided a visual comparison of our method and the baseline by showcasing the reconstruction results using the same samples. This demonstrates that our approach generally produces superior video reconstructions and is able to achieve high-quality cross-subject fMRI-to-video reconstruction. For more ablation study and statistical analysis, please refer to the Supplementary.

4.6 Within-subject fMRI-Video Reconstruction

In order to compare our method with the state-of-the-art fMRI-to-Video pipeline MinD-Video [6], we also conducted within-subject decoding on both datasets. As shown in Table. 3, our method exhibits comparable performance to MinD-Video [6], clearly surpassing LEA [25].

5 Conclusion

In this paper, we implemented a cross-subject fMRI-to-video pipeline, by combining fMRI-PTE [26] with a pretrained video generator [36]. To further enhance cross-subject decoding, we introduced fMRI Global-Local Functional Alignment (GLFA), which aligns fMRI from multiple subjects, improving cross-subject decoding performance and enabling high-quality video reconstruction. GLFA reduces cross-subject differences, improves correlations, and captures meaningful information. Additionally, we curated a large fMRI-video dataset. This mitigates the scarcity in fMRI-video decoding and expands foundation for studying cross-subject tasks.

Acknowledgments

The computations in this research were performed using the CFFF platform of Fudan University. Thanks for the support from the Shanghai Technology Development and Entrepreneurship Platform for Neuromorphic and AI SoC and, in part, the Shanghai Research and Innovation Functional Program under Grant 17DZ2260900. Yanwei Fu is with School of Data Science, FudanISTBI—ZJNU Algorithm Centre for Brain-inspired Intelligence, Fudan University,Shanghai Key Lab of Intelligent Information Processing, and TechnologyInnovation Center of Calligraphy and Painting Digital Generation, Ministry ofCulture and Tourism, China. Email: yanweifu@fudan.edu.cn

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1728–1738 (2021)
2. Bazeille, T., DuPre, E., Richard, H., Poline, J.B., Thirion, B.: An empirical evaluation of functional alignment using inter-subject decoding. NeuroImage **245**, 118683 (2021). https://doi.org/https://doi.org/10.1016/j.neuroimage.2021.118683, https://www.sciencedirect.com/science/article/pii/S1053811921009563
3. Bazeille, T., Richard, H., Janati, H., Thirion, B.: Local optimal transport for functional brain template estimation. In: Information Processing in Medical Imaging (2019), https://api.semanticscholar.org/CorpusID:162169103
4. Chau, W., McIntosh, A.R.: The talairach coordinate of a point in the mni space: how to interpret it. NeuroImage **25**(2), 408–416 (2005). https://doi.org/https://doi.org/10.1016/j.neuroimage.2004.12.007, https://www.sciencedirect.com/science/article/pii/S1053811904007554
5. Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: 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. pp. 22710–22720 (2023)
6. Chen, Z., Qing, J., Zhou, J.H.: Cinematic mindscapes: High-quality video reconstruction from brain activity. arXiv preprint arXiv:2305.11675 (2023)
7. Conroy, B., Singer, B., Haxby, J., Ramadge, P.J.: fmri-based inter-subject cortical alignment using functional connectivity. Advances in neural information processing systems **22** (2009)
8. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
9. Gao, J.S., Huth, A.G., Lescroart, M.D., Gallant, J.L.: Pycortex: an interactive surface visualizer for fmri. Frontiers in neuroinformatics **9**, 23 (2015)
10. Gao, J., Fu, Y., Wang, Y., Qian, X., Feng, J., Fu, Y.: Mind-3d: Reconstruct high-quality 3d objects in human brain. arXiv preprint arXiv:2312.07485 (2023)
11. Gao, Y., Zhang, Y., Cao, Z., Guo, X., Zhang, J.: Decoding brain states from fmri signals by using unsupervised domain adaptation. IEEE Journal of Biomedical and Health Informatics **24**(6), 1677–1685 (2020). https://doi.org/10.1109/JBHI.2019.2940695

12. Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J.W., Yacoub, E., Uğurbil, K., Andersson, J.L.R., Beckmann, C.F., Jenkinson, M., Smith, S.M., Essen, D.C.V.: A multi-modal parcellation of human cerebral cortex. Nature **536**, 171 – 178 (2016), https://api.semanticscholar.org/CorpusID:205249949
13. Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., et al.: The minimal preprocessing pipelines for the human connectome project. Neuroimage **80**, 105–124 (2013)
14. Gong, Z., Zhang, Q., Bao, G., Zhu, L., Liu, K., Hu, L., Miao, D.: Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. arXiv preprint arXiv:2404.12630 (2024)
15. Han, K., Wen, H., Shi, J., Lu, K.H., Zhang, Y., Fu, D., Liu, Z.: Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. NeuroImage **198**, 125–136 (2019). https://doi.org/https://doi.org/10.1016/j.neuroimage.2019.05.039, https://www.sciencedirect.com/science/article/pii/S1053811919304318
16. Haxby, J., Guntupalli, J., Connolly, A., Halchenko, Y., Conroy, B., Gobbini, M., Hanke, M., Ramadge, P.: A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron **72**(2), 404–416 (2011). https://doi.org/https://doi.org/10.1016/j.neuron.2011.08.026, https://www.sciencedirect.com/science/article/pii/S0896627311007811
17. Huang, S., Shao, W., Wang, M.L., Zhang, D.Q.: fmri-based decoding of visual information from human brain activity: A brief review. International Journal of Automation and Computing **18**(2), 170–184 (2021)
18. Jiang, Y.G., Wu, Z., Wang, J., Xue, X., Chang, S.F.: Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(2), 352–364 (2018). https://doi.org/10.1109/TPAMI.2017.2670560
19. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning (2023), https://api.semanticscholar.org/CorpusID:256390509
20. Li, W., Liu, M., Chen, F., Zhang, D.: Graph-based decoding model for functional alignment of unaligned fmri data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 2653–2660 (2020)
21. Lin, S., Sprague, T., Singh, A.K.: Mind reader: Reconstructing complex images from brain activities. Advances in Neural Information Processing Systems **35**, 29624–29636 (2022)
22. Liu, Y., Ma, Y., Zhou, W., Zhu, G., Zheng, N.: Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding from fmri. arXiv preprint arXiv:2302.12971 (2023)
23. Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.R., Griffanti, L., Douaud, G., Okell, T.W., Weale, P.J., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M.: Multimodal population brain imaging in the uk biobank prospective epidemiological study. Nature neuroscience **19**, 1523 – 1536 (2016), https://api.semanticscholar.org/CorpusID:1018393
24. Ozcelik, F., VanRullen, R.: Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion. arXiv preprint arXiv:2303.05334 (2023)

25. Qian, X., Wang, Y., Fu, Y., Sun, X., Xue, X., Feng, J.: Joint fmri decoding and encoding with latent embedding alignment (2023), https://api.semanticscholar.org/CorpusID:259076476
26. Qian, X., Wang, Y., Huo, J., Feng, J., Fu, Y.: fmri-pte: A large-scale fmri pretrained transformer encoder for multi-subject brain activity decoding. arXiv preprint arXiv:2311.00342 (2023)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
29. Ramírez, F.M., Revsine, C., Merriam, E.P.: What do across-subject analyses really tell us about neural coding? Neuropsychologia **143**, 107489 (2020). https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2020.107489, https://www.sciencedirect.com/science/article/pii/S0028393220301603
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
32. Sabuncu, M.R., Singer, B.D., Conroy, B., Bryan, R.E., Ramadge, P.J., Haxby, J.V.: Function-based Intersubject Alignment of Human Cortical Anatomy. Cerebral Cortex **20**(1), 130–140 (05 2009). https://doi.org/10.1093/cercor/bhp085, https://doi.org/10.1093/cercor/bhp085
33. Tong, F., Pratte, M.S.: Decoding patterns of human brain activity. Annual review of psychology **63**, 483–509 (2012)
34. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Advances in Neural Information Processing Systems (2022)
35. Wang, C., Yan, H., Huang, W., Li, J., Wang, Y., Fan, Y.S., Sheng, W., Liu, T., Li, R., Chen, H.: Reconstructing rapid natural vision with fMRI-conditional video generative adversarial network. Cerebral Cortex **32**(20), 4502–4511 (01 2022). https://doi.org/10.1093/cercor/bhab498, https://doi.org/10.1093/cercor/bhab498
36. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
37. Wang, S., Liu, S., Tan, Z., Wang, X.: Mindbridge: A cross-subject brain decoding framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11333–11342 (2024)
38. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
39. Watson, J.D.G., Myers, R., Frackowiak, R.S.J., Hajnal, J.V., Woods, R.P., Mazziotta, J.C., Shipp, S., Zeki, S.: Area V5 of the Human Brain: Evidence from a Com-

bined Study Using Positron Emission Tomography and Magnetic Resonance Imaging. Cerebral Cortex **3**(2), 79–94 (03 1993). https://doi.org/10.1093/cercor/3.2.79, https://doi.org/10.1093/cercor/3.2.79

40. Wen, H., Shi, J., Zhang, Y., Lu, K.H., Cao, J., Liu, Z.: Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. Cerebral Cortex **28**(12), 4136–4160 (10 2017). https://doi.org/10.1093/cercor/bhx268, https://doi.org/10.1093/cercor/bhx268
41. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: 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. pp. 7623–7633 (2023)