A Multi-stage Validation Framework for Cross-Subject EEG-to-Video Reconstruction

Ahhyun Lucy Lee   Minchan Kim   Shakhnoza Khojimatova   Wooseok Lee

Seoul National University
{ahhyun724, mmm5373, shakhnoza1, andylws}@snu.ac.kr

Abstract

Decoding dynamic visual perception from EEG has recently drawn increasing attention, driven by the release of new datasets and benchmarks. However, current EEG-to-video reconstruction approaches lack systematic validation frameworks for cross-subject generalization, a critical requirement for practical applications. To address this limitation, we introduce a comprehensive three-stage validation framework that systematically evaluates cross-subject EEG-to-video reconstruction across critical dimensions. First, we perform dataset integrity assessment through statistical analyses including linear mixed-effects models and mutual information analysis to validate EEG-video paired data consistency. Second, we evaluate encoder robustness by testing existing architectures under cross-subject conditions and propose two alternative encoder designs: CBraMod [33] and triplet loss-based contrastive learning encoders specifically designed for subject-invariant representations. Third, we conduct reconstruction fidelity validation through step-by-step module assessment of existing EEG-to-video approaches. Our framework reveals significant cross-subject performance variations and establishes systematic evaluation protocols for future EEG-to-video research. These results highlight the importance of comprehensive validation in neural decoding and provide a foundation for developing more robust cross-subject reconstruction systems.

1. Introduction

Human perceptions become both mirrors and mosaics, when we gaze upon the same scene. On one hand, the rhythms of our neural activity echo each other-our brains light up in harmonious patterns as we process the world's images [9]. Yet, within this shared symphony, each mind weaves its own melody, layering personal memories, emotions, and association atop the common neural score [22]. This universality and uniqueness of human visual perception often hampers the generalization of neuroscience research findings across participants.

This study seeks to enhance cross-subject generalizability in decoding dynamic visual experiences from EEG signals. Decoding, in this context, refers to the prediction or reconstruction of sensory information-such as visual stimuli-from brain signals [15]. Vision is one of the most fundamental senses for humans [10]. Dynamic visual environments have played a crucial role in human evolution, as the ability to rapidly encode and interpret changing scenes was essential for survival. Videos, as stimuli, are particularly attractive for decoding studies because they closely reflect the dynamic nature of real-world visual experiences.

To capture brain responses while viewing dynamically changing videos, functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) can be used. While both of them are non-invasive, fMRI captures hemodynamic flow while EEG captures electrical signals generated in the brain. While fMRI has been widely used for vision decoding [2, 5, 26] [27], its low temporal resolution (typically around 0.5 Hz) limits its ability to capture rapid changes in visual perception. In contrast, EEG offers millisecond-level temporal resolution (up to 1000 Hz), making it well-suited for decoding fast, dynamic visual processes [16, 19]. Recent advances have begun to address the challenges of EEG-based video decoding. Notably, Liu et al. [19] introduced the SEED-DV dataset and the EEG2Video model, initiating the first exploration on reconstructing dynamic visual experiences from EEG data. However, current EEG-based decoding models face several significant limitations across three critical dimensions:

First, regarding dataset integrity challenges, a critical gap remains in validating whether EEG-video datasets can establish meaningful connections between EEG signals and visual content at the fundamental level. Without demonstrating that EEG-vision relationships can be reliably captured for basic visual stimuli, the validity of dynamic video reconstruction approaches remains questionable.

Second, encoder robustness limitations persist as poor cross-subject generalization continues to affect performance, where models trained on one participant's data exhibit substantial performance degradation when applied to others due to individual neural differences.

Third, reconstruction validation gaps exist as current approaches lack end-to-end architectural design coupled with insufficient validation of individual modular components, making it difficult to identify bottlenecks and optimize performance systematically.

This study aims to establish a systematic validation framework for cross-subject EEG-to-video reconstruction by (1) assessing dataset validity of EEG-video paired datasets, (2) evaluating existing encoder architectures and proposing alternative designs to improve cross-subject robustness, and (3) validating reconstruction quality through comprehensive performance assessment. Through this comprehensive validation approach, we seek to establish reliable evaluation protocols and identify key factors that influence cross-subject generalization in EEG-to-video reconstruction.

2. Related Works

2.1. EEG Decoding

EEG-conditioned visual decoding has evolved along two main lines: the *imagination-driven* generation—where subjects merely imagine a concept—and the *stimulus-driven* reconstruction—where an external image is observed.

In the imagination regime, ThoughtViz [32] achieves strong Top-1 accuracies on three 10-class datasets, substantially outperforming random baselines. EEG2Image [28] improves IS by 25% over ThoughtViz and doubles k-means cluster purity. GWIT [20] scales to 40 classes with significant accuracy improvements over BrainVis.

Stimulus-driven pipelines leverage explicit visual input and show more impressive results. NeuroGAN [21] reconstructs ImageNet subsets with improved IS and tighter class-diversity. BrainDecoder [3] attains high Top-1 accuracy on 50-class datasets, far exceeding random baselines. BrainDreamer [34] injects language guidance via triple contrastive learning, achieving substantial EEG classification improvements. LowDensityEEG [6] demonstrates that eight-channel headsets can achieve reasonable classification accuracy but performance drops significantly on unseen classes. Recent work extends to 3D object reconstruction [7] and large-scale zero-shot recognition [30].

Modern approaches like ATM-CLIP [15], Perceptogram [4], and DVRM [23] employ CLIP alignment and diffusion models to bridge the neural–visual gap. These designs echo the two-step structure of our own inflated 3-D diffusion UNet for video.

Performance varies significantly across datasets: models that excel on small, stimulus-rich benchmarks often underperform on complex imagination corpora. Some studies showcase convincing visuals yet yield modest quantitative metrics, exposing the gap between perceptual plausibility and semantic fidelity.

Nevertheless, the maturity of single-frame decoding suggests that tackling temporally coherent video reconstruction is a natural next step. Building on these insights, our work extends diffusion-based decoders to EEG-to-video synthesis, addressing the temporal dimension that has remained largely unexplored in neural signal decoding.

Recent EEG-to-video approaches have shown promising initial results. EEG2Video [19] introduced the first EEG-to-video generation model using a Seq2Seq-based architecture with dynamic noise adding, achieving reasonable semantic accuracy on the SEED-DV dataset. NEVER [12] proposed a dual-branch encoder with perception and semantic understanding modules, explicitly addressing cross-subject generalization challenges by demonstrating performance drops when tested across subjects. However, these early approaches remain limited in their cross-subject generalization capabilities, highlighting the need for more robust architectures that can effectively transfer knowledge across different individuals.

2.2. Cross-subject generalization Efforts

Various decoding models, regardless of modality have explored ways to closely align neural signals from different participants. GLFA [14] proposed a global-local functional alignment framework that maps fMRI signals into a shared latent space. This alignment, combined with spatiotemporal attention and latent diffusion decoding, improves decoding performance across subjects, enabling limited cross-subject generalization. In EEG domain, adaptive deep feature representation learning (ADFR) [17] was suggested as a method to enhance cross-subject generalization performance by adopting EEG feature representation substantiated with regularization technique. The purpose of intense regularization, used in this framework, is to minimize the discrepancy in the distribution between EEG signals from different individuals.

3. Methods

3.1. Preliminaries

EEG-to-video reconstruction fundamentally relies on the assumption that EEG signals contain sufficient information to decode visual content. However, the inherently low signal-to-noise ratio (SNR) of EEG presents challenges in establishing reliable brain-vision relationships, particularly when generalizing across different individuals.

Cross-subject variability represents one of the most significant challenges in EEG-based brain-computer interfaces and neuroscience research. This phenomenon occurs when brain signals recorded from different individuals exhibit substantial differences despite observing identical stimuli. These differences arise from anatomical variations in brain structure, electrode positioning discrepancies, and individual-specific neural processing patterns [11, 18].

To evaluate cross-subject generalization capabilities, Leave-One-Subject-Out (LOSO) cross-validation is commonly employed. In LOSO, models are trained on data from all subjects except one, which serves as the test set. This process is repeated with each subject serving as the test set once, enabling comprehensive assessment of generalization performance across individuals.

Lastly, we try reconstructing video via EEG2Video framework. Formally, the problem of EEG-based video reconstruction can be defined as learning a mapping function $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^{C\times T}$ represents the EEG signal space with C channels and T time samples, and $\mathcal{Y}$ represents the target video space. The challenge lies in making this mapping robust to cross-subject variations.

3.2. Dataset Integrity Assessment

3.2.1. Statistical Analysis

To evaluate the suitability of SEED-DV dataset for EEG-to-video reconstruction task, statistical analysis was performed using Linear Mixed-Effects (LME) modeling with post-hoc pairwise comparisons. Feature extraction was performed using Principal Component Analysis (PCA) of 10 principal components. LME models were fitted with video labels as fixed effects and participants as random effects accounting for individual variability: $\text{Neural Response} \sim \text{Video Label} + (1 \mid \text{Subject})$ This structure allows us to generalize category-related effects across participants while controlling for repeated measures within subjects. Post-hoc pairwise comparisons between all label combinations were conducted using independent t-tests, with effect sizes calculated using Cohen's d. N classification resulted in comparison of $\binom{N}{2}$ pairs. Multiple comparison corrections were applied using both Bonferroni and False Discovery Rate (FDR) methods to control for Type I error inflation across the large number of statistical tests. Statistical significance was assessed at $\alpha = 0.05$ for FDR-corrected comparisons.

3.2.2. Mutual Information

To further evaluate the adequacy of the SEED-DV dataset for EEG-to-video reconstruction, we employed the Mutual Information Neural Estimation (MINE) framework [1], which estimates a lower-bound on the mutual information (MI) between two high-dimensional random variables. MI quantifies the amount of shared information between variables, and thus serves as a principled measure of cross-modal dependency. To contextualize the absolute MI values, we further computed a normalized mutual information (NMI) score, defined as the ratio of estimated MI to the maximum differential entropy of the involved modalities. This normalization provides a scale-invariant measure that allows us to assess the relative amount of information in EEG signals that is predictive of the corresponding video content. Together, MI and NMI offer a quantitative lens for examining whether the EEG and video modalities in the dataset are sufficiently correlated to support downstream tasks such as generative modeling or reconstruction.

3.3. Encoder Robustness Evaluation

To systematically evaluate cross-subject generalization capabilities, we establish a comprehensive testing framework that can assess any EEG encoder architecture across different subject conditions. This framework provides standardized protocols for measuring cross-subject performance degradation and identifying architectural factors that influence generalization. We demonstrate the framework's utility by evaluating multiple encoder architectures: 1) **GLMNet** from the original EEG2Video framework [19], which combines global and local feature processing, 2) **CBraMod encoder** [33], and 3) our proposed **Cross-Subject Contrastive Learning Encoder**. Through systematic comparison across subject conditions using consistent evaluation metrics, we establish validity measurements for cross-subject robustness and provide a reusable evaluation protocol that can be applied to future encoder designs.

3.3.1. CBraMod encoder

The CBraMod encoder [33] is a foundation model architecture specifically designed for EEG signal processing. Unlike traditional EEG encoders that treat all EEG channels equally or separate them into predefined groups, CBraMod employs a criss-cross transformer architecture that models spatial and temporal dynamics separately through two parallel attention mechanisms. This approach captures the heterogeneous nature of EEG signals more effectively. CBraMod was pre-trained on the Temple University Hospital EEG Corpus (TUEG), the largest publicly available EEG dataset, using a masked EEG reconstruction objective. Despite the original model being trained on 19-channel clinical EEG data, it demonstrates strong transfer capabilities to diverse downstream tasks, varying in channel number, task type and length. CBraMod is expected to show superior cross-subject generalization ability since it was trained on a diverse population of subjects, learning to extract fundamental EEG patterns that are consistent across individuals. We evaluate the effectiveness of CBraMod as an encoder by employing transfer learning-freezing the pre-trained weights and fine-tuning only the final layers on the SEED-DV dataset to adapt to the video reconstruction task.

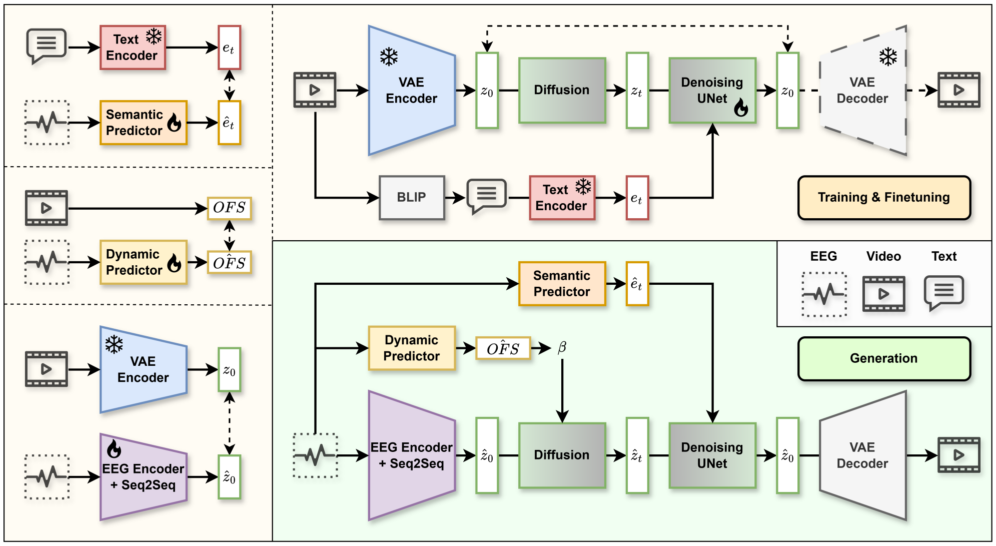

Figure 1. EEG2Video framework overview.

3.3.2. Cross-Subject Contrastive Learning Encoder

EEG signals are inherently subject-dependent, exhibiting significant inter-subject variability even when responding to the same visual stimulus. This poses a major challenge in cross-subject EEG classification tasks, where models trained on one set of subjects often fail to generalize to unseen individuals.

To mitigate this issue, we propose a contrastive learning strategy based on the triplet loss [25], designed to encourage the learning of subject-invariant and semantically meaningful EEG embeddings. Specifically, for each anchor EEG sample x^a, we construct a triplet (x^a, x^p, x^n), where:

- x^p is a **positive** sample from a *different subject* but with the *same class label* as x^a,
- x^n is a **negative** sample from the *same subject* but with a *different class label*.

The triplet loss is formulated as:

$$\mathcal{L}_{\text{triplet}} = \sum_{(x^a,x^p,x^n)\in\mathcal{T}} \Big[\|f(x^a) - f(x^p)\|_2^2 - \|f(x^a) - f(x^n)\|_2^2 + \alpha \Big]_+ \quad (1)$$

where $f(\cdot)$ is the EEG encoder, α is the margin, and $[\cdot]_+$ denotes the hinge function. This formulation promotes alignment of cross-subject embeddings sharing the same label while encouraging separation between embeddings from the same subject with different labels, effectively disentangling subject-specific noise from task-relevant features.

To ensure accurate classification while enforcing this structured embedding space, we jointly optimize a standard cross-entropy classification loss $\mathcal{L}_{\text{cls}}$ alongside the triplet loss. The total loss is given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{triplet}} \quad (2)$$

where λ controls the contribution of the contrastive regularization. This combined objective enables the model to retain strong discriminative ability while improving generalization across subjects.

3.4. Reconstruction Fidelity Validation

To assess reconstruction fidelity, we employ the EEG2Video framework [19] as a representative reconstruction pipeline for systematic evaluation. Specifically, the framework includes an EEG Encoder combined with a Seq2Seq model, a semantic predictor, a dynamic predictor, and a Denoising UNet.

EEG Encoder and Seq2Seq model are trained to generate latent representations ($\hat{z}_0$) from EEG signals. These representations are aligned with the corresponding video embeddings (z_0) obtained via a pretrained VAE encoder applied to the target video. The training objective minimizes the discrepancy between EEG-derived embeddings and video embeddings to ensure meaningful representation alignment. The pretrained Denoising UNet is finetuned on the SEED-DV dataset videos, enabling it to effectively reconstruct realistic video sequences during the diffusion-based generation process. Additionally, the semantic predictor, implemented as a simple multilayer perceptron (MLP), is trained to predict semantic embeddings ($\hat{e}_t$) from EEG data. The target semantic embeddings (e_t) are obtained from text descriptions processed by a pretrained text encoder. The semantic predictor's training objective minimizes the difference between EEG-derived embeddings and text embeddings, thus providing semantic guidance to the video reconstruction pipeline. The dynamic predictor, employing a GLMNet architecture, estimates the Optical Flow Score (OFS) from EEG signals. It is trained to produce OFS values closely matching the true optical flow derived from the corresponding videos. During generation, this dynamic predictor modulates the latent noise in accordance with the desired video dynamics, enabling control over the temporal realism of the generated videos. Ultimately, these trained and fine-tuned components collectively facilitate EEG-driven video generation: the EEG Encoder and Seq2Seq model produce initial latent representations for the diffusion process, the semantic predictor informs semantic aspects within the Denoising UNet, and the dynamic predictor dynamically adjusts latent noise, culminating in temporally and semantically coherent video reconstructions.

4. Experiments

4.1. Dataset and Benchmark

We utilize the SEED-DV (SJTU EEG Dataset for Dynamic Vision) and the EEG-VP benchmark [19] for video reconstruction and visual perception classification tasks.

4.1.1. SEED-DV

EEG signals were recorded from 20 healthy participants (10 males and females, mean age 21.75 years) using a 62-channel EEG cap, sampled at 200 Hz. During the experiments, subjects watched a series of color video clips while their EEG signals were recorded simultaneously. Each recording session consisted of 7 video blocks, where each block contained 200 two-second video clips spanning 40 semantic classes (5 clips per class). In total, the dataset includes 1,400 two-second video clips. Each video clip is additionally paired with a BLIP-generated text caption to support vision-language modeling. The SEED-DV dataset employs three types of EEG representations. Raw EEG signals consist of 62 channels over 400 timepoints (2-second segments at 200Hz), bandpass filtered (0.1–100Hz), and downsampled from 1000Hz. These are fed directly into temporal-spatial models without feature extraction. Power Spectral Density (PSD) features are computed using the autoregressive Burg method with a 256-sample Hamming window (1.28s, 50% overlap), capturing power in five frequency bands—delta (1–4Hz), theta (4–8Hz), alpha (8–12Hz), beta (12–31Hz), and gamma (31–99Hz)—across 62 channels and multiple windows per segment. Differential Entropy (DE) features, extracted using Butterworth bandpass filtering and Shannon entropy, also characterize these same five bands, providing a compact measure of signal complexity.

4.1.2. EEG-VP

The EEG-VP benchmark is designed to evaluate two main tasks:

- **EEG Visual Perception (EEG-VP) Classification Benchmark:** Seven classification tasks were established to probe different levels of visual information decoding from EEG signals: **40c**: 40-class classification of the fine-grained concept of the video clip; **9c** : 9-class classification of the course concept of the video clip; **Color**: 6-class classification of which color of the main object in the video clip; **Fast/Slow**: Binary classification distinguishing fast versus slow on the OFS; **Numbers**: 3-class classification of the number of the main objects in the video clip (one, two, or many); **Human Face**: Binary classification detecting the presence of a human face; **Human**: Binary classification detecting the presence of a human
- **EEG-to-Video Reconstruction Benchmark:** This task aims to reconstruct two-second video clips from EEG signals. The reconstructed outputs are evaluated using both frame-based and video-based metrics to assess visual quality and temporal coherence.

4.2. Implementation Detail

For all experiments, we followed the protocols established in [19] for fair comparison. Models were trained using the Adam optimizer with a learning rate of 0.001 for classification tasks and 0.0001 for contrastive learning. Training was performed for 100 epochs with early stopping (patience=10). Batch size was set to 256 for classification. For contrastive learning, the triplet loss margin and the loss weight was set to $\alpha = 0.2$ and $\lambda = 0.1$, respectively. To comprehensively evaluate the model's performance, we conducted two types of experiments: within-subject and cross-subject. For within-subject experiments, we used a 5/1/1 split of the 7 sessions per subject (cross-validated over 6 folds). For cross-subject (LOSO) evaluation, models were trained on 19 subjects and tested on the held-out subject, repeated for all 20 subjects.

4.3. Result

4.3.1. Dataset Integrity Assessment

Statistical Analysis Tab. 1 shows result of LME models applied to DE and PSD. Columns in the table; Sig., FDR Sig., Bonf Sig. denotes the number of comparisons between conditions(e.g. 40c: video category 1 vs. video category 2) that was significant. False Discovery Rate(FDR) is calculated as $FDP(\%) = \frac{\text{Sig}_{\text{Uncorrected}} - \text{Sig}_{\text{FDR}}}{\text{Sig}_{\text{Uncorrected}}} \times 100$. While Linear Mixed-Effects models suggested significant neural discrimination across multiple stimulus categories(40c, 9c, color, numbers in DE and 40c, 9c, color, numbers, human in PSD), rigorous pairwise comparisons with FDR correction revealed almost no statistically significant differences for any task after controlling for multiple comparisons. While some task categories remained significant after Bonferroni correction, it can hardly be considered as meaningful finding taking effect size(Cohen's d) into account. All effect sizes were negligible (**Cohen's d** $< \mathbf{0.1}$), with the largest

Table 1. **Result of Linear Mixed Effect model and Pair-wise comparison** Differential Entropy (DE); Power Spectral Density (PSD); Uncorrected Significance (Unc. Sig.) ; Significance (FDR Sig.); Bonferroni Significance (Bonf. Sig.); False Discovery Rate (False Dis. Rate)

		DE					PSD				
Task	**Classes (Comparisons)**	**Unc. Sig.**	**FDR Sig.**	**Bonf. Sig.**	**Max Cohen's d**	**False Dis. Rate**	**Unc. Sig.**	**FDR Sig.**	**Bonf. Sig.**	**Max Cohen's d**	**False Dis. Rate**
40c	40 (780)	193	61	0	0.099	16.9%	67	0	0	0.059	100%
9c	9 (36)	15	9	0	0.040	16.7%	5	0	0	0.025	100%
color	7 (21)	8	2	0	0.061	28.6%	1	0	0	0.018	100%
numbers	3 (3)	2	2	2	0.025	0%	1	0	0	0.013	100%
fast_slow	2 (1)	0	0	0	0.005	-	0	0	0	0.007	-
human	2 (1)	0	0	0	0.007	-	1	1	1	0.014	0%
face_human	2(1)	0	0	0	0.006	-	0	0	0	0.009	-

Table 2. **Average classification accuracy (%) and std across all subjects under within-subject setting** For each subject, five sessions were used for training, one for validation, and one for testing.

Methods	40-c top-1	40-c top-5	9-c top-1	9-c top-3	Color	Fast/Slow	Numbers	Human Face	Human
Chance level	2.50	12.50	11.11	33.33	20.57	50.00	65.64	62.25	71.43
Raw EEG Signals									
DeepNet [24]	3.37/0.76	14.83/1.76	14.13/1.87	40.04/2.86	18.31/3.54	50.80/1.73	59.89/4.39	77.26/2.92	64.79/2.94
TSConv [31]	3.53/0.73	15.62/1.91	14.74/1.52	39.87/2.35	19.32/2.41	51.20/1.84	55.80/1.87	75.58/1.18	63.77/1.96
EEGNet [13]	2.87/0.47	13.96/0.96	13.55/0.99	38.29/1.33	18.70/2.57	50.40/1.75	55.91/0.99	73.98/1.24	62.79/1.34
ShallowNet [24]	**4.07/0.94**	**17.31/2.58**	16.09/2.25	**43.68/3.73**	19.46/2.84	**52.82/2.22**	57.59/2.34	76.56/1.33	65.52/1.94
Conformer [29]	2.71/0.54	12.87/1.29	12.89/0.93	37.44/1.50	19.35/2.89	50.26/1.00	53.14/3.00	71.92/3.11	60.67/2.56
GLMNet [19]	2.84/0.39	13.58/0.96	13.14/1.18	37.53/2.34	17.96/2.86	50.51/1.38	55.69/1.63	75.76/1.01	63.34/1.93
CBraEnc (Ours) [33]	2.49/0.0	12.50/0.0	**17.50/0.0**	42.73/0.88	**20.82/0.69**	51.40/0.12	**65.64/0.0**	**82.14/0.0**	**71.42/0.0**
PSD Features									
MLP	**4.27/2.66**	**17.18/5.14**	**15.81/3.04**	**41.68/4.68**	**20.43/2.80**	51.39/1.20	**54.92/1.93**	63.93/3.02	**75.49/2.11**
GLMNet [19]	3.83/1.72	16.45/3.91	15.39/2.57	40.55/3.91	19.99/2.51	**51.51/1.38**	54.89/1.71	**75.01/1.52**	63.34/1.93
DE Features									
MLP	**4.38/2.60**	**17.63/5.63**	**15.99/3.50**	**42.27/5.03**	**20.89/2.95**	51.43/1.34	**56.31/1.85**	65.67/2.49	**77.26/1.37**
GLMNet [19]	4.10/1.82	16.65/4.00	15.39/2.57	41.02/4.01	20.38/2.30	**51.53/1.22**	56.23/1.25	**77.19/0.97**	65.15/1.81

effect observed in the 40-class task (**d = 0.099 in DE and 0.059 in PSD**). Uncorrected pairwise tests showed apparent significance in 1-193 comparisons per task, but almost no pairs exhibited significance after correction with high FDR. These results can be interperetated as either neural discrimination is below the detection threshold of this methodology, or the PCA features fail to capture task-relevant neural patterns, or the dataset is not sufficient to retain vision representation of EEG.

Mutual Information. We trained MINE [1] models to estimate the mutual information between EEG and video latent embeddings for each of the 20 subjects in the SEED-DV dataset. EEG signals were encoded using a pre-trained EEGNet [13], and video clips were processed through a pre-trained R3D-18 [8] network. The resulting latent vectors were concatenated and fed into MINE, which was trained to distinguish joint (aligned) from marginal (shuffled) pairs. The best mutual information (MI) lower-bound achieved across subjects averaged **0.72 bits**, with values ranging from 0.59 to 0.86 bits. To assess the relative informativeness, we computed the normalized mutual information (NMI) by dividing the estimated MI by the maximum of the differential entropies of the EEG and video representations. The average NMI across all subjects was **0.43%**, with none exceeding 0.52%.

These values indicate that the EEG and video modalities in SEED-DV share only a negligible amount of mutual information. Such low cross-modal dependency suggests that the two modalities may not be semantically or temporally aligned in a way that supports effective joint modeling. This concern is reinforced by our downstream results: the lack of mutual information coincides with poor performance in both EEG-to-video reconstruction and video classification from EEG signals. These observations collectively question the validity of SEED-DV as a benchmark dataset for multimodal generation and recognition tasks involving EEG and video.

4.3.2. Encoder Robustness Evaluation

We report results on the EEG-VP benchmark, focusing on the ability of different EEG encoders to extract semantic

Table 3. **Average classification accuracy (%) and std across all subjects under the Leave-One-Subject-Out (LOSO) setting**

Methods	40-c top-1	40-c top-5	9-c top-1	9-c top-3	Color	Fast/Slow	Numbers	Human Face	Human
Chance level	2.50	12.50	11.11	33.33	20.57	50.00	65.64	62.25	71.43
Raw EEG Signals									
DeepNet [24]	**6.42/1.71**	**24.82/4.56**	18.03/1.69	**46.65/2.43**	22.50/2.90	50.96/1.44	64.82/0.34	81.95/0.00	71.31/2.06
TSConv [31]	5.01/1.12	20.34/3.12	17.83/1.75	45.74/2.02	22.80/2.78	51.09/1.36	64.71/0.37	81.85/0.22	71.46/0.96
EEGNet [13]	5.94/1.24	23.00/3.43	17.47/1.58	46.15/2.14	23.06/3.55	**52.22/2.10**	**64.99/1.60**	**81.97/0.42**	71.14/1.10
ShallowNet [24]	5.87/1.32	22.67/3.51	**18.33/1.88**	46.23/2.81	**25.70/2.36**	49.98/0.00	64.11/2.41	81.46/1.68	71.53/1.14
Conformer [29]	4.20/1.10	18.10/2.08	17.30/0.61	44.67/0.79	20.50/2.25	51.31/1.43	64.86/0.19	81.92/0.09	71.76/0.13
GLMNet [19]	2.90/0.52	13.78/1.18	16.17/1.26	43.24/1.95	20.26/2.60	50.17/1.01	64.74/1.48	81.71/1.18	70.66/1.28
GLMNet *w/ CL* (Ours)	4.63/0.97	19.65/2.66	17.43/0.84	45.28/1.74	20.86/2.25	50.98/1.23	64.64/1.27	81.95/0.01	71.72/0.42
CBraEnc (Ours) [33]	2.50/0.18	12.54/0.42	17.43/0.00	42.35/2.15	20.31/0.0	50.03/0.30	64.92/0.0	81.95/0.0	**71.82/0.0**
PSD Features									
MLP	**3.19/0.36**	14.63/0.75	16.34/1.27	43.21/2.22	21.19/1.46	51.25/1.10	65.20/0.62	**82.03/0.22**	70.68/1.72
GLMNet [19]	**3.19/0.55**	**15.02/1.43**	**16.45/1.25**	**43.34/1.67**	**22.05/2.17**	**51.40/1.61**	**65.34/0.70**	**82.03/0.29**	**71.32/0.17**
DE Features									
MLP	**3.20/0.43**	**14.64/0.88**	15.79/1.19	42.35/1.30	**21.76/1.80**	51.33/1.34	65.44/0.32	**82.10/0.12**	71.26/0.42
GLMNet [19]	2.99/0.49	14.54/1.31	**16.02/1.05**	**43.49/1.56**	21.37/1.59	**51.62/1.38**	**65.58/0.14**	81.96/0.78	**71.41/0.04**

Table 4. **Average classification accuracy (%) and std of Latent Classification**

Methods	**40-c top-1**	**40-c top-5**	**9-c top-1**	**9-c top-3**	**Color**	**Fast/Slow**	**Numbers**	**Human Face**	**Human**
				DE Features					
Chance level (6th block)	2.50	12.50	17.50	45.00	26.00	54.00	69.00	80.50	71.50
Seq2Seq model	2.62/0.92	12.62/1.48	16.70/1.40	43.67/1.62	17.80/1.40	51.23/3.02	69.00/0.00	80.50/ 0.00	71.50/0.00
Chance level (7th block)	2.50	12.50	17.50	45.00	23.20	53.00	67.00	79.50	75.50
Semantic Predictor	2.80/0.65	13.03/2.09	16.26/1.85	42.30/2.30	18.79/4.97	57.15/7.84	79.79/0.44	87.21/0.73	86.74/1.14

and perceptual information from EEG signals. We evaluated each model using the three aforementioned types of EEG input representations: raw EEG signals, PSD features, and DE features. **CBraMod encoder** was only evaluated with raw EEG input, as it does not accept preprocessed features such as PSD or DE. GLMNet trained with contrastive learning(**GLMNet w/ CL**) was only tested under the LOSO protocol given its design for learning subject-invariant representations.

Tab. 2 shows classification performance of each EEG encoders across different classification tasks and EEG input representations. CBraMod encoder showed higher performance than GLMNet in all tasks excluding 40 class classification. CBraMod encoder showed below or equal chance level performance for 40 classs classification task for top1 accuracy and top5 accuracy. However, in 9class classification, CBraMod encoder showed 33.18% increase in 9 class top1 accuracy and 13.85% increase in top3 accuracy comapred to GLMNet. While not exceeding chance level significantly for Color, Fast/Slow, Numbers, and Human classification, classification performance of Human Face exhibited 31.94% increase compared to chance level and 8.42% increase compared to GLMNet classification performance.

In order to assess the reproducibility of GLMNet under the within-subject setting, the model was re-implemented and evaluated using the same task configurations with previous EEG2video. While the overall trend in relative model performance was consistent with the ones reported in EEG2Video [19], our reproduced results demonstrated a general degradation in absolute accuracy. It is noteworthy that the model exhibited substandard performance in a range of conditions, including multi-class classification and semantic binary tasks such as Color, Fast/Slow, and Numbers. The performance decline frequently exceeded 5%, and in certain instances, it exceeded 10%. Conversely, the Human Face category was the sole condition in which our model exhibited consistent superiority over the original implementation across all feature types, demonstrating enhancements exceeding 10%. Intriguingly, this gain was accompanied by a decline in performance for the closely related Human category, suggesting a potential issue with label alignment or category separation. While not all reproduced results exceeded the corresponding chance levels, the model generally maintained meaningful performance across most conditions, reaffirming its architectural validity. When tested with PSD and DE features, the overall performance seemed higher compared to when tested with raw EEG signals. This might be due to adequate feature extraction method used to filter useful and meaningful aspects of EEG data that might convey proxy of core vision perception, effectively managing noise in EEG data.

In the Leave-One-Subject-Out (LOSO) condition, the results in Table 3 reveal that contrastive learning marginally improves multi-class classification with raw EEG signals. GLMNet w/ CL outperforms its vanilla counterpart in 40-class and 9-class tasks, yet it does not achieve the highest accuracy in any individual metric. Instead, baseline models deliver surprisingly strong results across several tasks. These results suggest that achieving more pronounced cross-subject generalizability may require additional techniques beyond the use of a simple triplet loss. CBraMod encoder, while not so remarkably showing performance compared to GLMNet w/ CL, exhibited superior performance in Human classification task. However, GLMNet which was expected to extract visual and whole brain activities didn't show robust superiority over MLP both in within-subject condition and LOSO condition, casting doubt on its capability as an optimal EEG encoder.

Contrary to our initial expectations, the standard deviation of the results in LOSO condition(i.e. performance variability across subjects' model) was not always large compared to within-subject condition. Unlike within-subject condition that trained a seperate model for each individual, LOSO condition trains a single model using data from all participants except one, who is held out for testing. As a result, the model may not have specialized in individual-level prediction, but rather learned a general representation that reflects group-level trends. This suggests that the model is likely optimizing toward the group mean rather than capturing person-specific nuances-an interpretation that may be more appropriate given the current results.

4.3.3. Reconstruction Fidelity Validation

Latent Prediction As illustrated in Table 4, freezing the reconstruction-oriented encoders and training only a shallow classifier leaves performance essentially at chance. On the 40-class benchmark the Seq2Seq latent model attains $2.62\,\%$ top-1 accuracy, whereas the Semantic predictor reaches $2.80\,\%$; both are close to the $2.50\,\%$ chance level, and the corresponding top-5 scores improve by less than one percentage point. Accuracy on *Color* decreases from $26\,\%$ to $17.8\,\%$ for Seq2Seq and $18.8\,\%$ for the Semantic predictor, while *Fast/Slow* drops from $54\,\%$ to $51.2\,\%$. In contrast, the Semantic predictor shows marked gains for attributes that rely on coarse conceptual cues: *Numbers* improves to $79.79\,\%$, *Human Face* to $87.21\,\%$, and *Human* to $86.74\,\%$. These findings indicate that mapping EEG signals into a text-embedding space can capture high-level semantics, yet the resulting representations remain insufficiently separable for fine-grained object categorization. Directly re-using the frozen features in the diffusion decoder may therefore propagate semantic ambiguity during video reconstruction, although the text-style embedding strategy itself appears promising and warrants further study through joint optimization or partial fine-tuning.

Table 5. Quantitative evaluation of reconstructed videos on the SEED-DV dataset. We report Top-1 accuracy (%) for 2-way and 40-way classification at both image and video level, and SSIM for perceptual similarity.

Metric	Mean	Std	Type
Image-Level 2-way Acc	72.30%	23.05%	ViT
Image-Level 40-way Acc	9.77%	23.02%	ViT
Video-Level 2-way Acc	78.00%	21.19%	VideoMAE
Video-Level 40-way Acc	11.69%	23.08%	VideoMAE
Image SSIM	0.0782	0.0345	Structural Similarity

Quantitative Results We evaluate the fidelity and semantic consistency of reconstructed videos using both vision-language-based classification metrics and perceptual similarity scores. Following prior work, we report 2-way and 40-way top-1 accuracy under both frame-level and video-level settings using ViT and VideoMAE classifiers, respectively. Additionally, we report SSIM to assess pixel-wise similarity between generated and ground truth frames.

The high 2-way accuracy (78.00% video-level, 72.30% image-level) suggests that the generated videos preserve coarse semantic information from EEG. However, performance drops significantly in the 40-way setting (11.69% and 9.77%), indicating difficulty in capturing fine-grained class distinctions. The low SSIM (0.0782) further reflects limited visual similarity. This gap likely arises from relying solely on semantic embeddings, without latent-level conditioning, which was disabled due to instability.

Qualitative Results We present several qualitative examples in Figure 2. While some reconstructions capture the general scene layout or motion patterns, the semantic alignment with ground truth is mostly imperfect. In particular, object identity, textures, or scene context may diverge noticeably from the original video. This highlights the limitations of relying solely on EEG-derived semantic embeddings for guiding the generation.

References

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018. 3, 6

[2] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[3] Minsuk Choi and Hiroshi Ishikawa. Braindecoder: Style-based visual decoding of eeg signals. *arXiv preprint arXiv:2409.05279*, 2024. 2

[4] Teng Fei, Abhinav Uppal, Ian Jackson, Srinivas Ravishankar,

David Wang, and Virginia R. de Sa. Perceptogram: Reconstructing visual percepts from eeg, 2025. 2

[5] Zixuan Gong, Guangyin Bao, Qi Zhang, Zhongwei Wan, Duoqian Miao, Shoujin Wang, Lei Zhu, Changwei Wang, Rongtao Xu, Liang Hu, Ke Liu, and Yu Zhang. Neuroclips: Towards high-fidelity and smooth fMRI-to-video reconstruction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[6] Sven Guenther, Nataliya Kosmyna, and Pattie Maes. Image classification and reconstruction from low-density eeg. *Scientific Reports*, 14(1):16436, 2024. 2

[7] Zhanqiang Guo, Jiamin Wu, Yonghao Song, Jiahui Bu, Weijian Mai, Qihao Zheng, Wanli Ouyang, and Chunfeng Song. Neuro-3d: Towards 3d visual decoding from eeg signals. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23870–23880, 2025. 2

[8] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017. 6

[9] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664):1634–1640, 2004. 1

[10] W. R. Hendee. The perception of visual information. *Radiographics: a review publication of the Radiological Society of North America, Inc*, 7(6):1213–1219, 1987. 1

[11] Gan Huang, Zhiheng Zhao, Shaorong Zhang, Zhenxing Hu, Jiaming Fan, Meisong Fu, Jiale Chen, Yaqiong Xiao, Jun Wang, and Guo Dan. Discrepancy between inter- and intra-subject variability in eeg-based motor imagery brain-computer interface: Evidence from multiple perspectives. *Frontiers in Neuroscience*, Volume 17 - 2023, 2023. 3

[12] Shuai Huang, Yongxiong Wang, and Huan Luo. Never: A multi-stream framework for high-fidelity dynamic video reconstruction from eeg signals. arXiv preprint SSRN:5167418, 2025. Preprint, not peer-reviewed. 2

[13] 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. 6, 7

[14] 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 *Computer Vision – ECCV 2024*, pages 353–369, Cham, 2025. Springer Nature Switzerland. 2

[15] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2

[16] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion, 2024. 1

[17] S. Liang, L. Li, W. Zu, W. Feng, and W. Hang. Adaptive deep feature representation learning for cross-subject eeg decoding. *BMC Bioinformatics*, 25(1):393, 2024. 2

[18] Xu Lichao, Xu Minpeng, Ke Yufeng, An Xingwei, Liu Shuang, and Ming Dong. Cross-dataset variability problem in eeg decoding with deep learning. *Frontiers in Human Neuroscience*, Volume 14 - 2020, 2020. 3

[19] Xuanhao 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. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2, 3, 4, 5, 6, 7

[20] Eleonora Lopez, Luigi Sigillo, Federica Colonnese, Massimo Panella, and Danilo Comminiello. Guess what i think: Streamlined eeg-to-image generation with latent diffusion models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2

[21] Rahul Mishra, Krishan Sharma, Ranjeet Ranjan Jha, and Arnav Bhavsar. Neurogan: image reconstruction from eeg signals via an attention-based gan. *Neural Computing and Applications*, 35(12):9181–9192, 2023. 2

[22] Thomas Naselaris, Emily Allen, and Kendrick Kay. Extensive sampling for complete models of individual brains. *Current Opinion in Behavioral Sciences*, 40:45–51, 2021. Deep Imaging - Personalized Neuroscience. 1

[23] Hongguang Pan, Zhuoyi Li, Yunpeng Fu, Xuebin Qin, and Jianchen Hu. Reconstructing visual stimulus images from eeg signals based on deep visual representation model, 2024. 2

[24] 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. 6, 7

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 4

[26] Paul Steven Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Cohen Ethan, Aidan James Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[27] Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. Mindeye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *ICLR 2024 Workshop on Representational Alignment*, 2024. 1

[28] Prajwal Singh, Pankaj Pandey, Krishna Miyapuram, and Shanmuganathan Raman. Eeg2image: image reconstruction

from eeg brain signals. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[29] Yifan Song, Qiuqiang Zheng, Bing Liu, and Xiaojie Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022. 6, 7

[30] 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. 2

[31] Yifan Song, Bing Liu, Xiaoyi Li, Ningyuan Shi, Yuhang Wang, and Xiaojie Gao. Decoding natural images from eeg for object recognition. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 6, 7

[32] Praveen Tirupattur, Yogesh Singh Rawat, Concetto Spampinato, and Mubarak Shah. Thoughtviz: Visualizing human thoughts using generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 950–958, 2018. 2

[33] Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding, 2025. 1, 3, 6, 7

[34] Ling Wang, Chen Wu, and Lin Wang. Braindreamer: Reasoning-coherent and controllable image generation from eeg brain signals via language guidance. *arXiv preprint arXiv:2409.14021*, 2024. 2

A. Additional Results

B. Supplementary Figures

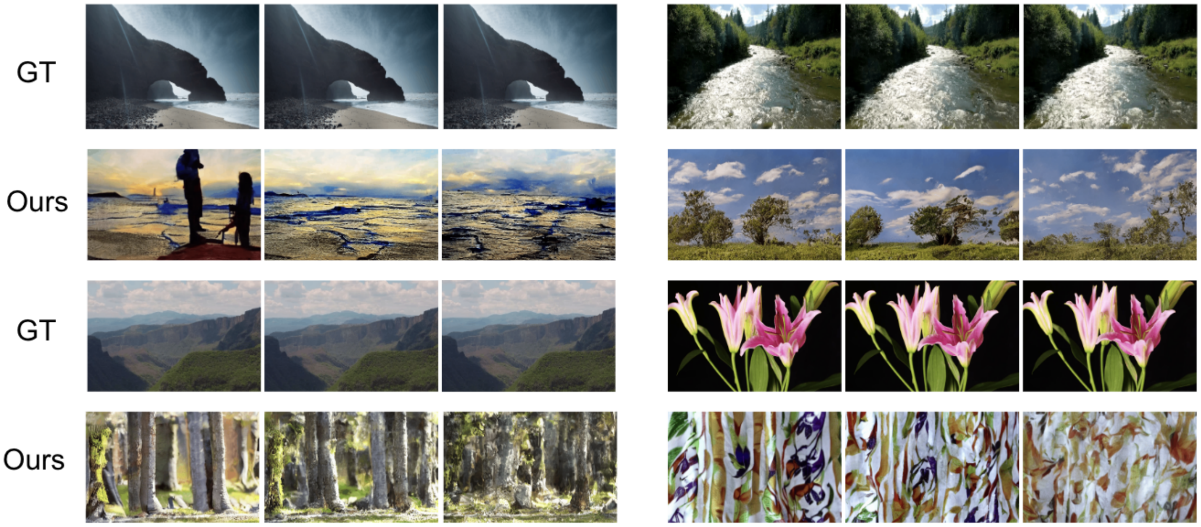

Figure 2. Qualitative reconstruction results from EEG. Each triplet shows 3 sampled frames from a video clip: the top row is ground truth (GT), and the bottom row shows the model output (Ours).

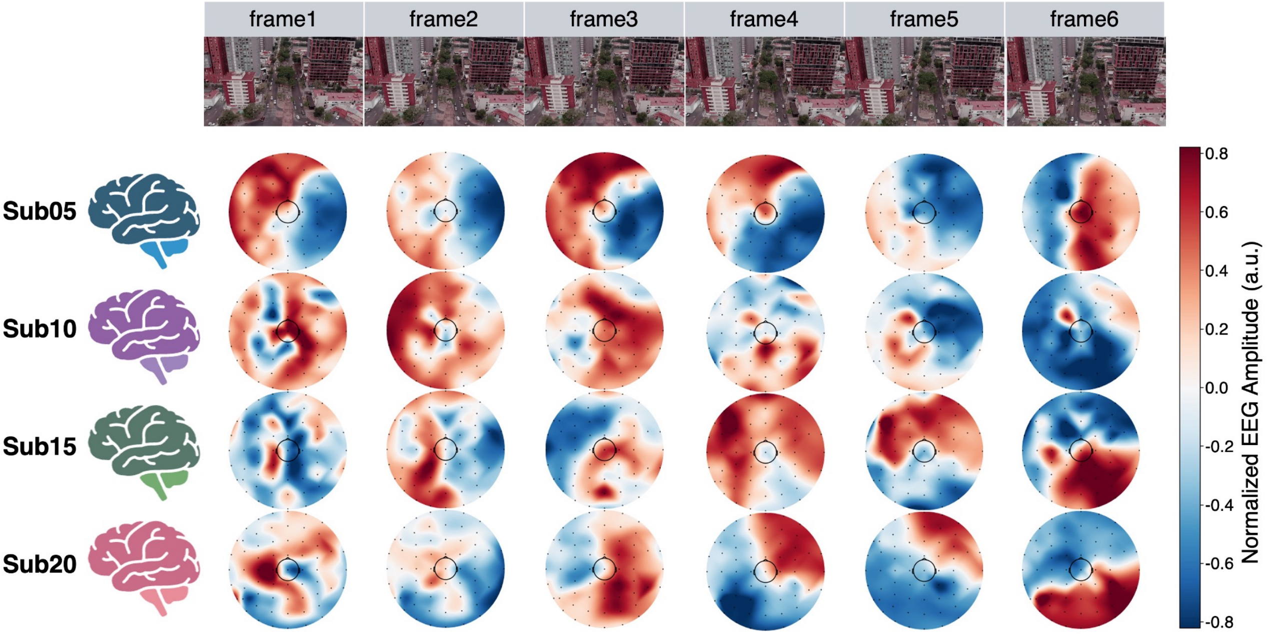

Figure 3. Inter-Subject variability in neural signals: EEG Amplitudes extracted from SEED-DV dataset at same timepoint while watching identical stimuli