Supplementary of CineBrain: A Large-Scale Multi-Modal Audiovisual Brain Dataset for Brain-Conditioned Video Generation

Jianxiong Gao[1], Yichang Liu[1], Baofeng Yang[1], Jianfeng Feng[1], Yanwei Fu[1,2,†]
[1]Fudan University [2]Shanghai Innovation Institute
{jxgao22,ycliu24}@m.fudan.edu.cn, {bfyang,jffeng,yanweifu}@fudan.edu.cn

1. Detailed Information on CineBrain

1.1. Detailed ROIs Selected in the Experiments

The regions of interest (ROIs) used in our study are summarized below. For the visual cortex, we include the following ROIs: *V1, V2, V3, V3A, V3B, V3CD, V4, LO1, LO2, LO3, PIT, V4t, V6, V6A, V7, V8, PH, FFC, IP0, MT, MST, FST, VVC, VMV1, VMV2, VMV3, PHA1, PHA2, PHA3*. For the auditory-related cortex, we select ROIs corresponding to: *4, 7AL, 7Am, 7m, 7PC, 7PL, 7Pm, 8Ad, 8Av, 8BM, 8C, 9a, 9p, 10d, 10v, 44, 45, 46, 47m, A1, IPS1, p32, PGp, s32, STGa, STSda, STSdp, STSva, STSvp, TPOJ1, TPOJ2, TPOJ3*. In total, the selected visual ROIs include 8,405 voxels, while the auditory ROIs include 10,541 voxels.

1.2. Prompts for Generating Text Descriptions

Textual descriptions play an important role in our training pipeline. The prompts used to generate these descriptions with Qwen2.5-VL [1] are provided in Fig. 1.

1.3. EEG Experimental Device

We illustrate the electrode montage of a 64-channel EEG cap configured according to the GSN-HydroCel-64_1.0 layout in Fig. 2.

2. Additional Experimental Results

2.1. Implementation Details

Due to GPU memory constraints, Our proposed CineSync is trained using a two-stage approach. In the first stage, only the multimodal fusion encoder is trained using contrastive loss with a batch size of 16 on a single H100 GPU. To enhance the effectiveness of contrastive learning, we augment the training dataset with diverse textual captions generated by Qwen2.5-VL-7B [1]. This pretraining phase lasts approximately 50 epochs. In the second stage, the pretrained encoder is integrated into the full model, which is finetuned for 5000 steps on 4 H100 GPUs using a batch size of 2 per

Users

Please watch the video and provide a detailed description of it, focusing on the main subjects, their actions, and the background scenes.

Qwen2.5-VL

The video depicts a lively office party scene with two men as the main focus. The setting is an office break room or similar communal area, characterized by a casual yet festive atmosphere. The background features a variety of people mingling, some holding drinks and others engaged in conversations, contributing to the social vibe. In the foreground, one man is dressed in a mustard-yellow corduroy blazer over a patterned sweater, paired with a white shirt. He appears to be in mid-conversation, his body language suggesting he is actively participating in the dialogue. His facial expression indicates he might be reacting to something the other man has said or done. The second man, wearing a gray suit jacket over a light blue shirt, is holding a glass of orange juice and seems to be in the middle of speaking. His posture and hand gestures suggest he is explaining something or making a point. The interaction between these two men is central to the scene, capturing a moment of friendly banter or discussion. The background reveals more details about the setting: a vending machine labeled \"COLD\" and \"drinks\" is visible, indicating the availability of refreshments for the attendees. A counter with various food items, including fruits and snacks, is also present, further emphasizing the party-like environment. The walls are adorned with colorful posters and artwork, adding a vibrant touch to the space. The overall ambiance suggests a relaxed and enjoyable gathering, typical of a workplace celebration or informal event.

Figure 1. Prompt utilized for generating video descriptions.

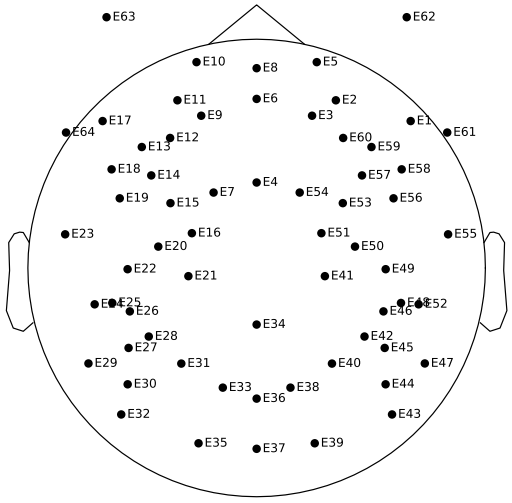

Figure 2. Electrode montage of a 64-channel EEG cap using the GSN-HydroCel-64_1.0 layout. Sensor positions are annotated with their corresponding channel labels.

†: Corresponding author.

Table 1. **Ablation study on multimodal alignment in CineSync.** We report the average performance across all subjects. CineSync* indicates the experiment that includes audio-related ROIs in fMRI.

Methods	Semantic-level					Perceptual-level			
	Video-based			Frame-based		Video-based		Frame-based	
	2-way↑	50-way↑	FVD↓	2-way↑	50-way↑	DTC↑	CTC↑	SSIM↑	PSNR↑
w/o Vision	0.863	0.275	52.06	0.858	0.378	0.895	0.738	0.272	11.69
w/o Text	0.891	0.294	50.15	0.887	0.394	0.908	0.949	0.285	11.95
w/o Across	0.873	0.279	51.29	0.864	0.382	0.902	0.943	0.274	11.74
CineSync*	**0.926**	**0.336**	**44.77**	**0.954**	**0.423**	**0.921**	**0.953**	**0.297**	**12.18**

Figure 3. **Video Reconstruction Results for Subjects 1, 2, and 6.** We compare the reconstructed frames from Subjects 1, 2, and 6 with the corresponding ground-truth (GT) frames. The consistent semantic alignment and visual fidelity across subjects demonstrate the robustness and strong cross-subject generalization ability of **CineSync**.

GPU. The fMRI and EEG transformers each consist of 12 layers with a transformer dimension of 2048 and a token length of 227 (226 spatial tokens plus one class token). The LoRA configuration uses a rank of 64 and a scaling factor α=64. Input data consist of 4-second multimodal brain signals, yielding standard inputs of 5×8405 fMRI voxels and 64×4000 EEG data points.

2.2. Ablation Study on Multimodal Alignment

To quantify the contribution of each alignment component in **CineSync**, we conduct ablations on the full model trained with both audio- and vision-related ROIs in fMRI. Specifically, *w/o Vision*, *w/o Text*, and *w/o Across* correspond to removing (i) the vision–fMRI alignment branch, (ii) the text–fMRI alignment branch, and (iii) the cross-modal alignment between EEG and fMRI, respectively. The averaged results across all subjects are reported in Tab. 1. Removing any of these alignment pathways leads to clear and consistent performance drops across both semantic- and perceptual-level metrics, demonstrating that rich cross-modal alignment is critical for achieving high-quality brain-to-video reconstruction.

2.3. Detailed Results for Each Subject

To further verify that our model remains robust across individuals, we report per-subject quantitative performance in Tab. 2. As shown, the results exhibit only minor variations across the six subjects, indicating that **CineSync** generalizes well to different brain patterns. Moreover, the averaged scores closely match the overall CineSync* results, further confirming subject-invariant performance.

2.4. Video Reconstruction Across Subjects

To better illustrate the robustness of our model, we visualize the reconstructed videos from different subjects under the same visual stimuli. Representative results are shown in

Figure 4. **Video Reconstruction Results for Subjects 3, 4, and 5.** Reconstructed frames from Subjects 3, 4, and 5 are shown alongside the GT frames at matched timestamps. The results further verify that **CineSync** maintains stable reconstruction quality across different individuals.

Table 2. **Performance of each subject in CineBrain.** CineSync* indicates the experiment that includes audio-related ROIs in fMRI.

METHODS	Semantic-level					Perceptual-level			
	Video-based			Frame-based		Video-based		Frame-based	
	2-way↑	50-way↑	FVD↓	2-way↑	50-way↑	DTC↑	CTC↑	SSIM↑	PSNR↑
Subject 1	0.921	0.332	46.12	0.951	0.419	0.918	0.949	0.294	12.11
Subject 2	0.928	0.337	44.03	0.956	0.426	0.922	0.954	0.298	12.20
Subject 3	0.923	0.334	45.87	0.952	0.421	0.917	0.950	0.295	12.14
Subject 4	0.927	0.338	43.95	0.957	0.427	0.923	0.956	0.299	12.23
Subject 5	0.929	0.339	44.62	0.955	0.425	0.920	0.952	0.296	12.17
Subject 6	0.924	0.335	44.97	0.953	0.420	0.919	0.951	0.297	12.22
CineSync*	**0.926**	**0.336**	**44.77**	**0.954**	**0.423**	**0.921**	**0.953**	**0.297**	**12.18**

Fig. 3 and Fig. 4. Since Subjects 1, 2, and 6 share the same train–test split, and Subjects 3, 4, and 5 share another split, we compare the reconstruction quality within each group accordingly. Specifically, Fig. 3 presents the reconstructions of Subjects 1, 2, and 6, while Fig. 4 shows those from Subjects 3, 4, and 5. These results indicate that our model achieves stable performance across different individuals.

2.5. More Results of Video Reconstruction

To further showcase the quality, semantic fidelity, and temporal coherence of the reconstructed videos, we present additional 12-frame examples in Fig. 5, Fig. 6, and Fig. 7. Across diverse scenes, our method consistently produces reconstructions that are semantically aligned with the stimuli and temporally smooth.

3. Limitations and Future Work

Although our CineSync successfully leverages the spatial resolution of EEG to compensate for the temporal resolution limitations of fMRI, substantially improving video and audio reconstruction performance, our dataset currently does not explicitly provide additional fMRI data. This limitation restricts broader applications of our dataset and represents an area for future exploration.

Furthermore, our CineBrain dataset supports synchronized audiovisual stimuli. However, we have independently evaluated our primary contributions solely through separate video and audio reconstruction tasks. If we aim to expand into more complex applications such as embodied intelligence, it will be necessary to reconstruct multiple modalities simultaneously. Therefore, joint audiovisual reconstruction represents another significant direction for our future research.

Figure 5. **More Results of CineSync:** We present 12 frames with timestamps compared with the ground truth (GT).

Figure 6. **More Results of CineSync:** We present 12 frames with timestamps compared with the ground truth (GT).

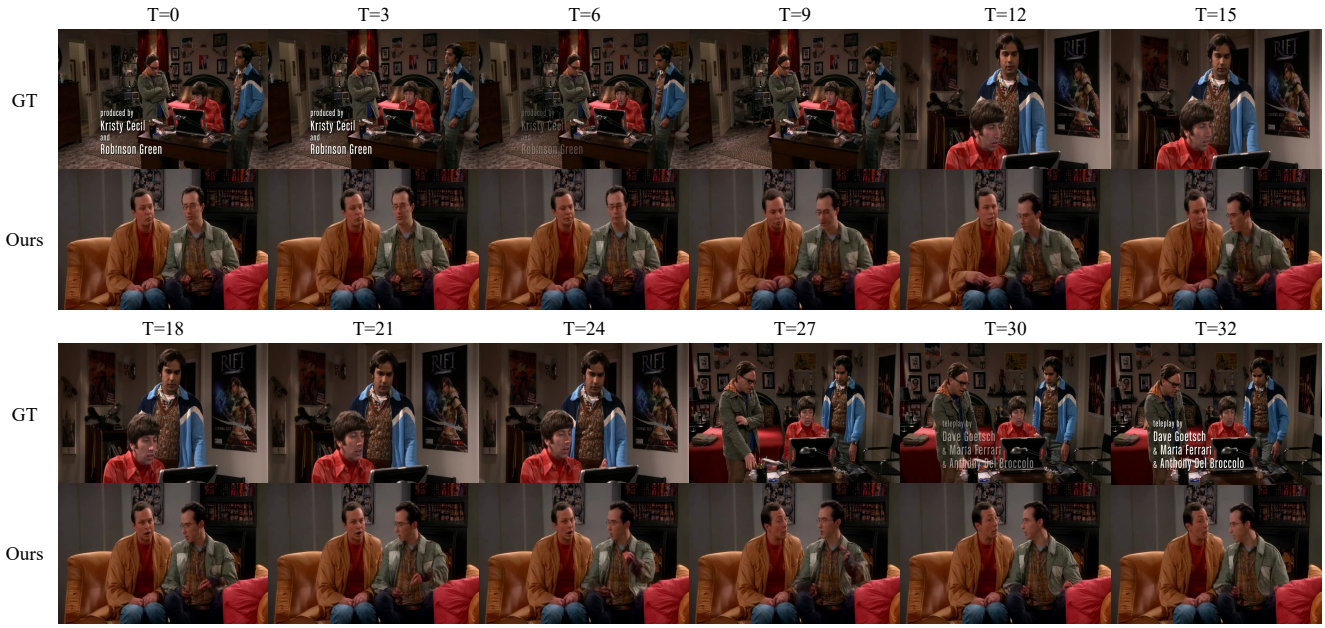

Figure 7. **More Results of CineSync:** We present 12 frames with timestamps compared with the ground truth (GT).

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1