A. SemMiner

A.1. The Implementation Details

Since the objective of *SemMiner* is to utilize a Multimodal Large Language Model (MLLM) to generate multi-level semantic descriptions of a video from diverse perspectives, directly prompting the MLLM for this complex task often results in hallucinations or inaccurate content. To address this, we decompose the overall objective into a sequence of ordered subtasks—forming a prompt chain—that guides the MLLM progressively from simpler prompts to more complex ones. This step-by-step prompting strategy enhances the model's reasoning ability and output accuracy, while significantly reducing the risk of generating false or unreliable information [6–8, 23].

SemMiner adopts a two-stage pipeline prompting strategy. In Stage 1, a basic prompt P_{basic} is used to produce a concise core event summary C_{basic}, which serves as an anchoring "rein" to constrain and guide subsequent generation. Without this focused summary, the following caption generation process is prone to semantic drift. In Stage 2, *SemMiner* uses three specialized prompts in conjunction with C_{basic} to generate three decoupled semantic descriptions: a static anchor description (C_{anchor}) of the initial frame, a motion-focused narrative (C_{motion}) describing the dynamic content, and a holistic summary (C_{holi}) that integrates both static and dynamic elements into a coherent spatio-temporal account. The specific prompt templates used in our experiments are illustrated in Figure.1.

A.2. The Results of SemPrompt

In this section, we randomly sample a video from the CC2017 dataset to qualitatively evaluate the descriptions generated by *SemMiner*. As shown in Figure.1 and Figure.2, we first present the semantic caption produced by BLIP2[20], a widely-used video description model in prior fMRI-based video reconstruction works[4, 10]. It is evident that BLIP2 generates a coarse, static scene description and fails to capture the temporal dynamics of the video. In contrast, the multi-level descriptions generated by *SemMiner* —namely C_{anchor}, C_{motion}, and C_{holi}—are significantly more fine-grained and informative. The C_{anchor} description accurately captures the visual content of the initial frame, identifying not only salient objects and scene types (*e.g.*, "a young woman standing in a wheat field") but also detailing spatial layout and appearance features (*e.g.*, "wearing a short-sleeved blouse of a distinct mustard-yellow color and gently cradling a small bundle of golden-brown wheat stalks in her hands"). Similarly, C_{motion} and C_{holi} provide coherent and semantically rich accounts of the dynamic processes and holistic scene context, respectively, demonstrating the effectiveness of *SemMiner* in capturing multi-faceted video semantics.

Additionally, we quantitatively compare the average number of tokens in the descriptions generated by BLIP2 and *SemMiner*, as shown in Table 1. It is evident that BLIP2 produces relatively short captions, with an average of only 10.6 tokens. In contrast, *SemMiner* generates significantly richer descriptions, with averages of 32.8, 41.1, and 52.8 tokens for C_{anchor}, C_{motion}, and C_{holi}, respectively. These results highlight the ability of *SemMiner* to provide more detailed and semantically informative supervision signals, which in turn facilitate higher-quality neural decoding and video reconstruction.

caption type	BLIP2	C_{anchor}	C_{motion}	C_{holi}
number	10.6	32.8	41.1	52.8

Table 1. Average number of tokens in the descriptions generated by BLIP2 and *SemMiner* on CC2017 Datasets.

A.3. Caption Similarity Analysis

To quantitatively assess the semantic gap between the captions generated by *SemMiner*, and to verify that each description indeed captures a distinct perspective of the video stimulus, we evaluate the semantic similarity among them using a set of standard language similarity metrics [5, 12]. As shown in Table 2, the results reveal clear semantic divergence across the generated descriptions (C_{anchor}, C_{motion}, and C_{holi}), confirming that *SemMiner* effectively extracts complementary and perspective-specific semantic information from the input video.

Pair	BLEU-1	BLEU-4	METEOR	ROUGE-1	ROUGE-L	CLIP-S
C_{anchor} v.s. C_{motion}	0.171	0.022	0.186	0.275	0.207	0.655
C_{anchor} v.s. C_{holi}	0.197	0.028	0.222	0.304	0.225	0.665
C_{motion} v.s. C_{holi}	0.270	0.067	0.292	0.400	0.293	0.744

Table 2. Semantic Similarity Experiments of different caption pairs. Lower values across all metrics indicate greater semantic dissimilarity.

A.4. The Impact of Different VLLM

To evaluate the generalizability of *SemMiner*, we integrate three different vision-language large models (VLLMs)—Video-LLaMA [21], Qwen2.5-VL [1], and LLaVA-Video [22]—to generate semantic targets, using the same implementation details described in Appendix A.1.

The comparison results, presented in Table 3, show that while the optimal performance on individual metrics varies slightly across different VLLMs, all configurations achieve consistently strong and competitive reconstruction quality. The consistently high results confirm that the effectiveness of *SemMiner* does not rely on any specific VLLM. This robustness is rooted in its core design — a hierarchical seman-

Seen Video: 

BLIP2 Generated Output: a girl in a yellow shirt is standing in a field of wheat.

SemMiner - Stage 1

P_{basic} : Please provide a concise summary of the core events in this video. Your description must be strictly under20 words.

C_{basic} : A young woman in a wheat field, holding stalks of wheat, looks up to face the camera.

SemMiner - Stage 2

P_{anchor}:

Conditioned on the provided summary, $\{C_{basic}\}$, generate a detailed textual description of the video's initial frame. The description must detail key objects, their attributes (e.g., color, shape), and the overall scene setting. The generated text should answer the implicit question: 'What visual semantics would a human observer perceive from this stimulus?' The final output must be clear and precise, suitable for use as an input to a text-to-image model.

C_{anchor}:

The initial frame presents a photorealistic, sun-drenched medium close-up of a young woman standing in a wheat field. The central figure has long, straight, dark brown hair; her head is slightly tilted down with her eyes closed. She is wearing a short-sleeved blouse of a distinct mustard-yellow color and gently cradles a small bundle of golden-brown wheat stalks in her hands. The setting is an outdoor field where the background is softly out of focus, showing a mix of green and golden foliage under bright, natural sunlight.

P_{motion}:

Conditioned on the provided summary, $\{C_{basic}\}$, generate a description centered on the motion and dynamic actions of the principal objects in the video. This description should detail attributes such as movement direction, speed, posture changes, and interaction timing. The final output must be a detailed narrative, strictly constrained to a maximum length of 50 tokens.

C_{motion}:

The woman's head, initially tilted down with eyes closed, lifts in a slow and smooth motion. As her head rises, her eyes open, and her gaze deliberately shifts upward from a downward glance to look directly forward into the camera. This single, graceful action concludes with direct eye contact at the end of the clip.

P_{holi}:

Conditioned on the preceding summary, $\{C_{basic}\}$, generate a holistic description of the full video scene. This description must integrate the static background with the dynamic actions of the primary objects, detailing how they move through space and interact with the environment. The output should be a detailed narrative of approximately 60 words, presented without any introductory or concluding remarks.

C_{holi}:

In a sun-drenched wheat field, a young woman is captured in a medium close-up against a soft-focus golden and green background. Initially still and contemplative, her head is tilted down with her eyes closed as she cradles a bundle of wheat stalks. A slow, deliberate motion begins as she gracefully lifts her head. Her eyes open simultaneously with this fluid movement, her gaze traveling upward from the wheat to meet the camera lens directly, establishing a clear line of engagement.

Figure 1. Prompt templates used in *SemMiner* and qualitative comparison of captions generated by BLIP2 and *SemMiner*. *SemMiner* employs four distinct prompt templates to generate hierarchical semantic descriptions for a given video stimulus. The process begins with the Basic Prompt P_{basic}, which establishes a core semantic summary to anchor subsequent generations. This is followed by three specialized prompts in stage 2: (i) P_{anchor}, which produces a detailed description of the initial frame, optimized for guiding text-to-image models; (ii) P_{motion}, which focuses on extracting dynamic information such as object motion, direction, speed, and posture changes; and (iii) P_{holi}, which integrates static and dynamic elements to form a coherent, high-level narrative of the entire video. Compared to BLIP2, which tends to generate short and static descriptions, *SemMiner* provides richer and more perspective-specific captions.

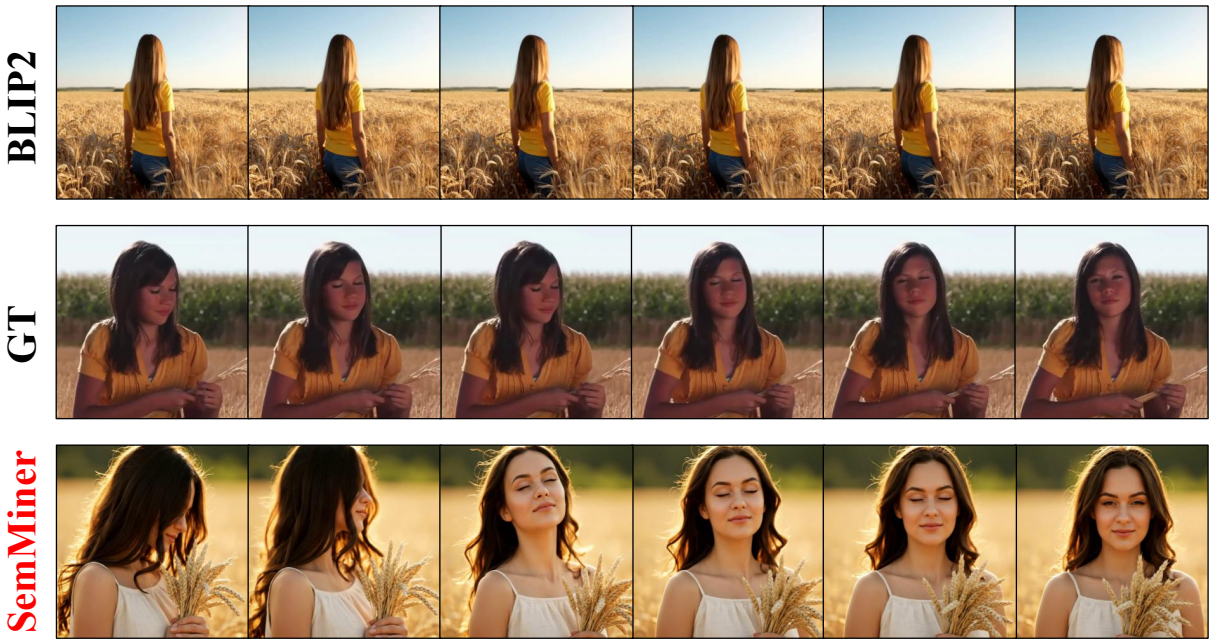

Figure 2. Video generated from BLIP2 caption(upper) and SemMiner(lower). The middle is the target video.

tic prompting strategy that reliably guides diverse VLLMs to generate rich, structured, and multi-perspective video descriptions, thereby providing high-quality semantic supervision for fMRI-based video reconstruction.

B. Evaluation Metrics

We devised a multi-level evaluation protocol to comprehensively assess the performance of our video reconstruction model. This protocol systematically quantifies the fidelity of reconstructed videos to the original visual stimuli across three key dimensions: semantic, pixel, and spatiotemporal.

B.1. Semantic-level

Semantic-level metrics quantitatively assess the conceptual alignment between the reconstructed and original videos. We employ two methods for this purpose: the N-way Top-1 accuracy test and the VIFI-Score. **N-way Top-1 Accuracy** measures the precision of semantic content reconstruction using a dual strategy: (1) Dynamic Content. To evaluate the semantic fidelity of dynamic events, a VideoMAE [17] model pretrained on the Kinetics-400 dataset classifies both the original and reconstructed videos, after which the consistency of their predicted class labels is compared. (2) Static Content. To assess the accuracy of static elements, a ViT-For-Image-Classification model [3] pretrained on ImageNet performs frame-by-frame classification.

VIFI-Score:This metric is calculated using a VIFI-CLIP [13] model that has been fine-tuned on video datasets. The model extracts high-level feature embeddings from both the original and reconstructed videos. The final VIFI-Score is the cosine similarity between these two embedding vectors, where a higher value indicates greater semantic congruence.

B.2. Pixel-level

Pixel-level metrics quantitatively assess low-level visual fidelity by directly comparing the reconstructed video to the original visual stimulus. These metrics provide a foundational measure of reconstruction quality by focusing on the accuracy of pixel values and the preservation of fundamental image characteristics.

Peak Signal-to-Noise Ratio (PSNR): A widely adopted metric for quantifying reconstruction error, PSNR is derived from the Mean Squared Error (MSE) between the pixel values of the original and reconstructed images and measures distortion on a logarithmic scale.

Structural Similarity Index (SSIM): This metric evaluates the preservation of structural information between frames. Unlike purely pixel-wise error metrics, SSIM is designed to model perceptual similarity by jointly considering luminance, contrast, and structure.

Hue Similarity [16]: This metric specifically assesses color reproduction accuracy by computing the cosine similarity between the hue channel vectors of the original and reconstructed frames. The hue component is isolated because it represents pure color, independent of saturation and brightness. A high score therefore indicates that the model has successfully reproduced the chromatic ambiance of the original scene, a critical factor for the overall visual experi-

MLLM	Semantic-Level					Pixel-Level			ST-Level	
	2-way-I↑	2-way-V↑	50-way-I↑	50-way-V↑	VIFI-score↑	SSIM↑	PSNR↑	Hue-pcc↑	CLIP ↑	EPE ↓
VideoLlama	**0.826**	0.860	**0.211**	0.239	0.590	**0.330**	**9.626**	0.841	**0.502**	4.768
Qwen2.5-vl	0.821	**0.863**	0.203	**0.242**	0.589	0.299	9.03	0.858	0.496	**4.40**
Llava-Video	0.815	0.855	0.207	0.228	**0.591**	0.309	8.96	**0.861**	0.493	4.65

Table 3. Ablation study on the CC2017 dataset investigating performance variations when using different VLLMs within the *SemMiner* framework.

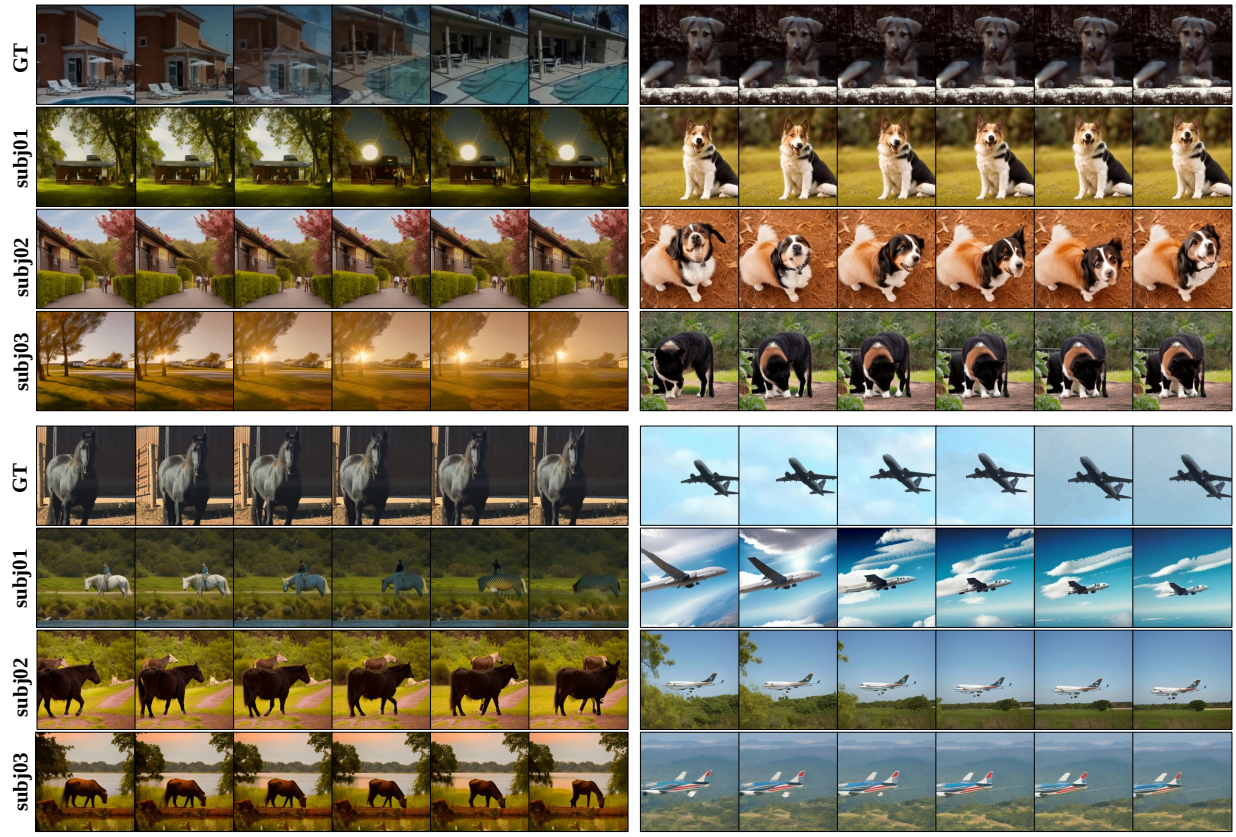

Figure 3. More reconstruction results on the CC2017 dataset from three subjects.

ence.

B.3. Spatiotemporal (ST)-level

Spatiotemporal (ST) metrics are designed to evaluate the dynamic aspects of a reconstructed video, focusing on its temporal coherence and motion representation accuracy. These metrics are crucial for assessing a model's ability to generate realistic and continuous sequences of events.

CLIP-pcc [19]: This metric assesses the temporal smoothness and semantic continuity of a video sequence. It is calculated by using a CLIP (ViT-B/32) model to extract feature embeddings for each frame and then averaging the cosine similarity across all adjacent frame pairs.

Endpoint Error (EPE [2]): This metric provides a precise quantification of motion reconstruction fidelity. It is defined as the mean Euclidean distance between the optical flow vectors of the reconstructed and original videos. By directly comparing these predicted motion fields, EPE serves as a critical indicator of the model's capability to accurately reproduce the trajectories and velocities of objects within the scene.

C. Implementation Details

C.1. Hyperparameter Settings

For the Semantic Alignment Decoder, we utilized the AdamW [9] optimizer. The training was conducted for 100 epochs with a batch size of 96. A cosine annealing schedule was employed to manage the learning rate, with a maximum value set to 1e-4. λ_{prior} and $\lambda_{\text{SoftCLIP}}$ are set to 0.1 and 0.5,

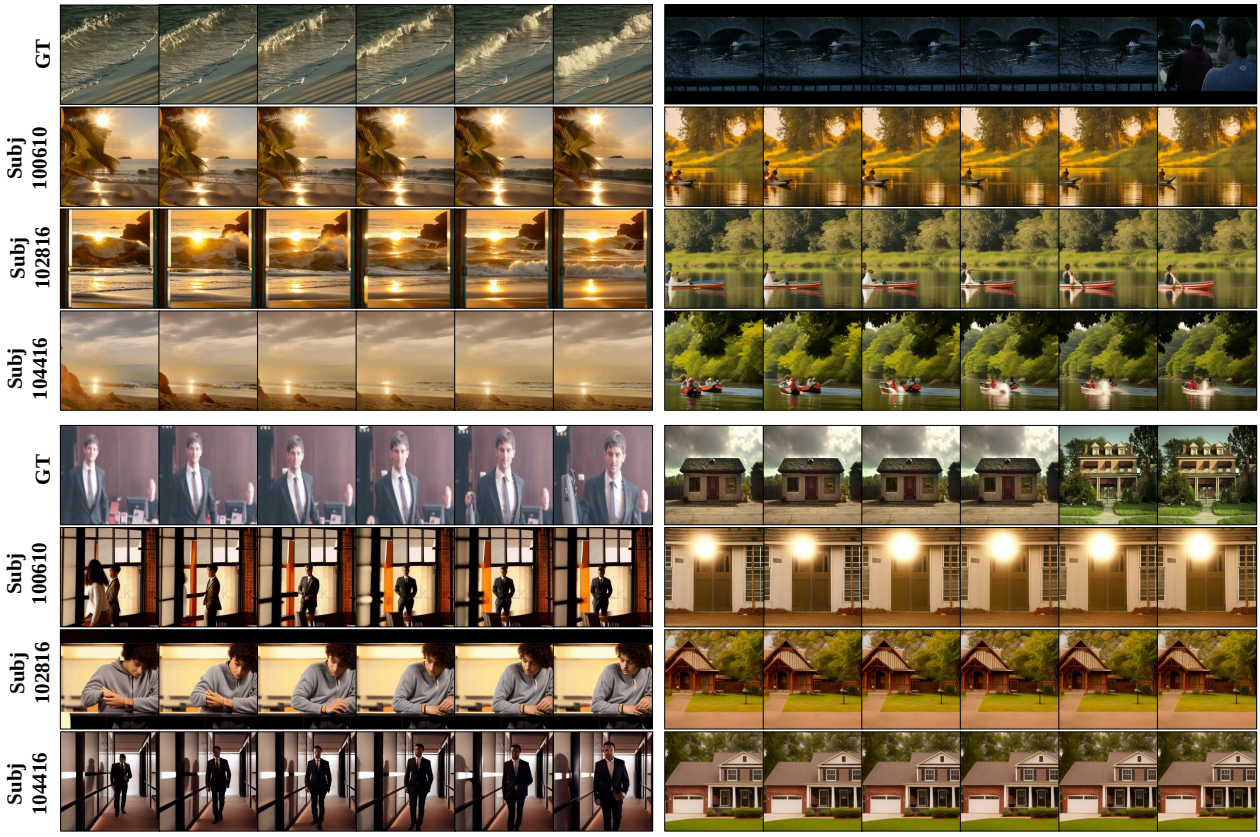

Figure 4. More reconstruction results on the HCP dataset from three subjects.

respectively. Furthermore, to mitigate the risk of overfitting arising from the scarcity of fMRI-video data pairs, we incorporated the Mixco data augmentation technique.

$$x_{mix_{i,j}} = \alpha x'_i + (1-\alpha)x'_j, \tag{1}$$

where $\alpha \sim \text{Beta}(0.15, 0.15)$ is a mixing coefficient sampled from the Beta distribution, and x'_i is the i-th latent representation to the x'-encoded batch.

For the Motion Adaptation Decoder, the AdamW optimizer was also employed, and the model was trained for 100 epochs with a batch size of 20. The learning rate was governed by a OneCycleLR [15] schedule, with the maximum learning rate configured to 3e-4. Notably, λ_{spat} and λ_{temp}, were not pre-defined hyperparameters. Instead, they were implemented as learnable parameters that are automatically optimized during the training process.

C.1.1. Model Architecture

For Semantic Alignment Decoder, we use a linear layer to implement the subject-specific mapper $f^{\theta_m}_{\text{SAD}}$, which maps voxels $x \in \mathbb{R}^{D_m}$ to the shared space $\mathbb{R}^{4096}$. f^{MLP}_{SAD} is implemented as a four-layer perceptual machine with output channels specified as $\{4096, 4096, 4096, 77 \times 768\}$, using the GELU nonlinear activation function. The (f^{Refine}_{SAD}) module uses a Transformer architecture with 4 layers, each containing 8 attention heads of dimension 64.

For the Motion Adaptation Decoder, f^{proj}_{MAD} is composed of a linear layer and a four-layer Multi-Layer Perceptron with output channels specified as $\{256, 6 \times 4096, 6 \times 4096, 4096, 6 \times (64 * 7 * 7)\}$.The fusion attention module is a lightweight decoder with a UNet-like [14] architecture, consisting of one UNetMidBlock2D and three cascaded AttnUpDecoderBlock2D upsampling blocks. The architecture takes a 64-channel, 7x7 feature map as input and progressively upsamples it to a 4-channel, 28x28 latent representation.

D. More Results

We present further examples of video reconstruction across multiple subjects from both the CC2017 [18] and HCP [11] datasets, as illustrated in Figure.3 and Figure.4.

D.1. Neural Interpretability

To further validate the neural interpretability of the proposed framework, we generated voxel-wise importance

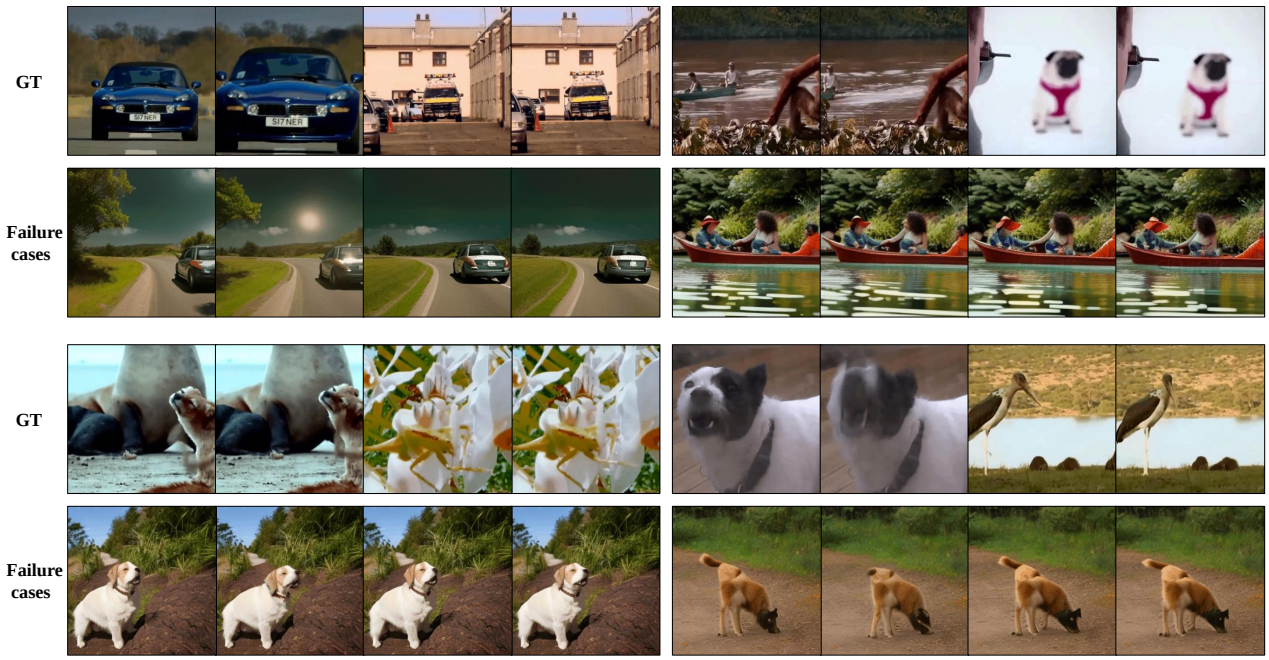

Figure 5. Reconstruction failure cases.

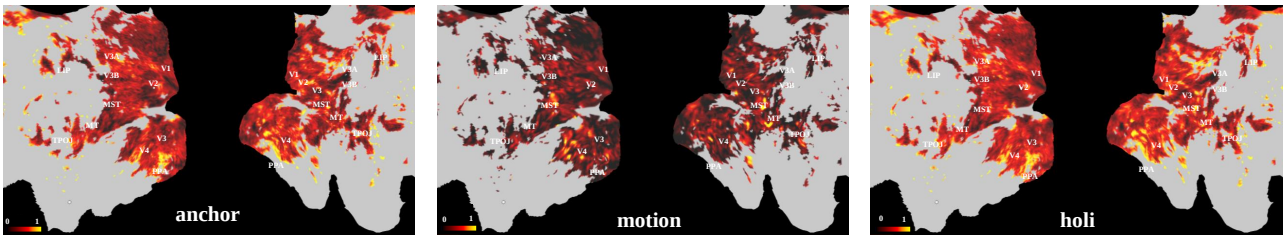

Figure 6. Importance visualization of different ROIs for Subject 2, based on the fitted weights from the first layer of the SFD. Weights from each module are averaged and normalized to the $[0, 1]$ range for comparison.

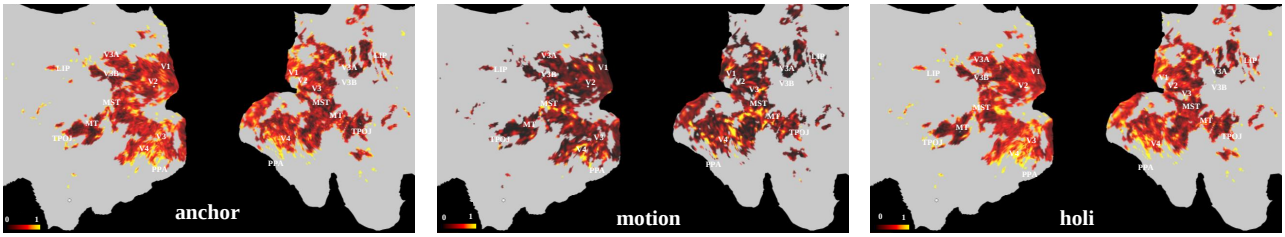

Figure 7. Importance visualization of different ROIs for Subject 3, based on the fitted weights from the first layer of the SFD. Weights from each module are averaged and normalized to the $[0, 1]$ range for comparison.

maps of the cerebral cortex for Subjects 2 and 3 from the CC2017 dataset, as depicted in Figure.6 and Figure.7. The experimental results for these subjects reveal voxel importance distribution patterns that are highly similar to those observed for Subject 1: The anchor component's activations were predominantly concentrated in visual cortical areas. The motion component exhibited stronger activations in brain regions specialized for motion processing. The holistic component displayed a balanced activation pattern across both of these functional regions. This hierarchical semantic mapping enables *SemVideo* to generate more perceptually faithful reconstructions and similar activation distributions across subjects demonstrate stable feature capture for both static semantics and dynamic behaviors.

D.2. Failure cases

To facilitate a comprehensive and objective assessment of $SemVideo$, we analyze several of its reconstruction failure cases in Figure.5. A primary cause of these failures stems from the inherent data acquisition paradigm. Specifically, the experimental data is created by evenly segmenting longer videos, which introduces abrupt content transitions at the boundaries of the resulting clips. Our model, $SemVideo$, struggles to replicate these sudden shifts in the reconstructed outputs.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[2] John L. Barron, David J. Fleet, and Steven S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12:43–77, 1994. 4

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *CVPR*, 2021. 3

[4] 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. In *NeurIPS*, 2024. 1

[5] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 1

[6] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 2025. 1

[7] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *ICLR*, 2023.

[8] Do Xuan Long, Duong Ngoc Yen, Anh Tuan Luu, Kenji Kawaguchi, Min-Yen Kan, and Nancy F. Chen. Multi-expert prompting improves reliability, safety, and usefulness of large language models. In *EMNLP*, 2024. 1

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019. 4

[10] Yizhuo Lu, Changde Du, Chong Wang, Xuanliu Zhu, Liuyun Jiang, and Huiguang He. Animate your thoughts: Decoupled reconstruction of dynamic natural vision from slow brain activity. In *ICLR*, 2025. 1

[11] Daniel S Marcus, John Harwell, Timothy Olsen, Michael Hodge, Matthew F Glasser, Fred Prior, Mark Jenkinson, Timothy Laumann, Sandra W Curtiss, and David C Van Essen. Informatics and data mining tools and strategies for the human connectome project. *Frontiers in Neuroinformatics*, 5:4, 2011. 5

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 1

[13] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *CVPR*, 2023. 3

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 5

[15] Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. *arXiv preprint arXiv:1708.07120*, 2018. 5

[16] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991. 3

[17] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 3

[18] 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, 2017. 5

[19] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023. 4

[20] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 1

[21] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 1

[22] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2025. 1

[23] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *ICLR*, 2023. 1