MINDSHOT: MULTI-SHOT VIDEO RECONSTRUCTION FROM FMRI WITH LLM DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing dynamic videos from fMRI is important for understanding visual cognition and enabling vivid brain-computer interfaces. However, current methods are critically limited to single-shot clips with video-level alignment and reconstruction, failing to address the multi-shot nature of real-world experiences. To bridge this gap, we propose MindShot, a novel shot-level framework that effectively reconstructs multi-shot videos from fMRI via a divide-and-decode strategy. Specifically, our framework consists of three stages: (1) Shot Decomposition: We first identify shot boundaries within fMRI, then decompose the mixed signals into distinct, shot-specific segments. (2) Keyframe Decoding: Each segment is decoded into a textual description representing the keyframe of its corresponding shot. (3) Video Reconstruction: The final video is generated from these keyframe captions, effectively mitigating noise from fMRI redundancy. Addressing the lack of data for multi-shot reconstruction, we construct a large-scale multi-shot fMRI-video dataset, synthesized from existing datasets. Experimental results demonstrate our framework outperforms state-of-the-art methods in both single-shot and multi-shot reconstruction fidelity. Ablation studies confirm the critical role of shot-level reconstruction in multi-shot video reconstruction, with decomposition significantly improving decoded caption CLIP similarity by 71.8%. This work establishes a new paradigm for multi-shot fMRI reconstruction, enabling accurate recovery of complex visual narratives through explicit decomposition and semantic prompting.

1 INTRODUCTION

Functional magnetic resonance imaging (fMRI) is a powerful, non-invasive tool for studying the human brain, particularly the visual system, through indirect measurement of neural activity(Horikawa & Kamitani, 2017). Reconstructing dynamic visual sequences from fMRI data is critical not only for advancing our understanding of dynamic visual perception and cognition, but also for developing next-generation brain-computer interfaces (BCIs) capable of more vivid and dynamic "mind-reading" applications (Wen et al., 2018; Fang et al., 2020; 2023). However, existing video reconstruction research mainly focuses on short-duration, single-shot videos (depicting a single, continuous scene or event) (Sun et al., 2025; Chen et al., 2023; Li et al., 2024; Lu et al., 2025), ignoring the multi-shot visual experiences that characterize real-world cognition, such as watching films or recalling episodic memories.

Reconstructing multi-shot video presents substantial challenges beyond single-shot reconstruction, especially for existing video-level paradigm. Whether aggregating signals over longer sequences or decoding short, fixed-length clips, these approaches attempt to reconstruct entire video clips from corresponding brain activity segments. This paradigm, however, critically neglects that natural videos are often composed of multiple semantic events. Consequently, when the fMRI signals correspond to multiple scenes, the video-level approach leads to temporal mixing of semantically distinct neural patterns. This not only confines current methods to effectively handling only single-shot videos but also introduces significant ambiguity and noise into the reconstructions, making it challenging to disentangle and accurately reconstruct the separate visual events (Figure 1).

To address this limitation, we propose a paradigm shift from video-level to shot-level reconstruction. Rather than reconstructing from temporally mixed neural representations, our framework adopts the

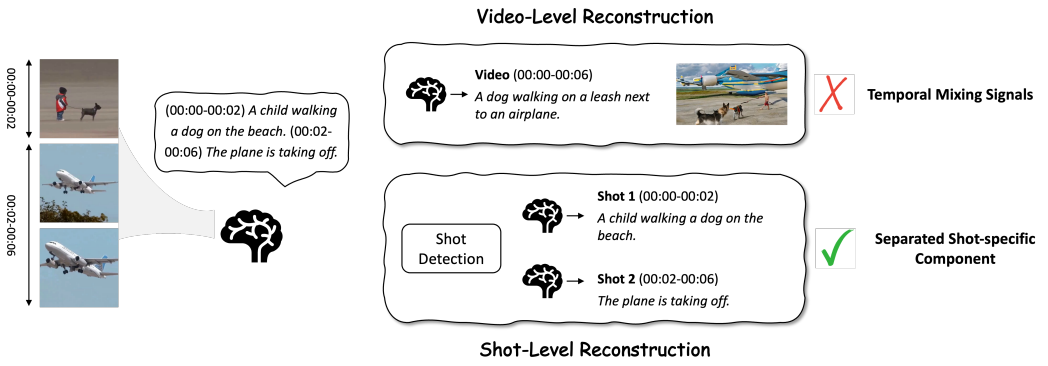

Figure 1: Illustration of the key innovation of our proposed shot-level paradigm. The conventional video-level framework decodes from temporally mixed fMRI signals, leading to semantically entangled results. In contrast, our shot-level paradigm explicitly decomposes the signals into shot-specific components before decoding, enabling clean and coherent reconstruction of each shot.

shot as the basic decoding unit. Specifically, we segment the continuous fMRI time series into shot-specific segments that align with corresponding video shots. This segmentation yields clean, semantically coherent neural signals, each corresponding to a single visual context. By grounding the reconstruction process in shot-level neural representations, our approach effectively mitigates the temporal mixing problem and enables accurate reconstruction of complex, multi-shot videos.

Based on this shot-level paradigm, we propose MindShot, a novel divide-and-decode framework for multi-shot video reconstruction. Inspired by text-to-video generation paradigms where a narrative is segmented into prompts corresponding to individual shots (Zhao et al., 2024; Wu et al., 2025), we introduce a shot boundary predictor for fMRI segmentation. Instead of aggregating the entire fMRI data, this shot boundary predictor learns to segment it into shot-specific components corresponding to individual shots, enabling explicit and independent reconstruction of each shot. To address the limitations imposed by the fMRI-video temporal resolution mismatch, we propose to decode keyframe captions from fMRI data to achieve semantically precise reconstruction. For each segmented shot-specific fMRI, we decode a textual caption describing the keyframe using Large Language Models (LLMs). This leverages the observation that humans remember salient events at a semantic level, which is more robust to temporal blurring. The decoded caption then provides a precise semantic prompt for the subsequent video generation stage. To overcome data scarcity, we develop novel synthesis strategies to construct a large-scale multi-shot fMRI-video dataset. Leveraging existing publicly available fMRI-video datasets, including the benchmark CC2017 (Wen et al., 2018) and the dataset by (Chen et al., 2023), we synthesize 20k sample pairs for each dataset, enabling effective training of our proposed model. Our contributions in this work can be summarized as follows:

- We introduce a new shot-level paradigm for fMRI-video reconstruction by establishing the shot as the fundamental unit of decoding. This shot-level paradigm enables the explicit reconstruction of complex, multi-shot videos for the first time.
- We design a learnable shot boundary predictor that automatically segments fMRI time series into shot-specific components, effectively mitigating the temporal mixing problem without manual intervention.
- We develop novel synthesis strategies to create large-scale multi-shot training data from existing datasets, facilitating model development for multi-shot video reconstruction.

2 Related Work

2.1 fMRI-to-Image Reconstruction

Benefiting from large-scale datasets like the Natural Scenes Dataset (NSD) (Scotti et al., 2023), generative vision models conditioned on fMRI signals have demonstrated unprecedented performance

in reconstructing static images from brain responses. Existing research primarily focuses on enhancing reconstruction fidelity through improved semantic alignment, such as contrastive learning techniques that align fMRI embeddings with image or text representations (Xia et al., 2024), or by incorporating low-level image features to preserve visual detail consistency (Wang et al., 2024). Additional efforts have developed subject-unified methods to address cross-subject alignment and model generalization (Scotti et al., 2024). Despite significant progress, reconstructing dynamic video sequences presents substantially greater challenges than static images.

2.2 fMRI-to-Video Reconstruction

As a pioneering work of fMRI-to-video reconstruction, MindVideo (Chen et al., 2023) achieves notable fidelity by aligning fMRI features to CLIP (Radford et al., 2021) space for latent diffusion model prompting. Subsequent studies enhance temporal modeling in fMRI encoders (Sun et al., 2025) or explore cross-subject alignment via fMRI projection (Li et al., 2024). Crucially, most of existing methods are confined to single-shot scenarios, neglecting the multi-shot dynamics inherent in real-world cognition. While NeuroClips (Gong et al., 2024) generates multiple shots by fusing semantically similar keyframes, it relies on post-hoc processing rather than intrinsic fMRI signal decomposition, failing to optimize encoders for disentangling mixed shot information within fMRI windows. Moreover, contrastive alignment in video reconstruction may be challenging due to the temporal resolution mismatch between fMRI and video. In contrast to prior work, we propose to explore the multi-shot video reconstruction by shot-specific fMRI segmentation and keyframe caption decoding for semantically precise reconstruction, circumventing contrastive alignment constraints.

3 Method

Our method can be divided into three main stages, as shown in Figure 2. In the first stage, the shot boundary predictor partitions fMRI into shot-specific components. Each segmented fMRI is then decoded to shot-specific keyframe caption via direct interaction with a LLM. These captions serve as precise semantic prompts input to a text-to-video diffusion model for final video synthesis.

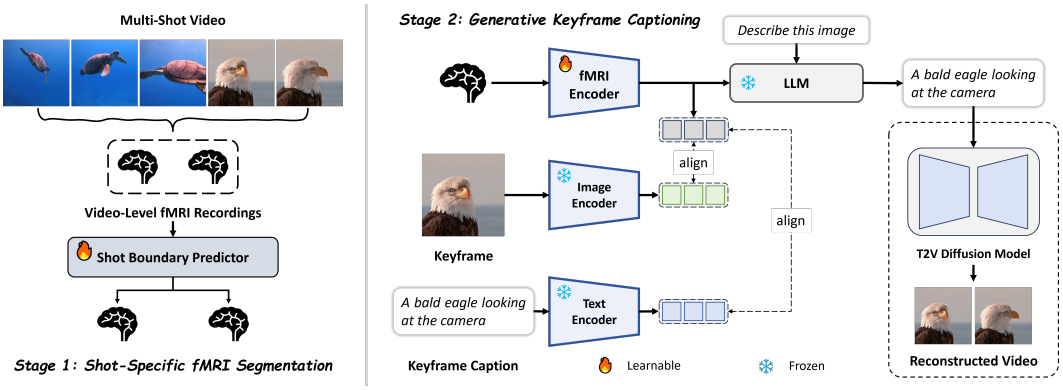

Figure 2: Overview of our proposed method, consisting of three main stages: Shot-Specific fMRI Segmentation, Generative Keyframe Captioning, and Shot-Centric Video Reconstruction.

3.1 Dataset Synthesis

Given the absence of publicly available multi-shot fMRI-video datasets, we develop novel synthesis strategies to enable effective model training. Our approach leverages two public resources: CC2017 (Wen et al., 2018) and Li et al.'s fMRI-WebVid dataset (Li et al., 2024).

Synthesis from fMRI-WebVid. Using 2,000 single-shot training clips and 400 test clips from fMRI-WebVid (each 4s video with 5 fMRI scans), we generate multi-shot videos by randomly concatenating distinct clips while preserving 4-second duration. To enhance data diversity, we vary shot duration ratios across samples, including partial-shot allocations (e.g., 2:3 fMRI frame split) and

degenerate cases reducing to single-shot. For each resulting shot, we extract the middle frame of its video segment as shot-specific keyframes and generate captions using BLIP-2 (Li et al., 2023). This yields 20,000 training and 1,000 test samples (fMRI-WebVid-Syn) with aligned fMRI sequences, video keyframes, and keyframe captions.

Synthesis from CC2017. From 1,440 training and 400 test clips in CC2017 (each 6s video with 3 fMRI scans), we synthesize two-shot videos by concatenating distinct clips while maintaining 6-second duration. The number of synthesized fMRI scans is set to 4, considering about the duration of each shot. To ensure single-shot sources, we first apply SceneSeg (Rao et al., 2020) to decompose original videos into constituent shots at scene boundaries. For each synthesized video, we randomly select two shots from different source clips, temporally cropping each according to sampled duration ratios (e.g., 3:1) before concatenation. Following the fMRI-WebVid-Syn protocol for keyframe extraction and caption generation, we produce 20,000 training and 1,000 test samples (CC2017-Syn) with aligned fMRI sequences, video keyframes, and keyframe captions.

3.2 Shot-Specific fMRI Segmentation

A primary challenge in multi-shot video reconstruction is the temporal mixing of neural signals across different shots. Achieving semantically precise reconstruction thus requires decomposing video-level fMRI into shot-specific components. The most intuitive approach is to detect shot boundaries that occur across successive fMRI volumes (TRs), transforming the problem into sequence boundary prediction for subsequent fMRI separation and aggregation. Therefore, we propose to detect boundaries by introducing a shot boundary predictor. It is important to note that due to the temporal integration inherent in fMRI signal acquisition within a single TR, resolving scene transitions that occur entirely within one TR presents a fundamentally different and more challenging problem. Consequently, this work focuses on establishing a robust framework for shot-level reconstruction by addressing inter-TR segmentation, which captures the majority of scene transitions in conventionally sampled fMRI data, while leaving the challenge of intra-TR decomposition for future research with advanced acquisition techniques.

Leveraging the fMRI encoder from (Li et al., 2024), we get fMRI embeddings $emb_f \in \mathbb{R}^{M\times c}$, where M is the number of fMRI scans, and $c = 1024$ represents the embedding dimension. Theses embeddings are then processed by our proposed shot boundary predictor, which comprises a two-layer bidirectional LSTM (Bi-LSTM) to model bidirectional temporal dependencies in fMRI signals, and a linear layer generating boundary probabilities.

Formally, given fMRI embeddings $emb_f \in \mathbb{R}^{M\times c}$, the boundary probabilities are computed as:

$$H = \text{Bi-LSTM}(emb_f) \tag{1}$$

where $H \in \mathbb{R}^{M\times d}$ are hidden states ($d = 512$), $P = WH + b \in [p_1, p_2, \ldots, p_{M-1}]$ denotes boundary probabilities and p_i represents the boundary probability of a boundary between fMRI scans i and $i + 1$.

The model is optimized via binary cross-entropy loss:

$$\mathcal{L}_{sbp} = -\frac{1}{M-1}\sum_{i=1}^{M-1}[y_i \log p_i + (1-y_i)\log(1-p_i)] \tag{2}$$

where $y_i \in \{0, 1\}$ indicates ground-truth boundaries, and the true number of shots N satisfies

$$N = 1 + \sum_{i=1}^{M-1} y_i \tag{3}$$

At inference, by binarizing p_i with a threshold τ, we get

$$o_i = \begin{cases} 1 & \text{if } p_i > \tau \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\tau = 0.5$ in this work. Using the binarized boundaries o_i, we partition the fMRI sequence into $\tilde{N}$ segments, where $\tilde{N} = 1 + \sum_{i=1}^{M-1} o_i$ is the predicted number of shots. The shot-specific embeddings are then aggregated as $emb_f^s \in \mathbb{R}^{\tilde{N}\times c}$.

3.3 GENERATIVE KEYFRAME CAPTIONING

Beyond temporal signal mixing, the fMRI-video temporal resolution mismatch makes direct reconstruction via contrastive alignment challenging. However, human cognition encodes experiences through semantic abstractions of key events rather than continuous visual streams. We therefore reformulate the task as keyframe-centered semantic reconstruction, where decoding keyframe captions bypasses strict temporal alignment requirements. Specifically, we learn to generate keyframe captions directly from shot-specific fMRI signals using an LLM, and the keyframe captions are then used for final video synthesis.

Using ground-truth shot boundaries during training, we obtain shot-specific fMRI embeddings $emb_f^s \in \mathbb{R}^{N\times c}$ and concatenate them with an instruction prompt and input into a frozen LLM. Leveraging the multimodal understanding capabilities of the LLM, the dialogue format is structured as follows: System: [system message]. User: $<$ instruction $>$ $<$ fMRI embedding $>$. Assistant: $<$ answer $>$. The tag $<$ instruction $>$ denotes natural language query, while $<$ image $>$ is a placeholder for fMRI embedding. The model generates the response $<$ answer $>$ as predicted captions. The objective for optimizing this decoding process is to minimize the text modeling loss $\mathcal{L}_{caption}$, which evaluates the ability of LLM to generate target captions from fMRI embeddings. This loss is formally defined as the negative log-likelihood of the target captions given context:

$$\mathcal{L}_{caption} = -\sum_{k=1}^{T} \log P_\theta \left(t_k | t_{<k}, I; emb_f^s\right) \tag{5}$$

where T is the length of target text, t_k is the k-th token, $t_{<k}$ represents the preceding tokens, I is the input prompt ('Describe this image $<$ image $>$' in this work), and P_θ is the token probability distribution parameterized by LLM weights θ.

We empirically found that introducing contrastive alignment and noise prediction during training can improve the final results. Given {keyframe, keyframe caption} pairs, CLIP loss is calculated for fMRI-keyframe and fMRI-caption pairs. With fixed CLIP encoders, we obtain keyframe embedding $emb_i \in \mathbb{R}^{N\times c}$ and text embedding $emb_t \in \mathbb{R}^{N\times c}$. The contrastive alignment loss is:

$$\mathcal{L}_{align} = \frac{1}{2}\left(\mathcal{L}_{CLIP}(emb_f^s, emb_i) + \mathcal{L}_{CLIP}(emb_f^s, emb_t)\right) \tag{6}$$

We also freeze the U-Net of video diffusion model for noise prediction, with MSE loss:

$$\mathcal{L}_{mse} = \mathbb{E}_{emb_f^s, \epsilon_{gt}^t \sim \mathcal{N}(0,1), t}\left[\left\|\epsilon_{gt}^t - \epsilon_{pr}^t\right\|_2^2\right] \tag{7}$$

where emb_f^s is input as condition embeddings, $\epsilon_{pr}^t = \text{U-Net}(emb_i^t, emb_f^s, t)$ is predicted noise conditioned on shot-specific fMRI, and ϵ_{gt}^t is ground-truth noise.

The overall training loss combines all components:

$$\mathcal{L} = \mathcal{L}_{sbp} + \lambda_1 \mathcal{L}_{caption} + \lambda_2 \mathcal{L}_{align} + \lambda_3 \mathcal{L}_{mse} \tag{8}$$

where λ_1, λ_2, and λ_3 are learnable parameters for automatic optimization. Only the fMRI encoder and shot boundary predictor are trained while other modules remain frozen.

3.4 SHOT-CENTRIC VIDEO RECONSTRUCTION

Following the generation of keyframe captions, we reconstruct the final video by generating each shot individually and concatenating them according to their original duration. The video generation is conditioned only on the textual captions using a frozen text-to-video diffusion model. While incorporating the original fMRI embeddings could provide supplemental multimodal information, it may introduce substantial noise, particularly given the temporal mixing inherent in multi-shot fMRI data.

4 EXPERIMENT AND RESULTS

4.1 EXPERIMENTAL SETTING

Dataset. We evaluated our method on both synthesized and original datasets, including CC2017 and fMRI-WebVid. CC2017 (Chen et al., 2023) contains three subjects with fMRI frames acquired using a 3T scanner (TR=2s), where each sample includes a 6s video and 3 fMRI scans. fMRI-WebVid (Li et al., 2024) involves five subjects with fMRI data acquired using a 3T scanner sampled at 1 frame per 0.8s. Stimuli videos (596×336) are sourced from WebVid (Bain et al., 2021), with each sample containing a 4s video and 5 fMRI scans. For synthesized datasets, we balanced samples across different duration ratios to avoid data bias. CC2017-Syn used fMRI ratios of [(1,3), (2,2), (3,1)] for 4 synthesized fMRI scans, while fMRI-WebVid-Syn used ratios of [(0,5), (2,3), (3,2)]. Synthesized training data originated only from original training data with no test overlap.

fMRI Preprocessing Following (Qian et al., 2023), each fMRI scan was projected to 32k_fs_LR brain surface space through anatomical structure and transformed to a 256×256 single-channel image, where only early and higher cortical regions retained values. fMRI data were averaged across multiple runs for the same video in both datasets.

Evaluation Metrics For video reconstruction, we utilized N-way top-K accuracy for semantic evaluation and SSIM for pixel-level assessment. Shot-specific fMRI segmentation employed segmentation accuracy, normalized mutual information (NMI), and adjusted rand index (ARI) following video scene segmentation research (Mahon & Lukasiewicz, 2024). For evaluating LLM-decoded captions, we used the CLIP text score to measure semantic alignment between generated and ground-truth descriptions.

Implementation Details For original CC2017, we processed 3 fMRI scans to generate 6s videos at 3 FPS. CC2017-Syn used 4 fMRI scans for 6s/6 FPS output. Original fMRI-WebVid processed 5 fMRI scans into 4s/3 FPS videos, while its synthesized counterpart used 5 fMRI scans for 4s/6 FPS reconstruction. All videos were generated at dimensions of 576×320. Theoretically, any text-to-video diffusion model can be used for video generation based on the decode captions. In this work, ModelScopeT2V (Wang et al., 2023) was used as our video generator, performing inference with 30 DDIM steps and adopt a 6.0 classifier-free guidance score. The image encoder and text encoder were initialized using CLIP ViT-H/14 from OpenCLIP (Cherti et al., 2023), and Qwen3-0.6B (Yang et al., 2025) served as the LLM decoder.

4.2 COMPARISON RESULTS

We compare our method against three fMRI-to-video baselines: MindVideo (Chen et al., 2023), NeuroClips (Gong et al., 2024), and GLFA (Li et al., 2024). Visual comparisons are shown in Figure 3, and quantitative results are presented in Table 1.

According to Table 1, our method outperforms all baselines, particularly in semantic-level metrics, demonstrating the effectiveness of our approach. Specifically, on the original fMRI-WebVid dataset, our method achieves a 7.5% improvement in frame-based 2-way classification score compared to the best baseline, while the 50-way classification score shows a substantial 55.1% improvement. These results suggest that decoding keyframe captions provides a more effective solution for fMRI-to-video reconstruction.

The visual comparisons in Figure 3 reveal that our shot-specific fMRI division strategy significantly contributes to multi-shot video reconstruction quality. In contrast, other baselines exhibit obvious quality degradation and fail to effectively reconstruct coherent multi-shot sequences.

4.3 ABLATION RESULTS

4.3.1 IMPACT OF SHOT SEGMENTATION

To evaluate the effectiveness of our proposed shot-specific fMRI segmentation, we conducted ablation studies focusing on three key aspects: (1) segmentation performance, semantic caption decoding accuracy, and final video reconstruction quality. We compare our full method incorporating the Shot Boundary Predictor (w/ $\mathcal{L}_{sbp}$) against a baseline (w/o $\mathcal{L}_{sbp}$) where $\mathcal{L}_{sbp}$ loss component is disabled.

Table 1: Quantitative comparison of fMRI-to-video reconstruction methods across four datasets, including two original datasets (fMRI-WebVid and CC2017) and two synthesized datasets (fMRI-WebVid-Syn and CC2017-Syn).

Dataset	Model	Video-Based		Frame-Based		
		Semantic-Level		Semantic-Level		Pixel-Level
		2-way↑	50-way↑	2-way↑	50-way↑	SSIM↑
fMRI-WebVid	MindVideo	0.736±0.04	0.075±0.01	0.760±0.03	0.109±0.01	0.097
	GLFA	0.790±0.03	0.107±0.01	0.729±0.03	0.118±0.01	0.143
	ours	**0.790±0.03**	**0.135±0.01**	**0.817±0.03**	**0.183±0.02**	**0.145**
fMRI-WebVid-Syn	MindVideo	0.788±0.03	0.117±0.01	0.735±0.03	0.122±0.01	0.095
	GLFA	0.800±0.03	0.109±0.01	0.727±0.04	0.092±0.01	0.108
	ours	**0.819±0.03**	**0.122±0.01**	**0.803±0.03**	**0.138±0.01**	**0.129**
CC2017	MindVideo	0.853±0.03	0.202±0.02	0.792±0.03	0.172±0.01	0.171
	NeuroClips	0.834±0.03	0.220±0.01	**0.806±0.03**	0.203±0.01	0.211
	GLFA	0.871±0.03	0.219±0.02	0.715±0.04	0.096±0.01	0.083
	ours	**0.891±0.03**	**0.235±0.02**	0.800±0.03	**0.206±0.01**	**0.244**
CC2017-Syn	MindVideo	0.813±0.03	0.164±0.01	0.780±0.03	0.107±0.01	0.107
	GLFA	0.877±0.02	0.181±0.02	0.752±0.04	0.087±0.01	0.124
	ours	**0.889±0.02**	**0.235±0.02**	**0.781±0.03**	**0.140±0.01**	**0.196**

Figure 3: Qualitative comparison of fMRI-to-video reconstruction results on four datasets. Visual examples from four datasets demonstrate the superiority of our method compared to three baselines.

In the w/o $\mathcal{L}_{sbp}$ baseline, the entire fMRI sequence is processed as a single unit and the fMRI encoder is optimized using a weighted sum of the lossess $\mathcal{L}_{caption}$, $\mathcal{L}_{align}$, and $\mathcal{L}_{mse}$. Here, $\mathcal{L}_{caption}$ is trained to decode a single, video-level caption describing the entire multi-shot sequence.

Table 2: Ablation results of shot-specific fMRI segmentation on fMRI-WebVid-Syn dataset.

Shot Segmentation	Caption CLIP	Segmentation Metrics			Video Reconstruction Metrics	
		ACC↑	ARI↑	NMI↑	2-way↑	50-way↑
w/o $\mathcal{L}_{sbp}$	0.177	-	-	-	0.814±0.03	0.112±0.01
w/ $\mathcal{L}_{sbp}$	0.304	0.685	0.683	0.690	0.819±0.03	0.122±0.01

Table 3: Ablation results of different prompt settings for video diffusion model.

Prompt	Video-Based		Frame-Based		
	Semantic-Level		Semantic-Level		Pixel-Level
	2-way↑	50-way↑	2-way↑	50-way↑	SSIM↑
fMRI Only	0.810±0.03	0.097±0.01	0.790±0.03	0.130±0.01	**0.145**
Text Only	**0.822±0.03**	**0.147±0.013**	0.815±0.03	**0.181±0.01**	0.144
Dual-Modal	0.809±0.03	0.108±0.01	**0.821±0.03**	0.171±0.02	0.101

As shown in Table 2, the shot boundary predictor achieves a segmentation accuracy of 0.685, and scores of 0.683 in ARI, and 0.690 in NMI, demonstrating the capability of shot boundary predictor to effectively identify transitions between distinct visual shots within the fMRI signal.

We also evaluate the impact of segmentation on the semantic precision of decoded captions. For the w/o $\mathcal{L}_{sbp}$ baseline, the decoded caption represents the entire video. Therefore, CLIP similarity is calculated against the ground-truth video-level caption. In contrast, for w/ $\mathcal{L}_{sbp}$ method, CLIP similarity is computed between the ground-truth keyframe captions and decoded keyframe captions for each individual shot. Results in Table 2 show that introducing shot segmentation improves CLIP similarity by 71.8%. This substantial improvement validates that dividing fMRI signals into shot-specific components and decoding keyframe captions per shot yields more semantically precise descriptions than attempting to decode a single, aggregated video-level caption from the mixed fMRI signal. Qualitative examples in Figure 4 illustrate that captions from w/o $\mathcal{L}_{sbp}$ are often semantically imprecise and biased towards unrelated content instead of reconstructing any specific shot, whereas captions from w/ $\mathcal{L}_{sbp}$ accurately focus on the core content of each individual shot.

Finally, we assess the impact of segmentation on the ultimate video reconstruction. Quantitative metrics in Table 2 show improvements of w/ $\mathcal{L}_{sbp}$ compared to w/o $\mathcal{L}_{sbp}$. This demonstrates that the enhanced semantic precision achieved through shot segmentation and keyframe caption decoding also contributes to higher-fidelity dynamic video reconstructions.

In summary, the divide-and-decode strategy effectively mitigates the semantic ambiguity inherent in processing mixed fMRI signals from multi-shot sequences, enabling the final high-fidelity multi-shot video reconstruction.

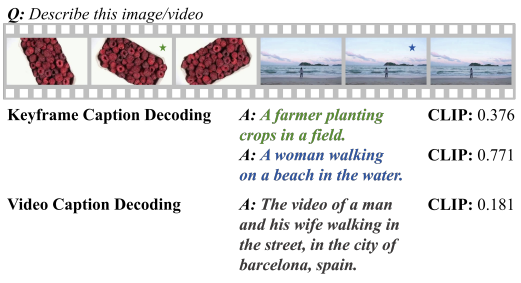

Figure 4: Comparison of decoded captions with and without shot-specific fMRI segmentation.

4.3.2 Impact of LLM Decoding

To evaluate the effectiveness of caption decoding against contrastive alignment for semantic extraction from fMRI, we conduct ablation experiments evaluating CLIP similarity under different training regimes. For contrastive alignment, since

there are no decoded captions, we calculate the CLIP similarity between fMRI embeddings and ground-truth caption embeddings.

As shown in Table 4, caption decoding ($\mathcal{L}_{caption}$ only) improves CLIP similarity by 6.7% over the alignment baseline ($\mathcal{L}_{align}$ only), demonstrating that decoding text descriptions better reconstruct semantics by mitigating temporal ambiguity. Although caption decoding outperforms the alignment baseline, semantic extraction is further enhanced by the multi-task framework that combines alignment, decoding, and reconstruction objectives. We ascribe this to the complementary information provided by different tasks, where alignment task helps preserve structural details while decoding primarily captures semantics.

4.3.3 Impact of Prompt Settings

Our method uses decoded keyframe captions as input prompts for the video generation model. To validate this design choice, we compare three prompt configurations: fMRI-only, text-only, and dual-modal. The fMRI-only setting uses fMRI embeddings directly as prompt embeddings for video diffusion model, while the dual-modal approach combines fMRI embeddings and text embeddings of decoded captions with equal weighting. As shown in Table 3, using only decoded keyframe captions as prompts achieves the best results, particularly on semantic-level metrics. Notably, dual-modal results are worse than text-only and even underperform fMRI-only on some metrics. We ascribe this degradation to the fact that combining embeddings may alter their representations in the latent space, contrary to our expectation that the combination would preserve visual details from fMRI while maintaining high semantic quality from decoded captions. Additionally, the inclusion of fMRI embeddings may introduce noise due to the inherent temporal signal mixing, which degrades reconstruction performance in multi-shot context.

Table 4: Ablation results on semantics extraction methods on fMRI-WebVid-Syn dataset.

Loss Function			Metric
$\mathcal{L}_{caption}$	$\mathcal{L}_{align}$	$\mathcal{L}_{mse}$	CLIP↑
-	✓	-	0.283
-	✓	✓	0.280
✓	-	-	0.302
✓	✓	-	0.313
✓	-	✓	0.300
✓	✓	✓	**0.336**

5 Limitation and Future Work

In this work, we establish a novel shot-level paradigm for multi-shot video reconstruction via shot-specific fMRI segmentation. However, several limitations remain to be addressed in future research. First, our current approach focuses on segmenting shot transitions that occur across different TRs. Decomposing intra-TR transitions presents a more challenging problem, as it requires modeling the underlying biomedical characteristics of fMRI signals in conjunction with acquisition protocols. Exploring this direction constitutes an important avenue for future work. Second, while our ablation studies indicate that using only textual captions for video generation yields better results than dual-modality conditioning, explore more sophisticated multimodal alignment and fusion mechanisms may further improve reconstruction quality.

6 Conclusion

In this work, we propose a novel divide-and-decode framework for reconstructing multi-shot videos from fMRI with high semantic fidelity. The shot-specific fMRI segmentation explicitly decouples mixed neural signals, providing cleaner shot-specific components for later semantics extraction and final video reconstruction. Decoding keyframe captions from shot-specific fMRI mitigates the temporal ambiguity caused by fMRI-video temporal resolution mismatch. By integrating these innovations, our framework achieves high-fidelity multi-shot reconstruction where prior methods fail. As one of the pioneering explorations in multi-shot fMRI decoding, we hope that our method can inspire future multi-shot video reconstruction endeavors.

REFERENCES

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

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.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2818–2829, 2023.

Tao Fang, Yu Qi, and Gang Pan. Reconstructing perceptive images from brain activity by shape-semantic gan. *Advances in Neural Information Processing Systems*, 33:13038–13048, 2020.

Tao Fang, Qian Zheng, Yu Qi, and Gang Pan. Extracting semantic-dynamic features for long-term stable brain computer interface. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5965–5973, 2023.

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.

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

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*, pp. 353–369. Springer, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 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.

Yizhuo Lu, Changde Du, Chong Wang, Xuanliu Zhu, Liuyun Jiang, Xujin Li, and Huiguang He. Animate your thoughts: Reconstruction of dynamic natural vision from human brain activity. In *The Thirteenth International Conference on Learning Representations*, 2025.

Louis Mahon and Thomas Lukasiewicz. Hard regularization to prevent deep online clustering collapse without data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14281–14288, 2024.

Xuelin Qian, Yun Wang, Jingyang Huo, Jianfeng Feng, and Yanwei Fu. fmri-pte: A large-scale fmri pretrained transformer encoder for multi-subject brain activity decoding. *arXiv preprint arXiv:2311.00342*, 2023.

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*, pp. 8748–8763. PmLR, 2021.

Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10146–10155, 2020.

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.

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.

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

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.

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*, pp. 11333–11342, 2024.

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.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent cot planning. *arXiv preprint arXiv:2503.07314*, 2025.

Weihao Xia, Raoul De Charette, Cengiz Oztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8226–8235, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655*, 2024.

A LLM USAGE

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text. It is important to note that the LLM was not involved in the research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis. The authors take full responsibility for the content of the manuscript. including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.