Multi-Granularity EEG Decoding: Bridging Object-Background Semantics and Temporal Motion for Video Reconstruction

Tian-Yi Zhou[1,*], **Zhenghao Xiao**[2,*], **Xuan-Hao Liu**[1], **Bao-Liang Lu**[1] & **Wei-Long Zheng**[1,†]
[1]School of Computer Science, Shanghai Jiao Tong University
[2]Carnegie Mellon University
[*]Equal Contribution, [†] Corresponding Author

Abstract

Reconstructing video from brain signals is a pivotal task in neural decoding. However, current frameworks struggle by directly aligning noisy EEG with coarse global captions and VAE latents, inevitably leading to semantic misalignment and temporal flickering. To bridge this gap, we propose M-STAR, a bio-inspired framework that mimics human cognitive decomposition. To our knowledge, M-STAR is the first framework to decompose dense captions into object- and environment-level semantics for robust alignment and employ an "Anchor-plus-Residual" strategy that anchors generation to a stable visual anchor and applies residual dynamics to maintain stability over time. Specifically, this is achieved via the Factorized Semantic Module (FSM) and the Anchor-Guided Dynamics Module (ADM), respectively. Extensive evaluations on the SEED-DV dataset demonstrate that M-STAR significantly outperforms state-of-the-art methods, generating high-fidelity videos with precise semantics and coherent motion. M-STAR establishes a new baseline for fine-grained visual decoding, marking a significant step towards high-fidelity visual brain-computer interfaces.

Keywords: EEG; Brain Visual Decoding; Video Generation;

Introduction

Decoding visual experiences from human brain activity stands at the forefront of cognitive neuroscience and computer vision (Guenther et al., 2024; Nishimoto et al., 2011). While early research was confined to decoding simple geometric shapes or static categories (Fang et al., 2020; Naselaris et al., 2009), the advent of generative models (e.g., Stable Diffusion (Rombach et al., 2022)) has propelled the field toward the complex domain of dynamic video reconstruction (J. Liu et al., 2025; Lu et al., 2025). Although fMRI has historically dominated this landscape due to its spatial fidelity (Chen et al., 2023; Gong et al., 2024), its low temporal resolution and reliance on cumbersome equipment limit practical application. In contrast, Electroencephalography (EEG) offers millisecond-level temporal precision and high portability, making it uniquely suited for capturing continuous, dynamic visual streams in real-world scenarios.

Despite recent advances in reconstructing visual content from EEG signals (Huang, Luo, et al., 2025; J. Liu et al., 2025; X.-H. Liu et al., 2024), significant limitations persist that impede the decoding of high-fidelity videos, as illustrated in Fig. 1. **(1) EEG-Text Modality Gap.** Natural videos inherently contain complex semantic hierarchies, comprising foreground objects, background environments, and temporal

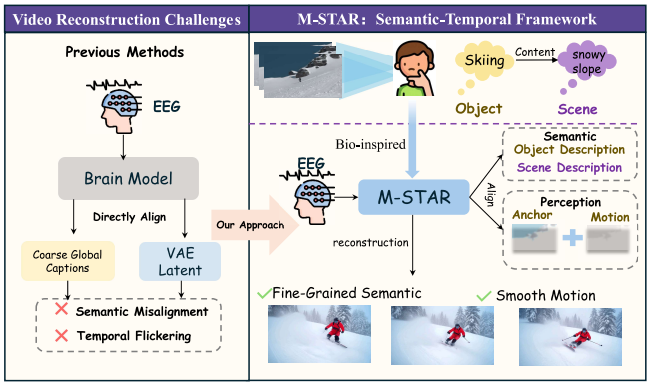

Figure 1: Motivation and Framework Overview. Unlike prior methods limited by semantic misalignment and temporal flicking, M-STAR (Right) leverages a bio-inspired multi-granularity strategy (Object/Scene) and Anchor-Guided dynamics. This design effectively ensures fine-grained semantic fidelity and smooth motion.

motion. Consequently, captions derived from BLIP-2 (Li et al., 2023) tend to be verbose and semantically saturated, creating a dense information space. In contrast, EEG signals reflect perceptual and cognitive processes rooted in neural dynamics, which are fundamentally disparate from the structured syntax of natural language. Compounded by the inherent low signal-to-noise ratio (SNR) of non-invasive recordings, EEG signals lack the granularity to fully address such dense textual semantics. Therefore, directly learning a mapping from low SNR EEG signals to semantic-dense texts is an ill-posed problem, making the model prone to severe overfitting and semantic misalignment. **(2) Spatiotemporal Inconsistency.** Existing approaches typically align EEG features directly with video VAE latents via MSE objectives (X.-H. Liu et al., 2024). However, such rigid point-wise constraints fail to explicitly model the underlying spatiotemporal structure of the visual content. By neglecting the inherent temporal continuity and treating feature alignment as a static matching problem, these methods struggle to capture valid motion priors. Consequently, the decoded videos frequently suffer from severe temporal flickering and disjointed motion trajectories, where the visual identity of objects drifts or disintegrates across frames.

To address these challenges, we introduce **M-STAR** (**M**ulti-granularity **S**emantic-**T**emporal **A**rchitecture for **R**econstruction). M-STAR employs two specialized mod-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

ules to bridge the modality gap and ensure consistency. First, inspired by the human visual attention mechanism, which prioritizes foreground subjects over backgrounds (Brandman & Peelen, 2017; Wischnewski & Peelen, 2021), the Factorized Semantic Module (FSM) explicitly decomposes dense captions into distinct object and scene descriptions. By aligning noisy EEG with these fine-grained semantic components instead of holistic sentences, FSM bridges the modality gap and guarantees robust text conditioning. Second, to resolve spatiotemporal inconsistency, the Anchor-Guided Dynamics Module (ADM) introduces a conditional "Anchor-plus-Residual" strategy. By grounding generation on a deterministic visual anchor and modeling temporal evolution as continuous residual deviations, ADM enforces coherent visual identity and effectively eliminates the flickering artifacts prevalent in prior methods.

In conclusion, our contributions can be summarized as follows:

1. We propose M-STAR, a bio-inspired framework designed to bridge the EEG-Text modality gap and resolve spatiotemporal inconsistencies, effectively addressing critical bottlenecks in video reconstruction.
2. To our knowledge, we are the first to introduce the Factorized Semantic Module (FSM) for decomposing dense captions and the Anchor-Guided Dynamics Module (ADM) with an "Anchor-plus-Residual" strategy to explicitly enforce motion consistency.
3. Experiments on the SEED-DV dataset demonstrate that M-STAR significantly outperforms state-of-the-art methods, achieving superior semantic fidelity and temporal smoothness.

Related Work

fMRI and MEG Visual Decoding

Early work mainly used fMRI for its high spatial resolution, evolving from using Bayesian decoding to deep generative models (deconvolutional neural networks (DeCNNs) (Wen et al., 2018), variational autoencoders (VAEs) (Kavasidis et al., 2017), and generative adversarial networks (GANs) (VanRullen & Reddy, 2019)), and more recently to latent diffusion models (LDMs) (Rombach et al., 2022), such as Stable Diffusion (Takagi & Nishimoto, 2023), for high-fidelity and semantically consistent reconstructions by aligning brain signals with CLIP-style multimodal embeddings(Miyawaki et al., 2008; Naselaris et al., 2009; Radford et al., 2021).

fMRI video reconstruction is hindered by low sampling rates (TR $\approx$ 2 s) and slow BOLD dynamics, so recent diffusion-based methods (e.g., MinD-Video, NeuroClips, Mind-Animator, and NeuralFlix) still cannot achieve truly frame-aligned decoding and usually map each fMRI window to numerous video frames (Chen et al., 2023; Gong et al., 2024; Lu et al., 2025; Sun et al., 2025). Meanwhile, MEG cannot match fMRI for detailed image reconstruction and has very limited deployability, which encourages a switch to EEG for dynamic reconstruction (Benchetrit et al., 2024; Gross, 2019; Kim et al., 1997; X.-H. Liu et al., 2024)

EEG Visual Decoding

EEG is more scalable than fMRI and MEG due to its low cost and portability (Gross, 2019; Nicolas-Alonso & Gomez-Gil, 2012), and its high temporal resolution makes it promising for reconstructing dynamic visual perception (J. Liu et al., 2025; X.-H. Liu et al., 2024).

Following the EEG2Video benchmark (X.-H. Liu et al., 2024), subsequent studies have significantly improved semantic fidelity and cross-subject generalization. For instance, DynaMind (J. Liu et al., 2025) and MindCross (X.-H. Liu et al., 2025) introduced explicit semantic-dynamic modeling and subject disentanglement, respectively, while MindCine (Zhou et al., 2026) leveraged the beyond-text-modality to reconstruct videos. Despite these advancements, a critical limitation persists: current diffusion-driven methods—including the LLM-enhanced MINDEV (Huang, Wang, et al., 2025)—predominantly rely on coarse global captions and implicit temporal structures. This dependency inherently restricts the preservation of fine-grained visual identity and motion smoothness. To bridge this gap, we propose M-STAR, which leverages dense-caption supervision and an Anchor-Guided Dynamics Module (ADM) to ensure both semantic fidelity and temporal coherence.

Method

In this section, we present M-STAR, a unified framework for reconstructing high-fidelity videos from non-invasive EEG signals. As illustrated in Fig. 2, M-STAR comprises two core components: Factorized Semantic Module (FSM) and Anchor-Guided Dynamics Module(ADM) .

Factorized Semantic Module

Since video captions are structurally complex and information-dense, directly learning a mapping between low signal-to-noise ratio EEG signals and semantic-dense texts is prone to overfitting. To mitigate this, the key insight is to decompose the complex texts into two simpler components: object descriptions and environment descriptions. By aligning EEG signals with these factorized prompts independently, we reduce the mapping ambiguity and facilitate robust feature learning. The FSM processes the Object and Environment streams in parallel. As shown in Fig. 2, each stream follows a two-stage components: a Semantic Encoder for feature alignment and a Condition Mapper for conditioning generative models.

Semantic Encoder The primary objective of the Semantic Encoder E_{sem} is to extract robust, discriminative features from noisy neural signals while preventing overfitting to specific caption patterns. To achieve this, we adopt SoftCLIP loss (Gao et al., 2024) to align the EEG embeddings $\mathbf{e}$ with their corresponding factorized text embeddings $\mathbf{t}$ from the

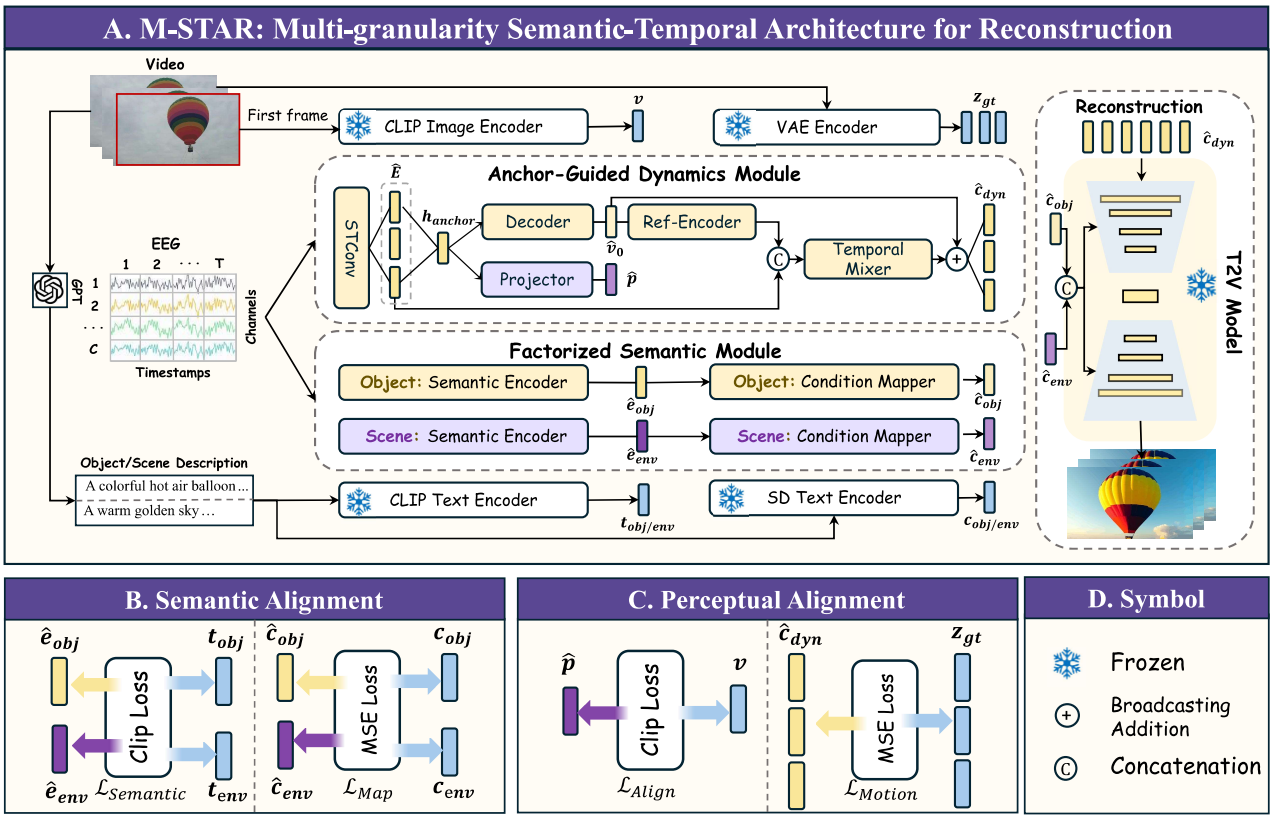

Figure 2: Overview of the M-STAR framework. The framework consists of the Factorized Semantic Module for fine-grained semantic grounding and the Anchor-Guided Dynamics Module for consistent motion reconstruction.

pre-trained CLIP.

$$\mathcal{L}_{\text{SoftCLIP}}(\mathbf{t}, \hat{\mathbf{e}}) = -\sum_{k=1}^{B}\sum_{l=1}^{B}\left[\frac{\exp(\mathbf{t}_k \cdot \mathbf{t}_l/\tau)}{\sum_{m=1}^{B}\exp(\mathbf{t}_k \cdot \mathbf{t}_m/\tau)} \cdot \log\left(\frac{\exp(\hat{\mathbf{e}}_k \cdot \mathbf{t}_l/\tau)}{\sum_{m=1}^{B}\exp(\hat{\mathbf{e}}_k \cdot \mathbf{t}_m/\tau)}\right)\right], \tag{1}$$

where $\mathbf{t}$ and $\hat{\mathbf{e}}$ are the ground truth text embedding and the EEG embedding in a batch of size B. τ is a learned temperature parameter. To guarantee comprehensive semantic understanding, we jointly optimize the object and environment streams. The total semantic alignment loss is formulated as the summation of the SoftCLIP objectives from both factorized subspaces:

$$\mathcal{L}_{\text{Semantic}} = \mathcal{L}_{\text{SoftCLIP}}(\mathbf{t}_{obj}, \hat{\mathbf{e}}_{obj}) + \mathcal{L}_{\text{SoftCLIP}}(\mathbf{t}_{env}, \hat{\mathbf{e}}_{env}), \tag{2}$$

where $\mathbf{t}_k$ and $\hat{\mathbf{e}}_k$ (for $k \in \{obj, env\}$) denote the ground-truth CLIP text embeddings and the predicted EEG embeddings for the respective semantic branches. Specifically, $\hat{\mathbf{e}}_k = E_k(X)$ represents the normalized semantic feature extracted from the raw EEG signal $\mathbf{X}$ by the branch-specific encoder E_k.

Condition Mapper To enable EEG-guided video generation, the aligned features must be adapted to match the conditioning space of the generative models. Therefore, lightweight Condition Mappers, M_{obj} and M_{env}, are employed to project the EEG embeddings into the conditioning space of the Latent Diffusion Model:

$$\hat{\mathbf{c}}_k = M_k(\hat{\mathbf{e}}_k), \quad k \in \{obj, env\}. \tag{3}$$

To ensure precise control, the text condition extracted by the pre-trained SD text encoder are utilized as ground-truth targets. A mapping constraint is enforced via the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\text{Map}} = \|\hat{\mathbf{c}}_{obj} - \mathbf{c}_{obj}\|_2^2 + \|\hat{\mathbf{c}}_{env} - \mathbf{c}_{env}\|_2^2, \tag{4}$$

where $\mathbf{c}_k$ denotes the ground-truth conditioning embedding generated by the frozen SD text encoder given the text prompt $\mathcal{T}_k$. Rather than merging these features into a unified vector—which risks collapsing distinct semantic attributes—a Late Fusion strategy is adopted via sequence concatenation to maintain semantic disentanglement. Let $[\cdot;\cdot]$ denote the concatenation operation along the sequence length dimension. The final condition $\mathbf{c}_{final}$, injected into the U-Net, is formulated as:

$$\hat{\mathbf{c}}_{final} = [\hat{\mathbf{c}}_{obj}\,;\,\hat{\mathbf{c}}_{env}]. \tag{5}$$

The overall objective for the FSM is the weighted sum:

$$\mathcal{L}_{\text{FSM}} = \mathcal{L}_{\text{Semantic}} + \lambda_{map}\mathcal{L}_{\text{Map}}. \tag{6}$$

Anchor-Guided Dynamics Module

To address spatiotemporal inconsistency, we reformulate video reconstruction not as independent frame generation, but as a factorized process comprising a static visual foundation (the initial state) and its subsequent temporal evolution (the motion). Based on this, the Anchor-Guided Dynamics Module is designed to generate a robust dynamic representation by explicitly modeling the visual anchor and motion dynamics sequentially. The process begins with temporal segmentation: the raw EEG signals $\mathbf{X} \in \mathbb{R}^{C\times T}$ are sliced into F consecutive segments $\mathbf{E} = [\mathbf{x}_1, \ldots, \mathbf{x}_F]$. Here, C and T denote the channel count and total time points, respectively, while F denotes the number of target video frames.

Visual Anchor Reconstruction The primary objective of this stage is to establish a deterministic Visual Anchor—a static representation corresponding to the initial frame of the video—which serves as the spatial foundation for subsequent motion generation.

Given the segmented EEG sequence $\mathbf{E}$, a lightweight Spatiotemporal (ST) Encoder Φ_{ST} is first employed to extract frame-level embeddings $\hat{\mathbf{E}} = \Phi_{ST}(\mathbf{E})$. Subsequently, these embeddings are temporally aggregated (e.g., via mean pooling) to yield a global visual representation $\mathbf{h}_{anchor}$:

$$\mathbf{h}_{anchor} = \frac{1}{F}\sum_{t=1}^{F}\hat{\mathbf{E}}^{(t)}, \quad \hat{\mathbf{v}}_0 = \mathcal{D}_{rec}(\mathbf{h}_{anchor}), \tag{7}$$

where F denotes the number of segments, $\mathcal{D}_{rec}$ is the reconstruction decoder, and $\hat{\mathbf{v}}_0$ represents the predicted latent of the first video frame. To enforce semantic consistency, we project the global anchor feature $\mathbf{h}_{anchor}$ into the CLIP latent space:

$$\hat{\mathbf{p}} = \mathcal{P}_{vis}(\mathbf{h}_{anchor}), \tag{8}$$

where $\mathcal{P}_{vis}$ is a learnable projector. We then align $\hat{\mathbf{p}}$ with the ground-truth CLIP image embedding $\mathbf{v}$ via a symmetric InfoNCE loss:

$$\begin{aligned} \mathcal{L}_{\text{Align}} &= \frac{1}{2}\left(\mathcal{L}_{\hat{\mathbf{p}}\to\mathbf{v}} + \mathcal{L}_{\mathbf{v}\to\hat{\mathbf{p}}}\right), \\ \text{where } \mathcal{L}_{x\to y} &= -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(x_i\cdot y_i/\tau)}{\sum_{j=1}^{B}\exp(x_i\cdot y_j/\tau)}. \end{aligned} \tag{9}$$

Here, τ denotes the temperature parameter, B is the batchsize.

Residual Motion Reconstruction This stage focuses on decoding the temporal dynamics. Specifically, the predicted anchor $\hat{\mathbf{v}}_0$ is re-encoded by a reference encoder Φ_{ref} to extract spatial guidance features. These features are concatenated with the original EEG sequence embeddings $\hat{\mathbf{E}}$ (extracted by the ST Encoder) to form a dual-conditioned representation. A temporal mixer $\mathcal{M}_{temp}$ then processes this sequence to predict motion residuals $\Delta\mathbf{V}$. The final dynamic input $\hat{\mathbf{c}}_{dyn}$ is synthesized by broadcasting the static anchor $\hat{\mathbf{v}}_0$ across the temporal dimension and superimposing the learned residuals:

$$\hat{\mathbf{c}}_{dyn}^{(t)} = \hat{\mathbf{v}}_0 + \Delta\mathbf{v}^{(t)}, \Delta\mathbf{V} = \mathcal{M}_{temp}(\hat{\mathbf{E}} \oplus \Phi_{ref}(\hat{\mathbf{v}}_0)). \tag{10}$$

Here, $\hat{\mathbf{c}}_{dyn}^{(t)}$ denotes the perceptual feature for the t-th frame, and $\oplus$ represents feature concatenation. This residual formulation ensures that the reconstructed video maintains the visual identity defined by $\hat{\mathbf{v}}_0$ while exhibiting plausible motion trajectories. Then, we align the motion-aware features with the ground-truth video latents $\mathbf{z}_{gt}$ (extracted by the frozen VAE encoder):

$$\mathcal{L}_{\text{Motion}} = \frac{1}{F}\sum_{t=1}^{F}\|\hat{\mathbf{c}}_{dyn}^{(t)} - \mathbf{z}_{gt}^{(t)}\|_2^2. \tag{11}$$

EEG-Guided Video Reconstruction

Leveraging the rich priors of pre-trained T2V models, we propose a dual-guidance reconstruction mechanism. Specifically, the structured semantic prompt $\hat{\mathbf{c}}_{final}$ (from FSM) is injected into the U-Net via Cross-Attention to determine the semantic content (i.e., "what" appears). Simultaneously, the motion-aware feature $\hat{\mathbf{c}}_{dyn}$ (from ADM) serves as the structural initialization, replacing standard Gaussian noise to guide the spatiotemporal dynamics. This synergy ensures both high semantic fidelity and temporal coherence in the reconstructed videos.

Experiments

Dataset

The experiments use an updated SEED-DV dataset (X.-H. Liu et al., 2024). Following the original protocol, the seven 10min videos are divided into 1,400 two-second clips (200 clips × 7 trials), excluding the 3-second hints. For each of the 1,400 clips, we uniformly sample 12 frames (every 4th frame at 24 fps) and pass them, in chronological order, as a single multi-image prompt to GPT-5.2 (OpenAI Responses API) (OpenAI, n.d.-a, n.d.-b) which returns a JSON response containing captions that capture videos' key features and cross-frame motion, then breaking each down into descriptions of (1) objects with attributes/actions and (2) background, and lastly confirm that the two groups are mutually exclusive and contain only their own associated details.

Evaluation Metrics

Following prior benchmarks (J. Liu et al., 2025; Zhou et al., 2026), we assess performance using pixel-level and semantic-level metrics. **(i) Pixel-level fidelity:** We report the average Structural Similarity Index Measure (SSIM) (Wang et al., 2004) and Peak Signal-to-Noise Ratio (PSNR) (Huynh-Thu & Ghanbari, 2008) per frame. **(ii) Semantic Accuracy (Frame & Video):** We employ the N-way top-1 accuracy protocol ($N\in\{2, 40\}$). A reconstruction is deemed correct if the ground-truth class ranks first among N candidates (GT + distractors). For frame-based metrics, we use a CLIP classifier (Radford et al., 2021); for video-based metrics, we use a VideoMAE classifier pretrained on Kinetics-400 (Tong et al., 2022). Scores are reported as mean±std over repeated trials.

Table 1: Quantitative comparison of brain decoding between M-STAR and other methods.

# Classes	Method	Video-based		Frame-based			
		Semantic-level		Semantic-level		Pixel-level	
		2-way	40-way	2-way	40-way	SSIM	PSNR
10	Ours	0.858±0.03	0.346±0.03	0.842±0.02	0.339±0.03	0.318±0.07	9.377±1.56
	DynaMind	0.847±0.01	0.330±0.03	0.833±0.02	0.308±0.01	0.309±0.02	—
	MindCine	0.867±0.03	0.331±0.02	0.826±0.03	0.288±0.02	0.333±0.10	9.568±1.42
	EEG2Video	0.852±0.02	0.340±0.01	0.798±0.03	0.232±0.02	0.300±0.03	9.058±1.24
40	Ours	0.824±0.03	0.215±0.03	0.800±0.03	0.200±0.03	0.289±0.08	8.989±2.14
	DynaMind	0.820±0.02	0.188±0.02	0.791±0.03	0.185±0.01	0.280±0.01	—
	MindCine	0.818±0.03	0.179±0.02	0.787±0.02	0.171±0.03	0.278±0.11	9.035±1.74
	EEG2Video	0.798±0.03	0.159±0.01	0.774±0.02	0.138±0.01	0.256±0.03	8.684±1.46

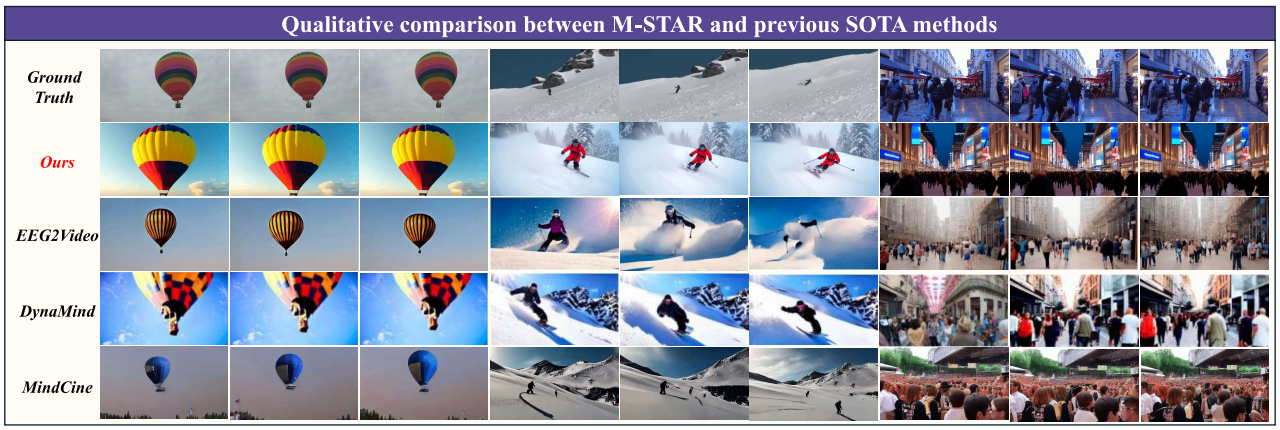

Figure 3: Comparison on EEG-to-video benchmark.

Results

Comparison with State-of-the-Art Methods

In this section, we compare M-STAR with previous video reconstruction methods on the SEED-DV dataset. As shown in Table 1, our proposed method demonstrates superior performance across most quantitative metrics. Notably, M-STAR delivers leading performance in semantic-level metrics (2-way and 40-way accuracy), indicating its exceptional capability in decoding high-level semantic information from EEG signals. In 40 classes full dataset, although M-STAR yields the second-best score in PSNR, it outperforms others in SSIM. It is worth noting that PSNR solely measures pixel-wise intensity differences (based on MSE) and may not fully reflect perceptual quality. In contrast, SSIM provides a better assessment of structural consistency and perceptual fidelity, further validating that our model generates frames with higher structural accuracy. These results validate that our method effectively bridges semantic and spatiotemporal feature learning, setting a new benchmark for video understanding tasks.

Ablation Study

In this section, we systematically evaluate the key components of M-STAR, as detailed in Table 2. Specifically, we investigate the impact of different semantic granularities, the effectiveness of the two proposed modules, and the contribution of different loss functions during the training phase.

Multi-Granularity Components Under the Only Bg setting, although the lack of foreground guidance limits semantic certainty, the model still achieves a 40-way score of 0.242; this is primarily attributed to the characteristics of the SEED-DV dataset, which contains certain scene-centric categories—such as forest and building —where foreground and background are inherently intertwined. While shifting to the Only Obj configuration yields a substantial boost in semantic metrics, it remains sub-optimal due to the absence of environmental context. Ultimately, by integrating both object and background semantics (Obj+Bg), the background information effectively compensates for these missing cues, resulting in reconstructed videos with significantly enhanced semantic precision and richness.

Table 2: Ablation study of key components in M-STAR on the 10-class subset.

Model	Video-based		Frame-based		
	2-way	40-way	2-way	40-way	PSNR
Multi-Granularity Components Ablation					
Only Bg	0.757±0.03	0.242±0.03	0.765±0.03	0.225±0.03	9.322±2.03
Only Obj	0.852±0.02	0.330±0.03	0.822±0.03	0.303±0.03	9.211±2.20
Obj+Bg	0.858±0.03	0.346±0.03	0.842±0.02	0.339±0.03	9.377±1.56
Module Ablation Performance					
w/o FSM	0.775±0.03	0.300±0.03	0.748±0.04	0.171±0.03	9.420±1.42
w/o ADM	0.835±0.03	0.335±0.03	0.828±0.02	0.292±0.03	9.098±1.42
Full Model	0.858±0.03	0.346±0.03	0.842±0.02	0.339±0.03	9.377±1.56
Loss Ablation Performance					
w/o $\mathcal{L}_{\text{Semantic}}$	0.841±0.03	0.291±0.03	0.788±0.03	0.249±0.03	9.388±1.94
w/o $\mathcal{L}_{\text{Align}}$	0.819±0.03	0.280±0.03	0.799±0.02	0.220±0.03	9.058±2.01

Module Ablation We further evaluate the contributions of the FSM and ADM branches. Removing FSM causes a sharp drop in semantic metrics, yet yields a higher PSNR. This occurs because, without explicit semantic guidance, the model tends to regress to generating conservative, blurry mean-value backgrounds; while this minimizes pixel-wise Mean Squared Error (thereby boosting PSNR), the resulting videos lack meaningful semantic content. Conversely, removing ADM leads to a significant decline in PSNR with only a minor impact on semantic scores, highlighting its role in maintaining perceptual fidelity. These results substantiate that the two modules are mutually indispensable: FSM ensures semantic fidelity, while ADM guarantees perceptual quality.

Loss Ablation Finally, we examine the influence of the training losses. Removing $\mathcal{L}_{\text{Semantic}}$ causes the model to degenerate into a coarse alignment framework similar to EEG2Video, where the lack of explicit factorized guidance leads to a substantial loss in semantic details. On the other hand, excluding $\mathcal{L}_{\text{Align}}$ significantly impairs the ADM branch. Without this alignment, the anchor features fail to match the ground-truth image embeddings and instead manifest as noise. This noisy guidance not only degrades perceptual fidelity, but also interferes with the semantic decoding process, resulting in an overall performance drop across all metrics.

Visualization

As illustrated in Fig. 3, M-STAR outperforms SOTA methods in semantic fidelity. Benefiting from the FSM, our model accurately reconstructs high-level semantics such as snowy slopes and specific objects (e.g., hot air balloons). Meanwhile, the ADM ensures superior temporal smoothness and realistic motion, as evidenced in the skiing sequence. To validate that our model truly learns the foreground and background semantics, we perform reconstructions using EEG embeddings

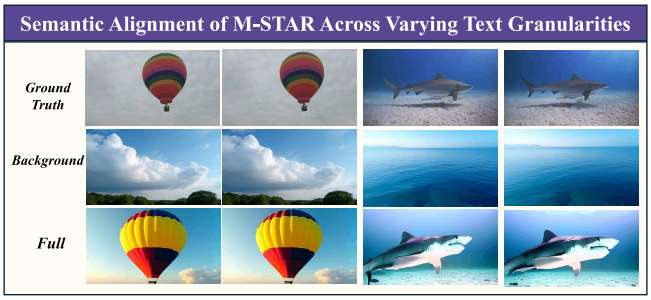

Figure 4: Qualitative results demonstrating the semantic alignment of M-STAR across different text granularities, from background-only to full object-background descriptions.

conditioned on varying semantic granularities. As illustrated in Fig. 4, when conditioned on background-only semantics (middle row), the model generates high-quality scenes (e.g., sky or ocean) that align with the background context but omit the specific objects. In contrast, when full object-background semantics are provided (bottom row), the model successfully reconstructs the salient objects while maintaining background consistency.

Conclusion

In conclusion, M-STAR addresses the critical challenges of semantic misalignment and temporal flickering in EEG-based video reconstruction. Driven by a bio-inspired cognitive strategy, our framework leverages the synergy between the Factorized Semantic Module and the Anchor-Guided Dynamics Module to effectively decouple complex semantics and enforce robust motion dynamics. Extensive experiments demonstrate its significant superiority over existing baselines, establishing a rigorous new standard for fine-grained, high-fidelity neural decoding.

Acknowledgements

This work was supported in part by grants from Brain Science and Brain-like Intelligence Technology-National Science and Technology Major Project (2025ZD0218900), National Key Research and Development Program of China (2024YFC3606800), National Natural Science Foundation of China (62376158), STI 2030-Major Projects+2022ZD0208500, Medical-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University "Jiao Tong Star" Program (YG2023ZD25, YG2024ZD25 and YG2024QNA03), Shanghai Jiao Tong University 2030 Initiative, the Lingang Laboratory (Grant No. LGL-1987), GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine, and Shanghai Jiao Tong University SCS-Shanghai Emotionhelper Technology Co., Ltd Joint Laboratory of Affective Brain-Computer Interfaces.

References

Benchetrit, Y., Banville, H., & King, J.-R. (2024). Brain decoding: Toward real-time reconstruction of visual perception [Conference paper. Yohann Benchetrit and Hubert Banville contributed equally.]. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2310.19812

Brandman, T., & Peelen, M. V. (2017). Interaction between scene and object processing revealed by human fmri and meg decoding. *The Journal of Neuroscience*, *37*(32), 7700–7710. https://doi.org/10.1523/JNEUROSCI.0582-17.2017

Chen, Z., Qing, J., & Zhou, J. H. (2023). Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/4e5e0daf4b05d8bfc6377f33fd53a8f4-Abstract-Conference.html

Fang, T., Qi, Y., & Pan, G. (2020). Reconstructing perceptive images from brain activity by shape-semantic gan. *Advances in Neural Information Processing Systems*, *33*, 13038–13048.

Gao, Y., Liu, J., Xu, Z., Wu, T., Zhang, E., Li, K., Yang, J., Liu, W., & Sun, X. (2024). Softclip: Softer cross-modal alignment makes clip stronger. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(3), 1860–1868.

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.

Gross, J. (2019). Magnetoencephalography in cognitive neuroscience: A primer. *Neuron*, *104*(2), 189–204. https://doi.org/10.1016/j.neuron.2019.07.001

Guenther, S., Kosmyna, N., & Maes, P. (2024). Image classification and reconstruction from low-density eeg. *Scientific Reports*, *14*, 16436. https://doi.org/10.1038/s41598-024-66228-1

Huang, S., Luo, H., Jing, H., Zhang, Q., Chang, L., Feng, Y., Lin, X., Qin, C., Chen, H., Jia, S., Sun, S., & Wang, Y. (2025). Need: Cross-subject and cross-task generalization for video and image reconstruction from eeg signals [NeurIPS 2025 (poster).]. *Advances in Neural Information Processing Systems (NeurIPS)*. https://openreview.net/forum?id=L3aEdxJMHl

Huang, S., Wang, Y., Luo, H., Jing, H., Qin, C., & Tang, J. (2025). Mindev: Multi-modal integrated diffusion framework for video reconstruction from eeg signals. *Proceedings of the 33rd ACM International Conference on Multimedia*, 3350–3359.

Huynh-Thu, Q., & Ghanbari, M. (2008). Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, *44*(13), 800–801.

Kavasidis, I., Palazzo, S., Spampinato, C., Giordano, D., & Shah, M. (2017). Brain2image: Converting brain signals into images. *Proceedings of the 25th ACM International Conference on Multimedia*, 1809–1817. https://doi.org/10.1145/3123266.3127907

Kim, S. G., Richter, W., & Uğurbil, K. (1997). Limitations of temporal resolution in functional MRI. *Magnetic Resonance in Medicine*, *37*(4), 631–636. https://doi.org/10.1002/mrm.1910370427

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

Liu, J., Lin, J., Li, J., & Li, J. (2025). Dynamind: Reconstructing dynamic visual scenes from eeg by aligning temporal dynamics and multimodal semantics to guided diffusion [Preprint]. https://doi.org/10.48550/arXiv.2509.01177

Liu, X.-H., Liu, Y.-K., Wang, Y., Ren, K., Shi, H., Wang, Z., Li, D., Lu, B.-L., & Zheng, W.-L. (2024). Eeg2video: Towards decoding dynamic visual perception from eeg signals. *Advances in Neural Information Processing Systems*, *37*. https://doi.org/10.52202/079017-2306

Liu, X.-H., Liu, Y.-K., Zhou, T., Lu, B.-L., & Zheng, W.-L. (2025). Mindcross: Fast new subject adaptation with limited data for cross-subject video reconstruction from brain signals. *arXiv preprint arXiv:2511.14196*.

Lu, Y., Du, C., Wang, C., Zhu, X., Jiang, L., Li, X., & He, H. (2025). Animate your thoughts: Reconstruction of dynamic natural vision from human brain activity [ICLR 2025 (Poster); introduces Mind-Animator]. *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=BpfsxFqhGa

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., Morito, Y., Tanabe, H. C., Sadato, N., & Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, *60*(5), 915–929. https://doi.org/10.1016/j.neuron.2008.11.004

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from

human brain activity. *Neuron*, *63*(6), 902–915. https://doi.org/10.1016/j.neuron.2009.09.006

Nicolas-Alonso, L. F., & Gomez-Gil, J. (2012). Brain computer interfaces, a review. *Sensors*, *12*(2), 1211–1279. https://doi.org/10.3390/s120201211

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641–1646. https://doi.org/10.1016/j.cub.2011.08.031

OpenAI. (n.d.-a). Images and vision — openai api documentation [Accessed: 2026-01-27].

OpenAI. (n.d.-b). Responses api — openai api reference [Accessed: 2026-01-27].

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, *139*, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695. https://doi.org/10.1109/CVPR52688.2022.01042

Sun, J., Li, M., & Moens, M.-F. (2025). Neuralflix: A simple while effective framework for semantic decoding of videos from non-invasive brain recordings. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(7), 7096–7104.

Takagi, Y., & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14453–14463. https://doi.org/10.1109/CVPR52729.2023.01389

Tong, Z., Fan, Y., Yang, J., Jiang, W., Li, M., Wang, Z., Xie, F., Zhang, J., Yuan, Z., Wang, Y., et al. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems (NeurIPS)*.

VanRullen, R., & Reddy, L. (2019). Reconstructing faces from fmri patterns using deep generative neural networks. *Communications Biology*, *2*, 193. https://doi.org/10.1038/s42003-019-0438-y

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., & Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, *28*(12), 4136–4160. https://doi.org/10.1093/cercor/bhx268

Wischnewski, M., & Peelen, M. V. (2021). Causal evidence for a double dissociation between object- and scene-selective regions of visual cortex: A preregistered tms replication study. *The Journal of Neuroscience*, *41*(4), 751–756. https://doi.org/10.1523/JNEUROSCI.2162-20.2020

Zhou, T.-Y., Liu, X.-H., Lu, B.-L., & Zheng, W.-L. (2026). Mindcine: Multimodal eeg-to-video reconstruction with large-scale pretrained models. *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7427–7431. https://doi.org/10.1109/ICASSP55912.2026.11460466