EVOKE: Efficient and High-Fidelity EEG-to-Video Reconstruction via Decoupling Implicit Neural Representation

Haodong Jing[1*], Panqi Yang[1*], Dongyao Jiang[1], Zhipeng Liu[2], Nanning Zheng[1†], Yongqiang Ma[1†]

[1]State Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center of Visual Information and Applications,
and Institute of Artificial Intelligence and Robotics, Xi'an Jiao Tong University
[2]School of software, Northeastern University
{jinghd, yangpq, jdy20020305}@stu.xjtu.edu.cn, 2310543@stu.neu.edu.cn, {nnzheng,musayq}@xjtu.edu.cn

Abstract

Visual neural decoding is an important research topic at the intersection of cognitive neuroscience and machine learning. While recent progress has been made in EEG-based neural decoding, reconstructing dynamic visual content remains challenging. In the field of EEG decoding, current models either utilize pre-trained encoders for feature extraction or employ graph neural networks to represent the spatio-temporal information embedding, resulting in poor model representation and high complexity. We propose ***EVOKE*** – an innovative framework for zero-shot decoding of high-fidelity videos from EEG signals. EVOKE employs *Implicit Neural Representations* to perform complete spatial modeling of EEG and continuously decouples information in the EEG-INR perceptual space. Additionally, we construct a *Hierarchical-aware Attention Module (HAM)* to decode EEG from three feature anchors: visual, semantic, motion, and progressively control task inference. The *Motion Attention Flow (MAF)* we developed overcomes the limitations of capturing motion features in dynamic stimuli, creating a more robust representation that enhances reconstruction consistency. Comprehensive experiments prove that SOTA performance of EVOKE (0.353 SSIM, 0.715 CLIP-pcc). We provide an effective method for converting brain activity into rich visual experiences and set a new benchmark for brain multimodal generation.

1 Introduction

Human visual cognition is a continuous and dynamic process, rather than a static image presentation (Makeig et al. 2002). For centuries, the brain's perception and processing of the ever-changing world have been a topic of intense research (Bartels 2005; Hafri et al. 2017). The human brain is capable of integrating information from the real world at an astonishing speed and efficiency, blending visual details, object motion, and semantic understanding into a coherent experience (Watson 2000). Unraveling the information within these complex brain activities is a significant challenge for cognitive neuroscience, various neuroimaging techniques have been widely applied, demonstrating the potential of cutting-edge developments in neuroscience and BCI research (Kriegeskorte 2018; Xu et al. 2021).

Recent advances in multimodal learning and generative AI have taken neural decoding to a new level, with outstanding work such as MindVis (Chen et al. 2023a), Mindeye (Scotti et al. 2023), NeuroCreat (Jing et al. 2025a), and so on (Liu et al. 2025b; Ma et al. 2024). However, extending this reconstruction capability from static to dynamic visual requires addressing challenges such as seamless temporal fusion and dynamic changes. Currently, mainstream dynamic visual reconstruction primarily relies on fMRI (Liu et al. 2025a), while fMRI offers advantages in spatial resolution, its temporal resolution is low, with acquisition intervals of at least 2 seconds per frame (Buckner 1998). This significant delay in neural activity-stimulus resolution limits dynamic visual reconstruction (Dubois and Adolphs 2016).

EEG signals, with their high temporal resolution and portability, show great potential in brain decoding and have been widely used in human vision-related research (Sabharwal and Rama 2024; Liu et al. 2024a; Jing et al. 2025b). Dynamic visual decoding based on brain signals involves capturing complex visual cues from neural signals, and the rich spatio-temporal information of EEG provides the decoding basis. Some researchers have attempted to construct a diffusion model for EEG-Video visual reconstruction and achieved positive results (Liu et al. 2024b; Le et al. 2022), validating the feasibility of this research.

One of the challenges of EEG signal decoding lies in its ***Insufficient spatial representation.*** EEG signals are typically recorded from 32 or 64 electrodes, each capturing activity from different brain regions. This limited spatial distribution makes it difficult to leverage the rich spatial information embedded in brain activity fully. While spatial graph neural networks have been used to model electrode correlations (Klepl et al. 2024; Graña et al. 2023; Demir et al. 2021), they often struggle with issues such as electrode misalignment and sparse layouts, limiting their effectiveness in dynamic decoding tasks.

Another challenge is the ***Missing multi-granularity feature in brain-visual alignment.*** Recent EEG decoding methods align brain signals directly with CLIP features (Radford et al. 2021) in latent space, overlooking the multi-level structure of visual perception (Watson 2000). The vi-

[*]These authors contributed equally.
[†]Corresponding author

Figure 1: **Left:** Inference comparison, the baseline model represents the CLIP mapping and decoding from EEG. EVOKE decouples visual features in a smoother EEG-INR perceptual space and accurately decodes dynamic features. **Right:** Compared to the baseline, EVOKE achieved high-quality, high-fidelity, and high-fluidity dynamic visual reconstruction.

sual cortex processes information hierarchically (Horikawa and Kamitani 2017), from low-level features to high-level semantics, progressive decoding of brain signals helps achieve higher alignment in visual learning, but existing models often use a flat representation paradigm (Chen et al. 2023a; Takagi et al. 2023) that cannot capture this neural relationship and multi-granularity visual information.

Based on the above insights, we propose **EVOKE**, a novel and flexible EEG-to-video reconstruction framework, as shown in Figure 1. We adopted implicit neural representation as a more flexible and robust basis for EEG spatial modeling and construct the **EEG-INR perceptual space**, its effective and smooth continuous spatial representation (Sitzmann et al. 2020; Grattarola and Vandergheynst 2022) supports the fine-grained decoupling of dynamic visual features of EEG. Furthermore, its scalability allows feature encoding to be infinitely extended to any modality. Following this, we introduce the **Hierarchical-aware Attention Module (HAM)**, which embeds multi-layer visual decoding features into the cross-attention layer, constrains the representation of objects through motion features, refines the core semantics, and integrates multi-level features across streams. Additionally, we constructed the **Motion Attention Flow (MAF)**, using EEG and video frame attention blocks to align visual motion features, enabling efficient learning of motion perception. During inference, the decoupled features controllably guide the enhanced representation of the generative model, jointly promoting motion consistency and semantically faithful video reconstruction.

We conducted comprehensive experiments on SEED-DV dataset, EVOKE consistently achieves SOTA performance.

In summary, our contributions are summarized as follows:

- We propose **EVOKE**, the first attempt to introduce implicit neural representations to break through representation limitations. It uses the continuous embedding of EEG-INR perceptual space to capture spatio-temporal features in EEG and finely decouple visual information.
- We innovatively proposed the Hierarchical-aware Attention Module, realized bottom-up multi-level EEG feature fusion and learning, and enhanced hierarchical alignment of visual features. We also designed Motion Attention Flow to achieve fine-grained motion feature perception.
- We have developed a unified inference stage, controllably decoupling features jointly promotes motion consistency and semantically faithful coherent video reconstruction, achieving zero-shot brain decoding.

2 Related Work

2.1 EEG-Based Brain Decoding

Due to its portability and low cost, EEG plays an important role in brain analysis and representation. Many researchers have been working to extract useful information from EEG, achieving breakthroughs in various decoding tasks such as movement, emotion, and text (Li et al. 2022; Altaheri et al. 2023; Duan et al. 2023). Given the similarity between brain voxel connections and GNN node communication, many studies have attempted to use GNN to construct brain functional connectivity matrices (Song et al. 2018; Li et al. 2021; Klepl et al. 2024) enabling the analysis of EEG data in a graph domain. This approach allows for the capture of complex brain network spatial information that other models cannot detect, and it have been applied to EEG decoding tasks such as emotion recognition, BCI, and neurological disorders. However, the GNN struggle to effectively represent the spatio-temporal dependencies of sparse EEG, and the dimension of the adjacency matrix grows exponentially with the number of time points in dynamic modeling, leading to computational overhead increase (Garg et al. 2020; Waikhom et al. 2023). Therefore, it is necessary to construct a *more flexible and robust EEG modeling space* to achieve smooth learning and constraints of spatial features.

2.2 Video Reconstruction from Brain Signals

Recently, visual decoding based on brain signals has attracted considerable attention, and has made significant progress in fields such as images (Scotti et al. 2024), videos (Chen et al. 2023b), and 3D reconstruction (Gao et al. 2024). The brain activity reconstruction primarily focuses on two types of data: fMRI and EEG. In fMRI decoding, Mindeye series (Scotti et al. 2023) have achieved high-precision image reconstruction, while Mind-Video (Chen et al. 2023b), NeuralFlix (Sun, Li, and Moens 2025) have driven improvements in video reconstruction. NeuroClips (Gong et al.

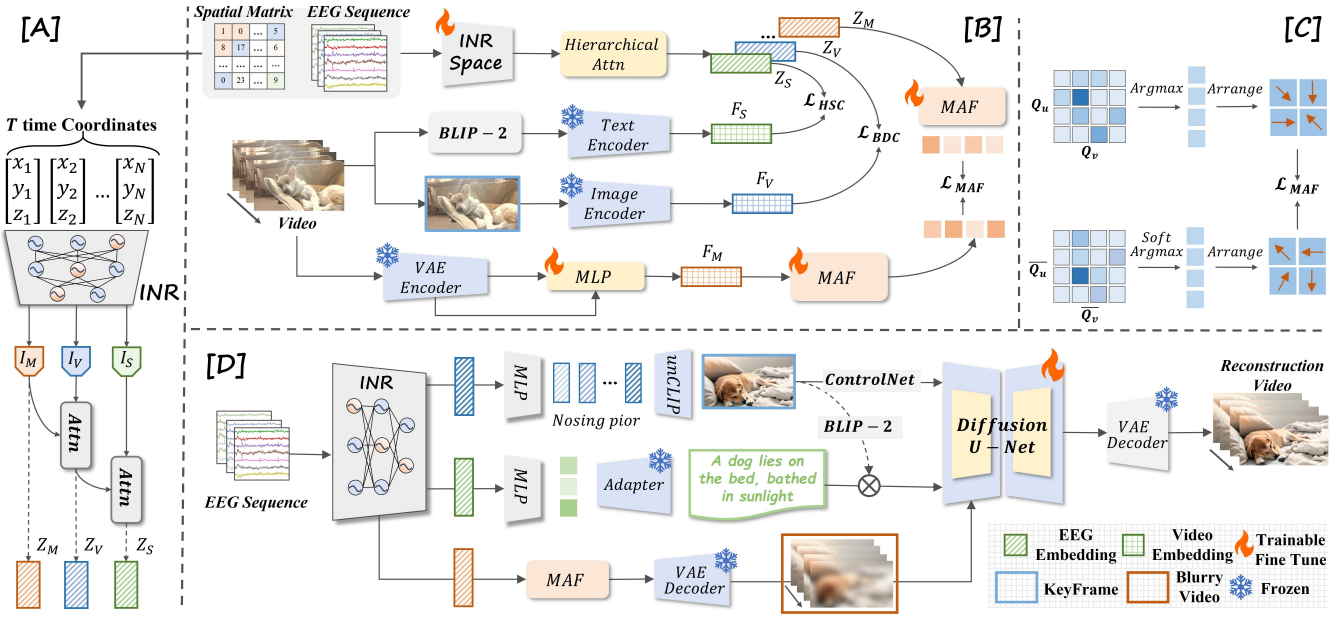

Figure 2: Framework of EVOKE. **[A]:** Use implicit neural representations to construct EEG-INR perceptual space, achieving smooth feature representation and decoupling. **[B]:** Hierarchical-aware Attention Module to fuse EEG features and achieve cross-modal contrastive learning of multi-layer features. **[C]:** Motion Attention Flow mechanism captures the motion changes. **[D]:** Enhanced video synthesis to reconstruct coherent and temporally consistent videos.

2024) and Mind-Animator (Lu et al. 2024) further enhance video reconstruction performance through semantic alignment and fine-grained dynamic perception. EEG offers millisecond-level temporal resolution, aligning more closely with dynamic visual decoding. Brain2Pix (Le et al. 2022) has achieved frame-level video reconstruction based on EEG first time, while EEG2Video (Liu et al. 2024b) extend it by a large-scale EEG-Video dataset and Seq2Seq architecture. However, it should be noted that existing methods have not yet achieved comprehensive capture of visual information, the *fine-grained fusion of EEG and dynamic visual features* remains a challenge that needs to be solved.

3 Methodology

EVOKE framework is shown in Figure 2. It contains: (1) EEG-INR Perceptual Space for continuous representation of spatio-temporal features. (2) Hierarchical-aware Attention Module for bottom-up integration of multi-level EEG features, and the Motion Attention Flow contained in it realizes the capture of motion features. (3) Video Synthesis pipeline.

3.1 Date Preprocess

The research objective is to reconstruct dynamic video sequences from EEG signals collected under continuous video stimulation. We denote $\{EEG, Video\}$ as $\boldsymbol{S} = (\boldsymbol{E}_i, \boldsymbol{V}_i)_{i=1}^{n}$.

The EEG data at each time point consists of a spatial information matrix of size $C_E \times 3$, which stores all the 3-dimensional spatial coordinates of the EEG signals. We treat the different EEG time points $t \in \{1, 2, 3, \cdots, T\}$ as additional time coordinates, enabling us to represent the entire EEG sequence as $\boldsymbol{E} \in \mathbb{R}^{C_E \times T}$ with C_E channels and T time points, each coordinate vector $\boldsymbol{E}|(d, \cdot)$ stores 3 spatial coordinates and a time index of the spot d. The goal is to reconstruct video sequence $\boldsymbol{V} \in \mathbb{R}^{H \times W \times 3 \times N}$ with height H, width W, RGB channels, and N frames.

We use the video sequence $\{\boldsymbol{V}_i\}$ as the base matrix to deconvolve the EEG, and we select the top-K EEG data points with high representation t-test. In addition, to improve the stability of data analysis, we normalize the EEG data. The normalized EEG expression matrix $\{\boldsymbol{X}_i\} \in \mathbb{R}^{C_N \times T}$ is:

$$\boldsymbol{X}_i|_{(d,n)} = \log\left(\frac{M\boldsymbol{E}_i|_{(d,n)}}{\sum_{n=1}^{N} \boldsymbol{E}_i|_{(d,n)}} + 1\right), \tag{1}$$

where M is a scaling factor.

3.2 Unified EEG-INR Perceptual Space

INR Network Here, we use INR to map the information from each EEG to the corresponding spatio-temporal features $\boldsymbol{X}_{INR} = \mathcal{M}_\theta(\boldsymbol{X})$, $\mathcal{M}_\theta(\cdot) : \mathbb{R}^T \rightarrow \mathbb{R}^{D_{TS}}$ is the INR that maps the three-dimensional information of each EEG point to the corresponding spatio-temporal representation vector, and $\boldsymbol{X}_{INR} \in \mathbb{R}^{C_N \times D_{TS}}$ represents the output dynamic reconstruction information. We use a sine-activated MLP as the INR framework (Sitzmann et al. 2020). Specifically, the INR structure constructed has the structure:

$$\mathcal{M}_\theta(\boldsymbol{X}) := W_D(\sin(W_{D-1} \cdots \sin(W_1 \boldsymbol{X}))), \tag{2}$$

where $\{\boldsymbol{W}_d\}_{d=1}^{D}$ is the weight matrix and D is the depth of INR. The output of INR is expected to be the features of the observed EEG data. At the same time, INR can be used to eliminate the inherent noise and sparsity of EEG data by utilizing the implicit bias of low-frequency and smooth components. The sinusoidal activation of INR can maintain a high

Lipschitz smoothness, which effectively ensures smoothness across electrodes and subjects in EEG decoding.

Therefore, INR implicitly encodes the correlation between the time dimension and the representation of different electrode spatial frequency domains in EEG data, and proves that the inherent smoothness of this representation is effective for EEG data denoising.

Decoupling Feature Representation After obtaining meaningful spatio-temporal features $\boldsymbol{X}_{INR}$ for decoding EEG signals, we use them for downstream tasks. We first use a spatio-temporal encoder $\mathcal{G}_\theta(\cdot) : \mathbb{R}^{D_{TS}} \rightarrow \mathbb{R}^b$ (two-layer MLP with sine-activation) to transform $\boldsymbol{X}_{INR}$ into a latent representation $\boldsymbol{I} = \{\boldsymbol{I}^T, \boldsymbol{I}^S\} \in \mathbb{R}^{C_N \times b}$. Notably, the dimension of the latent representation $\boldsymbol{I}$ is lower than EEG channels, and we use this latent representation as the decoupling feature extraction result, which helps to represent the spatio-temporal organizational structure. Due to the good representation of $\boldsymbol{X}_{INR}$, our encoder can more effectively extract the spatial interactions between EEG and video, enhancing downstream decoding tasks.

Since the input data $\boldsymbol{X}_{INR}$ is generated from the spatial coordinates and time-frequency domain coordinates of EEG via INR, the information decoded from electrodes in different brain regions can more clearly correspond to the frame matrix of video. Thus, we implicitly obtain EEG deconvolution results enhanced with spatial information. We further fuse decoupling features by combining direct feature concatenation with attention-weighted cross-attention:

$$\begin{aligned} \delta_{ts} &= \mathrm{softmax}\left(\frac{(\boldsymbol{I}^T\mathbf{W}_t)(\boldsymbol{I}^S\mathbf{W}_s)^\top}{\sqrt{d_k}} + \mathbf{G}_{\text{mask}}\right) \\ \boldsymbol{I}_{\text{out}} &= \mathrm{LayerNorm}(\mathrm{MLP}([\boldsymbol{I}^T; \boldsymbol{I}^S; \boldsymbol{I}^T \cdot \delta_{ts} \cdot \boldsymbol{I}^S])) \end{aligned}, \quad (3)$$

where $\mathbf{W}_t$ and $\mathbf{W}_s$ are learnable projection matrices for temporal and spatial features, d_k is the scaling factor $\mathbf{G}_{\text{mask}}$ is the mask bias used to ignore invalid positions. EEG are decoupled into three hierarchical features: $\boldsymbol{I}_S$, $\boldsymbol{I}_V$, and $\boldsymbol{I}_M$.

3.3 Hierarchical-aware Attention Module

In the visual cognition process, motion features can further enhance the representation ability of visual perception features, while visual perception features can drive fine-grained semantic understanding. Inspired by biological mechanisms, we propose the *Hierarchical-aware Attention Module (HAM)*, which integrates the visual perception, semantic, and motion features extracted by EEG-INR space. Through this strategy, we can reveal the causal contributions of different feature factors to decoding performance, providing theoretical insights for developing more neuro-interpretable brain-visual models.

First, we use the CLIP-ViT (Radford et al. 2021) to extract visual features F_V from video frame. We use BLIP-2 (Li et al. 2023) to generate caption prompts for the frame to extract semantic features F_S. And we use a pre-trained VAE to encode each frame of the video as the prior feature F_M for motion attention decoding.

Cross-Attention Integration In the stage of hierarchical integration of EEG features extracted from INR, we implement further integration of information from the bottom up. We apply a linear projection to transform EEG-INR features to generate the key, query, and value with the parameter matrices W^K, W^Q, W^V. We use the dot product between Q and K to calculate attention coefficients and compute the attention matrix. The output of multi-head self-attention is as follows:

$$Z_{eeg} = \mathrm{MHA}(K, Q, V) = (\|_{i=1}^{l} \mathrm{softmax}(\frac{Q^i {K^i}^{\mathrm{T}}}{\sqrt{d_j}}) V^i)\mathbf{W}^L + a, \quad (4)$$

where $d_j$is a hyperparameter, l is the number of attention heads, T^L is a linear transformation, a is bias term.

Motion-visual integration:

$$Z_V = \mathrm{LN}(I_V + \mathrm{Dropout}(\mathrm{MHA}(I_M,\ I_V, I_V))). \quad (5)$$

here, the Q and V are same.

Visual-semantic integration:

$$Z_S = \mathrm{LN}(I_S + \mathrm{Dropout}(\mathrm{MHA}(I_M,\ Z_V, I_S))). \quad (6)$$

And $Z_M = LN(I_M)$.

Motion Attention Flow The core of video reconstruction is the processing of contextual feature integration and dynamic changes. In the dynamic feature extraction part, we propose the *Motion Attention Flow (MAF)* to capture the motion information represented in the EEG. Specifically, we aim to guide EEG feature extraction by parsing the motion patterns of all patches within a video frame. Thanks to our previous EEG normalization to N dimensions, we use K to represent the key matrix of the index N_u and Q to represent the query matrix of the index N_v, so that we can calculate the *Motion Attention* $\mathbf{A}^*_{u,v}$. We first use argmax to process it, assigning each patch in N_u to the most attention-grabbing patch in N_v. We denote this result as $\bar{\mathbf{A}}^*_{u,v}$, where each entry $\bar{\mathbf{A}}^*_{u,v}(i,j)$ stores the assigned coordinate(i', j'). We utilize argmax to select a single index, which yields more reliable motion guidance. And we construct a patch displacement matrix $\mathbb{B}_{u,v}$ of size $H \times W \times 2$, where $\mathbb{B}_{u,v}(i,j) = (i'-i, j'-j)$. Finally, we aggregate the displacement matrices to construct *Motion Attention Flow*, serving as the motion signal:

$$\mathrm{MAF}(Z_M/F_M) = \{\mathbb{B}_{u,v} \mid u, v \in \mathbb{R}^N\}, \quad (7)$$

Cross-Modal Contrast Learning In the learning stage, we constructed a contrastive learning framework to align the EEG-INR decoupled features with the video embeddings. For contrast learning of visual features, we adopted the Bidirectional Dynamic Contrastive (BDC) (Lee et al. 2022) alignment mechanism as below:

$$\begin{aligned} \mathcal{L}_{BDC} = &- \sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij} \log \frac{\exp(\mathrm{sim}(Z_V^i, F_V^j)/\tau)}{\sum_k \exp(\mathrm{sim}(Z_V^i, F_V^k)/\tau)} \\ &+ \lambda \|\nabla_t Z_V - \nabla_t F_V\|_2^2, \end{aligned} \quad (8)$$

where w_{ij} represents learnable attention weights for feature importance, $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity, τ controls distribution sharpness, and ∇_t denotes temporal gradients.

Methods	Multi-class Classification				Binary Classification				
	40-class top-1	40-class top-5	9-class top-1	9-class top-3	Color	Fast/Slow	Numbers	Human Face	Human
▼ ***Methods using Pre-processed EEG Features*** *(PSD = Power Spectral Density Features, DE = Differential Entropy Features)*									
MLP (PSD)	5.85±2.95*	18.45±5.85*	21.15±2.95*	49.25±3.72*	21.65±3.22	54.65±1.18*	64.15±0.95	63.45±1.15	71.25±1.72
GLMNet (PSD) (2024b)	5.95±2.85*	18.55±5.55*	21.25±3.15*	49.55±4.05*	25.95±2.92*	54.95±1.28*	64.25±0.88	63.85±1.38	71.85±1.52
MLP (DE)	5.85±3.02*	18.65±5.65*	20.75±3.18*	48.95±4.88*	25.45±3.22*	54.25±1.22*	63.75±0.68	62.95±1.52	71.25±1.72
GLMNet (DE) (2024b)	5.85±3.12*	18.75±5.95*	20.95±3.28*	49.15±4.52	25.75±3.18*	54.65±1.18*	63.85±0.72	63.25±1.75	71.85±1.55
▼ ***Methods using Raw EEG***									
Conformer (2022)	4.65±1.52*	14.95±4.38*	20.45±1.02*	48.85±1.45*	26.95±1.42*	54.65±0.88*	65.25±0.32	64.35±1.18	72.65±0.82
BraVL (2023)	7.05±2.23*	19.05±3.75*	22.53±1.84*	51.07±2.41*	28.12±1.52*	57.88±1.64*	66.92±0.83	65.76±1.38	73.65±1.23
EEGPT (2024)	6.95±2.63*	18.85±4.11*	22.10±1.96*	50.55±2.33*	27.95±1.73*	57.10±1.85*	66.75±0.94	65.35±1.42	73.25±1.33
TSConv (2023)	4.58±1.05*	14.65±2.35*	19.65±1.05*	47.25±1.48*	26.35±1.78*	54.85±0.95*	64.95±0.38	63.85±1.42	72.25±0.72
EEG2Video (2024b)	5.85±2.95*	17.25±4.15*	21.45±1.82*	49.65±2.48*	26.85±1.42*	56.85±1.92*	65.85±0.88	64.65±1.42	72.95±1.28
EVOKE	**15.74**±2.35*	**26.33**±3.74*	**29.75**±2.12*	**59.18**±1.27*	**37.50**±1.75*	**68.12**±1.80*	**76.11**±0.94*	**73.06**±2.08*	**82.18**±1.15*
	+123.3%	**+38.2%**	**+32.0%**	**+15.9%**	**+33.4%**	**+17.7%**	**+13.7%**	**+11.1%**	**+11.6%**

Table 1: Average classification accuracy (%) and standard deviation across all subjects with different EEG classifiers. The star symbol (*) represents results above chance level with statistical significance (two-sample t-test: $p < 0.05$).

For contrast learning of semantic features, we use Hierarchical Semantic Contrastive (HSC) loss (Guo et al. 2022):

$$\mathcal{L}_{HSC} = \sum_{l=1}^{L} \alpha_l \cdot (-\log \frac{\exp(Z_S^l \cdot F_S^l/\tau)}{\sum_n \exp(Z_S^l \cdot F_{S_n}^l/\tau)}) + \beta \cdot \|Gram(Z_S) - Gram(F_S)\|_F, \quad (9)$$

where L represents the number of semantic levels (from concrete objects to abstract concepts), α_l are level-specific weights, and $Gram(\cdot)$ computes the Gram matrix capturing semantic structure, $\|\cdot\|_F$ denotes the Frobenius norm.

For contrastive learning of motion features, we minimize the element-wise Euclidean distance between the MAF displacement vectors of the frame and EEG:

$$\mathcal{L}_{MAF} = ||\text{MAF}(Z_M) - \text{MAF}(F_M)||_2^2. \quad (10)$$

3.4 Multi-modal Video Synthesis

As shown in Figure 2, during the inference stage, we use a carefully designed pre-trained T2V diffusion model for inference, guided by the visual, semantic descriptions, and dynamic video sequences decoded from EEG-INR features, to facilitate high-fidelity video generation.

Given the high cost of fully fine-tuning diffusion model, we choose to deploy our method within ControlNet (Zhang, Rao, and Agrawala 2023) while freezing the parameters of the Backbone model. We use an SDXL unCLIP (Scotti et al. 2024) decode Z_V to generate reconstructed key frames. Use the key frames' BLIP-2 captions and semantic descriptions generated by Z_S as text guidance, this dual-source caption fusion enhances semantic richness by integrating visual and neural information flows. And use the motion features Z_M to guide the generation of blurry and rough videos with high initial smoothness and good consistency. The integrated conditioning is computed as:

$$V_{re} = \text{ControlNet(Keyframe)} + \text{TextEmbed(Descrip)} + \text{ModtionVideo}. \quad (11)$$

And we implement an inflated 3D U-Net to maintain temporal consistency, combining 3D convolutions and self-attention mechanisms that operate across spatial and temporal dimensions, introducing a temporal consistency loss $\mathcal{L}_{\text{temp}}$ to maintain consistency between frames.

4 Experiments

4.1 Datasets & Evaluation

We conducted experiments on the **SEED-DV** dataset (Liu et al. 2024b), which contains EEG-Video data from 20 subjects. Each subject watched 7 blocks, totaling 1,400 video clips (40 visual categories). EEG signals were recorded using 64-channel electrodes at 1,000 Hz. The video clips were recorded at 4 frames per second with a resolution of 1920×1080. Preprocessing included bandpass filtering, downsampling, segmentation, and ICA artifact removal. The first 6 sessions for training, and the last session for testing.

We follow the **EEG-VP benchmark**. For semantic accuracy, we use a VideoMAE-based classifier trained on the Kinetics-400. For reconstruction quality, we use perception-based metrics (SSIM, PSNR) at the frame level. At the video level, we calculate the similarity and temporal consistency metrics (CLIP-pcc, FVD). Additionally, we use a classifier trained on ImageNet to compute N-way top-K accuracy.

4.2 Architecture & Implementation

For video feature extraction, we use the CLIP-ViT-G/14 module (Radford et al. 2021). HAM employs a multi-head attention mechanism (3 heads, dimension 256), combined with subject embedding (dimension 64), average pooling kernels 1×51 and 1×5, a dropout rate of 0.5, attention layer depth is 1. CLIP embeddings are normalized within the range [-1.5, 1.5]. The width of $\mathcal{M}_\theta$ is 400, depth is 4. The width of $\mathcal{G}_\theta$ is 200, depth is 6.

We implement EVOKE using PyTorch and conduct experiments on NVIDIA A800 GPUs. Training uses the AdamW (Loshchilov and Hutter 2017) optimizer (learning rate 1e-4 to 2e-5), with weight decay (0.01) and cosine scheduling. We train all tasks for 200 epochs with 128 batch size.

5 Results and Discussion

5.1 Visual Perception Performance

Table 1 presents the classification accuracy across multiple tasks. EVOKE significantly surpasses both raw and pre-processed EEG-based methods. Notably, EVOKE attains 15.74% top-1 accuracy in challenging 40-way classification

Dataset	Method	Video-based				Frame-based			
		2-way↑	40-way↑	CLIP-pcc↑	FVD↓	2-way↑	40-way↑	SSIM↑	PSNR↑
fMRI	MinD-Video (2023b)	0.839±0.03	0.197±0.02 (50-way)	0.408±0.46	452.1±8.7	0.796±0.03	0.174±0.03 (50-way)	0.171±0.08	8.662±1.52
	NeuroClips (2024)	0.834±0.03	0.220±0.01 (50-way)	0.738±0.17	213.1±5.1	0.806±0.03	0.203±0.01 (50-way)	0.390±0.08	9.211±1.46
	Mind-Animate (2024)	0.830±0.03	-	0.425±0.17	307.1±7.6	0.805±0.03	-	0.321±0.08	9.220±1.46
EEG	EEG2Video(2024b)	0.852±0.02	0.340±0.01	0.378±0.46	384.7±7.3	0.798±0.03	0.232±0.02	0.300±0.03	8.362±1.52
	EVOKE (w/o Motion)	0.836±0.02	0.320±0.02	0.558±0.28	231.8±8.7	0.719±0.02	0. 258±0.02	0.237±0.02	7.007±1.61
	EVOKE (w/o Semantic)	0.795±0.03	0.204±0.02	0.535±0.20	196.1±7.9	0.747±0.03	0.128±0.02	0.275±0.03	7.984±1.47
	EVOKE (w/o Visual)	0.738±0.02	0.279 ±0.03	0.637±0.24	285.6 ±9.5	0.670±0.02	0.223±0.02	0.212 ±0.02	5.559±1.13
	EVOKE (All)	**0.901**±0.02	**0.407**±0.01	**0.715** ±0.19	**198.1**±6.5	**0.841**±0.02	**0.294**±0.02	**0.353** ±0.02	**9.148**±1.75

Table 2: Comparison of video reconstruction performance across EEG, fMRI. We report 8 different metrics on Video & Frame to quantify the model's performance. We also showed the impact of different decoupling features.

Figure 3: Video reconstruction results (Subject-1). EVOKE decodes the most accurate video with semantic details.

tasks, outperforming previous approaches such as BraVL (improve 123.3%) and EEG2Video. For specialized recognition tasks, EVOKE demonstrates robust accuracy, achieving 76.11% for number recognition and 73.06% for human face detection, with particularly strong performance in binary classification (68.12% accuracy in fast/slow motion discrimination), highlighting its generalization across diverse categories. The superior performance of EVOKE can be attributed to EEG-INR perceptual space captures the inherent temporal dynamics and spatial relationships in raw EEG data, which are often lost during other decoding methods.

5.2 Video Reconstruction

We comprehensively evaluate the video reconstruction performance of EVOKE, comparing it to both EEG-based and fMRI-based baselines. As shown in Table 2, based on EEG signals, EVOKE achieves SOTA performance on all metrics and also surpasses advanced works based on fMRI on some metrics. In terms of perceptual quality, EVOKE achieves a 0.715 CLIP-pcc and 198.1 FVD, indicating superior semantic similarity and temporal consistency. The 0.353 SSIM and 9.148 PSNR demonstrate its ability to reconstruct visually coherent and high-fidelity frames. EVOKE's ability to accurately capture multi-dimensional features also leads to consistent improvements in N-way prediction accuracy.

The visualization of the video reconstruction results is shown in Figure 3. EVOKE's performance is impressive, achieving high consistency, whether in capturing detailed features, semantic features, or dynamic changes.

5.3 Discussion

Decoupling Feature Contribution Analysis In EVOKE, HAM treats EEG decoding as a process of learning features from decoupled modalities. We present the video reconstruction performance after different modalities ablation in Table 2. Removing motion features significantly reduces the FVD and PSNR metrics, as the spatial and motion consistency information makes a core contribution to reconstruction. Removing visual features leads to a significant decline in frame quality, while removing semantic features results in a decline in CLIP-pcc, confirming the indispensability of visual and semantic information.

Brain Contribution Analysis Figure 4 **[A]** shows the contribution of different brain regions. It can be seen, occipital lobe (primary visual cortex) alone can capture basic visual features (color, texture), but loses semantic features (the object 'car'). Adding temporal lobe regions improves the representation of semantic content, adding parietal lobe enhances the spatial distribution of objects, consistent with the two-stream hypothesis of vision. The frontal lobe is involved in human motor, and its addition further improves the coherence of movement. EVOKE effectively utilizes the distributed visual processing of brain, with different functional brain regions corresponding to consistent decoding features.

5.4 Ablation

Ablation of Different Modules In Table 3, we present ablation results for different modules in EVOKE. Removing

Method	Trainable Parameters	Semantic (Video)		ST-level		Semantic (Frame)		Pixel-level		Classification	
		2-way↑	40-way↑	CLIP-pcc↑	FVD↓	2-way↑	40-way↑	SSIM↑	PSNR↑	40-c top-1↑	9-c top-1↑
w/o HAM	55.1 M	0.748	0.3095	0.439	296.7	0.692	0.243	0.259	6.580	8.63	17.90
w/o MAF	72.6 M	0.852	0.344	0.566	221.5	0.731	0.263	0.268	7.119	10.28	20.95
w/o Condition Integrate	77.5 M	0.835	0.337	0.598	301.8	0.750	0.256	0.220	6.283	13.78	26.85
EVOKE (CLIP Space)	112.4 M	0.871	0.353	0.597	247.8	0.825	0.270	0.332	8.605	12.87	25.15
EVOKE	84.9 M	**0.901**	**0.407**	**0.715**	**198.1**	**0.841**	**0.294**	**0.353**	**9.148**	**15.74**	**29.75**

Table 3: Ablations on the modules of EVOKE. This analysis examines the effect of INR, HAM, and MAF on brain decoding.

$\mathcal{M}_\theta$ width	$\mathcal{M}_\theta$ depth	$\mathcal{G}_\theta$ width	$\mathcal{G}_\theta$ depth	Semantic (Video)		ST-level		Semantic (Frame)		Pixel-level		Classification	
				2-way↑	40-way↑	CLIP-pcc↑	FVD↓	2-way↑	40-way↑	SSIM↑	PSNR↑	40-c top-1↑	9-c top-1↑
100	3	200	4	0.883	0.372	0.672	217.2	0.828	0.276	0.336	8.757	14.92	28.15
200	4	400	6	0.891	0.398	0.703	202.9	0.836	0.287	0.347	8.980	15.40	29.10
400	6	200	4	0.877	0.389	0.690	207.5	0.830	0.288	0.341	8.823	15.05	28.45
400	4	200	6	**0.901**	**0.407**	**0.715**	**198.1**	**0.841**	**0.294**	**0.353**	**9.148**	**15.74**	**29.75**

Table 4: Ablation study on the depth and width of the INR network.

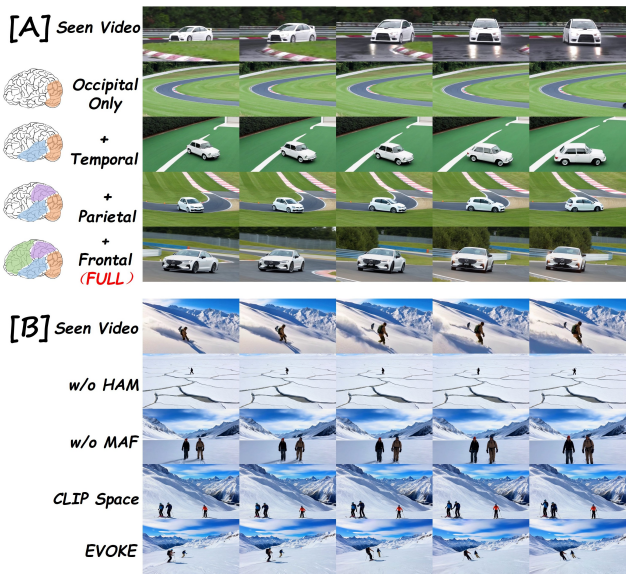

Figure 4: **[A]:** Reconstruction using different combination of brain regions. **[B]:** Results of different module ablation.

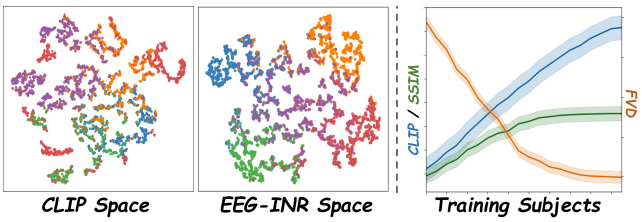

Figure 5: **Left:** Different encoding spaces t-SNE plots. **Right:** Impact of training data on model performance.

the HAM results in inadequate fusion of decoupling features, resulting in all metrics significant decline. Removing the MAF results in a loss of motion feature capture, preventing consistent video reconstruction. Removing the Condition Integrate during video synthesis phase leads to divergence in video reconstruction metrics, failing to achieve best performance. Replacing the INR with the CLIP, while also achieving comparable results, but cannot reach the performance of EVOKE's EEG-INR perceptual space and significantly increases trained parameters (+32.4%). We visualize the results of module ablation in Figure 4 **[B]**, which are consistent with the changes in quantitative metrics.

Ablation of INR setting In Figure 5 Left, we visualize the t-SNE of five video classes embedding distributions in CLIP and EEG-INR Space. The latent feature separation in CLIP Space is low, preventing accurate classification, whereas the INR-based Space exhibits significant feature separation. This demonstrates the importance of INR's smoothness representation in processing EEG spatio-temporal features. Table 4 also analyzes the width and depth parameter settings for $\mathcal{M}_\theta$ and $\mathcal{G}_\theta$ in INR. We see that is relatively robust to depth and width, achieving good results in different setting, demonstrating its superior design and task suitability.

Ablation of Training Data The Figure 5 Right shows the effect of using different amounts of training data. Increasing the amount of training data consistently improves the SSIM and CLIP metrics while reducing the FVD. Notably, EVOKE achieves comparable performance using only a small amount of data, and the rate of performance improvement slows as data increases, suggesting a threshold effect in capturing neural representations.

6 Conclusion

We propose a novel dynamic visual decoding method **EVOKE**, which solves the challenges of insufficient spatial modeling and inadequate neural-visual alignment in EEG-based visual reconstruction. Our approach maintains excellent performance across multiple tasks under zero-shot setting and has strong generalizability and interpretability. The deployment of EVOKE has achieved a consistent and unbiased dynamic visual experience, representing a significant advancement in neural decoding and opening up new possibilities for highly accurate and scalable BCI.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NO. 62088102, No. 62436005), Brain Networks and Brain-Inspired Intelligence Science Breakthrough Pilot Project, and STI2030-Major Projects (NO. 2022ZD0208801).

References

Altaheri, H.; Muhammad, G.; Alsulaiman, M.; Amin, S. U.; Altuwaijri, G. A.; Abdul, W.; Bencherif, M. A.; and Faisal, M. 2023. Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review. *Neural Computing and Applications*, 35(20): 14681–14722.

Bartels, A. 2005. Brain dynamics during natural viewing conditions—a new guide for mapping connectivity in vivo. *Neuroimage*, 24(2): 339–349.

Buckner, R. L. 1998. Event-related fMRI and the hemodynamic response. *Human brain mapping*, 6(5-6): 373–377.

Chen, Z.; Qing, J.; Xiang, T.; Yue, W. L.; and Zhou, J. H. 2023a. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22710–22720.

Chen, Z.; Qing, J.; Zhou, J. H.; and Wang, Y. 2023b. Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*, 36: 24841–24858.

Demir, A.; Koike-Akino, T.; Wang, Y.; Haruna, M.; and Erdogmus, D. 2021. EEG-GNN: Graph neural networks for classification of electroencephalogram (EEG) signals. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1061–1067. IEEE.

Du, C.; Fu, K.; Li, J.; and He, H. 2023. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10760–10777.

Duan, Y.; Zhou, J.; Wang, Z.; Wang, Y.-K.; and Lin, C.-t. 2023. Dewave: Discrete encoding of eeg waves for eeg to text translation. *Advances in Neural Information Processing Systems*, 36: 9907–9918.

Dubois, J.; and Adolphs, R. 2016. Building a science of individual differences from fMRI. *Trends in cognitive sciences*, 20(6): 425–443.

Gao, J.; Fu, Y.; Wang, Y.; Qian, X.; Feng, J.; and Fu, Y. 2024. Mind-3d: Reconstruct high-quality 3d objects in human brain. In *European Conference on Computer Vision*, 312–329. Springer.

Garg, V.; Jegelka, S.; Jaakkola, T.; and Patgiri, R. 2020. Generalization and representational limits of graph neural networks. In *International conference on machine learning*, 3419–3430. PMLR.

Gong, Z.; Bao, G.; Zhang, Q.; Wan, Z.; Miao, D.; Wang, S.; Zhu, L.; Wang, C.; Xu, R.; Hu, L.; et al. 2024. NeuroClips: Towards High-fidelity and Smooth fMRI-to-Video Reconstruction. *Advances in Neural Information Processing Systems*, 37: 51655–51683.

Graña, M.; Morais-Quilez, I.; Ahissar, M.; and Haruna, M. 2023. A review of Graph Neural Networks for Electroencephalography data analysis. *Neurocomputing*, 562: 126901.

Grattarola, D.; and Vandergheynst, P. 2022. Generalised implicit neural representations. *Advances in Neural Information Processing Systems*, 35: 30446–30458.

Guo, Y.; Xu, M.; Li, J.; Ni, B.; Zhu, X.; Sun, Z.; and Xu, Y. 2022. Hcsc: Hierarchical contrastive selective coding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9706–9715.

Hafri, A.; Trueswell, J. C.; Epstein, R. A.; and Zeki, S. 2017. Neural representations of observed actions generalize across static and dynamic visual input. *Journal of Neuroscience*, 37(11): 3056–3071.

Horikawa, T.; and Kamitani, Y. 2017. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1): 15037.

Jing, H.; Jiang, D.; Ma, Y.; Hua, H.; Huang, B.; and Zheng, N. 2025a. Beyond Brain Decoding: Visual-Semantic Reconstructions to Mental Creation Extension Based on fMRI. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19258–19268.

Jing, H.; Ma, Y.; Yang, P.; Hua, H.; and Zheng, N. 2025b. Pinpointing Visual Content: Disentangled Features in Multimodal Model for EEG Representation Learning and Decoding. *Knowledge-Based Systems*, 114212.

Klepl, D.; Wu, M.; He, F.; and Wetzstein, G. 2024. Graph neural network-based eeg classification: A survey. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32: 493–503.

Kriegeskorte, N. 2018. Cognitive computational neuroscience. *Nature neuroscience*, 21(9): 1148–1160.

Le, L.; Ambrogioni, L.; Seeliger, K.; Güçlütürk, Y.; van Gerven, M.; and Güçlü, U. 2022. Brain2pix: Fully convolutional naturalistic video frame reconstruction from brain activity. *Frontiers in Neuroscience*, 16: 940972.

Lee, G.; Eom, C.; Lee, W.; Park, H.; and Ham, B. 2022. Bi-directional contrastive learning for domain adaptive semantic segmentation. In *European Conference on Computer Vision*, 38–55. Springer.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, S.; Pan, J.; and Wang, F. 2021. Cross-subject EEG emotion recognition with self-organized graph neural network. *Frontiers in Neuroscience*, 15: 611653.

Li, X.; Zhang, Y.; Tiwari, P.; Song, D.; Hu, B.; Yang, M.; Zhao, Z.; Kumar, N.; and Marttinen, P. 2022. EEG based emotion recognition: A tutorial and review. *ACM Computing Surveys*, 55(4): 1–57.

Liu, A.; Jing, H.; Liu, Y.; Ma, Y.; and Zheng, N. 2024a. Hidden States in LLMs Improve EEG Representation Learning and Visual Decoding. In *ECAI*, volume 392, 2130–2137.

Liu, P.; Dong, G.; Guo, D.; Li, K.; Li, F.; Yang, X.; Wang, M.; and Ying, X. 2025a. A Survey on fMRI-based Brain Decoding for Reconstructing Multimodal Stimuli. *arXiv preprint arXiv:2503.15978*.

Liu, X.-H.; Liu, Y.-K.; Wang, Y.; Ren, K.; Shi, H.; Wang, Z.; Li, D.; Lu, B.-L.; and Zheng, W.-L. 2024b. EEG2video: Towards decoding dynamic visual perception from EEG signals. *Advances in Neural Information Processing Systems*, 37: 72245–72273.

Liu, Y.; Ma, Y.; Zhu, G.; Jing, H.; and Zheng, N. 2025b. See Through Their Minds: Learning Transferable Brain Decoding Models from Cross-Subject fMRI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5730–5738.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, Y.; Du, C.; Wang, C.; Zhu, X.; Jiang, L.; and He, H. 2024. Animate Your Thoughts: Decoupled Reconstruction of Dynamic Natural Vision from Slow Brain Activity. *arXiv preprint arXiv:2405.03280*.

Ma, Y.; Zhang, W.; Du, M.; Jing, H.; and Zheng, N. 2024. Hierarchical bayesian causality network to extract high-level semantic information in visual cortex. *International Journal of Neural Systems*, 34(01): 2450002.

Makeig, S.; Westerfield, M.; Jung, T.-P.; Enghoff, S.; Townsend, J.; Courchesne, E.; and Sejnowski, T. J. 2002. Dynamic brain sources of visual evoked responses. *Science*, 295(5555): 690–694.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Sabharwal, Y.; and Rama, B. 2024. Comprehensive Review of EEG-to-Output Research: Decoding Neural Signals into Images, Videos, and Audio. *arXiv preprint arXiv:2412.19999*.

Scotti, P.; Banerjee, A.; Goode, J.; Shabalin, S.; Nguyen, A.; Dempster, A.; Verlinde, N.; Yundler, E.; Weisberg, D.; Norman, K.; et al. 2023. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36: 24705–24728.

Scotti, P. S.; Tripathy, M.; Villanueva, C. K. T.; Kneeland, R.; Chen, T.; Narang, A.; Santhirasegaran, C.; Xu, J.; Naselaris, T.; Norman, K. A.; et al. 2024. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*.

Sitzmann, V.; Martel, J.; Bergman, A.; Lindell, D.; and Wetzstein, G. 2020. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33: 7462–7473.

Song, T.; Zheng, W.; Song, P.; and Cui, Z. 2018. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3): 532–541.

Song, Y.; Liu, B.; Li, X.; Shi, N.; Wang, Y.; and Gao, X. 2023. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*.

Song, Y.; Zheng, Q.; Liu, B.; and Gao, X. 2022. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31: 710–719.

Sun, J.; Li, M.; and Moens, M.-F. 2025. Neuralflix: A simple while effective framework for semantic decoding of videos from non-invasive brain recordings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7096–7104.

Takagi, Y.; Nishimoto, S.; Amin, S. U.; and Altuwaijri, G. A. 2023. 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*, 14453–14463.

Waikhom, L.; Patgiri, R.; Barron, J.; and Ng, R. 2023. A survey of graph neural networks in various learning paradigms: methods, applications, and challenges. *Artificial Intelligence Review*, 56(7): 6295–6364.

Wang, G.; Liu, W.; He, Y.; Xu, C.; Ma, L.; and Li, H. 2024. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37: 39249–39280.

Watson, J. D. 2000. The human visual system. In *Brain mapping: The systems*, 263–289. Elsevier.

Xu, L.; Xu, M.; Jung, T.-P.; and Ming, D. 2021. Review of brain encoding and decoding mechanisms for EEG-based brain–computer interface. *Cognitive neurodynamics*, 15(4): 569–584.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.