Bridging Brain and Semantics: A Hierarchical Framework for Semantically Enhanced fMRI-to-Video Reconstruction

Yujie Wei[1,*], Chenglong Ma[1,*], Jianxiong Gao[1], Chenhui Wang[1], Shiwei Zhang[2], Biao Gong[3], Shuai Tan[3], Hangjie Yuan[2], Hongming Shan[1,†]

[1] Fudan University [2]Alibaba Group [3]Ant Group

yjwei22@m.fudan.edu.cn, clma24@m.fudan.edu.cn

Abstract

*Reconstructing dynamic visual experiences as videos from functional magnetic resonance imaging (fMRI) is pivotal for advancing the understanding of neural processes. However, current fMRI-to-video reconstruction methods are hindered by a semantic gap between noisy fMRI signals and the rich content of videos, stemming from a reliance on incomplete semantic embeddings that neither capture video-specific cues (*e.g.*, actions) nor integrate prior knowledge. To this end, we draw inspiration from the dual-pathway processing mechanism in human brain and introduce* CINENEURON*, a novel hierarchical framework for semantically enhanced video reconstruction from fMRI signals with two synergistic stages. First, a bottom-up semantic enrichment stage maps fMRI signals to a rich embedding space that comprehensively captures textual semantics, image contents, action concepts, and object categories. Second, a top-down memory integration stage utilizes the proposed Mixture-of-Memories method to dynamically select relevant "memories" from previously seen data and fuse them with the fMRI embedding to refine the video reconstruction. Extensive experimental results on two fMRI-to-video benchmarks demonstrate that* CINENEURON *surpasses state-of-the-art methods across various metrics.*

1. Introduction

Understanding how the human brain processes visual information has long been a central goal in cognitive neuroscience [1, 11, 15, 50, 90]. Among various efforts, reconstructing visual experiences from neural signals, particularly functional magnetic resonance imaging (fMRI), offers valuable insights into the neural representations underlying visual perception. While earlier studies [2, 9, 53, 80] have made substantial progress in reconstructing static images from fMRI signals, extending this success to continuous visual stimuli (*e.g.*, videos) remains an open challenge.

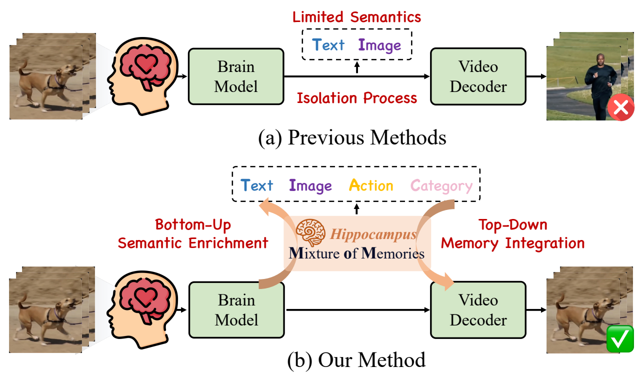

Figure 1. **Comparison with previous methods.** (a) Previous methods often align fMRI embeddings with limited semantics in an isolation process, relying only on the current stimulus and yielding semantically inaccurate results. (b) Our method enriches the fMRI embeddings with comprehensive video semantics and introduces Mixture-of-Memories to dynamically select and fuse prior knowledge, producing semantically coherent videos.

The difficulty mainly stems from accurately recovering both spatial content and temporal dynamics of videos from raw fMRI signals that have a low signal-to-noise ratio and limited temporal resolution [3]. The inherent sparsity and noise of fMRI further blur the high-level meaning of the underlying neural responses [73, 113], making it difficult to extract comprehensive semantics from fMRI data alone. Prior works [10, 25, 66] have attempted to enrich semantics by aligning fMRI embeddings with additional images and text representations. However, as shown in Fig. 1, these methods still suffer from low semantic coherence and reconstruct videos with incorrect objects for two reasons: **1)** They ignore the rich semantics embedded in video stimuli,

*Equal Contribution †Corresponding Author
Yujie Wei, Chenglong Ma, and Hongming Shan are with Institute of Science and Technology for Brain-inspired Intelligence, MOE Frontiers Center for Brain Science, Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, State Key Laboratory of Brain Function and Disorders, Fudan University, Shanghai, China

such as actions and object categories, which are essential for faithful video reconstruction; **2)** They treat the fMRI reconstruction in isolation, using only the current stimulus. In fact, the human brain interprets the visual world by incorporating both incoming stimuli and memories of learned semantic concepts [83, 84]. Therefore, *enriching the semantic capacity of fMRI embeddings and integrating prior knowledge are both crucial for bridging the semantic gap between brain activity and reconstructed videos.*

This hierarchical design principle is also consistent with the brain's dual processing pathways [12, 22]. Specifically, a bottom-up pathway extracts high-level semantics by accumulating sensory evidence from early to higher-order visual regions, while a top-down pathway delivers semantic predictions and integrated memories from hippocampal systems back to the sensory cortex to refine stimulus perception. This synergistic interaction is fundamental to how the brain constructs a coherent perception of the real world.

Drawing inspiration from this dual-pathway mechanism, we present CINENEURON, a hierarchical framework for semantically enhanced fMRI-to-video reconstruction by unifying two synergistic stages: **1)** *bottom-up semantic enrichment*, which enriches the fMRI representations with multimodal semantics extracted from the current video stimuli, and **2)** *top-down memory integration*, which incorporates "memories" from the semantic embeddings of previously seen data to supplement and refine the fMRI representations, ensuring a coherent video reconstruction.

Specifically, in the first stage, we decompose the semantics hidden in the video stimuli into multiple dimensions and extract corresponding embeddings using heterogeneous pre-trained encoders. Beyond aligning fMRI only with image and text modalities, we design explicit video tasks using action concepts and object categories. Then, we train a Brain Model to map fMRI signals into a space capturing these comprehensive semantics, creating a semantically rich foundation for fMRI-to-video reconstruction.

In the second stage, we introduce a Mixture-of-Memories (MoM) module to refine the Stage-1 fMRI embedding by fusing it with prior "memories," *i.e.*, multimodal embeddings from previously seen data. Analogous to hippocampal processing, MoM proceeds in two coordinated steps: retrieval and integration. The retrieval step uses a modality-aware router to dynamically assign weights to multimodal memories and selects the most relevant ones via weighted similarity. The integration step then aggregates these selected memories with the fMRI embedding, producing a fused representation that incorporates prior knowledge while preserving current neural evidence. This final representation is then passed to a video decoder to reconstruct temporally consistent and semantically coherent videos.

We evaluate the effectiveness of CINENEURON on two challenging fMRI-to-video reconstruction datasets, cc2017 [110] and CineBrain [20], where it outperforms state-of-the-art methods across metrics and reconstructs semantically accurate and visually coherent videos.

Our contributions are summarized as follows: **1)** We present CINENEURON, a novel hierarchical framework for semantically enhanced fMRI-to-video reconstruction by synergizing the bottom-up semantic enrichment with the top-down memory integration. **2)** We propose to enrich the semantics of fMRI signals by learning explicit tasks for video stimuli, capturing comprehensive semantics with image, text, action, and category concepts. **3)** We propose a novel Mixture-of-Memories method that dynamically selects and fuses relevant multimodal embeddings with the fMRI embedding to refine video decoding. **4)** Extensive experimental results on two fMRI-to-video benchmarks demonstrate the superior performance of CINENEURON over state-of-the-art methods.

2. Related Work

2.1. Brain Decoding

Decoding visual stimuli from brain activity, particularly fMRI signals, has attracted growing attention in recent years [4, 9, 14, 18, 19, 36–41, 59, 60, 65, 68, 75, 80, 97, 102]. While previous works have shown promising results in decoding static images from fMRI signals [26, 111, 115, 119, 120], the human visual perception is inherently dynamic and continuous. This gap highlights the need for fMRI-to-video reconstruction, which enables a deeper understanding of how the brain processes real-world stimuli. However, reconstructing video content from fMRI signals presents greater challenges due to the low temporal resolution and signal-to-noise ratio of the measurements [17, 58, 66, 96, 98].

Recent fMRI-to-video reconstruction methods typically learn semantic embedding from raw fMRI signals and then drive a video generator [66, 98, 116]. For instance, Mind-Video [10] aligns a self-supervised fMRI encoder to CLIP [76] and adapts an inflated Stable Diffusion [79] model for video reconstruction. NeuroClips [25] advances this further by encoding semantic keyframes from low-level perceptual flows, improving the smoothness of generated videos. Despite these advances, previous methods capture only shallow semantics as they rely primarily on image-text aligned spaces and treat reconstruction as an isolated process, using only the current fMRI stimulus and ignoring prior learned knowledge. In contrast, our work explicitly enriches fMRI with comprehensive semantics by learning tasks over text, image, *action*, and *category* concepts and integrating multimodal memories for refined decoding.

2.2. Diffusion Models for Video Generation

Diffusion models have become the dominant paradigm for image and video synthesis [28, 30–32, 108, 118], primarily due to their superior generation fidelity and flexible conditioning mechanisms [51, 52, 85, 87–89, 105, 106]. Early works like VDM [33] pioneer this direction by extending image diffusion to videos, while subsequent text-to-video models [6, 7, 99] integrate spatiotemporal modules for higher temporal consistency. Driven by the rapid progress of Transformers [93] in understanding [62, 63] and generation [61, 91, 107] tasks, Diffusion Transformers (DiT) [74] establishes a new paradigm that further enhances generation quality [45, 109]. Modern video generation models push long-range coherence with advanced DiT architectures. For example, CogVideoX [114] introduces a spatiotemporal VAE [8, 44] and a multimodal DiT with full attention. Wan [94] adopts a cross-attention DiT trained on large-scale data, enhancing video coherence. In this work, we adopt Wan2.1 1.3B as our default video decoder backbone, but our framework is model-agnostic and can be readily integrated into other models like CogVideoX.

3. Method

The overall framework of CineNeuron is illustrated in Fig. 2, which consists of two main learning stages. *1)* In the bottom-up semantic enrichment stage (Sec. 3.1), a Brain Model is designed to map fMRI signals to an embedding space that captures comprehensive semantics. *2)* In the top-down memory integration stage (Sec. 3.2), the proposed Mixture-of-Memories module first utilizes fMRI embeddings to retrieve relevant text, image, and action embeddings from an established memory pool, and then integrate them into the fMRI embeddings to serve as the input condition for the video decoder.

3.1. Bottom-Up Semantic Enrichment

Extracting reliable semantic cues from noisy raw fMRI signals remains a bottleneck in reconstructing dynamic videos, which necessitates transforming the fMRI signals into more compact, informative representations using multimodal semantic priors [21, 67]. To address this, we enrich the fMRI embeddings with comprehensive semantics, including text, image, action, and object categories, all of which are crucial for video reconstruction. This is driven by a suite of dedicated learning tasks: *image-text semantic alignment*, *action semantic alignment*, and *category semantic learning*.

To begin with, we build on the previous works [10, 20, 49] to design a transformer-based Brain Model that produces two outputs: a global fMRI token summarizing the fMRI signal and an fMRI embedding that retains finer contextual patterns. This dual output allows for functional decoupling, where the global fMRI token supports alignment with external semantic representations and the fMRI embedding facilitates downstream feature fusion and video reconstruction in the subsequent stage. Below, we detail the semantic enrichment within each task.

Image-Text Semantic Alignment. Given a video clip $\boldsymbol{y} \in \mathbb{R}^{B\times F\times C\times H\times W}$ and the paired fMRI signals $\boldsymbol{x} \in \mathbb{R}^{B\times N}$, we first feed $\boldsymbol{x}$ into the Brain Model f to obtain the fMRI global token $\boldsymbol{f^c} \in \mathbb{R}^{B\times D}$ and the fMRI embedding $\boldsymbol{f^e} \in \mathbb{R}^{B\times L\times D'}$. The video $\boldsymbol{y}$ and its caption c are processed by the CLIP encoder to produce multi-frame image embeddings $\boldsymbol{e}^{\text{img}} \in \mathbb{R}^{BF\times D}$ and text embeddings $\boldsymbol{e}^{\text{txt}} \in \mathbb{R}^{B\times D}$, respectively. Then, a projection head φ_v aggregates the multi-frame image embeddings into a consolidated image embedding $\hat{\boldsymbol{e}}^{\text{img}} = \varphi_v(\boldsymbol{e}^{\text{img}}) \in \mathbb{R}^{B\times D}$. Finally, we employ the InfoNCE loss [71] to align the global fMRI token with image and text embeddings:

$$\mathcal{L}_{\text{clip}} = \mathcal{L}_{\text{info}}(\boldsymbol{f^c}, \hat{\boldsymbol{e}}^{\text{img}}) + \mathcal{L}_{\text{info}}(\boldsymbol{f^c}, \boldsymbol{e}^{\text{txt}}), \tag{1}$$

$$\mathcal{L}_{\text{info}}(\boldsymbol{f^c}, \boldsymbol{e}) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\boldsymbol{f}_i^{\boldsymbol{c}} \cdot \boldsymbol{e}_i/\tau\right)}{\sum_{j=1}^{B} \exp\left(\boldsymbol{f}_i^{\boldsymbol{c}} \cdot \boldsymbol{e}_j/\tau\right)}, \tag{2}$$

where τ is the temperature.

Action Semantic Alignment. Apart from image and text representations, actions such as walking and swimming are vital to understanding video content and preventing semantic discrepancies for the fMRI-to-video reconstruction. To this end, we introduce an action alignment task using ViCLIP [103], a video understanding model pretrained on a large-scale dataset of 10 million video-text pairs. ViCLIP exhibits robust zero-shot capabilities for video understanding and action recognition, thereby effectively capturing the action and temporal information in the input videos [35]. We process the video $\boldsymbol{y}$ with ViCLIP to obtain action embeddings $\boldsymbol{e}^{\text{act}}$, and map the fMRI global token into the action space using an action head φ_a, yielding $\boldsymbol{f^a} = \varphi_a(\boldsymbol{f^c})$. Consequently, we compute the action contrastive loss $\mathcal{L}_{\text{action}}$ to align the projected fMRI token and the action embedding, formulated as $\mathcal{L}_{\text{action}} = \mathcal{L}_{\text{info}}(\boldsymbol{f^a}, \boldsymbol{e}^{\text{act}})$.

Category Semantic Learning. While contrastive learning ensures semantic alignment at the embedding level, explicitly identifying objects within a video helps the Brain Model capture category semantics. To achieve this, we introduce a multi-label classification task. Specifically, we utilize Qwen2.5-VL [101] to extract object categories from video captions based on the MSCOCO category list [54], and train a classification head φ_c for category prediction using binary cross-entropy (BCE) loss. However, directly training the Brain Model with this objective faces two main challenges: *1)* the large number of categories complicates the task, impeding effective category semantic learning; and *2)* the class imbalance can bias predictions towards frequent classes. To address these issues, we simplify the categories by filtering and merging infrequent classes into a reduced

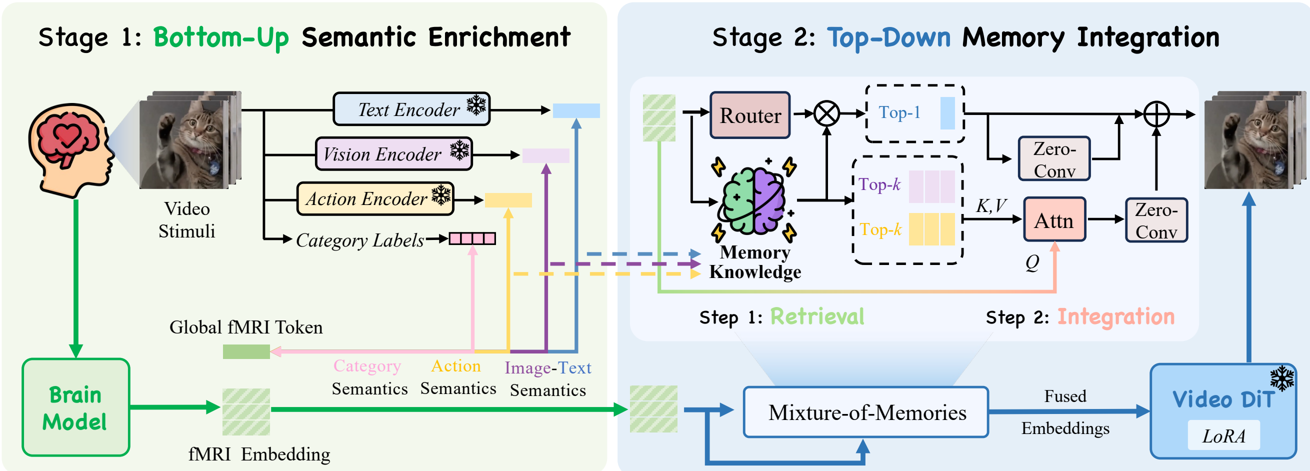

Figure 2. **Overview of the proposed CINENEURON.** In stage 1, given an input fMRI-video pair, the fMRI signals are first embedded by a Brain Model and enriched with the text, image, action, and category semantics extracted from the video. In stage 2, the proposed Mixture-of-Memories method dynamically selects multimodal embeddings from previously seen data via a router and fuses them with the fMRI embeddings via a fusion mechanism, guiding the Video DiT for semantically-enhanced and high-quality video reconstruction.

set of superclasses, and apply Focal Loss [55] to dynamically reweight samples within each category. The final classification loss $\mathcal{L}_{\text{cls}}$ is a combination of BCE and focal loss. Finally, the loss for this stage is given by:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{clip}} + \lambda_1 \mathcal{L}_{\text{action}} + \lambda_2 \mathcal{L}_{\text{cls}}. \tag{3}$$

3.2. Top-Down Memory Integration

The first stage equips fMRI embeddings with multimodal semantic information, yet further refinement is necessary to enhance and supplement semantic and perceptual details in the sparse fMRI signals. Inspired by the hippocampal memory replay, we introduce Mixture-of-Memories (MoM), a two-step retrieval-and-integration method. MoM first replays and retrieves multimodal embeddings from a memory pool built from seen training videos, and then integrates these memories into the fMRI embedding to guide the decoder, refining reconstructed videos.

Retrieval Step. The first step of MoM is to retrieve the most relevant memories using the fMRI embeddings from the bottom-up semantic enrichment stage. A straightforward solution is to compare the text embeddings in the memory pool with the fMRI embeddings, obtain the most similar one as a condition, and feed it into the subsequent Video DiT model for reconstruction. However, such naïve strategy is sensitive to imprecise retrieval due to noise or ambiguity in the fMRI signals.

To this end, we incorporate multimodal embeddings as memories and adaptively weighs the contributions of each modality. Let $\mathcal{M}$ denote the memory pool with N entries, where each entry i is a tuple of tri-modality embeddings $(\boldsymbol{e}_i^{\text{txt}}, \boldsymbol{e}_i^{\text{img}}, \boldsymbol{e}_i^{\text{act}})$ extracted from the same training video. A routing network R first computes instance-specific retrieval weights of three modalities using fMRI embedding $\boldsymbol{f}^{\boldsymbol{e}}$:

$$W_r = [w^{\text{txt}}, w^{\text{img}}, w^{\text{act}}] = \text{softmax}(R(\boldsymbol{f}^{\boldsymbol{e}})). \tag{4}$$

Instead of retrieving from each modality independently, we compute a *mixture score* S_i for each memory entry through a weighted sum of its multimodal similarities to $\boldsymbol{f}^{\boldsymbol{e}}$:

$$S_i = \sum_{m \in \mathcal{M}_{\text{modal}}} w_m \cdot \text{sim}(\boldsymbol{f}^{\boldsymbol{e}}, \boldsymbol{e}_i^m), \quad i \in \{1, \dots, N\}, \tag{5}$$

where $\mathcal{M}_{\text{modal}} = \{\text{txt}, \text{img}, \text{act}\}$ and $\text{sim}(\cdot,\cdot)$ denotes cosine similarity. We then re-rank all memory entries based on their mixture scores S_i. Specifically, let $\{i_1, i_2, \dots, i_K\}$ be the indices of the memory entries corresponding to the top-K highest scores, we select the top-1 text embedding $\boldsymbol{e}_{\text{txt}}^{\text{mem}} = \boldsymbol{e}_{i_1}^{\text{txt}}$ from the highest-scoring entry, along with the sets of top-K image and action embeddings $\boldsymbol{e}_{\text{img}}^{\text{mem}} = \{\boldsymbol{e}_{i_k}^{\text{img}}\}_{k=1}^{K}$ and $\boldsymbol{e}_{\text{act}}^{\text{mem}} = \{\boldsymbol{e}_{i_k}^{\text{act}}\}_{k=1}^{K}$. This mixture weighting strategy allows the model to dynamically prioritize the most relevant modality for each input instance, enhancing retrieval precision.

Integration Step. The second step of MoM is to fuse retrieved multimodal embeddings from the memory pool into the fMRI embedding. Specifically, we design a fusion mechanism using two cross-attention layers and zero-convolution layers. In cross-attention layers, the fMRI embedding $\boldsymbol{f}^{\boldsymbol{e}}$ functions as the query, while K retrieved image embeddings $\boldsymbol{e}_{\text{img}}^{\text{mem}}$ and action embeddings $\boldsymbol{e}_{\text{act}}^{\text{mem}}$ serve as keys and values, respectively. This attention mechanism injects visual and action cues into the fMRI embedding, producing an enhanced representation $\hat{\boldsymbol{f}}^{\boldsymbol{e}}$:

$$\begin{aligned} \hat{\boldsymbol{f}}^{\boldsymbol{e}} &= \text{CrossAttention}(\boldsymbol{Q}^{\boldsymbol{e}}, \boldsymbol{K}^{\text{img}}, \boldsymbol{V}^{\text{img}}) \\ &+ \text{CrossAttention}(\boldsymbol{Q}^{\boldsymbol{e}}, \boldsymbol{K}^{\text{act}}, \boldsymbol{V}^{\text{act}}), \end{aligned} \tag{6}$$

where $\boldsymbol{Q}^{e} = \boldsymbol{W_Q}\boldsymbol{f}^{e}$, $\boldsymbol{K}^{\text{img}} = \boldsymbol{W_K}\boldsymbol{e}^{\text{mem}}_{\text{img}}$, and $\boldsymbol{V}^{\text{img}} = \boldsymbol{W_V}\boldsymbol{e}^{\text{mem}}_{\text{img}}$. $\boldsymbol{K}^{\text{act}}$ and $\boldsymbol{V}^{\text{act}}$ are computed similarly.

We then employ a dual-stream structure to fuse $\hat{\boldsymbol{f}}^{e}$ with the retrieved top-1 text embedding $\boldsymbol{e}^{\text{mem}}_{\text{txt}}$. Concretely, $\hat{\boldsymbol{f}}^{e}$ is first passed through a normalization layer [86, 112] to scale it comparably to the text embeddings, facilitating effective integration. Following normalization, a zero-conv layer $\mathcal{Z}_{\text{fMRI}}$ is applied to $\hat{\boldsymbol{f}}^{e}$ to ensure stability in training by initializing outputs to zero. To balance the contributions of both modalities, another zero-conv layer $\mathcal{Z}_{\text{txt}}$ is added for text embedding $\boldsymbol{e}^{\text{mem}}_{\text{txt}}$ in a residual manner, ensuring the initial fused embedding equates to $\boldsymbol{e}^{\text{mem}}_{\text{txt}}$:

$$\boldsymbol{z_f} = \mathcal{Z}_{\text{fMRI}}(\text{Norm}(\hat{\boldsymbol{f}}^{e})), \quad \boldsymbol{z_t} = \mathcal{Z}_{\text{txt}}(\boldsymbol{e}^{\text{mem}}_{\text{txt}}) + \boldsymbol{e}^{\text{mem}}_{\text{txt}}. \tag{7}$$

Finally, the embeddings are then summed to produce the fused embeddings $\boldsymbol{f}^{\text{fuse}}$ as follows:

$$\boldsymbol{f}^{\text{fuse}} = \boldsymbol{z_t} + \alpha * \boldsymbol{z_f}, \tag{8}$$

where α is the weighting factor for balance. The final fused embedding $\boldsymbol{f}^{\text{fuse}}$ are fed into the Video DiT ϵ_θ, trained with LoRA [34] modules using diffusion loss [32, 56, 57]:

$$\mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{y},\epsilon,\boldsymbol{f}^{\text{fuse}},t}\left[\left\|\epsilon - \epsilon_\theta(\boldsymbol{y_t}, \boldsymbol{f}^{\text{fuse}}, t)\right\|_2^2\right], \tag{9}$$

where $\boldsymbol{y_t}$ is the noisy video at diffusion timestep t, and ϵ is the random Gaussian noise. In total, the loss for the Top-Down Memory Integration stage is formulated as:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{stage1}} + \mathcal{L}(\theta). \tag{10}$$

3.3. Inference

In contrast to the complex inference processes of previous methods [25, 98], which necessitate components like ControlNet [117], keyframe reconstruction for the first frame of the video, and additional condition generation, our *end-to-end* pipeline is straightforward and efficient. Users simply input fMRI signals into our Brain Model, and the Mixture-of-Memories then produces fused embeddings that condition the video decoder for seamless reconstruction.

4. Experiment

4.1. Experimental Setup

Datasets. We conduct experiments on two publicly available fMRI-to-video datasets: cc2017 [110] and CineBrain [20]. The cc2017 dataset comprises fMRI data from three subjects who viewed various natural videos; the stimuli are split into 18 eight-minute training segments (2 repeats) and 5 eight-minute test segments (10 repeats), yielding around 11.5 hours of data per subject. The CineBrain dataset provides synchronized EEG and fMRI recordings from six healthy participants as they watch and listen to 20 episodes of "The Big Bang Theory" (720p), totaling approximately 6 hours of audiovisual stimuli per subject. The fMRI data were sampled at 1.25 Hz. We preprocess the cc2017 fMRI data following NeuroClips [25]. For the CineBrain dataset, which provides visual ROIs, we follow the same pipeline to extract the additional hippocampus ROI for our experiments. More details about preprocessing are provided in the Supplementary Materials A.1.

Evaluation Metrics. We evaluate the reconstructed videos at semantic, spatiotemporal, and pixel levels [20, 25]. For semantic-level evaluation, we compute N-way top-K accuracy to assess whether the generated videos semantically match the ground-truth (GT) clips, using a VideoMAE [92]-based classifier on 400 video classes from the Kinetics-400 dataset [42], following prior work [20, 25, 98]. For spatiotemporal-level evaluation, we use CLIP temporal consistency (CLIP-pcc) [77] and DINO [72] temporal consistency [20, 35] (DTC) to measure the spatiotemporal coherence of the generated videos. Additionally, we employ the Motion Smoothness (MS) and Dynamic Degree [35] to assess the smoothness and magnitude of movements, along with the Mean End-Point Error (EPE) to assess the motion consistency compared with GT videos. Apart from video-based metrics above, we also assess frame-wise image quality at both semantic and pixel levels; see Supplementary Materials A.2 and C.2 for more details and results.

Implementation Details. We use the AdamW [64] optimizer with the OneCycle learning rate schedule [82]. In the first stage, the Brain Model is trained for 8,000 steps with a batch size of 144 and a learning rate of 1×10^{-4}. In the second stage, the Brain Model along with the router and fusion module in Mixture-of-Memories are trained for 20 epochs with a batch size of 32 and a learning rate of 1×10^{-6}. Following [10, 20], the transformer-based Brain Model consists of 24 layers, each with a hidden dimension of 2048 and a token length of 513 (512 fMRI embedding tokens plus one global fMRI token). We use a LoRA rank of 16 with a scaling factor of 16. We adopt Wan2.1 1.3B [95] as our video decoder. On the cc2017 dataset, we generate 57-frame videos at 624×624, much longer and higher resolution than NeuroClips (16 frames at 256×256) and MindVideo (6 frames at 256×256). On the CineBrain dataset, we generate 33 frames at 720×480 resolution, matching the source videos. More implementation details are provided in the Supplementary Materials A.3.

4.2. Main Results

Results on cc2017 Dataset. Fig. 3 reveals that the baselines suffer from semantic errors and low reconstruction quality. For example, in the left and right panels of Fig. 3, both MindVideo and NeuroClips fail to reconstruct the yoga and petting actions; in the middle panel, MindVideo incorrectly decodes the dog as a person, while NeuroClips hallu-

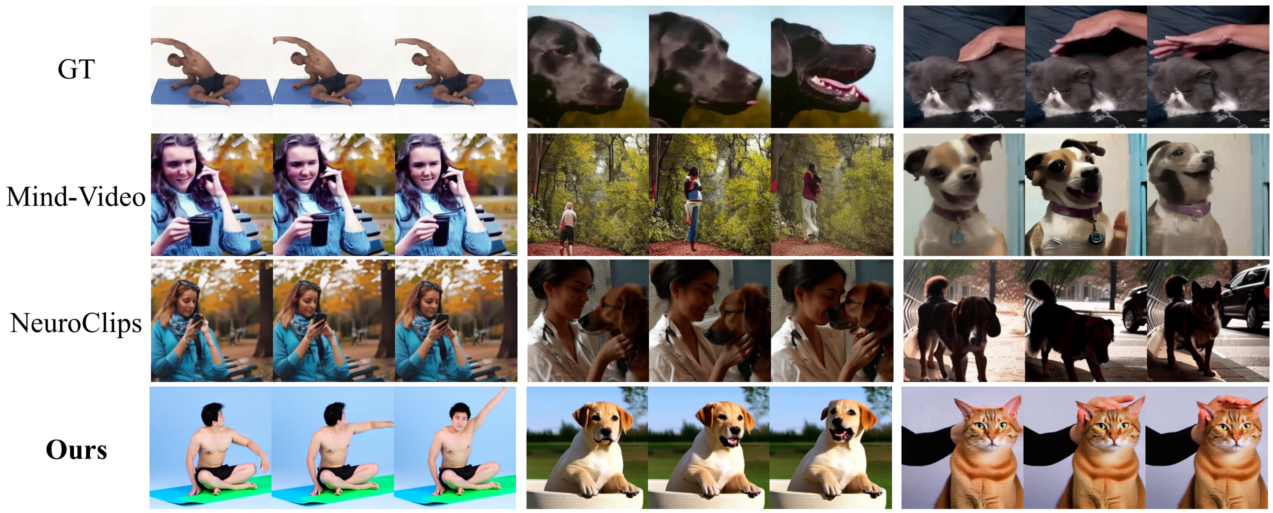

Figure 3. **Qualitative comparison of CineNeuron and baselines on the cc2017 dataset**.

Table 1. **Quantitative comparison results of all subjects on the cc2017 and CineBrain datasets.** Results for the cc2017 dataset are quoted from [25, 66], except for the DTC and MS (Motion Smoothness) metrics. Results for the CineBrain dataset are quoted from [20], except for the MS metric. "*" denotes methods reimplemented using the same decoder model and fMRI input as CineNeuron.

Dataset	Methods	Semantic-level		Spatiotemporal-level			Pixel-level	
		2-way	50-way	CLIP-pcc	DTC	MS	SSIM	PSNR
cc2017	Kupershmidt [47]	0.771±0.03	-	0.386±0.47	-	-	0.135±0.08	8.761±2.22
	MinD-Video [10]	0.839±0.03	0.197±0.02	0.408±0.46	0.884±0.08	0.901±0.05	0.171±0.08	8.662±1.52
	Mind-Animator [66]	0.830	-	0.425	-	-	0.321	9.220
	NeuroClips [25]	0.834±0.03	0.220±0.01	0.738±0.17	0.926±0.05	0.955±0.01	**0.390**±0.08	9.211±1.46
	CineNeuron	**0.850**±0.02	**0.240**±0.03	**0.972**±0.01	**0.954**±0.01	**0.966**±0.02	0.375±0.06	**9.476**±0.23
CineBrain	GLFA [49]	0.801	0.167	0.735	0.706	-	0.123	7.526
	CineSync [20]	0.893	0.307	0.945	0.907	0.974±0.01	0.240	11.92
	CineSync*	0.933±0.02	0.324±0.02	0.977±0.01	0.942±0.01	0.973±0.01	0.267±0.04	**16.04**±2.35
	CineNeuron	**0.937**±0.02	**0.393**±0.03	**0.988**±0.01	**0.975**±0.01	**0.975**±0.01	**0.271**±0.05	16.02±2.38

cinates spurious objects (a woman) and background context (a room) that are absent from the GT video. In contrast, our CineNeuron faithfully reconstructs complex videos with coherent action, category, and context semantics, demonstrating its superior performance.

Tab. 1 shows that our method achieves the best scores on semantic and spatiotemporal metrics, showing the effectiveness of our comprehensive semantic enrichment and the Mixture-of-Memories design. At the pixel level, we obtain the highest PSNR and competitive SSIM. Although NeuroClips reports slightly higher SSIM, its reconstructions exhibit inferior semantics and temporal coherence, indicating a trade-off between fidelity and semantic alignment. In contrast, our method strikes a better balance. Note that Tab. 1 reports the average scores across all subjects; per-subject results are provided in Supplementary Materials C.1.

Results on CineBrain Dataset. Tab. 1 shows that our method outperforms all baselines across most metrics while achieving comparable PSNR with only marginal differences, confirming the broad effectiveness of our framework. Notably, our method also outperforms CineSync*, an enhanced baseline that uses the same decoder and additional hippocampal fMRI input as CineNeuron. Per-subject results are provided in Supplementary Materials C.1.

Fig. 4 shows that strong baselines (*e.g.*, CineSync and CineSync*) still struggle to accurately reconstruct visual details or capture action semantics. In contrast, videos reconstructed from our method are temporally consistent and semantically accurate.

Comparison on Motion Consistency. Tab. 2 shows that our method achieves the lowest (best) endpoint error (EPE) on both cc2017 and CineBrain datasets, indicating the closest match to GT motion trajectories. We further assess motion magnitude using the Dynamic Degree metric [35], reporting the absolute difference from the GT (lower is better). Our method achieves the smallest discrepancy, closely matching the GT motion magnitude.

Human Evaluation Results. We conduct a comprehensive human evaluation to further evaluate the video reconstruction performance. Twenty participants evaluate 360 video groups, comparing videos from four anonymous methods (MindVideo, NeuroClips, Mind-Animator, and our

Figure 4. **Qualitative comparison of CINENEURON and baselines on the CineBrain dataset.**

Table 2. **Quantitative comparison of motion consistency on the cc2017 and CineBrain datasets**. "ΔDynamic Degree" denotes the absolute difference from the GT videos' Dynamic Degree; lower values indicate motion amplitudes closer to the GT.

	Test Set: cc2017			**Test Set: CineBrain**	
METRIC	MINDVIDEO	NEUROCLIPS	CINENEURON	CINESYNC	CINENEURON
EPE (↓)	9.045	4.432	**1.628**	3.258	**2.126**
ΔDynamic Degree (↓)	0.1155	0.2133	**0.0200**	0.2611	**0.0649**

Table 3. **Human evaluation results.**

METHOD	Semantic Alignment	Temporal Consistency	Visual Quality	Overall Fidelity
MindVideo	8.31%	7.74%	9.08%	7.79%
NeuroClips	16.30%	18.13%	13.79%	15.80%
Mind-Animator	11.62%	8.23%	6.46%	8.82%
Ours	**63.77%**	**65.90%**	**70.67%**	**67.59%**

Table 4. **Quantitative ablation of each component on Subject 1 of the cc2017 dataset.**

KEY COMPONENTS				Semantic-level		Spatiotemporal-level		
$\mathcal{L}_{\text{clip}}$	$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{\text{action}}$	MoM	2-way	50-way	CLIP-pcc	DTC	MS
✓				0.824±0.01	0.217±0.03	**0.973**±0.01	**0.960**±0.03	0.961±0.01
✓	✓			0.829±0.01	<u>0.228±0.03</u>	0.965±0.01	0.951±0.01	0.956±0.02
✓	✓	✓		<u>0.835±0.02</u>	0.223±0.04	<u>0.970±0.01</u>	0.957±0.01	<u>0.963±0.01</u>
✓	✓	✓	✓	**0.846**±0.02	**0.237**±0.02	**0.973**±0.01	<u>0.959±0.01</u>	**0.967**±0.02

CINENEURON) to the GT video over four dimensions: 1) Semantic Alignment, 2) Temporal Consistency, 3) Visual Quality, and 4) Overall Fidelity. The results in Tab. 3 show that CINENEURON significantly outperforms competing methods across all dimensions.

4.3. Ablation Studies

We conduct ablation studies to evaluate each component's effectiveness on cc2017 dataset, and the impact of using additional hippocampus fMRI data on CineBrain dataset.

Ablation on Each Proposed Component. Quantitative (Tab. 4) and qualitative (Fig. 6) ablation results highlight the effectiveness of each component. For ablation baselines without Mixture-of-Memories (MoM), we condition the video decoder on the retrieved most similar (top-1) text embedding for reconstruction. We observe that the baseline using only $\mathcal{L}_{\text{clip}}$ struggles to generate semantically accurate videos. Introducing the classification loss $\mathcal{L}_{\text{cls}}$ improves object recognition (*e.g.*, identifying the object as two persons, as in Fig. 6), but still fails to capture correct actions. Adding the action contrastive loss $\mathcal{L}_{\text{action}}$ further helps the model grasp action concepts (*e.g.*, detecting that a person is running), though details remain inaccurate. Finally, incorporating the MoM module allows the model to accurately perceive both object categories and actions, and effectively reconstruct fine-grained details.

Ablation on the Input fMRI Data. By default, on the CineBrain dataset, CINENEURON takes fMRI data from both the visual area and hippocampus as input for video reconstruction. As shown in the second row of Tab. 5, removing the hippocampus fMRI input causes a performance drop in semantic-level metrics while spatiotemporal metrics remain comparable. This indicates that the hippocampus provides crucial semantic information, validating our top-down memory integration design. Furthermore, we test the effect of adding fMRI signals from the mPFC region, which is known to collaborate with the hippocampus in memory processing. As shown in the third row of Tab. 5, including mPFC inputs yields further performance gains, especially on semantic metrics. This result reinforces our model's de-

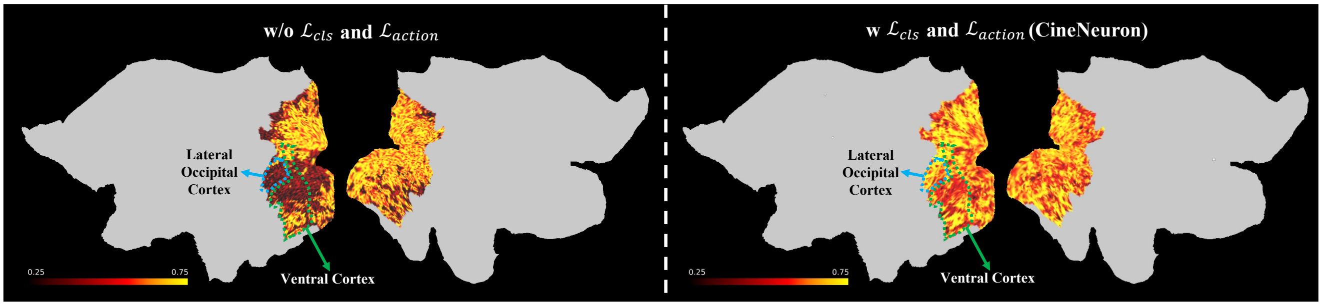

Figure 5. **Visualization of voxel weights from the first regression layer.** Voxel weights are averaged and normalized to [0, 1], displayed with a 0.25 to 0.75 colorbar. The blue and green dotted lines indicate the lateral occipital and ventral cortex, respectively.

Figure 6. **Qualitative ablation of each proposed component.**

sign, which integrates retrieved memories (long-term) with current fMRI features (working memory).

Ablation on Integration Step in MoM. Tab. 6 ablates the MoM integration step and shows it is the main contributor to spatiotemporal quality gains. Prior work like MindVideo directly conditions the generator on raw fMRI features, which are misaligned with the pretrained model's text-anchored conditioning space, leading to poor spatiotemporal quality. In contrast, our integration step in MoM uses residual fusion with retrieved multimodal memories to supplement details missing from fMRI while preserving the stability of the generator's feature space, producing higher-quality videos.

4.4. Interpretation Results

To assess the neural interpretability of our model, we visualize voxel weights within the visual cortex on a brain flat map, comparing models trained without and with the combination of category semantic learning $\mathcal{L}_{\text{cls}}$ and action semantic alignment $\mathcal{L}_{\text{action}}$ on the CineBrain dataset, as shown in Fig. 5. Incorporating these tasks significantly increases voxel weights in the lateral occipital and ventral visual cortices—regions that are critically related to object recognition, shape analysis, and complex visual processing [27, 46]. Notably, weight increases also appear in visual motion-sensitive areas (*e.g.* V5/MT, MST, and FST) within the lateral occipital cortex, highlighting the effectiveness of the designed tasks in enhancing both object identification and visual motion capture [5, 16].

Table 5. **Quantitative ablation of hippocampus and mPFC fMRI inputs on Subject 5 of the CineBrain dataset.**

METHODS	Semantic-level		Spatiotemporal-level		
	2-way	50-way	CLIP-pcc	DTC	MS
CINENEURON	0.938±0.01	0.401±0.02	0.990±0.01	0.978±0.01	0.976±0.01
– hippo. input	0.935±0.02	0.358±0.03	0.993±0.01	0.974±0.01	0.975±0.01
+ mPFC	0.945±0.01	0.442±0.03	0.993±0.01	0.981±0.01	0.976±0.01

Table 6. **Quantitative ablation of MoM's integration step on Subject 1 of the cc2017 dataset.**

METHODS	Semantic-level		Spatiotemporal-level		
	2-way	50-way	CLIP-pcc	DTC	MS
w/o Integration	0.830±0.02	0.224±0.05	0.802±0.01	0.920±0.01	0.961±0.01
w/ Integration	**0.846**±**0.02**	**0.237**±**0.02**	**0.973**±**0.01**	**0.959**±**0.01**	**0.967**±**0.02**

5. Conclusion

In this paper, we present CINENEURON, a hierarchical framework for semantically enhanced fMRI-to-video reconstruction that combines bottom-up semantic enrichment with top-down memory integration. We enrich fMRI embeddings with comprehensive video semantics from text, image, action, and category modalities. Building on these embeddings, our Mixture-of-Memories adaptively retrieves multimodal memories and fuses them into the fMRI representation to refine video reconstruction. Extensive experiments on two datasets show that CINENEURON surpasses prior methods on both qualitative and quantitative results. Furthermore, we provide an interpretable analysis demonstrating functional alignment between the human visual cortex and our model, offering insights into neural processing.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China (Nos. T2596013 and 62471148), STI2030-Major Projects (No. 2021ZD0200204), and Shanghai Center for Brain Science and Brain-inspired Technology.

References

[1] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022. 1

[2] Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. DreamDiffusion: Generating high-quality images from brain EEG signals. *arXiv preprint arXiv:2306.16934*, 2023. 1

[3] Johan Baijot, Stijn Denissen, Lars Costers, Jeroen Gielen, Melissa Cambron, Miguel D'Haeseleer, Marie B D'hooghe, Anne-Marie Vanbinst, Johan De Mey, Guy Nagels, et al. Signal quality as Achilles' heel of graph theory in functional magnetic resonance imaging in multiple sclerosis. *Scientific Reports*, 11(1):7376, 2021. 1

[4] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812*, 2023. 2

[5] Richard T Born and David C Bradley. Structure and function of visual area MT. *Annu. Rev. Neurosci.*, 28(1):157–189, 2005. 8

[6] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 3

[7] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 3

[8] Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *Forty-second International Conference on Machine Learning*, 2025. 3

[9] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. 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*, pages 22710–22720, 2023. 1, 2

[10] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*, 36:24841–24858, 2023. 1, 2, 3, 5, 6, 4

[11] Ian Daly. Neural decoding of music from the EEG. *Scientific Reports*, 13(1):624, 2023. 1

[12] Floris P De Lange, Micha Heilbron, and Peter Kok. How do expectations shape perception? *Trends in Cognitive Sciences*, 22(9):764–779, 2018. 2

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 2

[14] Alou Diakite, Cheng Li, Yousuf Babiker M Osman, Zan Chen, Yiang Pan, Jiawei Zhang, Tao Tan, Hairong Zheng, and Shanshan Wang. Dual-uncertainty guided multimodal mri-based visual pathway extraction. *IEEE Transactions on Biomedical Engineering*, 72(6):1993–2000, 2025. 2

[15] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116, 2019. 1, 2

[16] Dominic H Ffytche, CN Guy, and Semir Zeki. The parallel visual motion inputs into areas V1 and V5 of human cerebral cortex. *Brain*, 118(6):1375–1394, 1995. 8

[17] Camilo Fosco, Benjamin Lahner, Bowen Pan, Alex Andonian, Emilie Josephs, Alex Lascelles, and Aude Oliva. Brain netflix: Scaling data to reconstruct videos from brain signals. In *European Conference on Computer Vision*, pages 457–474. Springer, 2024. 2

[18] Jianxiong Gao, Yuqian Fu, Yun Wang, Xuelin Qian, Jianfeng Feng, and Yanwei Fu. Mind-3d: Reconstruct high-quality 3d objects in human brain, 2023. 2

[19] Jianxiong Gao, Yanwei Fu, Yuqian Fu, Yun Wang, Xuelin Qian, and Jianfeng Feng. Mind-3d++: Advancing fmri-based 3d reconstruction with high-quality textured mesh generation and a comprehensive dataset, 2025. 2

[20] 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, 3, 5, 6, 4

[21] Yuan Gao, Tao Tan, Xin Wang, Regina Beets-Tan, Tianyu Zhang, Luyi Han, Antonio Portaluri, Chunyao Lu, Xinglong Liang, Jonas Teuwen, et al. Multi-modal longitudinal representation learning for predicting neoadjuvant therapy response in breast cancer treatment. *IEEE Journal of Biomedical and Health Informatics*, 2025. 3

[22] Charles D Gilbert and Wu Li. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363, 2013. 2

[23] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013. 2

[24] 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

[25] Zixuan Gong, Guangyin Bao, Qi Zhang, Zhongwei Wan, Duoqian Miao, Shoujin Wang, Lei Zhu, Changwei Wang, Rongtao Xu, Liang Hu, et al. NeuroClips: Towards high-fidelity and smooth fmri-to-video reconstruction. *Advances in Neural Information Processing Systems*, 37: 51655–51683, 2024. 1, 2, 5, 6, 4, 7

[26] Zixuan Gong, Qi Zhang, Guangyin Bao, Lei Zhu, Rongtao Xu, Ke Liu, Liang Hu, and Duoqian Miao. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14247–14255, 2025. 2

[27] Kalanit Grill-Spector and Kevin S Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.*, 15(8):536–548, 2014. 8

[28] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[29] Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198: 125–136, 2019. 2

[30] Yujin Han, Hao Chen, Andi Han, Zhiheng Wang, Xinyu Liu, Yingya Zhang, Shiwei Zhang, and Difan Zou. Turning internal gap into self-improvement: Promoting the generation-understanding unification in mllms. *arXiv preprint arXiv:2507.16663*, 2025. 3

[31] Yujin Han, Andi Han, Wei Huang, Chaochao Lu, and Difan Zou. Can diffusion models learn hidden inter-feature rules behind images? *arXiv preprint arXiv:2502.04725*, 2025.

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 5

[33] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 3

[34] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[35] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 3, 5, 6

[36] Zhicheng Jiao, Haoxuan You, Fan Yang, Xin Li, Han Zhang, and Dinggang Shen. Decoding EEG by visual-guided deep neural networks. In *IJCAI*, pages 1387–1393. Macao, 2019. 2

[37] Haodong Jing, Dongyao Jiang, Yongqiang Ma, Haibo Hua, Bo Huang, and Nanning Zheng. Beyond brain decoding: Visual-semantic reconstructions to mental creation extension based on fmri. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19258–19268, 2025.

[38] Haodong Jing, Yongqiang Ma, Panqi Yang, Haibo Hua, and Nanning Zheng. Pinpointing visual content: Disentangled features in multimodal model for eeg representation learning and decoding. *Knowledge-Based Systems*, page 114212, 2025.

[39] Haodong Jing, Yongqiang Ma, Wenjie Gao, Dongyao Jiang, Shuai Huang, and Nanning Zheng. Mind-vad: Brain-inspired fmri-to-video precise reconstruction via cross-modal autoregressive diffusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

[40] Haodong Jing, Yongqiang Ma, Panqi Yang, Haoyu Li, Shuai Huang, Badong Chen, and Nanning Zheng. Damind: Zero-shot visual cross-domain alignment and representation for eeg decoding. *IEEE Transactions on Image Processing*, 2026.

[41] Haodong Jing, Panqi Yang, Dongyao Jiang, Zhipeng Liu, Nanning Zheng, and Yongqiang Ma. Evoke: Efficient and high-fidelity eeg-to-video reconstruction via decoupling implicit neural representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5539–5547, 2026. 2

[42] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[43] Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun. Mixco: Mix-up contrastive learning for visual representation. *arXiv preprint arXiv:2010.06300*, 2020. 4

[44] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[45] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3

[46] Zoe Kourtzi and Nancy Kanwisher. Representation of perceived object shape by the human lateral occipital complex. *Science*, 293(5534):1506–1509, 2001. 8

[47] Ganit Kupershmidt, Roman Beliy, Guy Gaziv, and Michal Irani. A penny for your (visual) thoughts: Self-supervised reconstruction of natural movies from brain activity. *arXiv preprint arXiv:2206.03544*, 2022. 6, 5

[48] Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N Apurva Ratan Murty, et al. Bold moments: modeling short visual events through a video fmri dataset and metadata. *bioRxiv*, pages 2023–03, 2023. 5, 6

[49] Chong Li, Xuelin Qian, Yun Wang, Jingyang Huo, Xiangyang Xue, Yanwei Fu, and Jianfeng Feng. Enhancing cross-subject fmri-to-video decoding with global-local functional alignment. In *European Conference on Computer Vision*, pages 353–369. Springer, 2024. 3, 6, 4

[50] 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. 1

[51] Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12112–12123, 2025. 3

[52] Quanhao Li, Zhen Xing, Rui Wang, Haidong Cao, Qi Dai, Daoguo Dong, and Zuxuan Wu. Flashmotion: Few-step controllable video generation with trajectory guidance. *arXiv preprint arXiv:2603.12146*, 2026. 3

[53] Yamin Li, Ange Lou, Ziyuan Xu, Shengchao Zhang, Shiyu Wang, Dario Englot, Soheil Kolouri, Daniel Moyer, Roza Bayrak, and Catie Chang. Neurobolt: Resting-state EEG-to-fMRI synthesis with multi-dimensional feature mapping. *Advances in Neural Information Processing Systems*, 37: 23378–23405, 2024. 1

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 3

[55] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international Conference on computer vision*, pages 2980–2988, 2017. 4

[56] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 5

[57] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 5

[58] 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. 2

[59] Yulong Liu, Yongqiang Ma, Wei Zhou, Guibo Zhu, and Nanning Zheng. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding. *arXiv preprint arXiv:2302.12971*, 2023. 2

[60] Yulong Liu, Yongqiang Ma, Guibo Zhu, Haodong Jing, and Nanning Zheng. See through their minds: Learning transferable brain decoding models from cross-subject fmri. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5730–5738, 2025. 2

[61] Zhihang Liu, Xiaoyi Bao, Pandeng Li, Junjie Zhou, Zhaohe Liao, Yefei He, Kaixun Jiang, Chen-Wei Xie, Yun Zheng, and Hongtao Xie. Showtable: Unlocking creative table visualization with collaborative reflection and refinement. *arXiv preprint arXiv:2512.13303*, 2025. 3

[62] Zhihang Liu, Chen-Wei Xie, Pandeng Li, Liming Zhao, Longxiang Tang, Yun Zheng, Chuanbin Liu, and Hongtao Xie. Hybrid-level instruction injection for video token compression in multi-modal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8568–8578, 2025. 3

[63] Zhihang Liu, Chen-Wei Xie, Bin Wen, Feiwu Yu, Jixuan Chen, Pandeng Li, Boqiang Zhang, Nianzu Yang, Yinglu Li, Zuan Gao, et al. Capability: A comprehensive visual caption benchmark for evaluating both correctness and thoroughness. *arXiv preprint arXiv:2502.14914*, 2025. 3

[64] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[65] Jingyu Lu, Haonan Wang, Qixiang Zhang, and Xiaomeng Li. A cognitive process-inspired architecture for subject-agnostic brain visual decoding. *arXiv preprint arXiv:2511.02565*, 2025. 2

[66] Yizhuo Lu, Changde Du, Chong Wang, Xuanliu Zhu, Liuyun Jiang, Xujin Li, and Huiguang He. Animate your thoughts: Decoupled reconstruction of dynamic natural vision from slow brain activity. *arXiv preprint arXiv:2405.03280*, 2024. 1, 2, 6

[67] Na Luo, Weiyang Shi, Zhengyi Yang, Ming Song, and Tianzi Jiang. Multimodal fusion of brain imaging data: Methods and applications. *Machine Intelligence Research*, 21(1):136–152, 2024. 3

[68] Yongqiang Ma, Wen Zhang, Ming Du, Haodong Jing, and Nanning Zheng. Hierarchical bayesian causality network to extract high-level semantic information in visual cortex. *International Journal of Neural Systems*, 34(01):2450002, 2024. 2

[69] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19023–19034, 2022. 5

[70] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011. 2

[71] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[72] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5

[73] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. 1

[74] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[75] Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 233–243, 2024. 2

[76] 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. PmLR, 2021. 2

[77] 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. PmLR, 2021. 5

[78] 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

[79] 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. 2

[80] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023. 1, 2, 4

[81] 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. 4

[82] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. 5

[83] Eleanor Spens and Neil Burgess. A generative model of memory construction and consolidation. *Nature human behaviour*, 8(3):526–543, 2024. 2

[84] Kaori Takehara-Nishiuchi. Entorhinal cortex and consolidated memory. *Neuroscience Research*, 84:27–33, 2014. 2

[85] Shuai Tan, Bin Ji, and Ye Pan. Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22146–22156, 2023. 3

[86] Shuai Tan, Biao Gong, Yutong Feng, Kecheng Zheng, Dandan Zheng, Shuwei Shi, Yujun Shen, Jingdong Chen, and Ming Yang. Mimir: Improving video diffusion models for precise text understanding. *arXiv preprint arXiv:2412.03085*, 2024. 5

[87] Shuai Tan, Bill Gong, Bin Ji, and Ye Pan. Fixtalk: Taming identity leakage for high-quality talking head generation in extreme cases. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3

[88] Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobin Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. In *International Conference on Learning Representations*, 2025.

[89] Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *European Conference on Computer Vision*, pages 398–416. Springer, 2025. 3

[90] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023. 1

[91] Genmo Team. Mochi 1. `https://github.com/genmoai/models`, 2024. 3

[92] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 5

[93] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[94] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3

[95] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 5

[96] Chong Wang, Hongmei Yan, Wei Huang, Jiyi Li, Yuting Wang, Yun-Shuang Fan, Wei Sheng, Tao Liu, Rong Li, and Huafu Chen. Reconstructing rapid natural vision with fmri-conditional video generative adversarial network. *Cerebral Cortex*, 32(20):4502–4511, 2022. 2, 5

[97] Haonan Wang, Jingyu Lu, Hongrui Li, and Xiaomeng Li. Zebra: Towards zero-shot cross-subject generalization for universal brain visual decoding. *arXiv preprint arXiv:2510.27128*, 2025. 2

[98] Haonan Wang, Qixiang Zhang, Lehan Wang, Xuanqi Huang, and Xiaomeng Li. NEURONS: Emulating the human visual cortex improves fidelity and interpretability in fmri-to-video reconstruction. *arXiv preprint arXiv:2503.11167*, 2025. 2, 5

[99] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 3

[100] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024. 5

[101] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

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

[103] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 3

[104] 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. 2

[105] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024. 3

[106] Yujie Wei, Shiwei Zhang, Hangjie Yuan, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Feng Liu, Zhizhong Huang, Jiaxin Ye, et al. Dreamvideo-2: Zero-shot subject-driven video customization with precise motion control. *arXiv preprint arXiv:2410.13830*, 2024. 3

[107] Yujie Wei, Shiwei Zhang, Hangjie Yuan, Biao Gong, Longxiang Tang, Xiang Wang, Haonan Qiu, Hengjia Li, Shuai Tan, Yingya Zhang, et al. Dreamrelation: Relation-centric video customization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12381–12393, 2025. 3

[108] Yujie Wei, Shiwei Zhang, Hangjie Yuan, Yujin Han, Zhekai Chen, Jiayu Wang, Difan Zou, Xihui Liu, Yingya Zhang, Yu Liu, et al. Routing matters in moe: Scaling diffusion transformers with explicit routing guidance. *arXiv preprint arXiv:2510.24711*, 2025. 3

[109] Yujie Wei, Xinyu Liu, Shiwei Zhang, Hangjie Yuan, Jinbo Xing, Zhekai Chen, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, et al. Dreamvideo-omni: Omni-motion controlled multi-subject video customization with latent identity reinforcement learning. *arXiv preprint arXiv:2603.12257*, 2026. 3

[110] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, 28(12):4136–4160, 2018. 2, 5

[111] Dian Xie, Peiang Zhao, Jiarui Zhang, Kangqi Wei, Xiaobao Ni, and Jiong Xia. BrainRAM: Cross-modality retrieval-augmented image reconstruction from human brain activity. In *ACM Multimedia 2024*, 2024. 2

[112] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. SANA: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. 5

[113] Yifan Yang, Yutong Mao, Xufu Liu, and Xiao Liu. Brain-MAE: A region-aware self-supervised learning framework for brain signals. *arXiv preprint arXiv:2406.17086*, 2024. 1

[114] 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. 3

[115] Jacob Yeung, Andrew F Luo, Gabriel Sarch, Margaret M Henderson, Deva Ramanan, and Michael J Tarr. Reanimating images using neural representations of dynamic stimuli. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5331–5343, 2025. 2

[116] Wenwen Zeng, Yonghuang Wu, Yifan Chen, Xuan Xie, Chengqian Zhao, Feiyu Yin, Guoqing Wu, and Jinhua Yu. Mindshot: Multi-shot video reconstruction from fmri with llm decoding. *arXiv preprint arXiv:2508.02480*, 2025. 2

[117] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international Conference on computer vision*, pages 3836–3847, 2023. 5

[118] Xulu Zhang, Xiaoyong Wei, Wentao Hu, Jinlin Wu, Jiaxin Wu, Wengyu Zhang, Zhaoxiang Zhang, Zhen Lei, and Qing Li. A survey on personalized content synthesis with diffusion models. *Machine Intelligence Research*, 22(5):817–848, 2025. 3

[119] Yuxiao Zhao, Guohua Dong, Lei Zhu, and Xiaomin Ying. Memory recall: Retrieval-augmented mind reconstruction for brain decoding. *Information Fusion*, page 103280, 2025. 2

[120] Xiao Zheng, Wanzhong Chen, Mingyang Li, Tao Zhang, Yang You, and Yun Jiang. Decoding human brain activity with deep learning. *Biomedical Signal Processing and Control*, 56:101730, 2020. 2