This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Cross-Subject EEG-to-Video Reconstruction and Beyond

Runduo Han Hongchen Tan†
Dalian University of Technology
hanrunduo@mail.dlut.edu.cn thc123@dlut.edu.cn

Abstract

Reconstructing video content from EEG (electroencephalogram) is a research task of significant scientific importance. However, due to inter-subject differences in physiological states and variations in signal acquisition configurations, this task faces the challenge of inconsistent cross-subject generation. To address this, we propose a Subject Adversarial and Mapping Network (SAM-Net). In SAM-Net, we first introduce a Hybrid Region-Temporal (HRT) Encoder to conduct inter-channel semantic interactions guided by brain regions and aggregate temporal semantics across different time scales. Secondly, we propose a Centered-progressive Subject Adversarial (C-SA) Mechanism to gradually narrow the metric distance between different subjects, thereby obtaining a unified and stable semantic representation. Thirdly, we design a New2Source Mapper to align the EEG distribution of new subjects with that of multiple known subjects. Finally, we adopt a keyframe-guided continuous semantic generation paradigm to drive the production of coherent and high-quality videos. Extensive experiments validate the competitive performance of our SAM-Net in cross-subject EEG-to-Video generation tasks, as well as its excellent performance in generation tasks involving new subjects.

1. Introduction

By decoding brain signals to reconstruct visual content, it becomes feasible to create a mental world, applying this approach to brain mechanism analysis, artistic creation, human - computer interaction, etc. Current research on reconstructing visual content from brain signals primarily employs functional magnetic resonance imaging (fMRI) [3, 5, 6, 9] or electroencephalography (EEG) [2, 4, 13, 26]. EEG recording equipment, compared to fMRI, is typically portable, cost-effective, and offers high temporal resolution. Given that humans typically perceive dynamic visual information, leveraging EEG signals proves more advantageous for reconstructing dynamic visual content. To this end, the dataset introduced in [17] pioneers video reconstruction from EEG, aligning EEG with descriptions and frames, and generate video using Tune-A-Video [25]. DynaMind [16] further improves dynamic consistency in EEG-to-Video (E2V) reconstruction by extracting temporal features from EEG time patches and integrating them into the diffusion process. However, generating videos from EEG faces the significant challenge of inter-subject EEG signal variability (i.e., the EEG semantic distributions of different subjects shown in Fig. 1-(a)), primarily due to physiological differences among subjects, variations in physical setups during signal acquisition, and interference from other spurious signals.

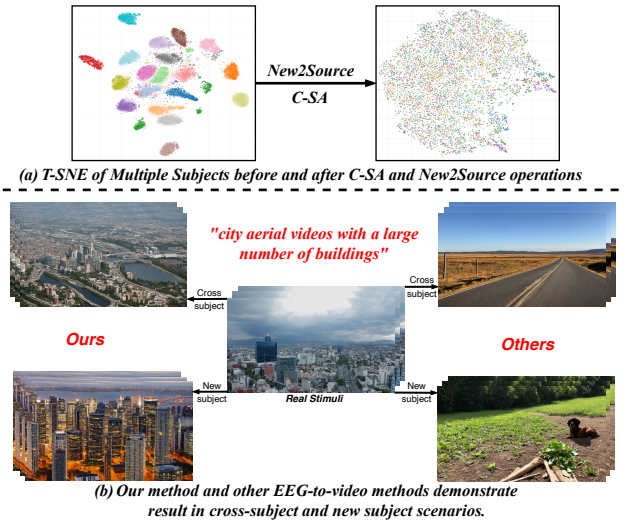

Figure 1. ***(a)***: T-SNE visualization of all subjects before and after using C-SA and New2Source. ***(b)***: Our method and other methods demonstrate result in cross-subject and new subject scenarios.

The key to addressing this challenge lies in establishing a unified EEG semantic representation for all subjects, especially in terms of its generalization to new subjects. Most existing E2V reconstruction methods [6, 18] overlook the perception and modeling of spatiotemporal semantics in EEG signals. ATM [13] divides EEG signals into tokens

†Corresponding author.

by channel and extracts spatiotemporal embeddings using a CNN-based method, while DreamDiffusion [2] randomly masks portions of EEG data and reconstructs them based on contextual cues. However, they often lack explicit modeling of brain priors, such as the functional specialization of brain regions and the multi-scale nature of neural dynamics, and struggle to capture fine-grained neural patterns, especially those distributed across specific brain regions and varying temporal scales. This limitation led us to design a multi-region, multi-timescale encoder that explicitly models neural activity from distinct brain areas and hierarchically integrates temporal dynamics. Accordingly, we attempt to partition brain regions and construct the aggregation of interactions between brain regions as well as multi-scale temporal semantics.

To enable EEG semantic representation to encompass multiple subjects, some methods use multiple subject-specific encoders [6], cyclic semantic reconstruction [22], and multi-expert and memory storage mechanisms [18] to retain heterogeneous information between subjects. While these strategies enhance cross-subject vision reconstruction, they increase storage overhead [18, 22] and impose significant computational costs for video generation tasks [22]. Moreover, they are often tailored for fMRI data and overlook the temporal semantics of EEG signals. Recent EEG-based emotion recognition tasks [14, 23, 27] have attempted to address the challenge of cross-subject scenarios. However, [14, 23] still assign a semantic modeling branch to each subject, posing challenges when dealing with a large number of subjects. [27] progressively reduces the distance between other subjects and a designated target subject. However, such target subject is determined either randomly or through direct assignment. In the event that an ambiguous subject is selected, deviations may occur in the unification of the overall subjects. To this, we attempt to combine multiple subjects to identify a reasonable central subject and drive other subjects to gradually approach the central subject from near to far.

Apart from conducting unified modeling for multiple known subjects, another key issue is how to enhance the generalization performance of subjects with minimal computational overhead when encountering new subjects. For new subjects, the model faces a significant distribution shift in EEG signals compared to known subjects, coupled with severe data scarcity due to the cumbersome and time-consuming nature of EEG data collection. As shown in Fig. 1-(b), methods lacking unique designs for cross-subject and new subject adaptation can cause semantic confusion during reconstruction, resulting in significant discrepancies with real stimuli. While [27] attempts to fine-tune the trained model using new subjects to achieve subject generalization capability, this approach risks perturbing the semantic representations of known subjects. Thus, we attempt to perform interactive mapping of EEG semantics between new and known subjects solely during the encoding stage to drive the model's adaptation to new subjects.

Beyond essential EEG semantic modeling, another key challenge in EEG-to-Video reconstruction is ensuring temporal coherence in visual semantics and spatial plausibility. A series of advanced Text-to-Video (T2V) generation models have made this concept feasible. Traditional T2V methods [7, 25], which generate videos solely from textual prompts, often produce spatially inconsistent or implausible scene layouts. Subsequently, SparseCtrl [8] enhances the controllability of T2V through the incorporation of an additional sparse conditional encoder and keyframes. Given EEG's high temporal resolution, the semantic gap between EEG and text makes it difficult to directly apply [8] for generating semantically coherent video content. Therefore, we aim to first bridge the modal gap between EEG and Vision/Text, and then model EEG semantic latent sequences alongside high-quality keyframes to guide [8] in producing high-quality content.

Above all, we propose a Subject Adversarial and Mapping Network (SAM-Net) for the cross-subject EEG-to-Video reconstruction. Within the framework of SAM-Net, we initially introduce a Hybrid Region-Temporal (HRT) Encoder. This encoder is designed to facilitate inter-channel semantic interactions under the guidance of brain regions, while also aggregating temporal semantics across diverse time scales. Subsequently, we put forward a Centered-progressive Subject Adversarial (C-SA) Mechanism. Its purpose is to incrementally reduce the metric distance between different subjects, ultimately achieving a unified and stable semantic representation. Moving forward, we devise a New2Source Mapper, which aims to align the distribution of new subjects with that of known subjects through a semantic interactive mapping approach. Lastly, we have designed a video content generation paradigm guided by keyframes and multi-type EEG semantics. This paradigm is intended to drive the generation of coherent and high-quality videos. Through extensive experiments, we have confirmed the competitive performance of our SAM-Net in cross-subject EEG-to-Video generation tasks. Moreover, it has demonstrated outstanding performance in generation tasks that involve new subjects.

2. Related Work

Brain Signal to Image Reconstruction. Reconstructing images from brain signals [2, 6, 9, 13, 20, 22, 26] is a foundational task in visual decoding. The primary goal is to generate images that are semantically and perceptually similar to the visual stimuli presented to the subject. Research in this area has evolved along two main tracks, primarily distinguished by the source of brain signals: the high spatial resolution of fMRI and the high temporal resolution of

EEG. **(i) fMRI-to-Image Reconstruction:** Early foundational studies, such as the work by Haxby *et al*. [9], demonstrated that distinct categories of visual stimuli evoke discriminable patterns of fMRI activity. Building on such insights, [6] uses multiple subject-specific encoders to generate cross-subject images. [22] adopts a cyclic semantic reconstruction approach to enhance the consistency between generated images and electroencephalogram (EEG) semantics. And [18] further employs multi-expert and memory storage mechanisms to retain heterogeneous information between subjects. Despite their outstanding performance in the image generation, these methods are often tailored for fMRI data and overlook the temporal semantics of EEG signals. Moreover, they increase storage overhead [18, 22] and impose significant computational costs for video generation tasks [22]. **(ii) EEG-to-Image Reconstruction:** Compared to fMRI, EEG provides a more temporally precise but spatially coarse signal, making visual reconstruction particularly challenging. Li *et al*. [13] proposed a zero-shot framework that incorporates a custom Adaptive Thinking Mapper (ATM) to encode EEG signals into a pre-trained text-to-image model. CognitionCapturer [26] argue that previous methods often underutilize EEG data by relying predominantly on image-EEG mutual information. Their method introduces depth information as an auxiliary supervisory signal to enrich the EEG feature representation and improve reconstruction quality.

Brain Signal to Video Reconstruction. The reconstruction of dynamic visual content [3, 5, 11, 12, 16, 17, 21] from brain activity represents a more complex and emerging frontier. This task requires capturing not only the static visual content of individual frames but also the temporal dynamics and motion patterns between them. Similarly, the brain to video is divided into two pathways: fMRI and EEG. **(i) fMRI-to-Video Reconstruction:** Cinematic Mindscapes [3] processes fMRI data via masked brain modeling and spatiotemporal multi-modal contrastive learning, feeding the representations into an augmented Stable Diffusion model for high-quality video synthesis. NeuroClips [5] injects both high-level semantics and low-level perceptual cues into a pre-trained diffusion model, ensuring smooth and semantically consistent video reconstruction. **(ii) EEG-to-Video Reconstruction:** EEG2Video [17] constructed the SEED-DV dataset—addressing the scarcity of EEG-video paired data—and pioneered EEG-to-video translation. DynaMind [16] achieves high visual fidelity and temporal coherence by explicitly modeling brain region interactions and temporal dynamics. Nevertheless, they have overlooked the exploration of cross-subject generation issues. Therefore, in contrast to them, we attempt to develop a cross-subject EEG-to-video reconstruction method.

3. Methodology

Overview. The pipeline is shown in figure 2. In the context of SAM-Net, the Hybrid Region-Temporal (HRT) Encoder is specifically engineered to enable inter-channel semantic interactions guided by brain regions, and simultaneously aggregate temporal semantics across various time scales; the Centered-progressive Subject Adversarial (C-SA) Mechanism is to gradually minimize the metric distance among different subjects, thereby attaining a unified and stable semantic representation; the New2Source Mapper is designed to make the EEG distribution of new subjects consistent with that of multiple known subjects; we adopt a keyframe-guided continuous semantic generation paradigm (i.e. Video Reconstruction Process) to spur the production of coherent and high-caliber videos.

3.1. HRT Encoder

The input to Hybird Spatio-Temporal (HRT) Encoder is the EEG signal $\mathcal{E} \in \mathbb{R}^{B\times C\times T}$, where B, C and T denote batch size, the number of EEG electrode channels and time steps, respectively. The output consists of embeddings aligned to text or latents aligned to video, denoted as $E \in \mathbb{R}^{B\times 77\times 768}$ and $L \in \mathbb{R}^{B\times \frac{H}{8}\times \frac{W}{8}}$ respectively, H and W are the height and width of the reconstructed video.

EEG Augmentation. A major challenge in cross-subject EEG decoding lies in the coexistence of target stimulus-evoked responses and subject-specific interference caused by non-target brain activity and biological noise (Detailed analysis in supplementary materials 5.2). To prevent the model from overfitting to these specific patterns, we augment EEG signals prior to formally extracting EEG semantics (as illustrated in Fig. 3). Specifically, we first inject Gaussian noise (akin to DiffMDD [24]) into the EEG signal $\mathcal{E}$, mimicking inter-subject variability to force the model to disregard subject-specific features and focus on robust, invariant neural correlates of visual stimuli common across individuals. Then, to ensure the model does not become overly sensitive to noise or specific electrode patterns, we apply random dropout $\mathcal{RD}$ to the electrode channels. The entire process is computed as $\mathcal{E}_{Augment} = \mathcal{RD}(\mathcal{E} + \mathcal{N}(\mu, \sigma^2))$.

Region-guided Semantic Perception. Research in neuroscience has shown that different regions of the brain activate in response to distinct types of visual stimuli. For example, calm natural scenery, architecture, and fast-moving objects, like cars, may activate different brain regions. And compared to electrode channels, perceiving and extracting semantics based on regions is more robust. To capture these region-specific responses, we propose the Region-guided Semantic Perception strategy, which learns to focus on the relevant brain regions for different types of visual stimuli. Specifically, we first apply electrode attention on $\mathcal{E}_{Augment}$ to measure the contribution of each electrode to the task at

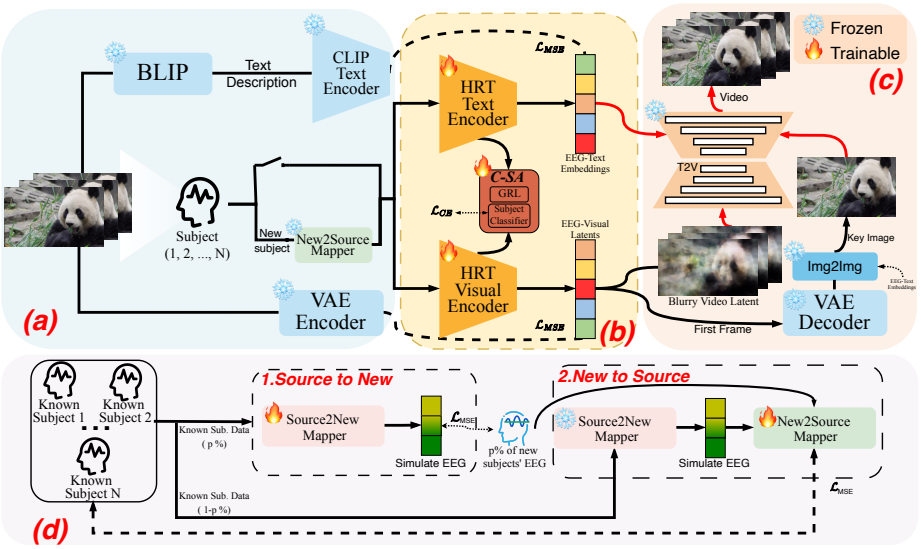

Figure 2. Overall architecture of Subject Adversarial and Mapping Network (SAM-Net). ***(a)*** EEG signal is encoded into the latent space using VAE [10] and the visual stimuli were understand as text descriptions using BLIP [15], and then encoded into the embeddings using the CLIP [19] text encoder. For the input EEG, if it comes from a new subject, it is then passed through New2Source Mapper before being input into the subsequent HRT encoder. ***(b)*** EEG input is fed into the HRT (as shown in Fig. 3) text and visual encoders for embedding and latent alignment, respectively. Meanwhile, we used the C-SA(as shown in Fig. 4) strategy for training on EEG signals from multiple subjects. ***(c)*** The embeddings and latents obtained from EEG, along with the keyframe obtained through *I2I* model from the first frame, are used as conditions for *T2V* model. ***(d)*** Simulating the scarce EEG of new subject using real EEG data from known subjects promotes the convergence of the new subjects' EEG distribution to the known subjects.

hand. This helps us prioritize the most informative electrode channels for each input.

$$\begin{aligned} W_{Electrode} &= \sigma(\mathbb{L}(ReLU(\mathbb{L}(\mathcal{AP}(\mathcal{E}_{Augment}))))) \\ \mathcal{E}_{EA} &= W_{Electrode} \otimes \mathcal{E}_{Augment} \end{aligned} \tag{1}$$

where σ is sigmoid, $\mathbb{L}$ is linear layer, $ReLU$ is activation function, $\mathcal{AP}$ is adaptive average pooling.

Next, we divide the EEG signals into five brain regions based on their anatomical locations: Frontal, Parietal, Central, Temporal, and Occipital. For each of these regions, we apply a dedicated extractor to capture the region-specific features.

$$\mathcal{E}_{EA} \xrightarrow{Divide} \mathcal{E}_{Fro.}, \mathcal{E}_{Par.}, \mathcal{E}_{Cen.}, \mathcal{E}_{Tem.}, \mathcal{E}_{Occ.} \tag{2}$$

After processing by the region-specific extractors, the features from each region are concatenated and passed through a linear layer to obtain the $\mathcal{E}_{Spatial}$.

$$\mathcal{E}_{Spatial} = \mathbb{L}(||_{i\in\{Fro,Par,...\}}[\mathcal{D}(BN(GELU(Conv(\mathcal{E}_i))))]) \tag{3}$$

where $\mathcal{D}$ is dropout, BN is batch normalization, $Conv$ is 1D CNN layer.

Multi-scale Temporal Dependency Perception. EEG signals frequently encompass temporal patterns with varying durations. For instance, the time spans occupied by a slowly walking individual and a rapidly moving vehicle within EEG signals differ significantly. In response to this, we propose the Multi-scale Temporal Dependency Perception Strategy in an attempt to perceive different types of motion events in visual scenes. Specifically, we first project $\mathcal{E}_{Spatial}$ to obtain the Key ($K = \mathbb{L}(\mathcal{E}_{Spatial})$) and Value ($V = \mathbb{L}(\mathcal{E}_{Spatial})$) matrices, while the Query (Q) is trainable parameters, similar to BoQ [1]. Then, we apply multi-scale 1D CNN at temporal dimension $K_{tem} = ||_{i\in\{5,11,21\}}[Conv_i(K)]$, where $||$ denotes concatenation. Finally, the K_{tem} are processed through an attention layer to capture semantic dependencies across different time steps:

$$\mathcal{E}_{temporal} = \bigcirc_{l=1}^{L} Transformer[Soft.(\frac{(Q)^T \cdot K_{tem}}{\sqrt{d}}) \cdot V] \tag{4}$$

where $Soft.$ is softmax, $\bigcirc_{l=1}^{L}$ represents a total of L layers. After that, we project features $\mathcal{E}_{temporal}$ to dimensions for downstream tasks (e.g., text embeddings, video latent), ensuring control over video generation.

3.2. C-SA Mechanism

A fundamental challenge in cross-subject EEG decoding is the significant distribution shift between subjects, where identical stimuli evoke subject-specific neural responses. This variability degrades model generalization across sub-

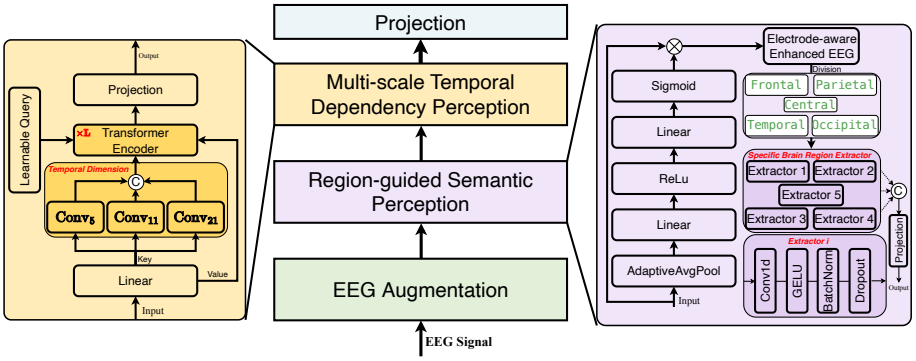

Figure 3. Overall architecture of our proposed Hybrid Region-Temporal (HRT) Encoder, which contains two novel designs: Region-guided Spatial Perception and Multi-scale Temporal Dependency Perception.

jects. To tackle this, we propose the Centered-progressive Subject Adversarial (C-SA) mechanism, which uses a progressive strategy: it selects the most representative central subject as a starting point and iteratively adds the nearest remaining subject to adversarial training. Specifically, for subjects $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_N\}$, where N represents the total number of subjects, we first need to find the most representative center subject as the starting point for progressive adversarial training. For the i-th subject $\mathcal{S}_i$, we calculate the average feature vector of its EEG data: $\mathbf{f}_i = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_{ij}$, n is the number of samples for the i-th subject, $\mathbf{x}_{ij}$ is the j-th EEG sample of the subject. Next, we calculate the distance metric between all pairs of subjects: $d(i,j) = 1 - \frac{\mathbf{f}_i \cdot \mathbf{f}_j}{\|\mathbf{f}_i\| \cdot \|\mathbf{f}_j\|}$. The central subject is defined as the subject with the highest average similarity to other subjects:

$$c^* = \arg\max_{i \in S} \left\{ \frac{1}{|S| - 1} \sum_{j \in S, j \neq i} \text{sim}(i,j) \right\} \tag{5}$$

Selecting the central subject as the training starting point is advantageous because its EEG feature distribution best reflects the common features of the whole subject group. Let C_t be the set of subjects selected in stage t, $R_t = S \setminus C_t$ is the remaining set of subjects.

Progressive Addition. *Initial stage* ($t = 0$): $C_0 = \{c^*\}$. *Selection process* ($t \geq 0$): For each remaining subject $r \in R_t$, calculate its minimum distance to the selected set C_t, and select the subject with the smallest distance to the selected set to join:

$$r_t^* = \arg\min_{r \in R_t} d_{\min}(r, C_t), d_{\min}(r, C_t) = \min_{c \in C_t} d(r,c) \tag{6}$$

Update selected set: $C_{t+1} = C_t \cup \{r_t^*\}$ Each time a new subject is added, C_t is input into the HRT encoder. Training ends when $C_t = S$, meaning all subjects have joined the training.

Subject Adversarial. Relying solely on Progressive Addition Process cannot bridge the gap between individuals

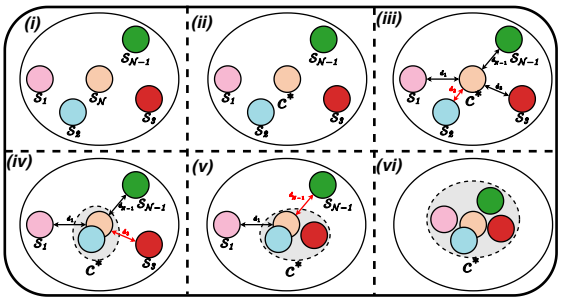

Figure 4. Process of Centered-progressive Subject Adversarial (C-SA). ***(i)*** Original multiple subjects distribution. ***(ii)*** Find the central subject that best represents all subjects. ***(iii)*** Find the subject closest to the center and perform subject adversarial learning. ***(iv)*** / ***(v)*** For the new center composed of multiple subjects, continue to progressively find the nearest subject for adversarial training. ***(vi)*** C-SA ends when all subjects have joined the training.

and thus fails to represent invariant EEG semantics across subjects. To facilitate the optimization of the HRT Encoder with Progressive Addition Process, we introduce a Gradient Reversal Layer (GRL) and a subject classifier. The aim to confuse HRT Encoder about which subject the features it obtains come from. The subject classifier receives the HRT encoder's output through a GRL and predicts the subject category $\hat{y}_{ik}$ to which the input sample belongs. The subject classifier uses cross-entropy loss:

$$\mathcal{L}_{\text{subject}} = -\sum_{k=1}^{K} y_{ik} \log(\hat{y}_{ik}) \tag{7}$$

where K is the number of subjects in the current phase.

The GRL enables adversarial training by acting as an identity function during forward propagation but reversing the gradient sign during backpropagation:

$$\frac{\partial \text{GRL}(f)}{\partial f} = -\lambda I \tag{8}$$

where f is the feature output of HRT, λ is the adversarial weight and I is the identity matrix. GRL's characteristics establish an adversarial relationship: During forward propagation, the subject classifier normally receives features from the encoder and calculates the loss $\mathcal{L}_{\text{subject}}$. During backpropagation, the subject classifier parameters are updated normally based on the gradients of $\mathcal{L}_{\text{subject}}$ to improve classification performance. However, the gradients flowing to the encoder first pass through the GRL and are multiplied by $-\lambda$. This means that the HRT's optimization objective effectively becomes maximizing $\mathcal{L}_{\text{subject}}$, which is the opposite of the subject classifier. This adversarial compels the HRT to learn subject-invariant representations. The process effectively minimizes the distributional discrepancy among subjects in the latent space, forcing the model to rely on robust embedding that are common across subjects.

3.3. New2Source Mapper

In practical applications, collecting EEG signals from new subjects is costly, and the neural patterns of different subjects often vary significantly. Directly applying EEG data from new subjects to existing models may yield suboptimal results due to these differences. To this end, we try to build the mapping relationship between the new subject and the known subjects, i.e. simulating the new subject using multiple known subjects and then combine it with the real new subject. Known subjects are denoted as $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N\}$, with the new subject represented by $\mathcal{S}_{new}$. Firstly, we used known subjects to simulate the new subject by

$$\begin{aligned} {\mathcal{S}^p_{new}}^* &= \text{Source2New}(\mathcal{S}^p) \\ \mathcal{L}_{\text{S2N}} &= \text{MSE}(\mathcal{S}^p_{new}, {\mathcal{S}^p_{new}}^*) \end{aligned} \tag{9}$$

where p represents the proportion of the current data relative to the complete dataset, $\mathcal{S}^p_{new}$ is $p\%$ EEG data from new subject and ${\mathcal{S}^p_{new}}^*$ is the corresponding $p\%$ simulated EEG data.

Secondly, we apply the trained Source2New Mapper to convert the remaining $(1-p)\%$ of the data from the known subjects into simulated EEG data for the new subject, denoted as ${\mathcal{S}^{(1-p)}new}^*$:

$${\mathcal{S}^{(1-p)}_{new}}^* = \text{Source2New}(\mathcal{S}^{(1-p)}) \tag{10}$$

Finally, we train the New2Source Mapper using both the simulated and real EEG data from the new subject, capturing more accurate inter-subject mappings.

$$\begin{aligned} \mathcal{S}^* &= \text{New2Source}(||[{\mathcal{S}^{(1-p)}_{new}}^*, \mathcal{S}^p_{new}]) \\ \mathcal{L}_{\text{N2S}} &= \text{MSE}(\mathcal{S}, \mathcal{S}^*) \end{aligned} \tag{11}$$

where $||$ is concatenation. This approach efficiently uses limited EEG data from new subjects, enabling pre-trained models to adapt without large-dataset retraining. The mapping strategy alleviates data scarcity and enhances cross-subject EEG model performance.

3.4. Video Reconstruction Process

In addition to fundamental semantic modeling, another major hurdle in EEG-to-Video generation lies in maintaining temporal coherence in visual semantics and ensuring spatial plausibility. To address this, we aim to leverage the SparseCtrl [8] to generate high-quality videos guided by EEG-based embeddings. [8] generate continuous video sequences from noise, guided only by keyframes. Therefore, we first construct keyframe from EEG embedings. Specically, we use the HRT encoder to obtain the latent representation of the first frame L_0, which is decoded by the VAE to generate a blurry first frame $BF = \text{VAE}(L_0)$. The motivation for this is to establish a stable visual foundation: the blurry frame BF provides a crucial structural prior, preserving the global composition and spatial layout inferred directly from the EEG signals. This coarse initial frame serves as a structural anchor, ensuring that the overall scene geometry remains consistent with the neural activity. Subsequently, the EEG-Text embeddings E and BF are fed into an image-to-image (*I2I*) model to produce a refined, high-quality keyframe $KF = \textit{I2I}(E, \text{VAE}(L_0))$. This step leverages the semantic richness of E to add precise details and clear textures onto the structural foundation provided by BF, effectively transforming a layout into a sharp and semantically consistent image.

However, the gap between the E2V (EEG-to-Video) and T2V (Text-to-Video) tasks prevents us from directly using random noise as the input for [8]. To address this, we employ L as a substitute for random noise to serve as the [8] input. Here, L can provide information such as layout and color. Additionally, we further utilize E as a guiding factor for semantic information to drive the generation process of the SparseCtrl. Finally, the refined keyframe KF, along with the full EEG latent sequence L and the embeddings E, are input into a T2V model from [8] to reconstruct the final video by $Video = \textit{T2V}(E, L, KF)$.

3.5. Loss and Training Summary

First, we train HRT Encoder so that it can represent invariant EEG semantics across subjects and align them with textual/visual modalities. Loss in this training stage is $\mathcal{L}_1 = \mathcal{L}_{\text{task}} + \lambda\mathcal{L}_{\text{subject}}$. where $\mathcal{L}_{\text{task}} = \text{MSE}(E, \text{HRT}(\mathcal{E})) + \text{MSE}(L, \text{HRT}(\mathcal{E}))$ is designed to align EEG semantics with text/visual semantics. The E and L are obtained by encoding the text description and video frames using BLIP [15] and VAE [10], respectively.

Secondly, we train the New2Source Mapper to facilitate interactive semantic representation between new subjects and known ones, thereby acquiring generalization capabili-

ties for new subjects. In this training stage: Source2New is trained using $\mathcal{L}_{S2N}$, then we use $\mathcal{L}_{N2S}$ to train the New2Source.

4. Experiments

Experimental details. Datasets: Our experiment was conducted on the **SEED-DV** [17] dataset, which includes EEG signals from 20 subjects watching video clips across 40 visual concept categories. **Corss-subject:** In our experimental setup, the first 15 subjects were known source subjects, and the last 5 subjects were new subjects. **Setup:** In addition to the full dataset with 40 concepts, we also selected {*Cat, Shark, Flower, Dancing, Face, Buildings, Road, Pizza, Guitar, Airplane*} to form a 10-class subset, and the first {*1-20*} classes and the first {*1-30*} classes. **New Subject:** We set p to 15 When training New2Source.

Cls	Sub.	Metrics	Video-based		Frame-based		
			Semantic-level		Semantic-level		Pixel-level
		Method	2-way	40-way	2-way	40-way	SSIM↑
10	SS	DynaMind	0.847±0.01	0.394±0.03	0.833±0.02	0.308±0.01	0.309±0.02
		EEG2Video	0.852±0.02	0.340±0.01	0.798±0.03	0.232±0.02	0.300±0.03
		Ours	0.870±0.02	0.380±0.03	0.887±0.03	0.390±0.02	0.312±0.02
	CS	Ours (Best)	0.899±0.01	0.312±0.01	0.888±0.02	0.388±0.02	0.307±0.03
		Ours (Average)	**0.864**	**0.250**	**0.855**	**0.321**	**0.324**
20	SS	DynaMind	0.835±0.01	0.345±0.01	0.818±0.02	0.277±0.02	0.290±0.02
		EEG2Video	0.813±0.02	0.273±0.03	0.785±0.04	0.184±0.02	0.242±0.03
		Ours	0.887±0.03	0.340±0.03	0.856±0.02	0.311±0.03	0.295±0.03
	CS	Ours (Best)	0.886±0.02	0.326±0.03	0.842±0.04	0.317±0.02	0.269±0.02
		Ours (Average)	**0.844**	**0.223**	**0.804**	**0.236**	**0.268**
30	SS	DynaMind	0.833±0.02	0.309±0.01	0.805±0.02	0.254±0.01	0.293±0.01
		EEG2Video	0.794±0.02	0.209±0.05	0.785±0.04	0.180±0.02	0.228±0.04
		Ours	0.878±0.03	0.320±0.02	0.841±0.02	0.307±0.02	0.284±0.02
	CS	Ours (Best)	0.874±0.02	0.312±0.03	0.855±0.01	0.334±0.02	0.275±0.04
		Ours (Average)	**0.842**	**0.229**	**0.805**	**0.247**	**0.278**
40	SS	DynaMind	0.828±0.02	0.284±0.02	0.807±0.03	0.241±0.01	0.280±0.01
		EEG2Video	0.798±0.03	0.159±0.01	0.774±0.02	0.138±0.01	0.256±0.03
		Ours	0.870±0.01	0.300±0.01	0.833±0.02	0.303±0.02	0.290±0.02
	CS	Ours (Best)	0.860±0.01	0.291±0.02	0.834±0.02	0.301±0.01	0.279±0.02
		Ours (Average)	**0.841**	**0.228**	**0.810**	**0.262**	**0.280**

Table 1. Reconstruction results are evaluated based on semantic/pixel metrics for different subsets of categories. DynaMind [16] and EEG2Video [17] use Single-Subject (SS) evaluation, while our method uses both Single-Subject (SS) and Cross-Subject (CS) evaluation. We perform cross-subject evaluation on the SEED-DV [17] using 15 subjects.

4.1. Comparison with State-of-the-art Methods

As shown in Table. 1, we compared our method with the SOTA method on the full dataset (40 classes) and subsets (10, 20 and 30 classes), we distinguished CS (Cross-Subject) and SS (Single-Subject). We evaluated the semantic-level at the video-based and frame-based using 2-way and 40-way, and pixel-level SSIM. The results demonstrate that our method outperforms existing methods on single-subject and also performs excellently on cross-subject. **SS:** In the 40-class single-subject setting, our video-based semantic accuracy reached 0.300, and our frame-based semantic accuracy reached 0.303, significantly outperforming the baseline performance 0.284 and 0.241. we also achieved a leading 0.290 in the SSIM metric. **CS:** In the challenging 40-class cross-subject setting, our video-based semantic accuracy reached 0.291, and our frame-based semantic accuracy reached 0.301. we also achieved a leading 0.279 in the SSIM metric.

Figure 5. Comparison of frames between GT and synthesis. The first row of each unit is the real frames, and the second row is the reconstructed frames. The red text is the text description generated by BLIP [15].

4.2. Reconstruct Video on New Subject

# Cls	Metrics	Video-based		Frame-based		
		Semantic-level		Semantic-level		Pixel-level
	Method	2-way	40-way	2-way	40-way	SSIM↑
10	Best	0.833±0.01	0.143±0.03	0.820±0.02	0.225±0.04	0.300±0.03
	Average	**0.838**	**0.175**	**0.776**	**0.178**	**0.296**
20	Best	0.839±0.04	0.163±0.05	0.730±0.05	0.123±0.04	0.266±0.04
	Average	**0.808**	**0.141**	**0.721**	**0.118**	**0.258**
30	Best	0.829±0.03	0.167±0.04	0.740±0.02	0.126±0.03	0.271±0.01
	Average	**0.815**	**0.147**	**0.729**	**0.127**	**0.266**
40	Best	0.826±0.04	0.162±0.04	0.745±0.03	0.136±0.02	0.257±0.03
	Average	**0.812**	**0.142**	**0.735**	**0.137**	**0.254**

Table 2. Video reconstruction results on new, unseen subjects. New subjects were defined as the last 5 subjects of SEED-DV [17].

Table. 2 shows the results of reconstructed video after processing new subjects using New2Source Mapper. We selected last 5 new subjects of SEED-DV [17] for the experiment. Under the most challenging 40-class setting, our method achieved a semantic accuracy of 0.142 on video-based and 0.137 on frame-based semantic accuracy, with SSIM also reaching 0.254. This is attributed to the fact that our New2Source can align the EEG distribution of new subjects with the EEG distribution of known subjects.

Figure 6. Reconstruction results under different input conditions during the generation process.

4.3. Ablation Study

Method	Video-based		Frame-based		
	2-way	40-way	2-way	40-way	SSIM↑
Ours	0.860±0.01	0.291±0.02	0.834±0.02	0.301±0.01	0.279±0.02
HRT Encoder					
A. w/o HRT	0.796±0.03	0.179±0.02	0.780±0.03	0.140±0.02	0.222±0.03
B. w/o Frontal	0.849±0.02	0.279±0.03	0.822±0.01	0.290±0.03	0.269±0.03
C. w/o Parietal	0.850±0.02	0.280±0.02	0.810±0.02	0.295±0.03	0.260±0.03
D. w/o Central	0.854±0.02	0.277±0.03	0.819±0.03	0.286±0.04	0.266±0.03
E. w/o Temporal	0.824±0.05	0.269±0.03	0.802±0.05	0.255±0.02	0.243±0.03
F. w/o Occipital	0.820±0.03	0.261±0.03	0.799±0.03	0.261±0.04	0.245±0.04
Conditions of Reconstruction					
G. w/o Latent	0.845±0.02	0.273±0.01	0.810±0.03	0.282±0.02	0.189±0.05
H. w/o Embedding	0.774±0.05	0.090±0.04	0.752±0.04	0.092±0.03	0.239±0.03
I. w/o KeyFrame	0.839±0.01	0.262±0.02	0.807±0.02	0.276±0.03	0.231±0.02

Table 3. Ablation of HRT Encoder and Conditional Input.

HRT Encoder. To verify HRT's design, we conducted the experiments shown in *A* to *F* of Table. 3. Experimental results show that, consistent with biological experience, the occipital region, which controls vision, has the greatest impact on video reconstruction, followed by the temporal region, which controls language.

Conditions of Reconstruction. To verify reconstruction framework, we conducted the experiments shown in *G* to *I* of Table. 3. We removed visual latents, semantic embeddings, and keyframe, respectively. Among them, row *H* was most affected, as the reconstruction framework completely lost semantic guidance, leading to a sharp drop in the 40-class classification results. In row *G*, we removed all latent from the reconstruction framework, which resulted in a significant drop in SSIM due to the absence of visual color and structure guidance. Finally, the absence of semantically and structurally rich KeyFrame resulted in a decline in both classification and visual similarity metrics, Fig. 7 shows more cases of KeyFrame. Fig. 6 clearly shows the impact of removing each condition on the final reconstruction result.

C-SA. To verify C-SA, we conducted the experiments shown in *A* to *C* of Table. 4. As can be seen, removing C-SA in row *A* has the greatest impact on reconstruction performance, with all metrics dropping sharply. In row *B*, we used a fixed subject, rather than the center subject of all subjects, as the starting point for progressive adversarial learning. In row *C*, instead of a progressive adversarial learning process from easy to difficult, we directly added all the subjects' data to the training process.

Figure 7. More cases of keyframe.

New2Source Mapper. To verify New2Source Mapper, we conducted the experiments shown in *D* of Table. 4. Because the new subject lacked effective alignment with multiple known subjects, the various metrics of video reconstruction dropped sharply.

Method	Video-based		Frame-based		
	2-way	40-way	2-way	40-way	SSIM↑
Ours	**0.841**	**0.228**	**0.810**	**0.262**	**0.280**
A. w/o C-SA	0.799	0.186	0.784	0.212	0.256
B. w/o Center	0.826	0.211	0.799	0.232	0.261
C. w/o Progressive	0.820	0.206	0.787	0.240	0.250
Ours	**0.826±0.04**	**0.162±0.04**	**0.745±0.03**	**0.136±0.02**	**0.257±0.03**
D. w/o New2Source	0.770±0.05	0.118±0.04	0.729±0.04	0.128±0.05	0.199±0.05

Table 4. Ablation of Centered-progressive Subject Adversarial(C-SA) and New2Source Mapper.

5. Conclusion

In this paper, we proposed a Subject Adversarial and Mapping Network (SAM-Net), for the cross-subject EEG-to-Video task. SAM-Net integrates multi-scale spatiotemporal features via a Hybrid Region-Temporal (HRT) Encoder, reduces inter-subject disparities using a Centered-progressive Subject Adversarial (C-SA) mechanism, and aligns new subjects' EEG distributions with known subjects through a New2Source Mapper. A keyframe-guided generation paradigm is employed to produce coherent videos. Experiments demonstrate the model's performance in both cross-subject and new-subject EEG-to-video generation tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China 62201020, the Fundamental Research Funds for the Central Universities DUT25RC(3)048, the General Program of the Natural Science Foundation of Liaoning Province 2025-MS-016.

References

[1] Amar Ali-Bey, Brahim Chaib-draa, and Philippe Giguère. Boq: A place is worth a bag of learnable queries, 2024. 4

[2] Yunpeng Bai, Xintao Wang, Yan-Pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: High-quality eeg-to-image generation with temporal masked signal modeling and CLIP alignment. In *ECCV*, pages 472–488. Springer, 2024. 1, 2

[3] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity, 2023. 1, 3

[4] Matteo Ferrante, Tommaso Boccato, Stefano Bargione, and Nicola Toschi. Decoding visual brain representations from electroencephalography through knowledge distillation and latent diffusion models. *Computers in Biology and Medicine*, 178:108701, 2024. 1

[5] Zixuan Gong, Guangyin Bao, Qi Zhang, Zhongwei Wan, Duoqian Miao, Shoujin Wang, Lei Zhu, Changwei Wang, Rongtao Xu, Liang Hu, et al. Neuroclips: Towards high-fidelity and smooth fmri-to-video reconstruction. *Advances in Neural Information Processing Systems*, 37:51655–51683, 2024. 1, 3

[6] Zixuan Gong, Qi Zhang, Guangyin Bao, Lei Zhu, Ke Liu, Liang Hu, and Duoqian Miao. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction, 2024. 1, 2, 3

[7] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[8] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2024. 2, 6

[9] James Haxby, Maria Gobbini, Maura Furey, Alumit Ishai, Jennifer Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science (New York, N.Y.)*, 293:2425–30, 2001. 1, 2, 3

[10] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 4, 6

[11] Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N Apurva Ratan Murty, et al. Modeling short visual events through the bold moments video fmri dataset and metadata. *Nature communications*, 15(1):6241, 2024. 3

[12] 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. Springer, 2024. 3

[13] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. In *Advances in Neural Information Processing Systems*, pages 102822–102864. Curran Associates, Inc., 2024. 1, 2, 3

[14] Jinpeng Li, Shuang Qiu, Yuan-Yuan Shen, Cheng-Lin Liu, and Huiguang He. Multisource transfer learning for cross-subject eeg emotion recognition. *IEEE Transactions on Cybernetics*, 50(7):3281–3293, 2020. 2

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 4, 6, 7

[16] Junxiang Liu, Junming Lin, Jiangtong Li, and Jie Li. Dynamind: Reconstructing dynamic visual scenes from eeg by aligning temporal dynamics and multimodal semantics to guided diffusion, 2025. 1, 3, 7

[17] 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 (NeurIPS)*, 2024. 1, 3, 7

[18] Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 233–243. IEEE, 2024. 1, 2, 3

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 4

[20] 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, 2023. 2

[21] Haonan Wang, Qixiang Zhang, Lehan Wang, Xuanqi Huang, and Xiaomeng Li. Neurons: Emulating the human visual cortex improves fidelity and interpretability in fmri-to-video reconstruction. *ArXiv*, abs/2503.11167, 2025. 3

[22] Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11333–11342, 2024. 2, 3

[23] Yiming Wang, Bin Zhang, and Yujiao Tang. Dmmr: Cross-subject domain generalization for eeg-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 628–636, 2024. 2

[24] Yilin Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Benyan Luo, Tao Li, and Gang Pan. Diffmdd: A diffusion-based deep learning framework for mdd diagnosis using eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:728–738, 2024. 3

[25] 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, 2

[26] Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14486–14493, 2025. 1, 2, 3

[27] Qi Zhu, Ting Zhu, Lunke Fei, Chuhang Zheng, Wei Shao, David Zhang, and Daoqiang Zhang. Multi-modal cross-subject emotion feature alignment and recognition with eeg and eye movements. *IEEE Transactions on Affective Computing*, 16(3):2102–2115, 2025. 2