CortexVideo: A Semantic-Spatial Dual-Anchor Framework for High-Fidelity fMRI-to-Video Reconstruction

Xiaoquan Shen[1,2,3], Jiajun Li(✉)[1,3,4], Jiaxuan Chen(✉)[1,3], and Gang Pan[1,3]

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
{xiaoquan_shen, jiajun_li, jiaxuan_chen}@zju.edu.cn
[2] Shanghai Innovation Institute, Shanghai, China
[3] State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China
[4] Liangzhu Laboratory, Zhejiang University School of Medicine, Hangzhou, China

Abstract. Recent advances in neural decoding have enabled the direct reconstruction of visual stimuli from non-invasive brain signals. However, reconstructing continuous video streams remains highly challenging due to the continuous spatial and semantic transformations inherent in dynamic scenes. Most methods rely on a "reconstruct-then-describe" paradigm, where captions are generated from reconstructed video keyframes, which are highly prone to semantic drift. To overcome these challenges, we propose a novel decoding framework, **CortexVideo**, to reconstruct video from functional magnetic resonance imaging (fMRI). Inspired by the visual dual-stream hypothesis, CortexVideo introduces dual-guidance decoding strategy. Specifically, we leverage subject-adaptive semantic information to guide video reconstruction, while concurrently integrating keyframe perceptual weights. Experiments demonstrate that CortexVideo provides a cognitive enhancement to existing hierarchical models along two critical dimensions: video semantic understanding and spatial localization. The reconstructed videos show greater fidelity to the visual stimuli in both semantic and spatial aspects, achieving state-of-the-art results.

Keywords: Neural Decoding · Cognitive Enhancement · Fine-grained Spatial Grounding · Dual-Anchor Reconstruction

1 Introduction

Decoding external stimuli or internal cognitive states [13, 31, 42] from neural activity has long been a central objective in neuroscience research [38, 41]. In recent years, the rapid advancement of multimodal and generative models in artificial intelligence, together with their deep integration with neuroscience, has driven remarkable progress in reconstructing high-fidelity visual stimuli from high-dimensional neural signals obtained through functional magnetic resonance imaging (fMRI) [7, 44, 48]. Nevertheless, researchers in this field continue to confront considerable challenges, as our understanding of how the brain simultaneously encodes semantic and spatial information remains limited.

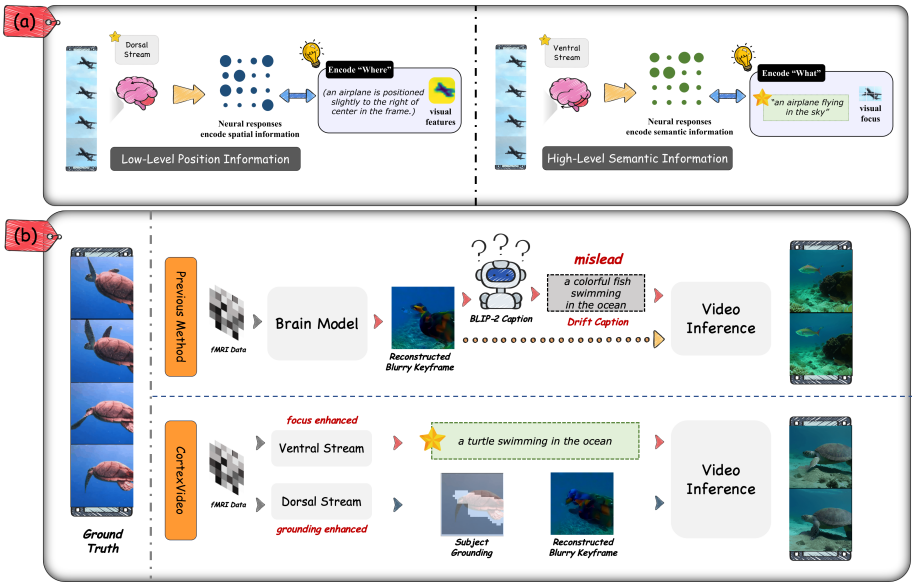

Fig. 1: Conceptual illustration of the dual-stream processing. (a)The Dorsal Stream encodes low-level spatial "where" information (e.g., object positioning), while the Ventral Stream captures high-level semantic "what" cues (e.g., identity and context). (b)Compared to previous methods that often suffer from drift captions, our CortexVideo leverages this dual-anchoring to achieve superior spatial localization and semantic fidelity in fMRI-to-video reconstruction.

Early studies [23, 40, 60] on fMRI-to-video reconstruction employed machine learning and deep learning techniques to map fMRI signals directly onto pixel space. However, these approaches often produced videos with ambiguous semantics and pronounced visual artifacts. With the advent of pre-trained diffusion models, recent works [10, 19, 20, 35, 59] have proposed more sophisticated methods that enhance the semantic clarity and spatial fidelity of reconstructed videos. Nonetheless, state-of-the-art frameworks such as NEURONS [59] and NeuroClips [20] rely on pre-trained vision–language models (e.g., BLIP-2 [32]) to extract high-level semantic information from blurred keyframes. This "reconstruct-then-describe" paradigm can cause error accumulation, leading to semantic drift and degrading reconstruction fidelity. Furthermore, existing models still struggle to capture low-level spatial details, particularly in accurately reconstructing the number of subjects and their spatial relationships in dynamic visual scenes. To overcome these challenges, effective biologically inspired guidance is essential for improving the fidelity of video reconstruction.

In cognitive neuroscience, the visual dual-stream hypothesis is a foundational theory [21, 36, 55]. It posits that when the brain is exposed to visual stimuli, it encodes not only "what" information (e.g., semantics) but also "where" information, including object positions [11,17,34,39,49,50]. For high-level semantics, the

ventral stream supports interpretive processing of visual content. Guided by selective attention, the brain does not encode all pixel-level information uniformly in complex scenes; rather, it prioritizes salient information that is task-relevant or personally meaningful [12]. For low-level perception, the dorsal stream encodes "where" information and encompasses regions including OPA, PPA, and SPL [14, 27, 28], which support the processing of spatial relationships, environmental layout, and motion. Earlier studies often assumed a strict semantic correspondence between stimulus images and the evoked brain activity, but this overlooks discrepancies between subjective neural representations and objective supervisory signals, leading to semantic misalignment [5,8] and spatial mismatch.

Building on the visual dual-stream hypothesis, we present CortexVideo, a framework that reconstructs high-fidelity videos from fMRI via a Semantic–Spatial Dual-Anchor paradigm. As showed in Fig. 1, our decoder integrates ventral-stream semantics with dorsal-stream spatial representations. This synergy improves both semantic fidelity and spatial accuracy in the reconstructions. We design an encoder to map fMRI signals to latent representations. Through cross-modal alignment, we embed high-level semantics and decode intermediate representations(e.g., keyframes) to guide subsequent video generation. In contrast to conventional methods, the core innovations of CortexVideo are threefold:

- **Adaptive Semantic Anchoring:** We design an adaptive semantic decoder that precisely captures subjects' attentional focus. Replacing caption generation from keyframes, it provides more reliable semantic guidance for subsequent video reconstruction and effectively mitigates semantic drift.

- **Spatial Perception Enhancement:** To improve spatial localization, we construct a feature-similarity heatmap by masking the primary subject with Grounded SAM [47] and computing similarity using DINO-v3 [53] features. A Spatial Perception Loss leverages this supervision to train the fMRI encoder to learn latent positional relationships, thereby improving spatial accuracy in the final video synthesis.

- **Dual-Anchor Guided Generation:** The final video reconstruction is jointly guided by three signals: trusted semantic cues, spatially informed reconstructed frames, and a spatially aware video stream, collectively enabling a more faithful restoration of the dynamic visual stimuli perceived by the brain.

2 Related Works

2.1 Visual Stimuli Reconstruction

In the study of decoding visual stimuli from fMRI, the reconstruction of static images and dynamic videos presents distinct challenges and trajectories of progress.

Static Image Reconstruction. After initial explorations [2, 15, 25], the research community in this area quickly converged on a successful paradigm [3, 4,

9,16,18,51,52]. This paradigm aims to address the information sparsity of fMRI signals. Its core strategy is to leverage the rich semantic and spatial representations of CLIP to align fMRI signals with common image/text modalities [6]. Subsequently, it utilizes the powerful generative capabilities of diffusion models to synthesize images that perform well at both the pixel and semantic levels.

Dynamic Video Reconstruction. More recently, new progress has been made with the application of multimodal (CLIP [44]) and advanced generative models. For instance, MinD-Video [10] first achieved video reconstruction with clear semantic content, but it lacked consideration for low-level visual details. NeuroClips [20] improved upon this by introducing dual semantic and perceptual reconstructors; however, its reliance on keyframe-derived captions for guidance can introduce semantic drift. NEURONS [59] further enhanced reconstruction quality and interpretability by decoupling the decoding process into four explicit hierarchical sub-tasks. Nevertheless, it still faces challenges in accurately reconstructing dynamic visual stimuli. Despite enhanced spatial localization, the method struggles to accurately reconstruct dynamic visual stimuli, particularly in scenes with complex object relationships.

2.2 Diffusion-based Generative Models

Diffusion Models. Diffusion models have recently become a research focus within the generative AI community. This wave began with DDPM [24], which demonstrated high-quality image synthesis capabilities comparable to GANs. In the static image generation domain, to enhance semantic controllability, DALL·E 2 [46] innovatively leveraged the joint representation space of CLIP. To address generation efficiency, Stable Diffusion [48] proposed executing the diffusion process within the latent space of a VQVAE [56]. Furthermore, to enable customized generation from pre-trained models, works such as ControlNet [62] and T2I-Adapter [37] have explored methods for adding external conditional control networks.

Extending this success to the more challenging domain of dynamic video synthesis, the mainstream technical route involves equipping existing image diffusion models with additional temporal modeling capabilities. For instance, AnimateDiff [22] introduced an innovative “plug-and-play" motion module that seamlessly transforms static image models into video generators. Similarly, Blattmann et al. [1] proposed a temporal autoencoder fine-tuning paradigm to generate high-resolution videos.

3 Methodology

The overall framework of CortexVideo framework is illustrated in Fig. 2, which is designed to reconstruct high-fidelity video from fMRI signals via a Semantic-Spatial Dual-Anchor paradigm inspired by the visual dual-stream hypothesis. Our model incorporates a hierarchical decoding structure, referencing classical

video reconstruction models, but is cognitively enhanced through two core modules: the Attentive Semantic Decoder(Sec. 3.2) and the incorporation of image feature similarity for fine-grained spatial perception modeling(Sec. 3.3), which provide more robust semantic guidance and accurate subject position reconstruction for subsequent video generation(Sec. 3.5).

3.1 Framework Overview

The overall model takes the fMRI time-series data $X \in \mathbb{R}^{T \times V}$ (where T is the number of time points and V is the number of voxels) as its sole input and outputs the reconstructed video clip ($\hat{y}_c$).For training purposes, the video stimulus is segmented into clips at the fMRI temporal resolution (two seconds). Each ground truth video clip y_c consists of $F = 6$ frames, denoted as $y_c \in \mathbb{R}^{B \times F \times C \times H \times W}$, with the corresponding fMRI signal clip being $\mathbf{x}_c$.

The fMRI signal X is simultaneously processed by two dedicated encoder streams:

Main Backbone Encoder E_f: Used for training all video and image reconstruction, spatial perceptio, and auxiliary tasks. It is responsible for capturing spatial and motion details.

Semantic Encoder E_f^{ASD}: Dedicated solely to the Attentive Semantic Decoding(ASD) task, responsible for extracting the purest subjective semantic representation y.

Let $X \in \mathbb{R}^{T \times V}$ be the fMRI signal. The main backbone encoder E_f outputs F_{global}; the semantic encoder E_f^{ASD} outputs y.

3.2 Ventral Stream Enhancement

To mitigate the semantic drift problem caused by the conventional "reconstruct-then-describe" paradigm, We adopt the latest brain-to-text approach, MindSA [5], as our ASD for improved semantic guidance. The ASD captures the subject's subjective focus of attention and provides reliable semantic support for subsequent video reconstruction, adhering to the function of the Ventral Stream (the "What" stream) in object recognition and semantic understanding.

Semantic Encoder. The ASD utilizes a specially trained fMRI Vision Transformer (ViT) encoder E_f^{ASD} to extract the semantic representation. This encoder, based on the ViT structure, is tailored for processing fMRI signals:

$$y = E_f^{ASD}(X), \quad y \in \mathbb{R}^{D_y} \tag{1}$$

where X is the fMRI signal input to this encoder. E_f^{ASD} processes the high-dimensional fMRI data into a sequence of embedded vectors, which are finally projected to the dense semantic representation y via a linear layer or MLP.

Self-adaptive Semantic Consensus. The core objective of this stage is to determine which regions of the visual stimulus image I are preferentially attended to by the subject's brain, given the fMRI representation y. We hypothesize that the pure fMRI semantic representation y should primarily align with

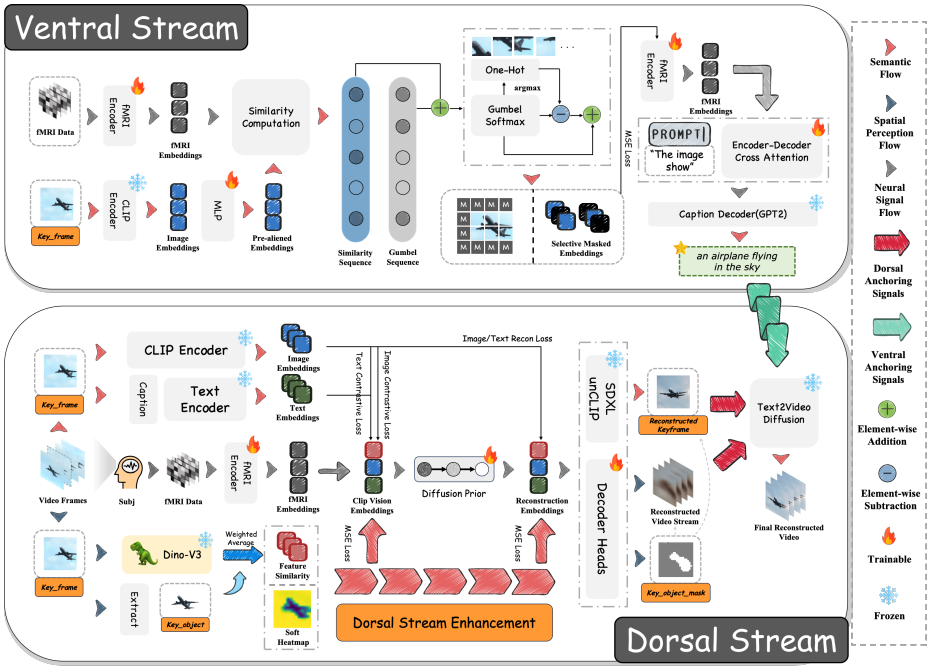

Fig. 2: The overall framework of CortexVideo integrates the Ventral Stream(Sec. 3.2) and Dorsal Stream(Sec. 3.3) for high-fidelity video reconstruction(Sec. 3.5). i)The Ventral Stream captures semantic priors from fMRI representations, dynamically forming visual focal points, and uses the captured fine-grained semantic cues to guide video caption generation. ii)The Dorsal Stream incorporates spatial reconstruction losses to enhance the model's ability to localize subjects, thereby improving the spatial fidelity of the output. iii)During the synthesis stage, the Ventral Stream offers semantic cues, and the Dorsal Stream provides motion features of the subjects. Together, they collaborate to faithfully reconstruct dynamic stimuli, ensuring high fidelity to the subject's brain activity.

the semantic features of these focal regions, thereby eliminating the semantic misalignment between the objective stimulus and subjective cognition.

We first divide the image I into N visual patches and extract their patch features $\{x_i\}_{i=1}^{N}$ using a CLIP visual encoder. These features are subsequently projected into the fMRI representation space via a shallow Multi-Layer Perceptron f_{MLP}, denoted as $\{f_{MLP}(x_i)\}_{i=1}^{N}$.Next, we use the Gumbel-Softmax trick [26] to enable differentiable and sparse attention selection. We compute a similarity score vector p between the fMRI representation y and all projected patch features:

$$p_i = \frac{\exp(f_{MLP}(x_i) \cdot y + \gamma_i)}{\sum_{j=1}^{N} \exp(f_{MLP}(x_j) \cdot y + \gamma_j)} \tag{2}$$

where $\{\gamma_i\}_{i=1}^N$ are i.i.d. noise terms sampled from a Gumbel(0, 1) distribution. We obtain an approximate one-hot assignment vector p^* through the Top-K operation and Straight-Through gradient estimation [56], selecting K patch features $\{\hat{x}_i\}_{i=1}^K$ that represent the brain's attentional focus.

The training objective for E_f^{ASD} at this stage is to minimize two losses to ensure the semantic focus of y:

Adaptive Alignment Loss: Compels the fMRI representation y to approximate the mean of these K selected patch features:

$$\mathcal{L}_{adaptive} = ||y - \frac{1}{K}\sum_{i=1}^{N} p_i^* \cdot f_{MLP}(x_i)||_2^2 \tag{3}$$

Pre-alignment Loss: Ensures that y maintains basic cross-modal alignment with the global semantic feature $x_{[CLS]}$ of the image:

$$\mathcal{L}_{prealign} = ||E_f^{ASD}(X_{sem}) - x_{[CLS]}||_2^2 \tag{4}$$

Text Reconstruction. The second component of the ASD is an autoregressive language decoder D_{sem} (e.g., GPT-2 [45]). This decoder integrates the purified fMRI representation y via a trainable multi-head cross-attention layer F_{CA} and generates the text sequence $\mathcal{T}_{direct}$. Its loss function $\mathcal{L}_{sem}$ is the standard Negative Log-Likelihood (NLL) Loss:

$$\mathcal{L}_{sem} = -\mathbb{E}_{X,\mathcal{T}\sim\mathcal{D}} \sum_{i=1}^{M} \log P(t_i | t_{<i}, F_{CA}(y)) \tag{5}$$

where t_i is the i-th token in the sequence, and M is length of the text sequence. The $\mathcal{T}_{direct}$ generated by this module is faithful to the subject's subjective focus and will be injected as the core Text Prompt during the final video synthesis stage, effectively intervening to mitigate the semantic drift caused by relying solely on keyframes.

3.3 Dorsal Stream Enhancement

Conventional fMRI-to-video reconstruction methods typically rely on the direct fitting of 0/1 binary segmentation masks for spatial localization tasks, which limits the model's understanding of fine spatial relationships and often leads to unstable optimization. To overcome this limitation and adhere to the function of the Dorsal Stream (the "Where" stream) in spatial localization, we introduce a novel Spatial Perception Loss ($\mathcal{L}_{spat}$). This loss is designed to provide fine-grained spatial distribution details during the training of the CortexVideo Main Backbone Encoder (E_f), significantly enhancing the model's capacity for positional understanding.

Soft Heatmap Generation. The core of $\mathcal{L}_{spat}$ is the use of a continuous, semantic-spatial fused "soft heatmap" as the supervision signal. We first

construct a Subject Dictionary by processing all keyframes in the training set offline, which is used to generate this ground truth H_{GT}.

For each frame image I in the video sequence, we first perform Key Subject Identification by following the rule-based key object discovery mechanism of NEURONS [59]. Specifically, primary objects detected by Qwen are mapped to a compact WordNet-based concept dictionary, and the dominant subject names in each frame are selected as textual prompts for grounding(details are illustrated in Appendix). Second, for Mask Generation, we process each frame using Grounded SAM [47], taking the identified key subject names as textual prompts. This yields the corresponding binary mask $M \in \{0,1\}^P$, where P is the number of image patches covered by the key objects. Third, for Feature Extraction and Subject Feature Calculation, we utilize DINO-V3 [53] to extract all image patch features $\Phi_{patches} \in \mathbb{R}^{P \times D_v}$. We then compute a weighted average of the features selected by the mask M to derive an average feature vector $F_{subject}$ that represents the image's core subject:

$$F_{subject} = \frac{\sum_{j=1}^{P} M_j \cdot \Phi_{patches}^{(j)}}{\sum_{j=1}^{P} M_j} \tag{6}$$

Finally, our supervision signal H_{GT} is defined as the Cosine Similarity (Feature Similarity) between $F_{subject}$ and all image patch features $\Phi_{patches}^{(k)}$. This similarity is spatially reorganized to form a smooth, continuous semantic-spatial soft heatmap H_{GT}:

$$H_{GT}^{(k)} = \text{sim}(F_{subject}, \Phi_{patches}^{(k)}) = \frac{F_{subject} \cdot \Phi_{patches}^{(k)}}{||F_{subject}|| \cdot ||\Phi_{patches}^{(k)}||}$$

The magnitude of H_{GT} values not only encodes spatial location but also reflects the semantic proximity of that location to the core subject.

Spatial Perception Loss and Regularization. $\mathcal{L}_{spat}$ serves as a spatial regularization loss applied to the features output by the CortexVideo main backbone encoder E_f. We train a lightweight decoder head D_{spat} to predict the heatmap H_{pred}. The objective of $\mathcal{L}_{spat}$ is to fit the model's prediction H_{pred} to H_{GT} using the Mean Squared Error (MSE) loss function:

$$\mathcal{L}_{spat} = \mathbb{E}_{X,I}||\sigma(H_{pred}) - H_{GT}||_2^2 \tag{7}$$

where $\sigma(\cdot)$ is the Sigmoid activation function used to normalize the predicted values to the range $[0,1]$.

By minimizing $\mathcal{L}_{spat}$, we compel the main backbone encoder E_f to learn precise, continuous spatial representations concurrently with visual and motion information. This stable regression objective significantly enhances the brain model's understanding of spatial relationships, thereby indirectly improving the quality and accuracy of the Key_object_mask generated by the key object segmentation module $\mathcal{D}_{vs}$.

3.4 Training for Dual-Anchor Prior Generation

Our training procedure is decoupled into two independent stages, targeting the two dedicated encoders: the Semantic Encoder (E_f^{ASD}) and the Main Backbone Encoder (E_f).

Training the Semantic Encoder. The Semantic Encoder E_f^{ASD} and its associated decoders (f_{MLP} and D_{sem}) are trained to generate the high-fidelity Semantic Anchor $\mathcal{T}_{\text{direct}}$, reflecting the Ventral Stream's function. The training objective for this module is solely defined by the combined loss functions introduced in Sec. 3.2:

$$\mathcal{L}_{\text{ASD}} = w_{\text{adapt}}\mathcal{L}_{\text{adaptive}} + w_{\text{prealign}}\mathcal{L}_{\text{prealign}} + w_{\text{sem}}\mathcal{L}_{\text{sem}} \tag{8}$$

where $\mathcal{L}_{\text{adaptive}}$ enforces alignment with the subject's attentional focus, $\mathcal{L}_{\text{prealign}}$ ensures basic cross-modal alignment, and $\mathcal{L}_{\text{sem}}$ is the standard NLL loss for text generation.

Training the Main Backbone Encoder. The Main Backbone Encoder E_f is trained using a two-stage process to generate the full suite of visual, motion, and the Spatial Anchor priors.

Pre-alignment. We use the pre-trained MindEyeV2 [52] architecture as the foundation for our Main Backbone Encoder(details are illustrated in Appendix) E_f and use the same Motion Projector $P_{vid}(\cdot)$ as in NEURONS [59] to map image embeddings $e_i \in \mathbb{R}^{B\times N\times C}$ into a spatio-temporal embedding $e_v \in \mathbb{R}^{B\times F\times N\times C}$. This stage focuses on establishing a robust foundation for cross-modal and spatial understanding. The model is initially trained using a combination of contrastive losses($\mathcal{L}_{CLIPv}$, $\mathcal{L}_{CLIPt}$)(details are illustrated in Appendix) and the spatial regularization loss (detailed in Sec. 3.3). The $\mathcal{L}_{spat}$ is applied continuously from this stage onward to enforce the learning of fine-grained positional understanding.

Enhanced Task Joint Training. In this stage, the backbone encoder E_f continues training jointly with all its auxiliary decoders. Crucially, the Spatial Perception Loss ($\mathcal{L}_{spat}$), detailed in Sec. 3.3, is continuously applied in this stage to enforce fine-grained spatial regularization, helping the model learn spatial perception more finely. The overall training objective $\mathcal{L}_{\text{total}}$ for the Main Backbone Encoder E_f is the weighted sum of all task losses applied in this stage. The total loss $\mathcal{L}_{\text{total}}$ includes the continuous alignment and spatial losses, along with the auxiliary task losses:

$$\begin{aligned}\mathcal{L}_{\text{total}} = &(w_{\text{CLIPv}}\mathcal{L}_{\text{CLIPv}} + w_{\text{CLIPt}}\mathcal{L}_{\text{CLIPt}}) \\ &+ w_{\text{spat}}\mathcal{L}_{\text{spat}} + w_{\text{prior}}\mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{aux}}\end{aligned} \tag{9}$$

where the auxiliary loss $\mathcal{L}_{\text{aux}}$ is defined as:

$$\mathcal{L}_{\text{aux}} = w_{\text{seg}}\mathcal{L}_{\text{seg}} + w_{\text{cat}}\mathcal{L}_{\text{cat}} + w_{\text{mot}}\mathcal{L}_{\text{mot}} + w_{\text{txt}}\mathcal{L}_{\text{txt}} \tag{10}$$

The specific formulations for the Contrastive Losses ($\mathcal{L}_{CLIPv}$, $\mathcal{L}_{CLIPt}$) and all the Conventional Auxiliary Losses ($\mathcal{L}_{\text{prior}}$, $\mathcal{L}_{\text{seg}}$, $\mathcal{L}_{\text{cat}}$, $\mathcal{L}_{\text{mot}}$, $\mathcal{L}_{\text{txt}}$) are consistent with prior fMRI-to-video reconstruction works [59]. All detailed loss formulations are provided in Appendix.

3.5 Video Inference

Following NeuroClips [20], we perform inference with a pretrained T2V diffusion model [22], guided by a control image, a blurred video stream, and textual descriptions. As shown in Fig. 2, the keyframe is decoded via unCLIP [46] and the high-level semantic guidance comes from the ventral stream. We decode the blurred video in the same manner as NEURONS [59], and to emphasize key objects, we rescale the predicted subject masks output by the dorsal-stream to [0.5, 1] and multiply them with both the control image and the blurred video to ensure their prominence.

4 Experiments

4.1 Dataset

We validate the fMRI-to-video reconstruction on the open-source cc2017 dataset [60]. This dataset contains MRI (T1 and T2-weighted) and fMRI data acquired using a 3-T MRI device, with a temporal resolution of 2 seconds.The dataset is divided into training and testing partitions, consisting of 18×8 minute and 5×8 minute video segments, respectively.For each subject, fMRI responses were acquired with a repetition presentation of 2 trials for the training segments and 10 trials for the testing segments. The results for the testing portion were obtained by trial-averaging. This yields a final paired fMRI-video sample size of 8,640 for training and 1,200 for testing.

4.2 Evaluation Metrics

The quantitative analysis of this study focuses on video-level metrics, covering two categories: Semantic and Spatio-Temporal measures. Semantic metrics are implemented by testing classification across the 400 classes of the Kinetics-400 dataset [29], utilizing a [54]-based video classifier. Spatio-temporal consistency metrics rely on the average cosine similarity of adjacent frame pairs, formed by extracting CLIP image embeddings from every frame of the predicted video. This metric is also known as CLIP-pcc and is widely adopted in video editing tasks [61].

Guidance Signal Quality Metrics are used to validate the superiority of our dual-stream design: The quality of the ASD-decoded $\mathcal{T}_{\text{direct}}$ is evaluated against GT text using standard NLP metrics, including BLEU-1 to 4 [43], ROUGE [33], and CIDEr [57]. The performance of the E_f-driven Key Object Mask prediction is quantified using standard segmentation metrics such as IoU and Dice. These metrics are computed by evaluating the mask frames with the highest scores for the key subject in each video segment, comparing them with the corresponding GT masks. High IoU and Dice scores confirm that the model generates precise spatial constraints, directly supporting the enhanced subject localization used in the final synthesis stage.

4.3 Comparison with State-of-the-Art

In this section, we compare CortexVideo with previous methods such as NEURONS [59], NeuroClips [20], and MindAnimator [35], focusing on high-level semantics and spatial perception.

Qualitative Analysis. First, we present qualitative results to demonstrate the superiority of CortexVideo in video generation. As shown in Fig. 3, qualitative analysis reveals that CortexVideo, through the design of reliable semantic guidance, can correct semantic errors when key frame content is incorrect. Additionally, by simulating the brain's spatial perception, it tends to restore the subject's position during decoding. The combination of these two factors enables a more accurate reconstruction of dynamic visual stimuli as perceived by the brain. In contrast, other methods struggle to simultaneously restore both semantic and spatial perception, highlighting that the use of dual streams in decoding significantly enhances reconstruction quality.

Fig. 3 illustrates reconstruction results and several subject mask predictions. While previous models often suffer from semantic drift when keyframes are blurred, CortexVideo achieves faithful restoration. As shown in Fig. 3, Our method precisely captures the quantity and positional relationships of the two closely spaced dogs, and accurately predicts both the semantics and trajectory of the turtle. Furthermore, our model effectively restores the structural topology of street scenes and landscape. Notably, the subject mask predictions generated by CortexVideo are more precise; they are neither too small (as observed in the diver example) nor overly divergent (as seen in the truck example). Instead, they capture the subject's extent as maximally as possible, providing a robust spatial foundation for the subsequent video generation process. More visualizations can be found in Appendix.

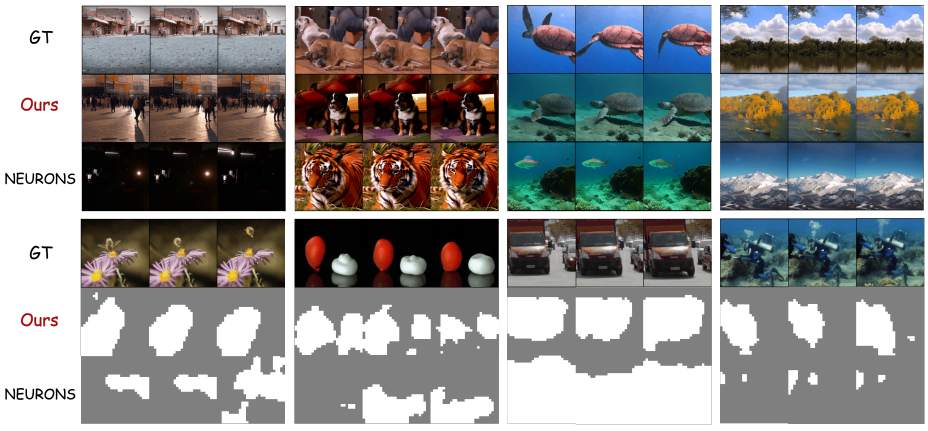

Fig. 3: Qualitative comparison between CortexVideo and previous SOTA method, NEURONS [59]. The upper rows show video reconstruction samples. The lower rows compare reconstructed subject masks. See Appendix for additional results.

Quantitative Results. Table 1 presents the semantic evaluation guiding video generation and the subject mask evaluation impacting spatial effects. Table 2 compares CortexVideo with state-of-the-art methods based on video-level metrics.

Table 1: Quantitative comparison on Caption Similarity Metrics and Grounding Metrics using the cc2017 dataset. The best and second-best performance are indicated by **bold** font and underline text, respectively. Results are reported for NeuroClips and NEURONS, with CortexVideo. The results for NeuroClips, and NEURONS are averaged across all three subjects.

Method	Dataset	Caption Similarity Metrics						Grounding Metrics	
		B@1	B@2	B@3	B@4	ROUGE	CIDEr	IoU	Dice
NeuroClips [20]	cc2017 dataset [60]	0.195	0.093	0.053	0.032	0.233	0.142	-	-
NEURONS [59]	cc2017 dataset [60]	0.220	0.092	0.044	0.025	0.227	0.174	0.353	0.474
CortexVideo	cc2017 dataset [60]	**0.247**	**0.111**	**0.057**	**0.033**	**0.236**	**0.240**	**0.362**	**0.485**
subject 1	cc2017 dataset [60]	0.243	0.111	0.059	0.036	0.231	0.244	0.370	0.493
subject 2	cc2017 dataset [60]	0.247	0.109	0.053	0.030	0.238	0.229	0.360	0.481
subject 3	cc2017 dataset [60]	0.250	0.112	0.058	0.034	0.239	0.246	0.357	0.481

CortexVideo outperforms previous methods in semantic captioning, achieving state-of-the-art performance across all metrics. For spatial mask metrics, CortexVideo achieves an IoU of 0.362 and a Dice score of 0.485, surpassing NEURONS. These results demonstrate superior semantic understanding and spatial localization, critical for high-quality fMRI-to-video reconstruction.

Table 2: Quantitative comparison of video reconstruction performance between CortexVideo and state-of-the-art methods. Baseline results are cited from [10, 20, 35, 59].

Method	Semantic-level		ST-level
	2-way ↑	50-way ↑	CLIP-pcc ↑
Wen [60]	-	$0.166_{\pm 0.02}$	-
Wang [58]	$0.773_{\pm 0.03}$	-	$0.402_{\pm 0.41}$
Kupershmidt [30]	$0.771_{\pm 0.03}$	-	$0.386_{\pm 0.47}$
MinD-Video [10]	$0.839_{\pm 0.03}$	$0.197_{\pm 0.02}$	$0.408_{\pm 0.46}$
MindAnimator [35]	$0.830_{\pm 0.03}$	-	$0.428_{\pm 0.03}$
NeuroClips [20]	$0.834_{\pm 0.03}$	$0.220_{\pm 0.01}$	$0.738_{\pm 0.17}$
NEURONS [59]	$0.863_{\pm 0.03}$	$0.262_{\pm 0.01}$	$\mathbf{0.934}_{\pm 0.17}$
CortexVideo	$\mathbf{0.874}_{\pm 0.02}$	$\mathbf{0.275}_{\pm 0.01}$	$0.926_{\pm 0.04}$
subject 1	$0.867_{\pm 0.02}$	$0.262_{\pm 0.02}$	$0.926_{\pm 0.04}$
subject 2	$0.879_{\pm 0.02}$	$0.282_{\pm 0.02}$	$0.928_{\pm 0.04}$
subject 3	$0.875_{\pm 0.02}$	$0.280_{\pm 0.02}$	$0.923_{\pm 0.04}$

In video-level evaluation, CortexVideo outperforms NEURONS, NeuroClips, and MindAnimator. With a 2-way accuracy of 0.874, it surpasses NEURONS (0.863) and MinD-Video (0.839). Additionally, CortexVideo leads in 50-way classification with 0.275, outperforming NEURONS by 5.0%.For spatio-temporal consistency, NEURONS scores 0.934 on CLIP-pcc, while CortexVideo follows closely with 0.926, indicating strong temporal consistency.

Due to our direct use of fMRI-derived semantics to mitigate the semantic drift caused by relying solely on key frames, when the derived semantics conflict with the key frame content, it may cause slight blurring that affects the CLIP-

pcc score. Despite this, we believe that a semantically correct but slightly blurred video is preferable to a temporally coherent video relying on incorrect key frames. This ensures the video remains faithful to the true fMRI data, even with minor spatio-temporal trade-offs.

4.4 Analysis of the dual-guidance decoding strategy

To better understand the distinct roles and contributions of the ventral stream and dorsal stream in CortexVideo, as well as their synergistic decoding when used together, we performed ablation studies by decoupling these dual-streams in the model. The qualitative results are shown in Fig. 4, and the quantitative results are shown in Table 3, with all ablation results reported for Subject 2.

As shown in Fig. 4, we tested the impact of disabling the ventral stream enhancement and dorsal stream enhancement on video generation. The results indicate that when the ventral stream enhancement is disabled, the model can still accurately recover the spatial positions of subjects. However, semantic information becomes heavily reliant on keyframe captions, leading to semantic drift (e.g., errors in keyframes exacerbate semantic inconsistencies, such as generating "man and a bird are standing next to each other"). On the other hand, when the dorsal stream enhancement is disabled, although the model retains correct semantic guidance, spatial recovery is compromised, resulting in distortions (e.g., despite generating the beach and subjects, there is ambiguity in the position and quantity of the objects).

By utilizing both the ventral stream and dorsal stream, CortexVideo effectively combines semantic and spatial information, resolving these issues and generating more accurate and consistent video results. As shown in the qualitative results, when using the full CortexVideo framework, the generated video not only maintains semantic consistency but also accurately restores spatial positions, resulting in a more natural and precise video reconstruction.

4.5 Ablation Studies

In this section, we quantitatively evaluate the dual-stream structure of CortexVideo through ablation studies, focusing on the contributions of semantic guidance and spatial regularization. Specifically, we examine the impact of removing these components on the model's performance.

As shown in Table 3, we compare the performance of CortexVideo with two variants: one without ventral stream enhancement and one without dorsal stream enhancement. The results indicate that removing semantic guidance leads to a slight reduction in 2-way accuracy (from

Table 3: Individual contributions of each stream. **VSE** and **DSE** denote Ventral and Dorsal Stream Enhancement, respectively.

Method	**VSE**	**DSE**	2-way ↑	50-way ↑	B@1 ↑	Dice ↑
NEURONS	✗	✗	0.860	0.252	0.222	0.472
w/o **VSE**	✗	✓	0.868	0.266	0.222	**0.481**
w/o **DSE**	✓	✗	0.870	0.269	**0.247**	0.472
CortexVideo	✓	✓	**0.879**	**0.282**	**0.247**	**0.481**

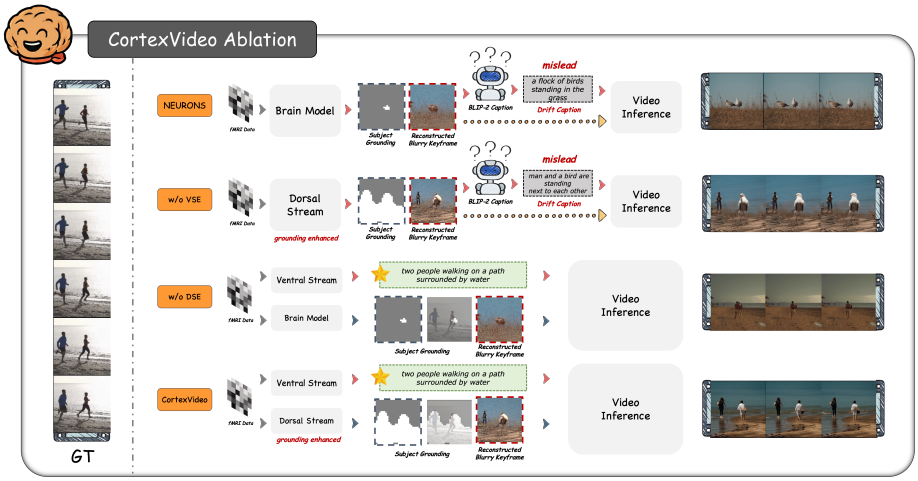

Fig. 4: Ablation study on CortexVideo, showing the individual contribution of the ventral and dorsal stream. The blue solid boxes indicate the predicted subject masks.

0.879 to 0.868) and 50-way accuracy (from 0.282 to 0.266). Similarly, when spatial regularization is removed, we observe a small drop in 2-way accuracy (from 0.879 to 0.870) and Dice score (from 0.481 to 0.472).

In contrast, when both components are included, CortexVideo achieves the highest performance in all metrics, demonstrating that both semantic guidance and spatial regularization contribute significantly to the model's ability to generate high-quality reconstructions. Specifically, CortexVideo outperforms the ablated versions in terms of 50-way accuracy, B@1, and Dice score, further confirming the importance of these two components in improving video generation.

These findings validate the effectiveness of the dual-stream design, where semantic guidance ensures accurate semantic content and spatial regularization improves the consistency and precision of spatial localization.

5 Conclusion

We propose the CortexVideo framework, which achieves high-fidelity video reconstruction from fMRI signals through an innovative Semantic-Spatial Dual-Anchoring strategy. This framework functionally simulates the human visual system's Dual-Streams. For the ventral stream, the Attentive Semantic Decoder dynamically captures specific semantic patterns in brain activity to extract more credible semantic descriptions. For the dorsal stream, the Spatial Regulation significantly improves the quality of key object segmentation masks, indirectly providing the generative model with more precise spatial constraints during the final phase. CortexVideo significantly improves semantic guidance and key object mask prediction, outperforming previous methods across all video-based semantic evaluation metrics. By strengthening the semantic and spatial information

flow during the fMRI decoding process, CortexVideo enhances reconstruction performance and offers a novel design paradigm for building high-fidelity visual decoding models.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 624B2127.

References

1. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
2. Chen, J., Qi, Y., Pan, G.: Rethinking visual reconstruction: Experience-based content completion guided by visual cues. In: International conference on machine learning. pp. 4856–4866. PMLR (2023)
3. Chen, J., Qi, Y., Wang, Y., Pan, G.: Bridging the semantic latent space between brain and machine: Similarity is all you need. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 11302–11310 (2024)
4. Chen, J., Qi, Y., Wang, Y., Pan, G.: Mind artist: Creating artistic snapshots with human thought. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27207–27217 (2024)
5. Chen, J., Qi, Y., Wang, Y., Pan, G.: Bridging the gap between brain and machine in interpreting visual semantics: Towards self-adaptive brain-to-text decoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21938–21948 (2025)
6. Chen, J., Qi, Y., Wang, Y., Pan, G.: Mindgpt: Interpreting what you see with non-invasive brain recordings. IEEE Transactions on Image Processing (2025)
7. Chen, J., Qi, Y., Wang, Y., Pan, G.: Human-like cognitive generalization for large models via mental representation-guided supervision. Nature Communications (2026)
8. Chen, J., Qi, Y., Wang, Y., Pan, G.: Reducing semantic mismatch in brain-to-text decoding through personalized multimodal masking. In: The Fourteenth International Conference on Learning Representations (2026)
9. Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: 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. pp. 22710–22720 (2023)
10. Chen, Z., Qing, J., Zhou, J.H.: Cinematic mindscapes: High-quality video reconstruction from brain activity. Advances in Neural Information Processing Systems **36**, 24841–24858 (2023)
11. Cloutman, L.L.: Interaction between dorsal and ventral processing streams: where, when and how? Brain and language **127**(2), 251–263 (2013)
12. Desimone, R., Duncan, J., et al.: Neural mechanisms of selective visual attention. Annual review of neuroscience **18**(1), 193–222 (1995)

13. Du, C., Zhou, Q., Liu, C., He, H.: Review of visual neural encoding and decoding methods in fmri(in chinese). Journal of Image and Graphics **28**(2), 372–384 (2023)
14. Epstein, R., Harris, A., Stanley, D., Kanwisher, N.: The parahippocampal place area: recognition, navigation, or encoding? Neuron **23**(1), 115–125 (1999)
15. Fang, T., Qi, Y., Pan, G.: Reconstructing perceptive images from brain activity by shape-semantic gan. Advances in Neural Information Processing Systems **33**, 13038–13048 (2020)
16. Fang, T., Zheng, Q., Pan, G.: Alleviating the semantic gap for generalized fmri-to-image reconstruction. Advances in Neural Information Processing Systems **36**, 15096–15107 (2023)
17. Freud, E., Plaut, D.C., Behrmann, M.: 'what'is happening in the dorsal visual pathway. Trends in cognitive sciences **20**(10), 773–784 (2016)
18. Gao, J., Fu, Y., Fu, Y., Wang, Y., Qian, X., Feng, J.: Mind-3d++: Advancing fmri-based 3d reconstruction with high-quality textured mesh generation and a comprehensive dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
19. Gao, J., Liu, Y., Yang, B., Feng, J., Fu, Y.: Cinebrain: A large-scale multi-modal audiovisual brain dataset for brain-conditioned video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 36224–36234 (2026)
20. Gong, Z., Bao, G., Zhang, Q., Wan, Z., Miao, D., Wang, S., Zhu, L., Wang, C., Xu, R., Hu, L., et al.: Neuroclips: Towards high-fidelity and smooth fmri-to-video reconstruction. Advances in Neural Information Processing Systems **37**, 51655–51683 (2024)
21. Goodale, M.A., Milner, A.D.: Separate visual pathways for perception and action. Trends in neurosciences **15**(1), 20–25 (1992)
22. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
23. Han, K., Wen, H., Shi, J., Lu, K.H., Zhang, Y., Fu, D., Liu, Z.: Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. NeuroImage **198**, 125–136 (2019)
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
25. Horikawa, T., Kamitani, Y.: Generic decoding of seen and imagined objects using hierarchical visual features. Nature communications **8**(1), 15037 (2017)
26. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016)
27. Kamps, F.S., Julian, J.B., Kubilius, J., Kanwisher, N., Dilks, D.D.: The occipital place area represents the local elements of scenes. Neuroimage **132**, 417–424 (2016)
28. Kanwisher, N., McDermott, J., Chun, M.M.: The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of neuroscience **17**(11), 4302–4311 (1997)
29. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
30. Kupershmidt, G., Beliy, R., Gaziv, G., Irani, M.: A penny for your (visual) thoughts: Self-supervised reconstruction of natural movies from brain activity. arXiv preprint arXiv:2206.03544 (2022)
31. Li, J., Qi, Y., Pan, G.: Phase-amplitude coupling-based adaptive filters for neural signal decoding. Frontiers in Neuroscience **17**, 1153568 (2023)

32. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
33. Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics. pp. 150–157 (2003)
34. Linsley, D., MacEvoy, S.P.: Encoding-stage crosstalk between object-and spatial property-based scene processing pathways. Cerebral Cortex **25**(8), 2267–2281 (2015)
35. Lu, Y., Du, C., Wang, C., Zhu, X., Jiang, L., Li, X., He, H.: Animate your thoughts: Decoupled reconstruction of dynamic natural vision from slow brain activity. arXiv preprint arXiv:2405.03280 (2024)
36. Milner, D., Goodale, M.: The visual brain in action, vol. 27. Oup Oxford (2006)
37. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
38. Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L.: Encoding and decoding in fmri. Neuroimage **56**(2), 400–410 (2011)
39. Nasr, S., Devaney, K.J., Tootell, R.B.: Spatial encoding and underlying circuitry in scene-selective cortex. Neuroimage **83**, 892–900 (2013)
40. Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L.: Reconstructing visual experiences from brain activity evoked by natural movies. Current biology **21**(19), 1641–1646 (2011)
41. Palazzo, S., Spampinato, C., Kavasidis, I., Giordano, D., Schmidt, J., Shah, M.: Decoding brain representations by multimodal learning of neural activity and visual features. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(11), 3833–3849 (2020)
42. Pan, G., Li, J.J., Qi, Y., Yu, H., Zhu, J.M., Zheng, X.X., Wang, Y.M., Zhang, S.M.: Rapid decoding of hand gestures in electrocorticography using recurrent neural networks. Frontiers in neuroscience **12**, 555 (2018)
43. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
45. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
46. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
47. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
48. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

49. Saleem, A.B., Diamanti, E.M., Fournier, J., Harris, K.D., Carandini, M.: Coherent encoding of subjective spatial position in visual cortex and hippocampus. Nature **562**(7725), 124–127 (2018)
50. Sathian, K., Lacey, S., Stilla, R., Gibson, G.O., Deshpande, G., Hu, X., LaConte, S., Glielmi, C.: Dual pathways for haptic and visual perception of spatial and texture information. Neuroimage **57**(2), 462–475 (2011)
51. Scotti, P., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Dempster, A., Verlinde, N., Yundler, E., Weisberg, D., Norman, K., 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)
52. Scotti, P.S., Tripathy, M., Villanueva, C.K.T., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K.A., et al.: Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. arXiv preprint arXiv:2403.11207 (2024)
53. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
54. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)
55. Ungerleider, L.G.: Two cortical visual systems. Analysis of visual behavior **549**, chapter–18 (1982)
56. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
57. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
58. Wang, C., Yan, H., Huang, W., Li, J., Wang, Y., Fan, Y.S., Sheng, W., Liu, T., Li, R., Chen, H.: Reconstructing rapid natural vision with fmri-conditional video generative adversarial network. Cerebral Cortex **32**(20), 4502–4511 (2022)
59. Wang, H., Zhang, Q., Wang, L., Huang, X., Li, X.: Neurons: Emulating the human visual cortex improves fidelity and interpretability in fmri-to-video reconstruction. arXiv preprint arXiv:2503.11167 (2025)
60. Wen, H., Shi, J., Zhang, Y., Lu, K.H., Cao, J., Liu, Z.: Neural encoding and decoding with deep learning for dynamic natural vision. Cerebral cortex **28**(12), 4136–4160 (2018)
61. Wu, J.Z., Li, X., Gao, D., Dong, Z., Bai, J., Singh, A., Xiang, X., Li, Y., Huang, Z., Sun, Y., et al.: Cvpr 2023 text guided video editing competition. arXiv preprint arXiv:2310.16003 (2023)
62. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)