DynaMind: Reconstructing Dynamic Visual Scenes from EEG by Aligning Temporal Dynamics and Multimodal Semantics to Guided Diffusion

Supplementary Material

This supplement presents technical specifications and extended experimental analysis for the DynaMind framework in EEG-to-video reconstruction tasks. We organize the content as follows:

- In Appendix 1, we introduce the SEED-DV and CineBrain datasets.
- In Appendix 2, we outline the training protocols for the RSM, TDA, and DGVR modules.
- In Appendix 3, we define the metrics for frame quality, semantic accuracy, and temporal dynamics.
- In Appendix 4, we list all comparison baselines.
- In Appendix 5, we analyze the computational complexity of our framework and other baselines.
- In Appendix 6, we detail the extra ablation studies.
- In Appendix 7, we display additional visualizations with failure case analysis.

1. Dataset

We validate our method using two public benchmarks: the SEED-DV dataset [8] and the CineBrain dataset [2].

1.1. SEED-DV dataset

SEED-DV targets the reconstruction of dynamic visual perception from EEG signals. It contains 1,400 EEG-video pairs collected from 20 participants. The data spans 40 distinct visual concepts and includes detailed meta-information for each video clip. The associated baseline, EEG2Video, employs a Seq2Seq architecture with a dynamic noising strategy. This setup supports high temporal resolution decoding of dynamic features, including color changes, object trajectories, and motion. Technically, the model uses an inflated diffusion model fine-tuned on the dataset to generate video content. Therefore, this dataset serves as a primary benchmark for evaluating neural representation learning and video reconstruction performance.

1.2. CineBrain dataset

CineBrain constitutes a large-scale multimodal dataset capturing brain responses to natural narrative stimuli, specifically synchronized fMRI and EEG recordings obtained while subjects viewed 20 episodes of *The Big Bang Theory*. The archive comprises approximately 6 hours of video content, corresponding to 162,000 fMRI frames and concurrent 64-channel EEG signals. Data acquisition utilizes 3T whole-brain fMRI and high-density EEG, incorporating defined visual and auditory Regions of Interest (ROIs). This resource supports tasks including video reconstruction, cross-modal alignment, and functional analysis, serving as a benchmark for multimodal decoding.

2. Implementation Details

The proposed framework consists of three primary modules, implemented with specific training protocols:

- **Regional-aware Semantic Mapper (RSM):** We spatially parcellate EEG data into $K = 4$ brain regions [3, 9]: Frontal, Parietal, Occipital, and Temporal. Electrode configurations within these regions vary based on the specific data acquisition system. Independent encoders process data from each region to yield 512-dimensional embeddings, which we fuse and project into a unified 1024-dimensional feature vector. We align this vector with the text and image feature spaces of a pretrained model (CLIP ViT-L/14 or T5-XXL), utilizing dataset labels for classification supervision. Training for the regional encoder and alignment mappings spans 300 epochs (learning rate 2×10^{-5}, batch size 32). We train a separate prior module for 1000 epochs (learning rate 1×10^{-5}, batch size 8).
- **Temporal-aware Dynamic Aligner (TDA):** We segment EEG signals into N overlapping time windows, where N corresponds to the video frame count. A temporal encoder converts these windows into a feature sequence. This module undergoes training for 300 epochs (learning rate 1×10^{-5}, batch size 32) to align outputs with VAE latents extracted from ground-truth videos via the specific VAE encoder used in DGVR.
- **Dual-Guidance Video Reconstructor (DGVR):** We adapt the DGVR module to different backbone generative models to align with dataset-specific benchmarks. For SEED-DV, we fine-tune Stable Diffusion v1.4 for 200 epochs (learning rate 3×10^{-5}), following the protocol in [16]. We utilize 6-frame video clips at a resolution of 512×288 (3 fps) for this process. For CineBrain, we adopt CogVideoX-5B-T2V as the decoder consistent with [2], generating 33-frame clips at 720×480 resolution. Given the strong pre-trained priors of CogVideoX, we employ the model without additional fine-tuning, demonstrating that the frozen backbone yields competitive performance. All training are conducted on an NVIDIA A800 GPU using a cosine annealing scheduler.

3. Evaluation Metrics

We adopt the evaluation protocols established by EEG2Video [8] and CineBrain [2] to maintain com-

parative consistency. At the frame level, we utilize the Structural Similarity Index (SSIM) [15] to measure pixel fidelity and N-way top-K accuracy via a CLIP-based classifier [10] to assess semantic alignment. To capture temporal dynamics beyond static frame analysis, we incorporate complementary video-level metrics. Specifically, a VideoMAE [14] classifier pre-trained on Kinetics-400 [4] evaluates semantic coherence. Finally, we quantify temporal consistency and motion smoothness using the Fréchet video distances [7] and feature-based consistency [1, 10].

3.1. Structural Similarity Index (SSIM)

The Structural Similarity Index (SSIM) evaluates the perceptual quality of reconstructed frames by comparing their structural information against the ground truth. This metric models visual perception by explicitly analyzing three components: luminance, contrast, and structure. Given two corresponding local image patches x (ground truth) and y (reconstruction), the index is calculated as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (1)$$

where μ_x and μ_y denote the mean pixel intensities, σ_x^2 and σ_y^2 represent the variances, and σ_{xy} is the covariance between x and y. The terms $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$ serve as stabilization constants, with L representing the dynamic range of pixel values. We report the final metric as the mean SSIM computed across all image windows.

3.2. Peak Signal-to-Noise Ratio (PSNR)

Peak Signal-to-Noise Ratio (PSNR) serves as a standard metric for quantifying image reconstruction quality. It measures the ratio of maximum signal power to corrupting noise power on a logarithmic scale, utilizing the Mean Squared Error (MSE) between the reconstruction and the ground truth. Given a ground truth image I and reconstructed image K of size $M \times N$, we compute the metric:

$$\text{MSE} = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [I(i,j) - K(i,j)]^2 \quad (2)$$

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{L^2}{\text{MSE}} \right), \quad (3)$$

where L denotes the maximum pixel intensity (*e.g.*, $L = 255$ for 8-bit formats). Measured in decibels (dB), higher values indicate superior reconstruction fidelity.

3.3. Fréchet Video Distances

Fréchet Video Distances (FVD) and Fréchet Video Motion Distances (FVMD) assess **video generation quality** by quantifying the distributional discrepancy between generated and real video feature sets. Assuming multivariate Gaussian distributions for these features, both metrics employ the **Fréchet distance**:

$$\text{FD} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}), \quad (4)$$

where μ and Σ denote the mean vectors and covariance matrices, respectively, while Tr represents the matrix trace. The metrics diverge in their feature extraction methods:

- FVD (Fréchet Video Distance) leverages spatio-temporal features from a pre-trained 3D CNN (*e.g.*, I3D) to assess general visual coherence and semantic consistency.
- FVMD (Fréchet Video Motion Distance) utilizes explicit motion features (*e.g.*, velocity and acceleration fields) to specifically isolate and validate motion dynamics.

For both metrics, lower scores indicate higher fidelity to the real data distribution.

3.4. N-way top-K Accuracy on Frames and Videos

This metric quantifies semantic accuracy across both individual frames and complete video sequences. For a given sample, we predict the class from a candidate pool S_N containing N categories: the ground truth and $N-1$ distractor classes. A prediction is deemed correct if the ground-truth label ranks within the top-K probabilities output by the classifier. Across a test set of magnitude M, we compute the accuracy as:

$$\text{Acc}_{N,K} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}(y_i \in \text{TopK}(\hat{P}_i(S_N))), \quad (5)$$

where y_i denotes the ground-truth label, while $\hat{P}_i(S_N)$ represents the predicted probability distribution over the class set S_N. The function $\text{TopK}(\cdot)$ extracts the K highest-probability classes, and $\mathbb{I}(\cdot)$ serves as the indicator function. Following the setup of SEED-DV and CineBrain, our experiments report 40-way/2-way top-1 accuracy scores for SEED-DV, and 50-way/2-way top-1 accuracy scores for CineBrain to gauge the semantic fidelity of the reconstructions. For frame-level analysis, we fine-tune a pre-trained CLIP encoder on ground-truth images for 40-way/50-way classification, adhering to the protocol in Gao et al. [2], Liu et al. [8]. Similarly, video-level assessment utilizes a VideoMAE model fine-tuned on ground-truth clips, maintaining consistency with the established benchmark [2, 8].

3.5. Feature-based Temporal Consistency

DINO Temporal Consistency (DTC) and **CLIP Temporal Consistency (CTC)** evaluate the semantic and structural fidelity of reconstructed frames across the temporal dimension. Unlike pixel-based metrics, these indices quantify the alignment of high-level feature representations between the reconstructed sequence and the ground truth. Let $\{x_t\}$ and $\{y_t\}$ denote the ground truth and reconstructed frame

sequences, respectively, and $\mathcal{F}$ represent a feature extractor (DINO [1] or CLIP [10]). We define the consistency at time step t as the cosine similarity:

$$\text{TC}_t = \frac{\mathcal{F}(x_t) \cdot \mathcal{F}(y_t)}{\|\mathcal{F}(x_t)\|\|\mathcal{F}(y_t)\|}. \quad (6)$$

The final metric reports the mean similarity score averaged over all frames in the sequence. DTC utilizes DINO's self-supervised features to capture dense, local structural details, whereas CTC employs CLIP's multi-modal embedding to assess global semantic alignment. In both cases, higher scores indicate superior feature fidelity to the ground truth.

4. Comparison Baselines

We benchmark our model against a diverse range of EEG decoding architectures. The selection spans classic convolutional networks, hybrid Transformer architectures, and recent generative frameworks.

ShallowNet and DeepNet [11] are CNN architectures for EEG decoding. **ShallowNet** utilizes temporal and spatial convolutional layers to decode frequency power information. In contrast, **DeepNet** is a VGG-inspired network extracting features through stacked convolutional blocks.

EEGNet [5] is a lightweight CNN architecture tailored for Brain-Computer Interfaces (BCIs). It employs depthwise and separable convolutions to extract spatiotemporal features with minimal parameters, facilitating generalization across paradigms.

Conformer [12] integrates Convolutional Neural Networks with Transformer mechanisms. The model utilizes convolutions for local feature extraction and multi-head self-attention to capture global temporal correlations.

TSConv (Temporal-Spatial Convolutional Network) [13] represents a specialized CNN framework. It employs distinct convolutional blocks to explicitly decouple temporal and spatial feature learning.

GLMNet [8] serves as a primary generative baseline. This network extracts high-level semantic features from EEG signals to drive visual reconstruction.

GLFA (General Linear Model Fitting Approach) [6] constitutes a standard linear baseline for neuroimaging tasks. It maps brain activity directly to visual features via linear regression. This method provides a non-deep learning reference point for evaluating neural network performance.

CineSync-EEG [2] is a uni-modal adaptation of the CineSync framework, which originally utilizes a Multi-Modal Fusion Encoder and a diffusion decoder. We isolate the EEG encoding branch to evaluate reconstruction capabilities that rely on high-temporal-resolution EEG features.

5. Computational Complexity

We analyze the computational efficiency of our model by comparing parameter counts and FLOPs against the respective baselines. In Tab. 1, we compare two dataset-specific configurations, **DynaMind-SEED** and **DynaMind-CINE**, with their respective benchmarks, EEG2Video and CineBrain:

- **DynaMind-SEED**: This configuration reduces the model size by approximately 19.53% compared to EEG2Video's 1903.24 M parameters. Despite this reduction, it achieves superior performance, showing 12.5% and 19.7% improvements in video-based accuracy and FVMD, respectively.
- **DynaMind-CINE**: This version achieves a parameter reduction of approximately 740 M and a FLOPS reduction of over 206 G. It maintains competitive performance, yielding a 4.28% increase in video-based accuracy and a 4.19% increase in FVD over the baseline.

Furthermore, the table details the computational complexity of individual modules. The results indicate that the RSM incurs notable computational costs due to the diffusion priors, while the DGVR accounts for the majority of parameters and computational resources, driven by the use of large-scale generative models for reconstruction. Overall, our model outperforms the baselines while maintaining superior computational efficiency.

Model	# Params (M)	GFLOPS
EEG2Video (Baseline)	**1903.24**	677.81
DynaMind-SEED (Ours)	1531.58	**789.60**
- *RSM*	428.93	114.29
- *TDA*	26.57	0.21
- *DGVR*	1076.08	675.10
Cinebrain (Baseline)	**6236.36**	**8214.55**
DynaMind-CINE (Ours)	5494.68	8008.54
- *RSM*	437.64	114.80
- *TDA*	28.60	0.24
- *DGVR*	5028.44	7893.50

Table 1. Comparison of total parameters and flops for optimized DynaMind versions against baselines (M: Millions).

6. Ablation Study

To evaluate the contribution of specific architectural components, we perform an ablation analysis on the SEED-DV dataset, isolating the impact of individual modules. As detailed in Table 2, we examined the model configurations under two restricted conditions: (1) removing the simulated neural pathway of "Dorsal Stream: Occipital → Parietal" and "Ventral Stream: Occipital → Temporal" (w/o Pathways), which serves as the gating network, and (2) excluding the dynamic temporal blueprint from the TDA (w/o TDA) into the generative backbone. We choose 40-

Method	Acc ↑	FVMD ↓	SSIM ↑
Full Model	**0.284**	**1637.55**	**0.280**
w/o Pathways	0.269	1688.51	0.275
w/o TDA	0.278	1966.30	0.269

Table 2. Ablation study on the effectiveness of key model components on the SEED-DV dataset. We report performance metrics by removing specific modules from our full model.

way video classification accuracy (Acc) for semantic correctness, Fréchet Video Motion Distance (FVMD) to measure temporal motion coherence, and SSIM to gauge the structural fidelity of the generated frames. The empirical results indicate that the Full Model achieves superior performance across all indices. Specifically, removing the pathways causes the accuracy to drop from 0.284 to 0.269, confirming that the neuroscience-prior-based gating mechanism aids in semantic decoding. Furthermore, the ablation of latent representations leads to a sharp increase in FVMD to 1966.30 and a decline in SSIM to 0.269, suggesting that the temporal alignment is essential for maintaining motion smoothness and image quality in the video reconstruction.

7. More Visualizations

To complement the quantitative analysis, we include supplementary qualitative comparisons that substantiate the reliability of our approach. These visualizations illustrate the model's capacity to generate semantically accurate and temporal coherent videos across a broader range of samples.

7.1. More Successful Examples on SEED-DV

As illustrated in Figs. 1 and 3, we present a qualitative evaluation comparing our reconstruction results (**Ours**) against the Ground Truth (**GT**) and the baseline method EEG2Video on the SEED-DV dataset. These supplementary examples cover a diverse range of semantic categories, including: These supplementary examples span a diverse array of semantic categories, specifically:

- **Human-centric Scenarios:** Including social interactions (*e.g.*, children walking) and specific skills (*e.g.*, musical performance, snowboarding).
- **Nature-related Scenarios:** Covering fauna (*e.g.*, pandas, jellyfish) and flora (*e.g.*, flower fields).
- **Man-made Environments:** Encompassing urban architecture (*e.g.*, city skylines), transportation (*e.g.*, moving vehicles), and objects like hot air balloons.

In complex static scenes, our model accurately captures the primary semantic objects; for instance, in the sequence depicting "pandas eating bamboo," the distinct black-and-white coat pattern and the feeding action are clearly preserved. Similarly, for the "musician playing guitar" sample, the model reconstructs the performer's posture and the instrument's structure with high fidelity. Regarding temporal dynamics, the results demonstrate superior motion consistency; in high-speed scenarios such as "cars driving on highways," our approach maintains the vehicle's trajectory without the severe blurring or flickering artifacts often observed in baseline methods.

Across these varied domains, the generated videos exhibit strong alignment with the ground truth in terms of color saturation (evident in the jellyfish sample) and structural integrity. This visual evidence, combined with the quantitative metrics reported earlier, validates the reliability of our model in decoding time-varying visual stimuli from non-invasive EEG signals, effectively bridging the gap between neural signals and pixel-level reconstruction.

7.2. More Successful Examples on CineBrain

As displayed in Figs. 2 and 4, we present a comparative evaluation of our reconstruction performance against the CineSync-EEG baseline on the CineBrain dataset. Derived from the sitcom *The Big Bang Theory*, this benchmark centers on character interactions and dialogue-heavy sequences. However, the reconstruction task is complicated by visually dense backgrounds (*e.g.*, crowded store shelves) and varying indoor illumination. In contrast to the baseline, which often suffers from severe identity loss and blurring artifacts in these cluttered settings, our method maintains higher fidelity. As evidenced in the samples, our model successfully recovers specific semantic attributes, such as the yellow graphic t-shirt and facial characteristics, ensuring that the reconstructed subjects remain recognizable despite the complex surroundings.

7.3. Some failure samples

Although our model performs well in most cases, analyzing its failure cases allows us to more clearly identify its limitations. As shown in Fig. 5, we categorize the limitations observed on both datasets into two distinct types:

- **Incoherent Object Motion:** In dynamic sequences such as the "motorcycle" example, the model successfully retrieves the correct semantic category (a motorbike) but struggles to align the object's physical orientation with its motion trajectory. This results in a reconstructed video where the vehicle appears to slide laterally or rotate unnaturally, indicating a disconnect between the semantic features and the temporal physics modeling.
- **Semantic Hallucination in Low-Texture Scenes:** The "skiing" case illustrates a specific type of semantic error where the model confuses visually similar activities. Due to the shared body pose between skiing and skateboarding, the decoder hallucinates a "skateboarding on a ramp" scenario. This suggests that when explicit background cues are sparse, the model over-relies on learned priors from the pre-training data rather than the EEG signal.

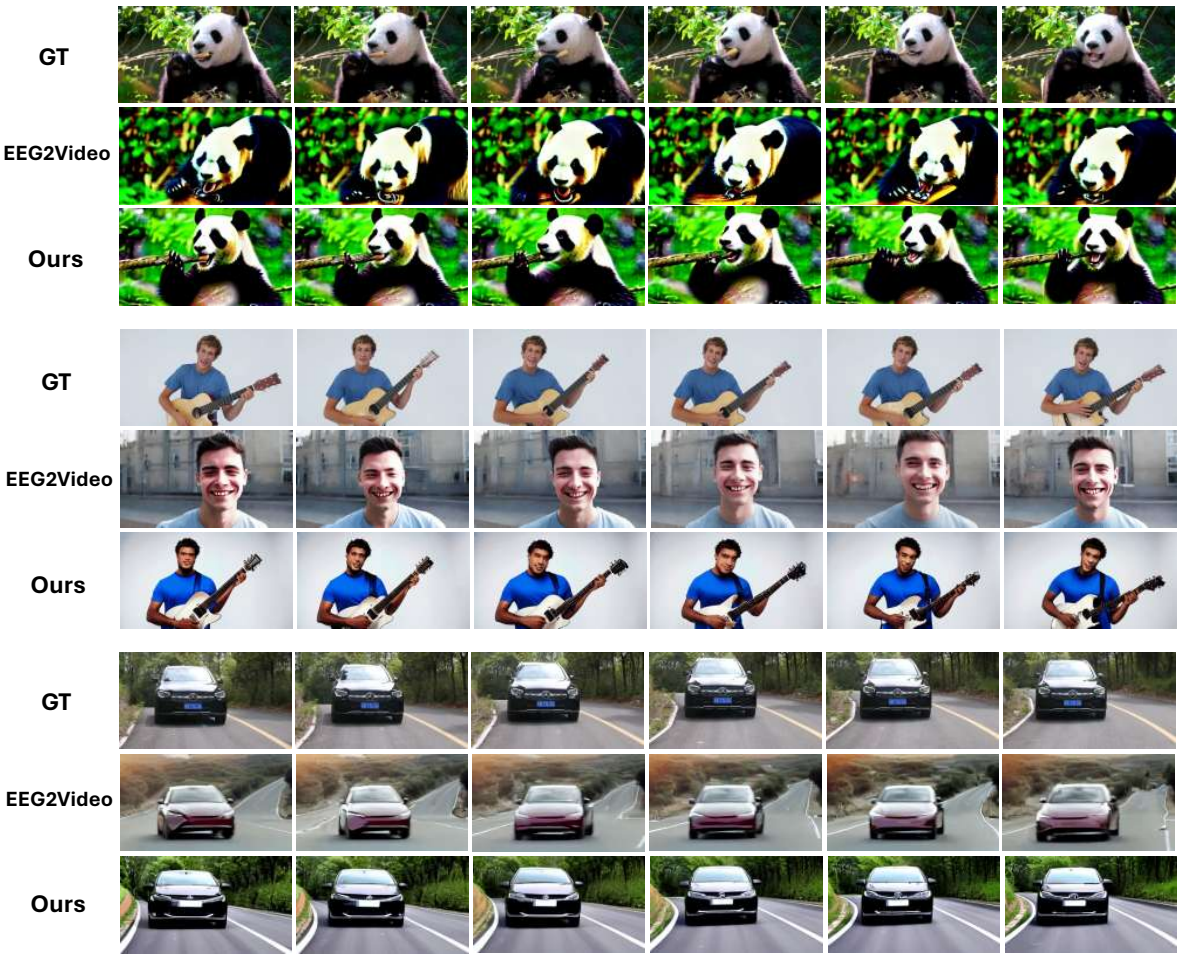

Figure 1. Reconstruction samples compared with EEG2Video on SEED-DV.

Figure 2. Reconstruction samples compared with CineSync-EEG on CineBrain.

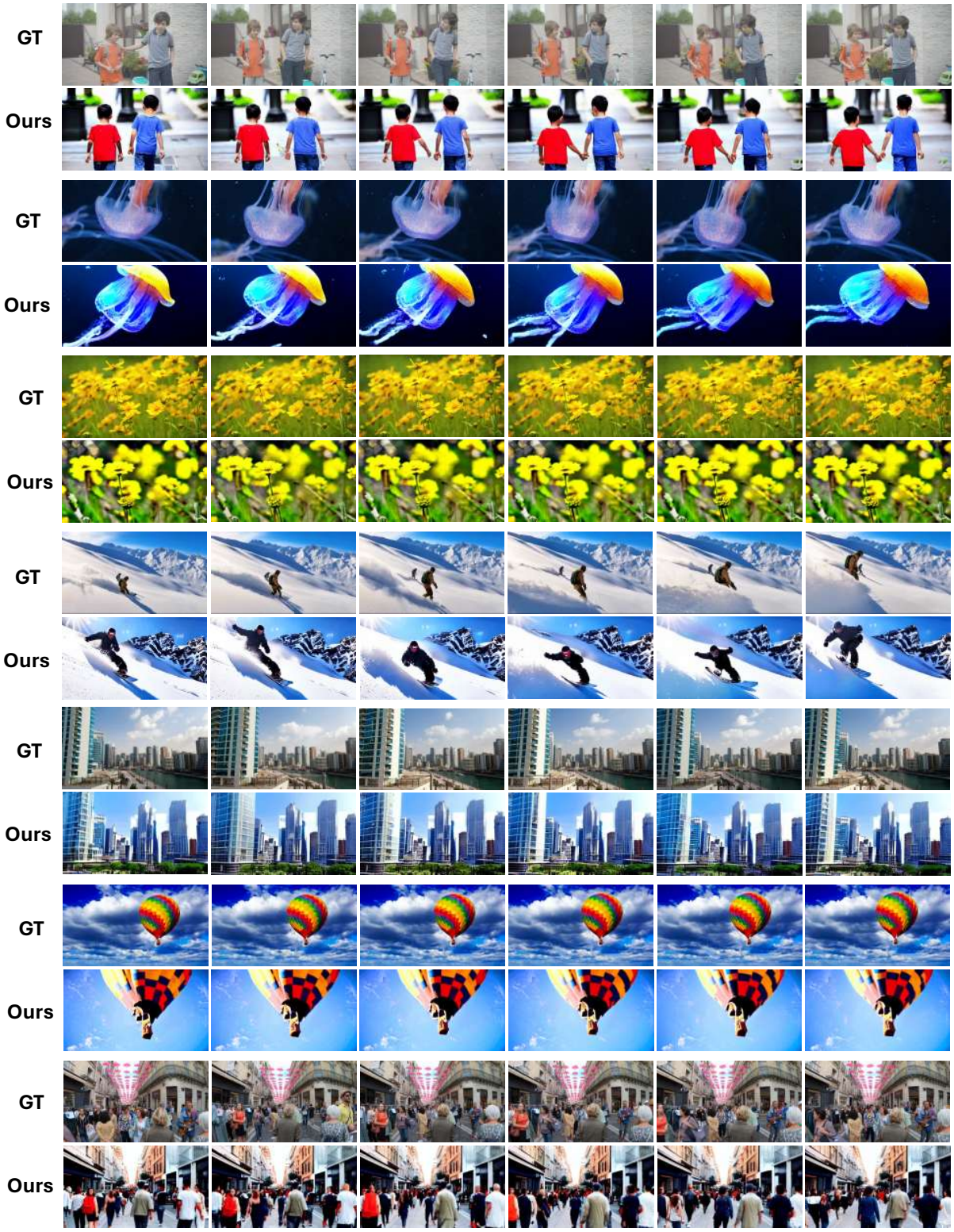

Figure 3. Various videos reconstruction samples on SEED-DV.

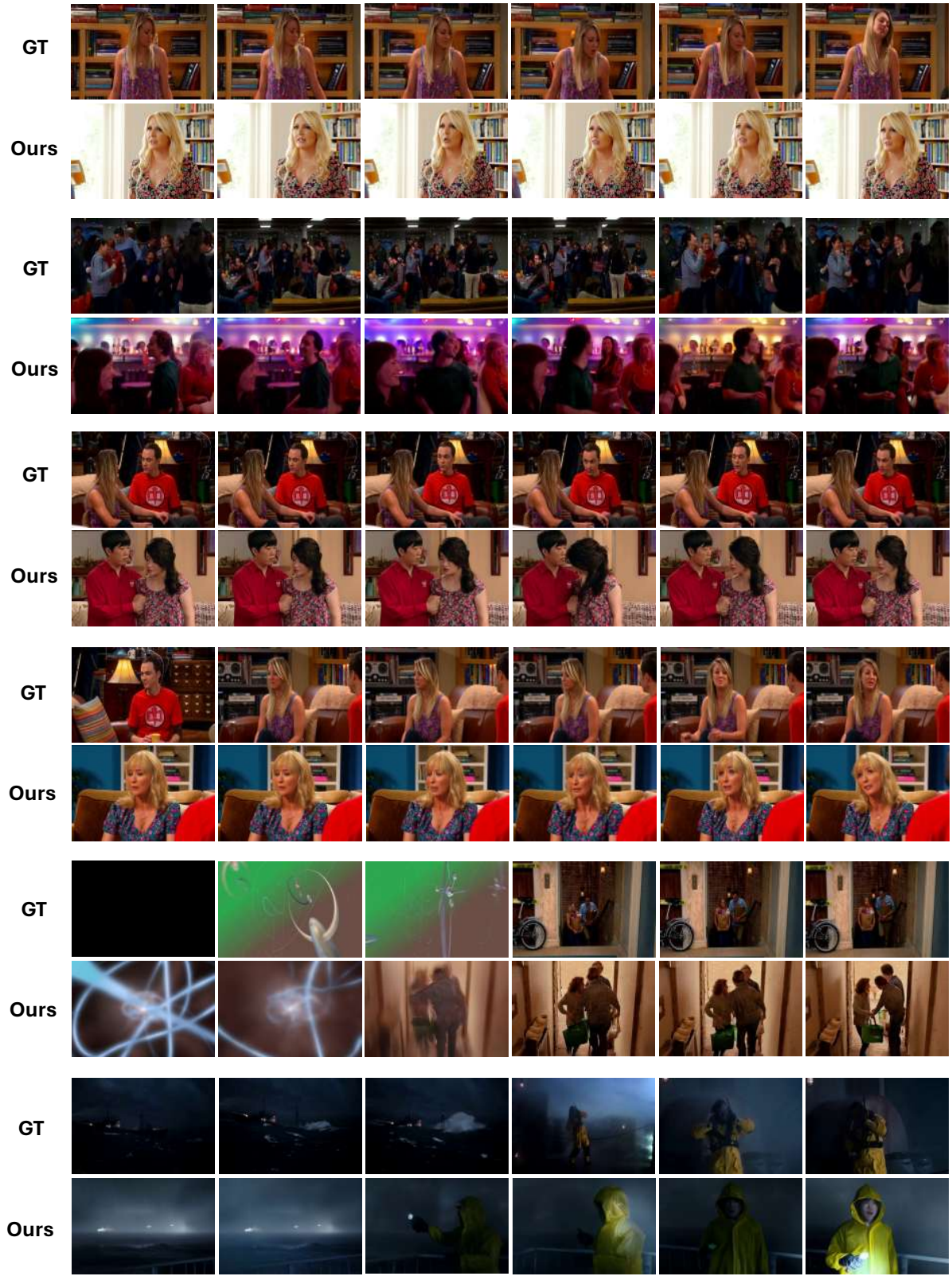

Figure 4. Various videos reconstruction samples on CineBrain.

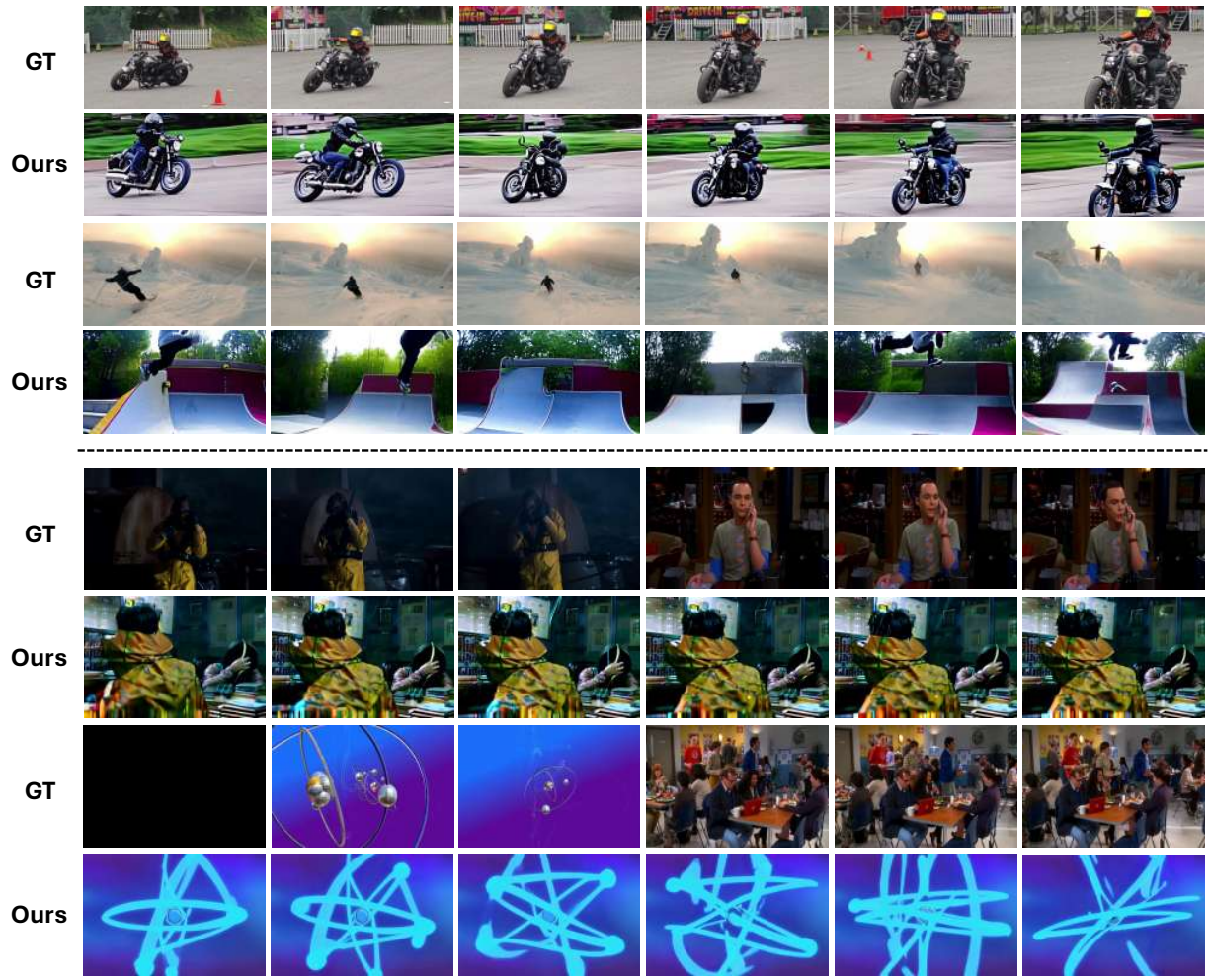

Figure 5. Some failure samples on both datasets.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 3

[2] Jianxiong Gao, Yichang Liu, Baofeng Yang, Jianfeng Feng, and Yanwei Fu. Cinebrain: A large-scale multi-modal brain dataset during naturalistic audiovisual narrative processing. *arXiv preprint arXiv:2503.06940*, 2025. 1, 2, 3

[3] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016. 1

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[5] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018. 3

[6] Chong Li, Xuelin Qian, Yun Wang, Jingyang Huo, Xiangyang Xue, Yanwei Fu, and Jianfeng Feng. Enhancing cross-subject fmri-to-video decoding with global-local functional alignment. In *European Conference on Computer Vision*, pages 353–369, 2024. 3

[7] Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fr\'echet video motion distance: A metric for evaluating motion consistency in videos. *arXiv preprint arXiv:2407.16124*, 2024. 2

[8] Xuan-Hao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-Liang Lu, and Wei-Long Zheng. Eeg2video: Towards decoding dynamic visual perception from eeg signals. *Advances in Neural Information Processing Systems*, 37:72245–72273, 2024. 1, 2, 3

[9] Xuyu Qian, Kyle Coleman, Shunzhou Jiang, Andrea J Kriz, Jack H Marciano, Chunyu Luo, Chunhui Cai, Monica Devi Manam, Emre Caglayan, Abbe Lai, et al. Spatial transcriptomics reveals human cortical layer and area specification. *Nature*, pages 1–11, 2025. 1

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021. 2, 3

[11] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017. 3

[12] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022. 3

[13] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*, 2023. 3

[14] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2

[15] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2

[16] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 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*, pages 7623–7633, 2023. 1