Enhancing Cross-Subject fMRI-to-Video Decoding with Global-Local Functional Alignment (Supplementary Material)

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

1 More Implementation Details

Pretraining In order to mitigate the impact of variations in fMRI data collected by different machines and individuals, we utilized a large-scale fMRI Pretrained Transformer Encoder (fMRI-PTE) [3], pretrained on the extensive UK Biobank dataset [2]. Subsequently, we performed additional fine-tuning on both the benchmark dataset [5] and our curated dataset, employing an autoencoder training framework [3]. More specifically, we processed the original 256×256 fMRI surface image using a patch size of 16. Each patch was then transformed into a 1024-dimensional embedding using 24 layers of transformer with 16 heads. The resulting features, with a dimensionality of 257×1024, served as the output of the encoder. Subsequently, only the [CLS] token was input into another set of 24 transformer layers for decoding and reconstructing the original fMRI surface image. Mean square error (MSE) was used as the loss function during this process.

This methodology allows the encoder to compress the each fMRI frame into a 1×1024 vector and then reconstruct the original fMRI, thereby acquiring a comprehensive understanding of fMRI data. Moreover, to prevent the encoder from accessing the test set data, unlike MinD-Video [1], we exclusively conducted fine-tuning on the training set data. In this phase, the batch size was set to 64 with a learning rate of 1e-4. To prevent overfitting, we applied a weight decay of 0.01, and the training was limited to 2000 steps.

Contrastive Learning During the contrastive learning phase, to mitigate overfitting, we applied data augmentation for image and text. Random cropping was employed for images, while text underwent synonym augmentation and random

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

Table 1: Hyperparameter settings for model training and inference on the two datasets.

	batch size	weight decay	learning rate	training steps
benchmark dataset	16	0.01	3e-5	18k
our dataset	12	0.01	1e-4	30k
	pretrained checkpoint	num_frame	guidance_scale	num_inference
benchmark dataset	modelscope T2V	6/16	9	25
our dataset	zeroscope_v2_576w	12	6	30

word swapping. Throughout the experiments, we used a batch size of 32 for the benchmark dataset [5] and 24 for our dataset. A dropout rate of 0.6 was applied to prevent over-reliance on specific features during training. Additionally, a weight decay of 0.01 was used to regularize the model and prevent overfitting. The models were trained for 12,000 steps, and a learning rate of 2e-5 was employed during the training process. These parameters were chosen to optimize the model's performance and ensure effective training.

Co-training with Video Generator The hyperparameter settings are presented in Table 1. Due to the extensive GPU memory requirements for video generation, the batch size was limited to 16/12. During the training phase, a learning rate of 3e-5/1e-4 and a weight decay of 0.01 were used, and the model was trained for 18k/30k steps. The pretrained modelscope T2V [4] is a multi-stage text-to-video generation diffusion model that utilizes the Unet3D architecture. The zeroscope_v2_576w model was fine-tuned based on the weights of modelscope T2V using a dataset comprising 9,923 video clips and 29,769 tagged frames, with a resolution of 576×320 at 24 frames. In fMRI-to-video generation, fMRI embeddings play a role similar to text embeddings in conditioning the generation of videos. Furthermore, GLFA module used kernel size of 13 for locally connected 2D layer, and consistency loss scale $\lambda = \frac{1}{2}$ for $\mathcal{L}_{GLFA}$ (Eq.(7)).

Inference Each video was generated utilizing 25/30 diffusion steps, guided by fMRI adversarial techniques. The fMRI adversarial guidance specifically employs the average fMRI from the training set as the negative guidance, with a guidance scale of 9/6.

2 Supplementary Results and Visualizations

2.1 Ablation study and statistical analysis

We report standard deviation and statistical significance (two-sample t-test) in Tab.2 for the cross-subject results (benchmark dataset, task "$1, 2 \rightarrow 3$"). Given the limited existing work on cross-subject fMRI-to-video generation, we modified fMRI-PTE-V and LEA from existing works for comparison. These methods inherently serve as the ablation study to directly show the significance of GLFA.

Table 2: Cross-subject decoding results and training stage ablation study (benchmark dataset, task "$1, 2 \rightarrow 3$"). Colors represent statistical significance. $p < 0.001$(**cyan**); $p < 0.05$(pink); $p > 0.05$(**gray**);

	t-test reference	video-based		frame-based		
		2-way	50-way	2-way	50-way	SSIM
LEA	-	.811±.03	.136±.01	.750±.03	.108±.01	0.139±.06
fMRI-PTE-V	LEA	.834±.03	.182±.01	.765±.03	.107±.01	0.161±.07
GLFA	fMRI-PTE-V	.847±.03	.193±.02	.777±.03	.116±.01	0.172±.08
training stage ablation on GLFA						
w/o pretraining	GLFA	.848±.03	.173±.01	.748±.03	.096±.01	.137±.05
w/o stage-1	GLFA	.846±.03	.169±.02	.753±.03	.096±.01	.153±.05

Table 3: Cross-dataset decoding results of our dataset. Due to inconsistent resolutions, SSIM cannot be computed for this task.

	Video-based		Frame-based	
	Semantic-level		Semantic-level	
	2-way↑	50-way↑	2-way↑	50-way↑
LEA	0.761	0.073	**0.813**	**0.108**
GLFA	**0.792**	**0.081**	0.802	0.107

2.2 Within-subject Decoding Results

We conducted within-subject decoding on subjects from both the benchmark dataset [5] and our dataset (WebVid). For the benchmark dataset, we generated 6 frames and displayed 3 frames for each of the three subjects (Figure 1). In the case of our dataset, we generated 12 frames, and displayed 6 frames for subject 1 (Figure 2).

2.3 Cross-dataset Decoding Results

To examine the cross-person and cross-dataset decoding capability of the model, we have chosen two distinct video datasets as stimuli for the fMRI signals. In detail, we trained on the data of three individuals from our dataset (WebVid stimulation), and subsequently tested on the data of three individuals from our dataset (FCVID stimulation). We demonstrated all metrics as depicted in the Table 3. The comparative results of GLFA and LEA reconstructions are illustrated in the figure 3.

3 Limitation

Although our curated dataset provides a substantial number of fMRI-stimulus paired samples across 8 subjects, achieving a more generalized cross-subject

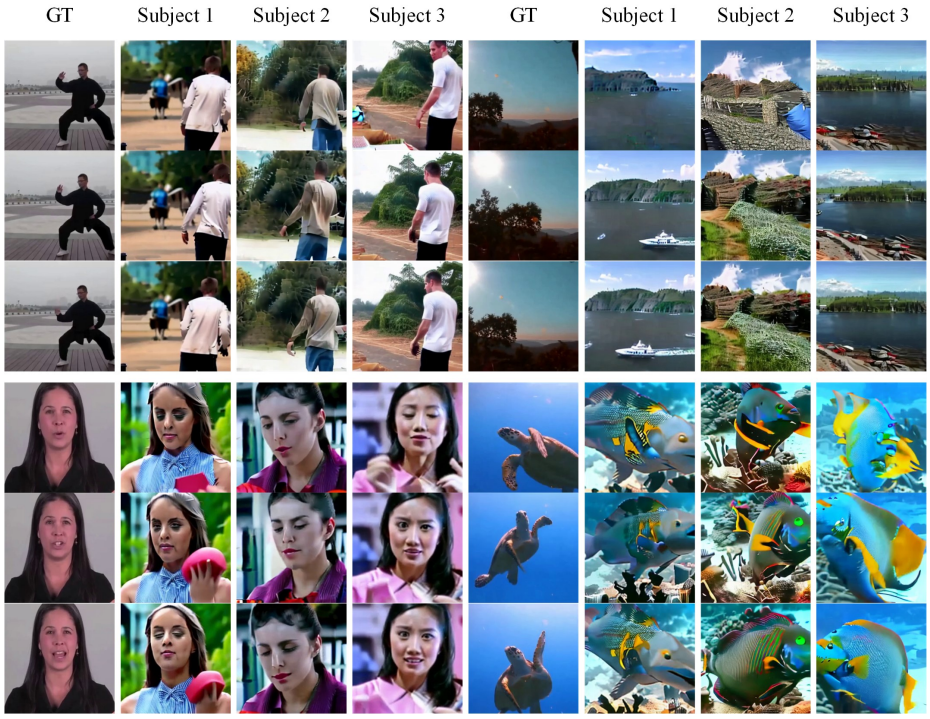

Fig. 1: Within-subject decoding results of benchmark dataset [5] .

video decoding model requires additional data support. Furthermore, this work directly trains cross-subject models by combining data paired across multiple subjects, which reduces the coverage of stimuli within each batch. Designing more effective training strategies to balance cross-subject alignment and decoding task learning is also a challenge that needs to be addressed.

References

1. Chen, Z., Qing, J., Zhou, J.H.: Cinematic mindscapes: High-quality video reconstruction from brain activity. arXiv preprint arXiv:2305.11675 (2023)
2. 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`
3. 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)
4. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)

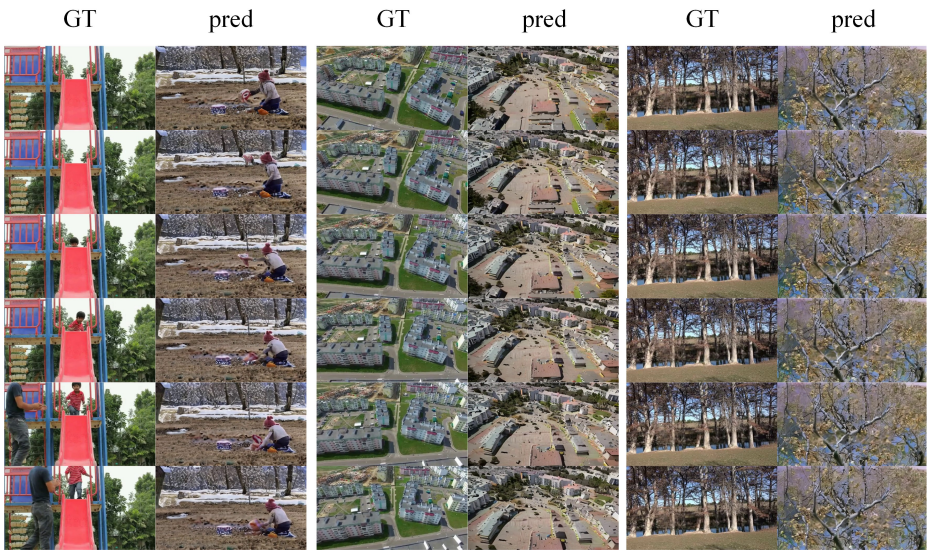

Fig. 2: Within-subject decoding results of our dataset.

5. 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

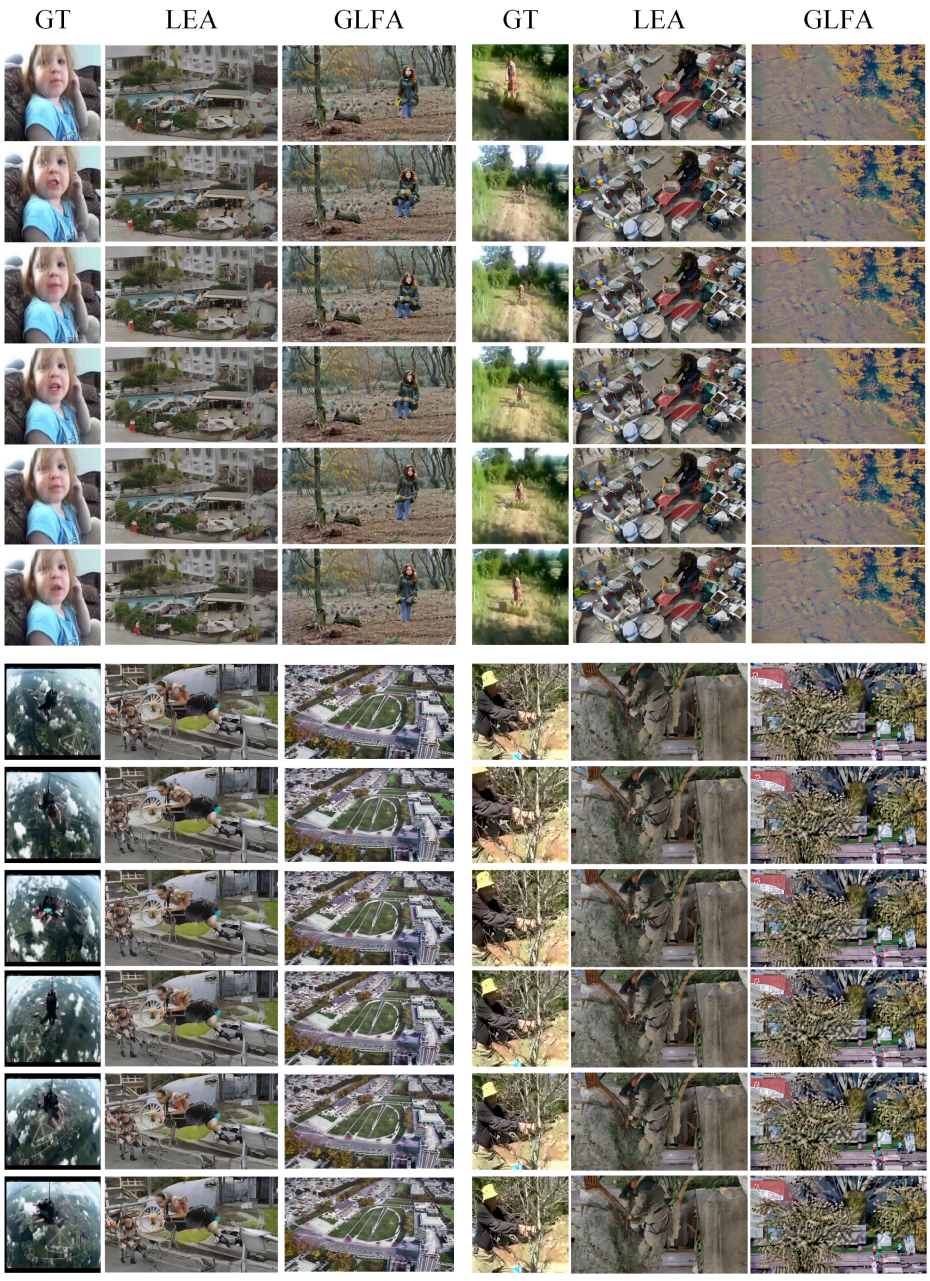

Fig. 3: Cross-dataset decoding results of our dataset.