Modeling Spatiotemporal Neural Frames for High Resolution Brain Dynamics

Supplementary Material

1. Data Preprocessing

For CineBrain dataset, the fMRI preprocessing follows the CineBrain protocol and is performed using the standard fMRIPrep pipeline. The main steps include motion correction, susceptibility distortion correction, slice-timing correction, co-registration to the T1-weighted structural image, and normalization to the fsLR-32k grayordinate space defined by the Human Connectome Project (HCP). After preprocessing, each fMRI frame is represented by 91282 grayordinates, consisting of 32492 cortical vertices per hemisphere and 26298 subcortical gray-matter voxels covering thalamus, striatum, hippocampus, and other deep gray structures.

EEG data are acquired using an MRI-compatible 64-channel cap at a sampling rate of 1000 Hz, with simultaneous recording of ECG signals. Synchronization between EEG and fMRI is ensured by logging the exact fMRI TR timings during acquisition. EEG preprocessing follows the CineBrain protocol, with modifications to accommodate our recording configuration. A multi-step artifact removal procedure is applied to suppress scanner-induced and physiological noise while preserving neural activity. The preprocessing pipeline includes bandpass filtering between 0.1 Hz and 75 Hz to remove baseline drift and muscle artifacts, and a 50 Hz notch filter to attenuate powerline interference. ECG artifacts are first reduced using QRS-based correction methods, followed by independent component analysis (ICA) to isolate and remove residual artifacts. The recorded ECG signals are further used to refine artifact rejection through correlation-based adaptive adjustment, resulting in clean EEG data suitable for subsequent analysis.

2. Subject-wise Dynamic fMRI Reconstruction

To complement the averaged results reported in the main paper, we provide a full breakdown of dynamic fMRI reconstruction performance for all six subjects in the CineBrain dataset. Tables 1 2 3 report the Mean Squared Error (MSE), Pearson correlation coefficient (r), and cosine similarity across three spatial regions: the whole brain (91282 vertices), the primary visual cortex V1 (8405 vertices), and the combined visual and auditory cortex V1+A1 (18946 vertices). Results are provided under three temporal window lengths of 3, 10, and 30 frames.

Furthermore, we report subject-wise performance for the InterRecon task. While the main paper provides results for Subject 1, here we include results for all six subjects.

Table 1. Subject-wise dynamic fMRI reconstruction performance (3-frame).

Subject	MSE	r	Cos
sub1	0.272 ± 0.07	0.839 ± 0.04	0.852 ± 0.03
sub2	0.189 ± 0.03	0.865 ± 0.03	0.890 ± 0.03
sub3	0.285 ± 0.03	0.843 ± 0.06	0.850 ± 0.06
sub4	0.346 ± 0.09	0.781 ± 0.08	0.815 ± 0.06
sub5	0.288 ± 0.08	0.812 ± 0.04	0.839 ± 0.05
sub6	0.314 ± 0.149	0.794 ± 0.06	0.834 ± 0.06
avg	0.282	0.822	0.847

Table 2. Subject-wise dynamic fMRI reconstruction performance (10-frame).

Subject	MSE	r	Cos
sub1	0.271 ± 0.05	0.840 ± 0.03	0.853 ± 0.03
sub2	0.188 ± 0.02	0.865 ± 0.03	0.890 ± 0.03
sub3	0.284 ± 0.09	0.843 ± 0.04	0.850 ± 0.04
sub4	0.342 ± 0.19	0.780 ± 0.07	0.816 ± 0.06
sub5	0.275 ± 0.05	0.819 ± 0.03	0.846 ± 0.05
sub6	0.303 ± 0.11	0.797 ± 0.04	0.838 ± 0.04
avg	0.277	0.824	0.849

Table 3. Subject-wise dynamic fMRI reconstruction performance (30-frame).

Subject	MSE	r	Cos
sub1	0.283 ± 0.03	0.832 ± 0.02	0.846 ± 0.02
sub2	0.196 ± 0.01	0.857 ± 0.02	0.883 ± 0.02
sub3	0.290 ± 0.07	0.841 ± 0.03	0.849 ± 0.04
sub4	0.344 ± 0.10	0.769 ± 0.04	0.807 ± 0.04
sub5	0.280 ± 0.04	0.814 ± 0.02	0.842 ± 0.04
sub6	0.294 ± 0.104	0.802 ± 0.03	0.841 ± 0.04
avg	0.281	0.819	0.845

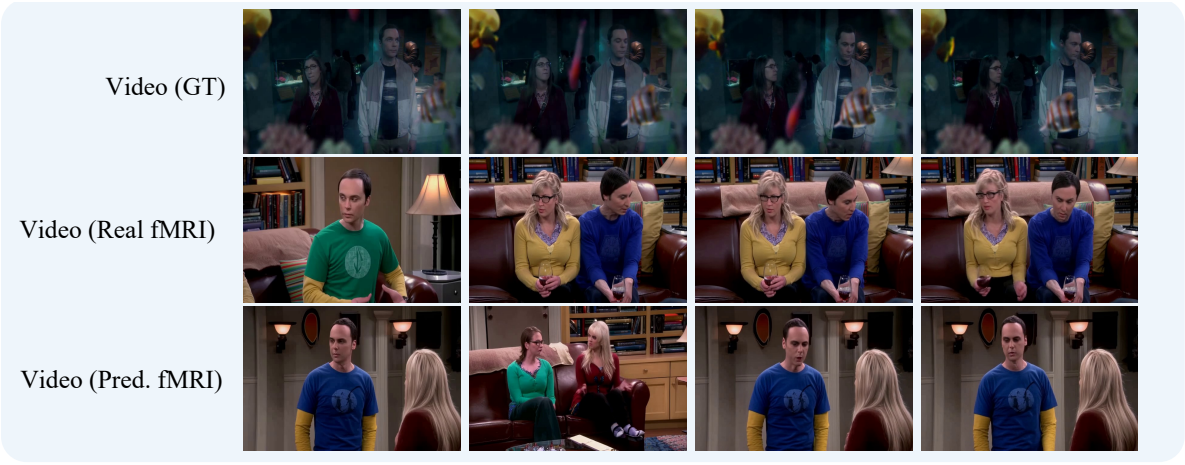

Figure 1. Functional validation through visual decoding. Comparison among the original stimulus video frames Video (GT), the frames decoded from ground-truth fMRI Video (Real fMRI), and the frames decoded from our reconstructed fMRI Video (Pred. fMRI) using the CineSync-f decoder. These three-way visual comparisons further illustrate the preservation of semantic content in the reconstructed fMRI.

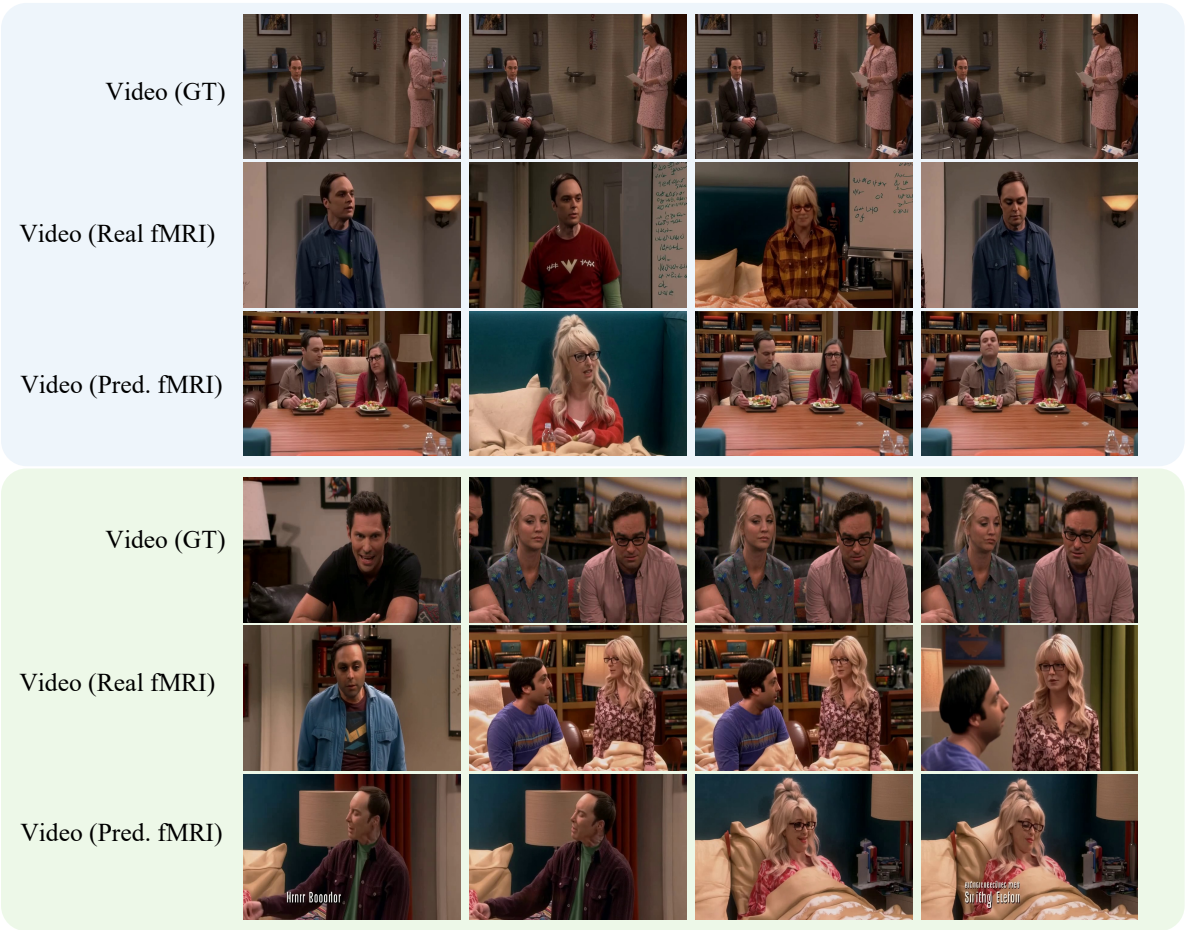

Figure 2. Functional validation through visual decoding. Comparison among the original stimulus video frames Video (GT), the frames decoded from ground-truth fMRI Video (Real fMRI), and the frames decoded from our reconstructed fMRI Video (Pred. fMRI) using the CineSync-f decoder. These three-way visual comparisons further illustrate the preservation of semantic content in the reconstructed fMRI.

Table 4. Ablation on latent dimensionality.

	LinearAE Reconstruction	EEGtofMRI		
dim	**MSE↓**	**MSE↓**	**r↑**	**Cos↑**
512	0.0045	0.288	0.814	0.830
1024	0.0043	**0.282**	**0.822**	**0.847**
2048	**0.0042**	0.285	0.820	0.843

3. Additional Video Reconstruction Results

As shown in fig 1 and fig 2, we include additional visual comparisons of the original videos, the videos decoded from real fMRI signals, and those decoded from the reconstructed fMRI. These supplementary results provide further support for the quantitative and qualitative analyses discussed in the main manuscript.

Across diverse scenes and character configurations, the videos decoded from our predicted fMRI closely match those decoded from real fMRI. As shown in the figures, the decoding results based on reconstructed fMRI faithfully recover the overall scene layout, camera viewpoint, and background structure, while also capturing key aspects of the characters' appearance, posture, interactions, and contextual semantics. Whether in indoor conversations, multi-person interactions, or rapidly changing scenes, the semantic content decoded from predicted fMRI is nearly equivalent to that obtained from real fMRI, indicating that our reconstructed fMRI successfully preserves the functional information required for neural visual decoding. In these varied scenarios, the EEG-driven fMRI reconstructions exhibit stable, high-level semantic representations that can be effectively read out by the visual decoder, yielding video outputs that are comparable to those produced using real fMRI. This functional consistency further demonstrates that the reconstructed fMRI carries meaningful representational value and can serve as a reliable substitute for real fMRI in visual cognition analyses.

4. Additional Ablation on Latent Dimensionality

To further examine whether the linear autoencoder introduces a bottleneck effect, we perform additional ablation experiments with latent dimensionalities of 512, 1024, and 2048. As shown in Table 4, the downstream EEG-to-fMRI reconstruction performance remains stable across different latent dimensions, indicating that the proposed method is robust to this design choice and that the linear autoencoder is not the dominant limiting factor.