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

Junxiang Liu Junming Lin Jie Zhou Wei Xiong
Jiangtong Li* Jie Li* Jie Zhuang Hongfei Ji
School of Computer Science and Technology, Tongji University
{junxiang_liu, 2432262, 2432263, xw1216,
jiangtongli, jieli, jiezhuang, hongfeiji}@tongji.edu.cn

Abstract

*Reconstructing dynamic visual scenes from electroencephalography (EEG) signals presents a significant challenge. Existing methods often yield temporally disjointed and inaccurate visual semantic reconstructions, struggling with poor dynamic timing alignment and lacking the integration of cognitive priors. In neuroscience, the dual-stream theory describes the physiological basis for the generation and transmission of visual neural signals, offering a valuable prior to guide the reconstruction process. To address these challenge, we follow the guidance of dual-stream theory and introduce **DynaMind**, a model that reconstructs video by jointly modeling neural dynamics and semantic features using three core modules: a Regional-aware Semantic Mapper (RSM), a Temporal-aware Dynamic Aligner (TDA), and a Dual-Guidance Video Reconstructor (DGVR). The **RSM models neural pathways** to capture detailed semantic information, using regional-aware encoders interconnected via channel-wise multiplicative gating. Meanwhile, the **TDA ensures temporal alignment**, generating a dynamic latent sequence (blueprint) that corresponds to the original neural recordings. The **DGVR uses both semantic and temporal guidance** to translate the aligned blueprint into a high-fidelity video reconstruction. On the SEED-DV and CineBrain datasets, **DynaMind** improves reconstructed video accuracy by 12.5% and 4.28%, respectively. Moreover, it boosts temporal consistency on both datasets, achieving score increases of 19.7% (FVMD) and 4.19% (FVD). Our Project Page: https://zl35-liu.github.io/DynaMind.*

1. Introduction

EEG signals are widely applied in cognitive related tasks, such as motor imagery [7] and emotion recognition [37]. As visual stimuli constitute a primary sensory input [28],

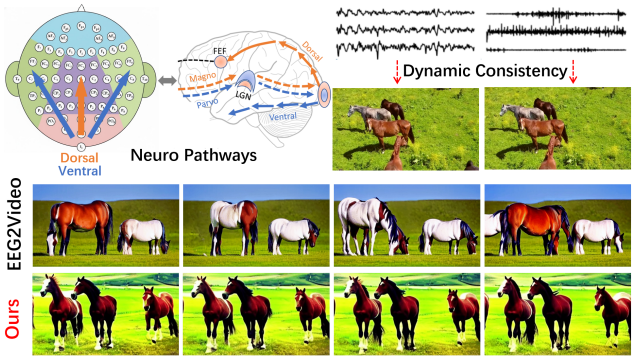

Figure 1. Our approach captures spatial features from diverse brain regions with pathway guided interaction, enforces temporal dynamic consistency between EEG and video, and generates videos with superior fidelity, temporal coherence, and semantic accuracy compared to EEG2Video.

decoding visual information from EEG has become a key area of cognitive research [33]. Pioneering research [31, 39] has established visual reconstruction from EEG as a notable emerging field. Contemporary methods employ advanced deep learning models such as CLIP [23] and Stable Diffusion (SD) [25]. The prevailing approach projects EEG into the CLIP latent space [2] for semantic representations and then controls generative models for image synthesis [41]. This method has achieved high image quality and semantic accuracy, establishing a reliable baseline. *This success in static image generation directs focus toward the next frontier: capturing the dynamic visual world from brain activity to generate coherent video.*

The nascent domain of video reconstruction from EEG involves substantial scientific and technical challenges. As shown in Fig. 1, existing methods like EEG2Video [18] often produce reconstructions that exhibit obvious intra-frame semantic errors and inter-frame temporal inconsistencies. These failures are essentially rooted in two primary issues: the difficulty of aligning dynamic EEG timing information

*Corresponding authors.

with corresponding video frames, and the insufficient integration of neuroscience priors on the reception of visual stimuli by human brain. *These limitations impede progress toward high-fidelity EEG-to-Video reconstruction.*

Neuroscience research confirms that visual processing involves distinct functional zones [11, 22], such as the occipital and parietal lobes, conceptually illustrated by the regional divisions in Fig. 1. However, visual processing relies on complex interactions [9] between these regions, not merely on isolated activity. The dual-stream theory [4, 5], for example, offers an established model for the generation and transmission of visual neural signals, detailing how different pathways process distinct aspects of visual stimuli. Integrating the physiological basis of these neural pathways into the feature extraction module is therefore essential for capturing accurate, high-level semantic information.

To fully apply the aforementioned neuroscience mechanisms and address the current challenges, we propose **DynaMind**, a novel framework that enriches semantic features using a brain-functional partitioning approach and enforces dynamic coherence with the video content. The DynaMind framework relies on three synergistic components: a Regional-aware Semantic Mapper (**RSM**), a Temporal-aware Dynamic Aligner (**TDA**), and a Dual-Guidance Video Reconstructor (**DGVR**). Guided by dual-stream theory, the **RSM** extracts features from distinct functional brain regions, models neuromorphic feature interactions using a channel-wise gating mechanism, employs multimodal constraints to enrich semantic diversity, and aggregates these features into a unified diffusion prior. Meanwhile, the **TDA** captures temporal dynamics by exploiting EEG features to generate a blueprint sequence, enforcing dynamic consistency between the neural recordings and the video stimulus. Finally, the **DGVR** uses the multimodal diffusion prior as semantic guidance to translate the coherent temporal blueprint into high-fidelity videos, exhibiting both temporal consistency and semantic accuracy.

Experiments on the SEED-DV [18] dataset show that our **DynaMind** framework not only improves direct EEG classification accuracy and the classification accuracy based on its reconstructed videos, but also yields a substantial improvement in pixel-level quality. Specifically, DynaMind achieves a 12.5% improvement in video-level semantic classification (40-way Top-1) and a 19.7% reduction in Fréchet Video Motion Distance (FVMD). To verify the generalizability of our framework and its ability to generate long sequence and more complex scene videos, we conduct further experiments on the challenging CineBrain [10] dataset. On CineBrain, our model also achieves significant performance improvements, with an 4.28% increase in video-based semantic accuracy (50-way Top-1) and a 4.19% improvement in video temporal consistency (FVD). Moreover, we conduct detailed ablation studies to analyze the contributions of different brain regions and the effect of multimodal feature integration. Overall, our results mark a notable advancement in the decoding of dynamic visual perception from EEG signals. Our main contributions are summarized as follows:

- We identify and analyze key limitations in EEG-based video reconstruction, namely the insufficient integration of regional brain information and the weak temporal consistency of existing methods.
- We propose DynaMind, a novel three-module framework designed to improve the dynamic coherence and semantic richness of video reconstructions from EEG signals.
- We evaluate DynaMind on the public SEED-DV dataset and CineBrain dataset, achieving significant improvement in both classification accuracy and reconstruction quality, thus providing a new benchmark for the field.

2. Related works

2.1. Decoding Visual Information from Brain

A substantial body of research has focused on decoding visual information from physiological signals, utilizing modalities such as functional magnetic resonance imaging (fMRI) [8, 35], magnetoencephalography (MEG) [1], and electroencephalography (EEG) [34]. To accomplish this, a variety of generative models have been explored, prominently featuring variational autoencoders (VAEs) [12] and generative adversarial networks (GANs) [33]. More recently, the advent of advanced diffusion models like Stable Diffusion (SD) has spurred significant progress in visual reconstruction [25, 29]. Generative schemes employing these models primarily rely on rich embeddings to guide the diffusion process toward specific semantic objectives. Therefore, numerous studies have attempted to derive such semantic information from EEG [16, 33].

The inherent high temporal resolution of EEG presents a unique and critical advantage for the more complex task of dynamic video reconstruction. However, the high signal-to-noise and low spatial resolution ratio of EEG still renders feature extraction difficult. Current methods often attempt to solve this by focusing exclusively on the relationship between EEG and image semantics [15], overlooking valuable multimodal information inherent in the brain signals [6] themselves. This limitation becomes particularly salient in EEG-to-video generation, where it may lead to videos that are conceptually approximate but exhibit deviations. While pioneering work such as EEG2Video [18] has ventured into this area, its resulting limitations in semantic accuracy and temporal coherence highlight the critical need for further research in EEG-to-video reconstruction.

2.2. Video Generation Models

Breakthroughs in Text-to-Image (T2I) generation have been largely driven by large-scale multimodal datasets composed

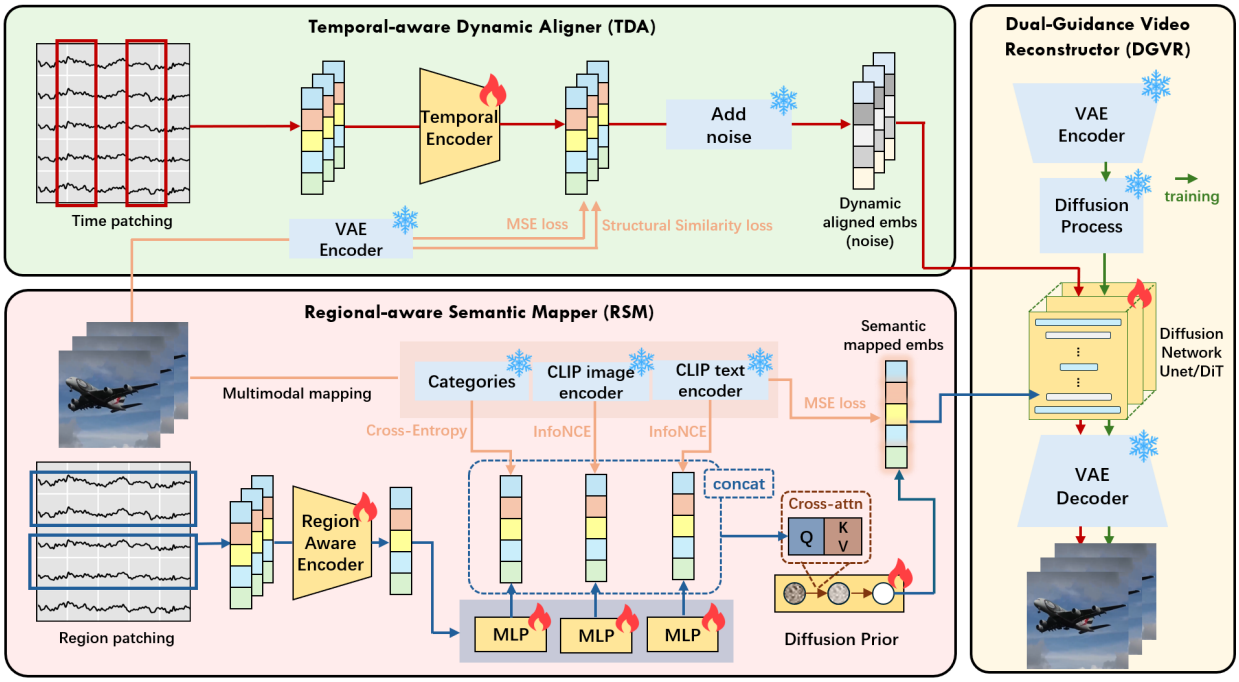

Figure 2. An overview of our DynaMind framework: A dual-guidance architecture for reconstructing high-fidelity videos from EEG signals. The RSM processes EEG signals based on their originating brain regions, using neuromorphic feature interactions to extract a semantically rich diffusion prior. In parallel, the TDA processes the global EEG sequence to generate a dynamic temporal blueprint that encodes the visual dynamics. Finally, the DGVR, a latent diffusion model, uses the semantic prior for conditional guidance and the temporal blueprint as initial latent variables to synthesize videos, ensuring both semantic accuracy and temporal coherence.

of billions of text-image pairs [20, 26]. To replicate this success in Text-to-Video (T2V) generation, recent studies [30, 36] have extended these space-only T2I models to the spatio-temporal domain. By training on large-scale text-video datasets such as WebVid-10M [3], these models have achieved promising results, laying a foundation for reconstructing visuals from other modalities.

However, prior research in brain signal-based reconstruction has typically relied on holistic semantics from the entire signal to directly guide the generative process [2]. Fundamentally, successful video generation hinges on preserving the continuous motion of consistent objects across frames [40]. As shown in the 2nd row of Fig. 1, this reliance on holistic guidance leads to reconstructions with significant temporal inconsistencies between consecutive frames.

3. Methodology

As shown in Fig. 2, our DynaMind model reconstructs high-fidelity video from EEG signals using three synergistic modules. The Regional-aware Semantic Mapper (RSM) generates a semantic diffusion prior by processing regional brain patterns with neuromorphic feature interactions. The Temporal-aware Dynamic Aligner (TDA) produces a coherent temporal blueprint to capture corresponding visual dynamics. The Dual-Guidance Video Reconstructor (DGVR), adapted from a pre-trained video diffusion model, uses both the semantic prior and temporal blueprint to generate videos with high semantic accuracy and temporal coherence.

3.1. Regional-aware Semantic Mapper (RSM)

3.1.1. Region-Aware EEG Representation Learning

Human cognitive processes are distributed across functionally specialized networks in the brain. For instance, the occipital lobe is primarily involved in processing visual stimuli, while the temporal and frontal lobes manage more abstract semantic and emotional contexts [19]. The well-established dual-stream theory for vision exemplifies this, modeling how information is processed in specialized pathways while also interacting between these functional areas. Therefore, the RSM is designed to model this neural structure, leveraging functional specialization to extract semantically rich and varied information from EEG signals.

Initially, the multi-channel EEG signals $\mathbf{E} \in \mathbb{R}^{C \times T}$, where C is the channel count and T is the time points, are partitioned into K distinct groups $\{\mathbf{E}_1, \mathbf{E}_2, \cdots, \mathbf{E}_K\}$ based on anatomical brain regions (*i.e.*, frontal, temporal, parietal, and occipital lobes). Each regional signal group $\mathbf{E}_i$ is then processed by a dedicated region-aware encoder Enc_i.

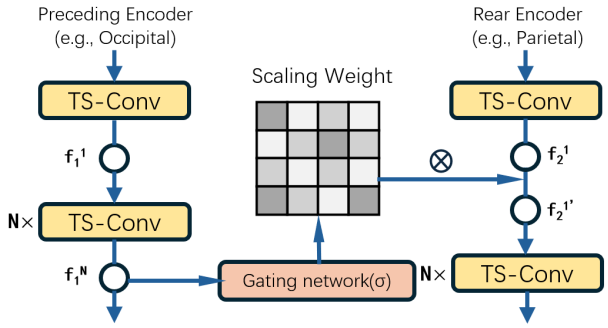

Figure 3. An overview of how to pass information between region encoders, using the prior knowledge of dual neural ways.

The Enc_i structure utilizes spatiotemporal convolutions to capture region-specific neural patterns.

Meanwhile, to model the hierarchical information flow observed in neural pathways (*e.g.*, the dual-stream theory), we implement channel-wise multiplicative gating between the Enc_i. Specifically, a feature from an upstream region is processed by a Gating Module to generate weights, which modulate the input feature maps of a downstream encoder, thus establishing a directed region-aware dependency. Based on the dual-stream neural pathways, our model implements the following directed interactions:

- **Dorsal Stream:** Occipital → Parietal.
- **Ventral Stream:** Occipital → Temporal.

Fig. 3 presents an example of the interaction between Occipital and Parietal features. Let Enc_1^i and Enc_2^i be the i-th layers of the Occipital feature encoder and Parietal feature encoder, respectively. Let $\mathbf{f}_1^i$ and $\mathbf{f}_2^i$ be their corresponding output features. The gating process is defined as:

$$\mathbf{W}_{O\to P} = \sigma(\mathbf{L}_{O\to P}(\mathbf{f}_1^N)); \quad \hat{\mathbf{f}}_2^1 = \mathbf{f}_2^1 \otimes \mathbf{W}_{O\to P}, \tag{1}$$

where N is the total number of layers in the Occipital encoder, $\mathbf{L}_{O\to P}(\cdot)$ is a learnable linear transformation, and $\sigma(\cdot)$ is the sigmoid function bounding the weights $\mathbf{W}_{O\to P}$. The term $\hat{\mathbf{f}}_2^1$ is the modulated output of the first Parietal layer, serving as the input for subsequent layers, and $\otimes$ denotes element-wise multiplication with broadcasting. This process produces regional feature embeddings $\mathbf{f}_i$, which are then concatenated to form the global feature $\mathbf{f}$.

$$\mathbf{f} = \text{Concat}(\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_K), \tag{2}$$

3.1.2. Semantic Enrichment via Multi-Modal Alignment

To align these fused embeddings with diverse semantic modalities, we project them into multiple latent spaces corresponding to different CLIP embeddings. Specifically, we use a set of mapping networks, M_i, to transform the fused EEG representation $\mathbf{f}$ into these target latent spaces.

$$\hat{\mathbf{c}}_\mathbf{I} = M_{\phi_I}(\mathbf{f}), \quad \hat{\mathbf{c}}_\mathbf{T} = M_{\phi_T}(\mathbf{f}), \quad \hat{\mathbf{c}}_\mathbf{Y} = M_{\phi_Y}(\mathbf{f}), \tag{3}$$

where the mapping networks are multi-layer perceptrons, and ϕ_I, ϕ_T, and ϕ_Y are their learnable parameters.

The alignment between the predicted embeddings ($\hat{\mathbf{c}}$) and ground-truth embeddings ($\mathbf{c}$) is optimized using the InfoNCE loss [21], defined as

$$\mathcal{L}_{\mathbf{a}\to\mathbf{b}} = -\log \frac{\exp(\text{sim}(\mathbf{a},\mathbf{b})/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(\mathbf{a},\mathbf{b_j})/\tau)}, \tag{4}$$

$$\mathcal{L}_{\text{info}}(\hat{\mathbf{c}},\mathbf{c}) = \frac{1}{2}(\mathcal{L}_{\hat{\mathbf{c}}\to\mathbf{c}} + \mathcal{L}_{\mathbf{c}\to\hat{\mathbf{c}}}), \tag{5}$$

where $\text{sim}(\cdot,\cdot)$ is cosine similarity and τ is a temperature hyperparameter. This symmetric formulation effectively aligns the two embedding spaces. Moreover, a cross-entropy (CE) classification loss is applied to the predicted embedding $\hat{\mathbf{c}}_Y$ after a linear projection to class logits $\hat{\mathbf{y}}$:

$$\mathcal{L}_{\text{category}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_i \cdot \log(\text{Softmax}(\hat{\mathbf{y}}_i))\right), \tag{6}$$

where $\hat{\mathbf{y}}$ and $\mathbf{y}$ represent the predicted logits and ground-truth labels. The total loss is a sum of the InfoNCE losses from image and text and the classification loss:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{info}}(\hat{\mathbf{c}}_\mathbf{I},\mathbf{c}_\mathbf{I}) + \mathcal{L}_{\text{info}}(\hat{\mathbf{c}}_\mathbf{T},\mathbf{c}_\mathbf{T}) + \mathcal{L}_{\text{category}}. \tag{7}$$

3.1.3. Generation of Semantic Diffusion Prior

The generation of high-fidelity video content depends on effective latent embeddings to condition the downstream generative model. We observe that embeddings within the CLIP text space serve as a more effective condition for the reconstruction process than features derived directly from EEG. Therefore, translating the EEG-derived features into the CLIP text embedding space is a critical step for improving the final reconstruction quality. Inspired by DALLE-2 [24], we employ a diffusion prior to perform this cross-modal translation. This diffusion prior model learns to generate the CLIP text embedding $\hat{\mathbf{c}}_\mathbf{T}$ conditioned on the other EEG-derived embeddings. For training this prior model, we use the same prior loss, $\mathcal{L}_{\text{prior}}$, as defined in DALLE-2 [24]:

$$\hat{\mathbf{c}}_{\text{diff}} = \text{Concat}(\hat{\mathbf{c}}_I, \hat{\mathbf{c}}_T, \hat{\mathbf{c}}_Y), \tag{8}$$

$$\mathcal{L}_{\text{prior}} = \mathbb{E}_{t\sim[1,T]}\left[\left\| P_\theta(\hat{\mathbf{c}}_\mathbf{T}^{(t)}, t, \hat{\mathbf{c}}_{\text{diff}}) - \hat{\mathbf{c}}_\mathbf{T}\right\|^2\right]. \tag{9}$$

The diffusion prior is trained in an independent stage, following the initial alignment training. Specifically, the feature extractor and multimodal mapping modules are frozen and are used to generate the conditioning vector $\hat{\mathbf{c}}_{\text{diff}}$.

3.2. Temporal-aware Dynamic Aligner (TDA)

Unlike the RSM, which focuses on semantic content, the TDA is specifically designed to address the key challenge of temporal coherence. This module generates a dynamic

temporal blueprint to ensure that the motion and flow of the reconstructed video remain synchronized with the underlying temporal dynamics of the EEG signals. Therefore, the TDA is vital for mitigating the inter-frame inconsistencies that often hamper prior EEG-to-video synthesis methods. To achieve this, the TDA operates in two main stages: the direct generation of a temporal blueprint and its subsequent contrastive alignment with ground-truth video features.

3.2.1. Extraction of Temporal Blueprint

The **TDA** takes the raw EEG signal sequence $E \in \mathbb{R}^{C\times T}$ as input. This sequence is first segmented along the temporal dimension into N overlapping windows of length T_w, $\{\mathbf{E}_1^t, \mathbf{E}_2^t, \cdots, \mathbf{E}_N^t\}$, where each window $\mathbf{E}_i^t \in \mathbb{R}^{C\times T_w}$. Each temporal segment $\mathbf{E}_i^t$ is then processed by a temporal network, denoted TN_{temp}. This architecture is employed for its ability to model time series features and is optimized to capture rhythmic and transitional patterns within the neural signals. The concatenated outputs from TN_{temp} across all segments form the dynamic temporal blueprint, $\mathbf{H}_{temp}$:

$$\mathbf{H}_{temp} = [\mathbf{h_1}, \mathbf{h_2}, \ldots, \mathbf{h_N}], \tag{10}$$

$$\mathbf{h}_i = \text{TN}_{temp}(\mathbf{E}_i^t). \tag{11}$$

This dynamic temporal blueprint is a sequence of latent vectors, where $\mathbf{h}_i$ encodes the anticipated motion and state changes for the corresponding i-th segment of the video.

3.2.2. Contrastive Alignment with Video Features

To ensure the temporal blueprint $\mathbf{H}_{\text{temp}}$ captures both the content and temporal dynamics of the corresponding video, we align it with video features using a dual objective that enforces both content-wise and structural consistency. First, we define the input features. The temporal blueprint from EEG signals is a feature sequence $\mathbf{H}_{\text{temp}} \in \mathbb{R}^{N\times D_{\text{eeg}}}$, where N is the number of frames. The corresponding video is processed by a VAE encoder to yield a sequence of frame-level visual features $\mathbf{V} \in \mathbb{R}^{N\times C\times H\times W}$. Next, both sequences are projected into a shared latent space of dimension D_{latent} using distinct projection heads, P_H and P_V:

$$\mathbf{Z}_H = P_H(\mathbf{H}_{\text{temp}}) \in \mathbb{R}^{N\times D_{\text{latent}}}, \tag{12}$$

$$\mathbf{Z}_V = P_V(\mathbf{V}') \in \mathbb{R}^{N\times D_{\text{latent}}}. \tag{13}$$

With the features projected into this common space, we apply two alignment losses. To enforce direct, frame-by-frame content similarity, we compute the Mean Squared Error (MSE) between the two projected sequences:

$$\mathcal{L}_{\text{HV}} = \frac{1}{N}\sum_{i=1}^{N} \|(\mathbf{Z}_H)_i - (\mathbf{Z}_V)_i\|^2. \tag{14}$$

Moreover, to align the relational dynamics between frames, we introduce a structural similarity loss. This loss compares the intra-modality similarity matrices, $\mathbf{S}_H \in \mathbb{R}^{N\times N}$ and $\mathbf{S}_V \in \mathbb{R}^{N\times N}$, which capture the internal structure of the two feature sequences $\mathbf{Z}_H$ and $\mathbf{Z}_V$. The matrices are computed using cosine similarity, *i.e.*, $(\mathbf{S}_H)_{ij} = \text{cos-sim}((\mathbf{Z}_H)_i, (\mathbf{Z}_H)_j)$. The loss then minimizes the MSE between these two structural representations:

$$\mathcal{L}_{\text{Struct}} = \frac{1}{N^2}\|\mathbf{S}_H - \mathbf{S}_V\|_F^2. \tag{15}$$

The final loss for the TDA module is the sum of the content and structural alignment losses:

$$\mathcal{L}_{\text{TDA}} = \mathcal{L}_{\text{HV}} + \mathcal{L}_{\text{Struct}}. \tag{16}$$

3.3. Dual-Guidance Video Reconstructor (DGVR)

The DGVR module accepts two distinct features, *i.e.*, the Semantic Prior from RSM and the Temporal Blueprint from TDA, to synthesize a high-fidelity video via a reverse diffusion denoising process. This dual-guidance is key to generating videos that are both semantically accurate and temporally coherent. Instead of initiating the reverse diffusion process from pure Gaussian noise, we structure the initial latent variable $\mathbf{x}_T$ using the temporal blueprint $\mathbf{H}_{temp}$. This ensures that the foundational spatio-temporal structure of the video is aligned with the expected dynamics from the outset. Specifically, we first project the blueprint to a latent representation $\mathbf{z}_B$ and define the initial latent as:

$$\mathbf{x}_T = \mathcal{E} + \alpha \cdot \mathcal{U}(\mathbf{z}_B), \tag{17}$$

where $\mathcal{E} \sim \mathcal{N}(0, \mathbf{I})$ is a random Gaussian noise tensor, $\mathcal{U}(\cdot)$ is an upsampling network matching the blueprint's dimensions to the latent space, and α is a hyperparameter controlling the temporal guidance strength. During each step of the reverse diffusion process (from $t = T$ down to 1), the semantic prior, $\hat{\mathbf{c}}_{\text{diff}}$, is injected into the model's U-Net architecture via cross-attention layers, while $\mathbf{z}_B$ serves as an additional condition. This dual-conditioning ensures dynamic and semantic coherence throughout the denoising process. The U-Net of the diffusion model (ϵ_θ) is trained to predict the noise ϵ from a noisy latent $\mathbf{x}_t$ at a given timestep t. The prediction is conditioned on $\mathbf{x}_t$, t, the semantic prior, $\hat{\mathbf{c}}_{\text{diff}}$, and the projected temporal blueprint $\mathbf{z}_B$. The objective function is the standard diffusion loss:

$$\mathcal{L}_{\text{DGVR}} = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \hat{\mathbf{c}}_{\text{diff}}, \mathbf{z}_B)\|^2\right], \tag{18}$$

where $\mathbf{x}_t$ is the noisy latent at timestep t, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. After the iterative denoising process is complete, a final decoder module transforms the clean latent representation $\mathbf{x}_0$ into the pixel-space video. By using both dynamic and semantic guidance, this dual-conditional approach reconstructs videos with superior fidelity and coherence. We note that the core generative mechanism can be instantiated using various advanced video diffusion models (*e.g.*, Tune-A-Video [40] and CogVideoX [42]).

Method	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
ShallowNet [27]	5.59/2.27*	16.93/4.66*	21.40/1.96*	49.62/2.34*	27.00/2.09*	56.62/1.77*	66.15/0.89	64.87/1.54	73.21/1.52
DeepNet [27]	4.56/1.52*	14.30/3.25*	20.27/1.25*	48.06/1.59*	26.37/1.95*	55.42/0.59*	65.71/0.24	61.58/3.93	72.86/0.40
EEGNet [13]	4.64/0.86*	14.25/1.87*	19.63/0.81*	47.04/1.45*	25.46/1.31*	51.99/2.00	64.67/0.60	61.37/1.31	72.38/0.98
Conformer [32]	4.93/1.57*	15.36/4.44*	20.92/0.98*	49.25/1.49*	**27.53/1.37***	55.02/0.83*	65.73/0.26	64.96/1.14	73.00/0.85
TSConv [33]	4.92/0.99*	15.05/2.31*	20.00/1.01*	47.76/1.51*	26.89/1.83*	55.32/0.99*	65.39/0.41	64.39/1.47	72.68/0.67
GLMNet [18]	6.20/3.02*	17.75/4.24*	21.93/1.87*	50.01/2.52*	27.33/1.45*	**57.35/1.98***	66.21/0.91	65.10/1.45	73.34/1.31
Ours	**8.27/2.85***	**22.38/4.80***	**22.73/3.07***	**51.93/4.73***	27.43/1.11*	56.44/1.51*	**67.56/0.94**	**82.47/0.81***	**73.64/1.70**

Table 1. Average classification accuracy (%) and standard deviation (std) across all subjects with different EEG classifiers on different tasks. Chance level is the percentage of the largest class. * indicates a result significantly above chance level (two-sample t-test: $p < 0.05$).

# Classes	Method	Video-based			Frame-based		
		Semantic-level		Pixel-level	Semantic-level		Pixel-level
		2-way	40-way	FVMD↓	2-way	40-way	SSIM↑
10	Ours	0.847±0.01	0.394±0.03	1601.84±0.02	0.833±0.02	0.308±0.01	0.309±0.02
	EEG2Video	0.852±0.02	0.340±0.01	1977.13±0.01	0.798±0.03	0.232±0.02	0.300±0.03
20	Ours	0.835±0.01	0.345±0.01	1587.91±0.01	0.818±0.02	0.277±0.02	0.290±0.02
	EEG2Video	0.813±0.02	0.273±0.03	1960.20±0.03	0.785±0.04	0.184±0.02	0.242±0.03
30	Ours	0.833±0.02	0.309±0.01	1661.06±0.03	0.805±0.02	0.254±0.01	0.293±0.01
	EEG2Video	0.794±0.02	0.209±0.05	2016.67±0.04	0.785±0.04	0.180±0.02	0.228±0.04
40	Ours	0.828±0.02	0.284±0.02	1637.55±0.01	0.807±0.03	0.241±0.01	0.280±0.01
	EEG2Video	0.798±0.03	0.159±0.01	2038.27±0.02	0.774±0.02	0.138±0.01	0.256±0.03

Table 2. Reconstruction results on SEED-DV across varying class numbers, evaluated on semantic/pixel metric at the video/frame level.

4. Experiments

4.1. Experimental Setup

Dataset and Baselines. Our experiments are conducted across two distinct EEG-video datasets to demonstrate the versatility of our proposed model. The **SEED-DV** dataset [18] contains EEG signals from 20 subjects viewing videos across 40 visual conceptual categories. We benchmark DynaMind against foundational EEG models (*i.e.*, ShallowNet [27], DeepNet [27], EEGNet [13], Conformer [32], TSConv [33]) and the SOTA approach GLMNet [18]. The **CineBrain** dataset [10] features EEG and fMRI recordings from 6 participants watching naturalistic audiovisual stimuli from *The Big Bang Theory*. We use the EEG-only results reported in the CineBrain as the baseline. These comparisons enable a rigorous evaluation of DynaMind on established protocols.

Implementation and Evaluation. Our framework utilizes three core modules: the RSM, the TDA, and the DGVR. However, the underlying generative model and evaluation protocol are adapted to each target dataset and its comparison standard. For the **SEED-DV**, the RSM aligns regional EEG embeddings with semantic features from a pre-trained CLIP ViT-L/14 model. The DGVR fine-tunes a Stable Diffusion V1-4 model to synthesize 6-frame video clips at 512×288 resolution (3 fps). We strictly follow the evaluation protocol established by EEG2Video [18] to ensure fair comparison, including classification accuracy (7 categories of videos), frame-level quality (SSIM [38] and N-way top-K accuracy based on CLIP features) and video-level quality (N-way top-K accuracy based on VideoMAE and FVMD [17]). For the **CineBrain**, the DGVR uses CogVideoX [42] to align with the generative architecture used by CineSync [10]. This allows our model to reconstruct higher-resolution and longer video clips (*e.g.*, 480×720 resolution, 33 frames per clip) from the EEG signals, and the RSM is trained without explicit classification supervision. The evaluation follows the Cine-Benchmark [10] protocol, which is more comprehensive for naturalistic human-centric videos, assessing semantic-level metrics (N-way top-K accuracy and FVD), and perceptual-level metrics (DTC, CTC, SSIM, PSNR) across both frame-based and video-based assessments. The core modules are trained independently, with hyperparameters tuned for each specific dataset. **More details about the implementation and evaluation are in supplementary material.**

4.2. Main Results

Classification Accuracy. We evaluate the features from the RSM module using classification accuracy on the SEED-DV dataset, with detailed results presented in Tab. 1. The results show our proposed model surpasses existing baselines

Method	Video-based					Frame-based			
	Semantic-level			Perceptual-level		Semantic-level		Perceptual-level	
	2-way ↑	50-way ↑	FVD ↓	DTC ↑	CTC ↑	2-way ↑	50-way ↑	SSIM ↑	PSNR ↑
GLFA [14]	0.801	0.167	128.76	0.706	0.735	0.847	0.225	0.123	7.526
CineSync-EEG [10]	0.891	0.304	53.75	0.899	0.937	0.918	**0.349**	**0.231**	**11.75**
Ours	**0.902**	**0.317**	**51.50**	**0.907**	**0.942**	**0.920**	**0.349**	0.225	10.98

Table 3. Reconstruction results on CineBrain (EEG-only), evaluated on semantic/perceptual metric at the video/frame level.

Method	SEED-DV					CineBrain						
	Cls.	Video-based		Frame-based		Video-based				Frame-based		
	40-c ↑	40-way ↑	FVMD ↓	40-way ↑	SSIM ↑	50-way ↑	FVD ↓	DTC ↑	CTC ↑	50-way ↑	SSIM ↑	PSNR ↑
Ours	**8.27**	**0.284**	1637.55	**0.241**	**0.280**	**0.317**	**51.50**	**0.907**	**0.942**	**0.349**	**0.225**	**10.98**
Brain Regions												
w/o Frontal	7.67	0.264	1677.42	0.220	0.267	0.300	57.41	0.891	0.926	0.341	0.216	10.40
w/o Parietal	7.65	0.263	**1631.21**	0.221	0.267	0.291	60.89	0.877	0.918	0.328	0.211	10.33
w/o Occipital	6.73	0.239	1712.11	0.189	0.253	0.254	70.88	0.868	0.903	0.310	0.208	10.11
w/o Temporal	7.22	0.252	1658.89	0.205	0.260	0.297	60.64	0.880	0.916	0.321	0.211	10.38
Features												
w/o Image	8.08	0.281	1649.77	0.235	0.277	0.302	57.33	0.891	0.929	0.341	0.220	10.67
w/o Text	7.56	0.252	1703.48	0.212	0.270	0.245	76.23	0.866	0.899	0.301	0.202	9.98
w/o Category	8.11	0.276	1644.07	0.232	0.278	-	-	-	-	-	-	-
Consistency												
w/o Consistency	-	0.279	1916.91	0.240	**0.280**	0.308	56.88	0.859	0.887	0.348	0.221	10.78

Table 4. Ablation study on the key components of our model, evaluated across both the **SEED-DV** (Classification, Video-based, and Frame-based tasks) and **CineBrain** (Video and Frame Reconstruction tasks) datasets. The consistent performance drop after removing individual brain regions, features, or the consistency module validates the crucial effectiveness of each component.

across most metrics, setting a new SOTA performance. In the challenging 40-class task, our model achieves 8.27% Top-1 accuracy, outperforming the runner-up, GLMNet (6.20%). Similarly, in the 9-class task, our model achieves the highest Top-1 accuracy of 22.73%. This consistency suggests our model captures richer and more discriminative semantic details from the EEG signals. The model's performance on fine-grained attribute recognition is particularly strong; for the "Human Face" task, it achieves 82.47% accuracy, over 17 percentage points above GLMNet (65.10%). This margin of improvement demonstrates our model's capacity for decoding specific, high-level semantic concepts. Furthermore, our model achieves leading results in other fine-grained tasks, including "Numbers" and "Human". In summary, these results verify our model's ability to learn highly discriminative and semantically meaningful features for the subsequent high-fidelity video reconstruction.

Video Reconstruction. We evaluate generated video quality on two datasets. On **SEED-DV**, we compare our model to the EEG2Video baseline on 10-40 class tasks. Tab. 2 assesses semantic and pixel metrics at video and frame levels. Our model substantially outperforms the baseline in all settings, especially as task difficulty increases (10-40 classes). In the 40-class task, our video semantic accuracy is 0.284 versus the 0.159 of the baseline. Our frame-based semantic accuracy is 0.241 versus the baseline's 0.138. This shows our model provides more precise semantic guidance. Pixel-level FVMD assesses quality and coherence (lower is better). Our model achieves an FVMD of 1637.55 in the 40-class task, markedly lower than the baseline's 2038.27. This indicates superior smoothness and temporal consistency, validating our TDA module. On **CineBrain**, we compare our model against GLFA and CineSync-EEG [10] for EEG-only video reconstruction. Tab. 3 details the semantic and perceptual metrics at video and frame levels. Our model outperforms both baselines (GLFA/CineSync-EEG) on nearly all metrics. Our model's superiority is most pronounced in video-based metrics. In the semantic-level evaluation,, our model achieves higher classification accuracy (2-way 1.23%, 50-way 4.28%), showing superior video-level semantic capture. Our model also obtains a lower FVD (a 4.19% reduction), indicating a closer feature space distribution. On the perceptual-level, our model shows clear progress on DTC and CTC, confirming its dynamic consistency. In conclusion, these results verify the effectiveness of our model across both semantic and pixel dimensions. Our model sets a new bench-

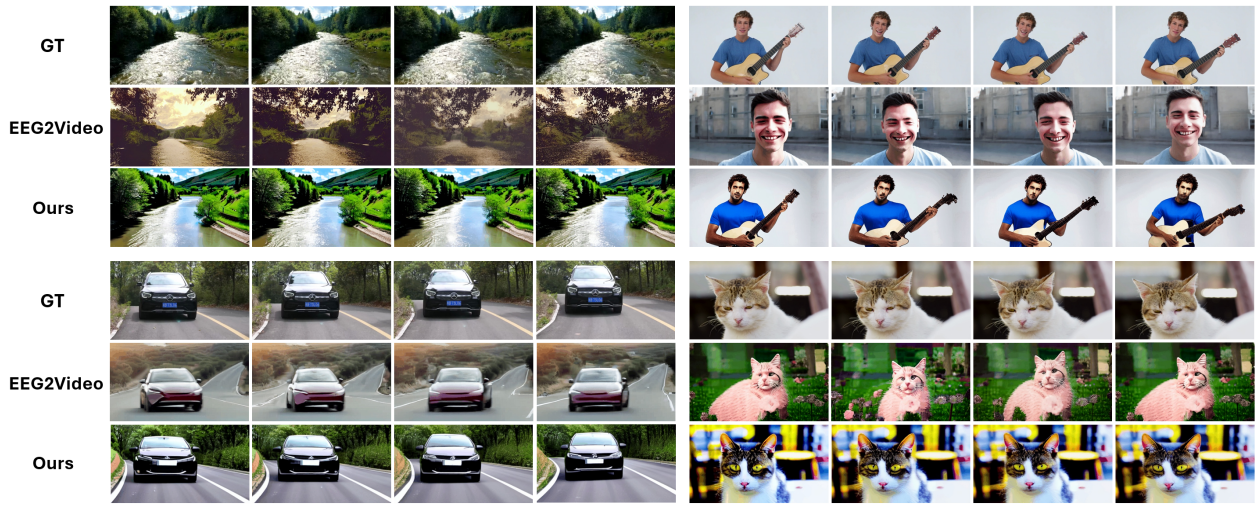

Figure 4. Video reconstruction results of ours and EEG2Video on SEED-DV dataset compared against the ground truth. The results of EEG2Video are from their openly released samples. **More result on SEED-DV and CineBrain dataset are found in supplementary.**

mark for EEG-based video reconstruction in both content accuracy and visual fidelity.

4.3. Ablation Study

We conduct ablation studies to validate our model's key components, as shown in Tab. 4. For each dataset, we perform ablations on most major evaluation metrics.

Effect of Brain Regions. The Occipital Lobe shows the largest performance impact, aligning with its primary role in visual processing. However, removing signals from any single brain region (*e.g.*, frontal, parietal, or temporal) also degrades performance. This result underscores the collective, distributed contribution of all regional features.

Effect of Features. Ablating modal features reveals that removing the text feature (w/o Text) causes the largest drop in semantic and classification, highlighting its critical role. Performance degradation from removing image or category features also validates our multimodal fusion strategy.

Effect of Consistency. Removing the structural consistency loss (**w/o Consistency**) markedly worsens the FVMD score (from 1637.55 to 1916.91). This sharp increase confirms the objective's key role in improving the temporal coherence of the generated video.

4.4. Visualization

The qualitative comparison results demonstrate our model's advantage in semantic fidelity. On the SEED-DV dataset, Fig. 4 shows an illustrative case with the reconstruction of a "cat". Our model successfully generates a cat with realistic morphology and texture, exhibiting high fidelity to the ground truth. The EEG2Video baseline fails to capture the correct semantics, generating a morphologically distorted object with severe content hallucination. This superiority is also evident in other examples: the "car" reconstructed by our model features a well-defined contour and accurate coloration, while the "river" scene is rendered with vivid colors and rich details. Beyond static content, our model also exhibits better temporal coherence. In the dynamic scenes of the "river" and "car", our generated videos feature smooth transitions, reducing the flickering and warping artifacts common in the baseline. This result validates our TDA in maintaining dynamic consistency. Moreover, when handling complex scenes containing a "human," our model reconstructs the scene composition and human poses, whereas the baseline struggles to produce coherent imagery.

5. Conclusion

This work addresses key challenges in EEG-to-video reconstruction: poor integration of distributed neural information and weak temporal coherence. Therefore, we propose DynaMind, a new model integrating multimodal semantics and temporal dynamics to reconstruct dynamic visual scenes. DynaMind employs three core modules: a Region-aware Semantic Mapper for extracting region-aware multimodal semantic priors, a Temporal-aware Dynamic Aligner for generating a dynamically coherent temporal blueprint, and a Dual-guided Video Reconstructor for synthesizing the final video using this dual guidance. Qualitative and quantitative experiments on the SEED-DV and CineBrain dataset show that DynaMind achieves superior performance, outperforming existing methods in both EEG classification and reconstructed video quality. Ablation studies confirm the necessity of each component, advancing the decoding of visual experiences from non-invasive brain recordings. Our findings offer a new path for high-fidelity neural decoding by integrating cognitive priors into generative models.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62402341, 62472319, in part by the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (CPSF) under Grant GZC20241225, and in part by the China Postdoctoral ScienceFoundation under Grant Number 2025M771513.

References

[1] Peyman Adjamian, Avgis Hadjipapas, Gareth R Barnes, Arjan Hillebrand, and Ian E Holliday. Induced gamma activity in primary visual cortex is related to luminance and not color contrast: An meg study. *Journal of Vision*, 8(7):4–4, 2008. 2

[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 *European Conference on Computer Vision*, pages 472–488. Springer, 2024. 1, 3

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 3

[4] Minkyu Choi, Kuan Han, Xiaokai Wang, Yizhen Zhang, and Zhongming Liu. A dual-stream neural network explains the functional segregation of dorsal and ventral visual pathways in human brains. *Advances in Neural Information Processing Systems*, 36:50408–50428, 2023. 2

[5] Cristina de la Malla, Eli Brenner, Edward HF de Haan, and Jeroen BJ Smeets. A visual illusion that influences perception and action through the dorsal pathway. *Communications biology*, 2(1):38, 2019. 2

[6] Ioannis Delis, Robin AA Ince, Paul Sajda, and Qi Wang. Neural encoding of active multi-sensing enhances perceptual decision-making via a synergistic cross-modal interaction. *Journal of Neuroscience*, 42(11):2344–2355, 2022. 2

[7] Yidan Ding, Chalisa Udompanyawit, Yisha Zhang, and Bin He. Eeg-based brain-computer interface enables real-time robotic hand control at individual finger level. *Nature Communications*, 16(1):1–20, 2025. 1

[8] Changde Du, Changying Du, Lijie Huang, and Huiguang He. Reconstructing perceived images from human brain activities with bayesian deep multiview learning. *IEEE transactions on neural networks and learning systems*, 30(8):2310–2323, 2018. 2

[9] Karolina Finc, Kamil Bonna, Xiaosong He, David M Lydon-Staley, Simone Kühn, Włodzisław Duch, and Danielle S Bassett. Dynamic reconfiguration of functional brain networks during working memory training. *Nature communications*, 11 (1):2435, 2020. 2

[10] Jianxiong Gao, Yichang Liu, Baofeng Yang, Jianfeng Feng, and Yanwei Fu. Cinebrain: A large-scale multi-modal brain dataset during naturalistic audiovisual narrative processing. *arXiv preprint arXiv:2503.06940*, 2025. 2, 6, 7

[11] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016. 2

[12] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1809–1817, 2017. 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

[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 *European Conference on Computer Vision*, pages 353–369, 2024. 7

[15] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024. 2

[16] Hanwen Liu, Daniel Hajialigol, Benny Antony, Aiguo Han, and Xuan Wang. Eeg2text: Open vocabulary eeg-to-text decoding with eeg pre-training and multi-view transformer. *arXiv preprint arXiv:2405.02165*, 2024. 2

[17] Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fr\'echet video motion distance: A metric for evaluating motion consistency in videos. *arXiv preprint arXiv:2407.16124*, 2024. 6

[18] Xuan-Hao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-Liang Lu, and Wei-Long Zheng. Eeg2video: Towards decoding dynamic visual perception from eeg signals. *Advances in Neural Information Processing Systems*, 37:72245–72273, 2024. 1, 2, 6

[19] M-Marsel Mesulam. From sensation to cognition. *Brain: a journal of neurology*, 121(6):1013–1052, 1998. 3

[20] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[22] Xuyu Qian, Kyle Coleman, Shunzhou Jiang, Andrea J Kriz, Jack H Marciano, Chunyu Luo, Chunhui Cai, Monica Devi Manam, Emre Caglayan, Abbe Lai, et al. Spatial transcriptomics reveals human cortical layer and area specification. *Nature*, pages 1–11, 2025. 2

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021. 1

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 4

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[26] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[27] 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

[28] Simon Scholler, Sebastian Bosse, Matthias Sebastian Treder, Benjamin Blankertz, Gabriel Curio, Klaus-Robert Muller, and Thomas Wiegand. Toward a direct measure of video quality perception using eeg. *IEEE transactions on Image Processing*, 21(5):2619–2629, 2012. 1

[29] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024. 2

[30] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[31] 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. 1

[32] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022. 6

[33] 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. 1, 2, 6

[34] Jingyuan Sun, Mingxiao Li, Zijiao Chen, Yunhao Zhang, Shaonan Wang, and Marie-Francine Moens. Contrast, attend and diffuse to decode high-resolution images from brain activities. *Advances in Neural Information Processing Systems*, 36:12332–12348, 2023. 2

[35] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14453–14463, 2023. 2

[36] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 3

[37] Xingyi Wang, Yuliang Ma, Jared Cammon, Feng Fang, Yunyuan Gao, and Yingchun Zhang. Self-supervised eeg emotion recognition models based on cnn. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1952–1962, 2023. 1

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[39] Yayun Wei, Lei Cao, Hao Li, and Yilin Dong. Mb2c: Multimodal bidirectional cycle consistency for learning robust visual neural representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8992–9000, 2024. 1

[40] 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. 3, 5

[41] Guangyu Yang and Jinguo Liu. A new framework combining diffusion models and the convolution classifier for generating images from eeg signals. *Brain Sciences*, 14(5):478, 2024. 1

[42] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 5, 6