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

Supplementary Material

Contents

In this supplementary material, we provide comprehensive details and additional experimental results to support the findings in the main text. First, Sec. A elaborates on the experimental setup, including data preprocessing pipelines, detailed metric definitions, and implementation specifics. Sec. B describes the construction of the category semantic learning task. Sec. C presents extensive quantitative and qualitative evaluation of CINENEURON. Sec. D presents further ablation studies on the EEG input, superclass processing, and different semantics incorporated in CINENEURON. Finally, Sec. E and Sec. F discuss the limitations and ethical considerations of our work, respectively.

A. Experimental Setup

In this section, we provide a comprehensive description of the experimental setup. We first detail the data preprocessing pipelines and voxel selection procedures for the cc2017 and CineBrain datasets in Sec. A.1. Next, we define the frame-based evaluation metrics used to assess semantic and pixel-level quality in Sec. A.2. Finally, we present additional implementation details and specific hyperparameter settings in Sec. A.3.

A.1. Details of Datasets and Data Preprocessing

cc2017 Dataset. We preprocess fMRI data in the cc2017 dataset using the pipelines described in [25, 98], following the procedures outlined in [23]. The pipeline consists of five stages: artifact removal, motion correction using six degrees of freedom, registration to MNI standard space, transformation to cortical surfaces, and subsequent coregistration to a cortical surface template as detailed in [24]. In line with [25], we select stimulus-activated voxels by evaluating voxel-wise correlations of training videos. These correlations are processed through Fisher z-transformation and assessed using a one-sample t-test. We then select voxels with significant activation (Bonferroni-corrected, $P < 0.05$), resulting in 13,447, 14,828, and 9,114 activated voxels in the visual cortex for Subjects 1, 2, and 3, respectively. The training videos are processed into 2-second clips to match the 2-second temporal resolution of the fMRI data. These clips have a resolution of $57 \times 624 \times 624$ to align with the input requirements of the video decoder model (Wan2.1). As a result, we obtain 8,640 (4,320×2) fMRI-video pairs for training and 1,200 pairs for testing. Additionally, we incorporate a 4-second delay in the BOLD signals to account for hemodynamic response latency when mapping movie stimulus responses, as suggested by [29, 70, 96].

CineBrain Dataset. We preprocess fMRI data from the CineBrain dataset utilizing the processes described in [20], specifically employing the widely adopted fMRIPrep pipeline [15]. The fMRI data are collected at a frequency of 1.25 Hz, corresponding to a temporal resolution of 0.8 seconds. The selected visual regions of interest (ROIs) in the CineBrain dataset are characterized using the parcellation provided by the Human Connectome Project Multi-Modal Parcellation (HCP-MMP) within the 32k_fs_LR space. The identified visual ROIs include areas such as "V1, V2, V3, V3A, V3B, V3CD, V4, LO1, LO2, LO3, PIT, V4t, V6, V6A, V7, V8, PH, FFC, IP0, MT, MST, FST, VVC, VMV1, VMV2, VMV3, PHA1, PHA2, PHA3", totaling 8,405 voxels in the visual cortex for each subject. Additionally, we select 1,559 voxels within the hippocampus using the same pipeline applied to the visual regions. Therefore, unless stated otherwise, our experiments on the CineBrain dataset utilize a total of 9,964 voxels from the visual cortex and hippocampus as fMRI data. The videos viewed by participants are standardized to 18 minutes and segmented into 4-second clips to align with the fMRI temporal resolution of 5 fMRI signals, each with a resolution of 0.8 seconds, as described in CineBrain. Consequently, we obtain 4,860 training samples and 540 testing samples of fMRI-video pairs with resolution $33 \times 480 \times 720$. In addition, fMRI signals underwent z-scoring across vertices to adjust for the inherent delay in BOLD responses, factoring in a 4-second lag.

A.2. Details of Frame-Based Metrics

In addition to the video-level metrics described in the main text, following [20, 25, 98], we evaluate the generated frames at both semantic and pixel levels. For semantic-level evaluation, we perform an N-way top-K accuracy classification test on 1,000 ImageNet [13] classes using an ImageNet classifier. A trial is successful if the ground truth (GT) class is among the top-1 probabilities in the generated frame results, selected from N random classes, including the GT class. For pixel-level evaluation, we employ the Structural Similarity Index (SSIM) [104] and Peak Signal-to-Noise Ratio (PSNR) to assess the image quality of video frames.

A.3. Implementation Details

In addition to Sec. 4.1, we provide more implementation details here. The temperature τ is set to 0.07. The loss weights λ_1 and λ_2 are set to 0.1 and 10, respectively, while α in Eq. (8) is set to 1. The routing network within the Mixture-of-Memories (MoM) is implemented as a linear layer. We train our model and conduct experiments on 4 A800 GPUs.

B. Classification Task Construction

This section outlines the construction of the category semantic learning task used to enrich fMRI representations. We describe the automated pipeline for extracting object nouns from video captions using Qwen2.5-VL and mapping

Instruction Prompt (User Input):

Given a caption of a certain video and a list of pre-specified categories, your task is to extract the objects in this caption, and classify each object into a class in the pre-specified categories below. ****Your response should be in JSON format****.

Pre-specified Categories:
["accessory", "animal", "appliance", "electronic", "food", "furniture", "indoor", "kitchen", "outdoor", "man", "woman", "sports", "vehicle", "crowd", "others"]

Output Format (JSON):
```json
{ <extracted object 0>: <corresponding category>, <extracted object 1>: <corresponding category>, ..., <extracted object n>: <corresponding category> }
```

Instructions:
- The objects to be extracted include both things (objects with a well-defined shape, e.g. car, person, cat) and stuff (amorphous background regions, e.g. grass, sky, water).
- The fields in the JSON output are the objects extracted from the caption, while the value in each field is a string for the matching class in the pre-specified categories.
- Do NOT add new categories beyond the Pre-specified Categories.
- Strictly follow the above JSON format, and directly respond with one JSON output in one go.
- ****IMPORTANT****: You should only respond with one JSON object formatted as above. Do NOT add any further note, explanation, discussion or correction.

Input Caption:
A woman in a black sweater bends over to pick up oranges from a crate on the ground. She is standing next to another woman in a green shirt who is watching her. There are cars parked on the street behind them.

Now, generate the classification results for the input caption, strictly following the above JSON structure.

Qwen 2.5-VL:

{ "woman": "woman", "oranges": "food", "cars": "vehicle", "street": "outdoor"}

Figure S1. The instruction prompts used for constructing the object category recognition task.

Table S1. **Quantitative results of all subjects on the cc2017 and CineBrain datasets.** Please note that, currently, the CineBrain dataset provides fMRI data only for subjects 1 and 5.

DATASET	METHODS	Semantic-level		Spatiotemporal-level		
		2-way	50-way	CLIP-pcc	DTC	MS
cc2017	Subject1	0.846±0.02	0.237±0.02	0.973±0.01	0.959±0.01	0.967±0.02
	Subject2	0.852±0.02	0.241±0.04	0.972±0.01	0.956±0.01	0.966±0.01
	Subject3	0.853±0.02	0.242±0.04	0.971±0.01	0.948±0.01	0.964±0.02
	Average	0.850±0.02	0.240±0.03	0.972±0.01	0.954±0.01	0.966±0.02
CineBrain	Subject1	0.936±0.02	0.384±0.03	0.986±0.01	0.971±0.01	0.974±0.01
	Subject5	0.938±0.01	0.401±0.02	0.990±0.01	0.978±0.01	0.976±0.01
	Average	0.937±0.02	0.393±0.03	0.988±0.01	0.975±0.01	0.975±0.01

them to a simplified set of superclasses to facilitate robust classification and address class imbalance.

Specifically, we generate category names for the key objects in each video clip to establish the object category recognition task. To streamline this process, we directly utilize video captions generated by Qwen2.5-VL to extract object nouns (*e.g.*, ['woman', 'car', 'oranges', ...]). We then use Qwen2.5-VL to classify these nouns into the pre-defined MSCOCO [54] categories, and simplify the categories by filtering and merging infrequent classes into a reduced set of 15 superclasses for both the cc2017 and CineBrain datasets. The output is formatted as a JSON object containing the names of superclasses. Detailed instruction prompts for category name generation, along with the superclass lists, are presented in Fig. S1.

C. More Experimental Results

In this section, we provide a comprehensive evaluation of our CINENEURON through various quantitative and qualitative analyses. We first present the fMRI-to-video reconstruction performance for each individual subject (Sec. C.1) and evaluate the model using frame-based semantic and pixel-level metrics (Sec. C.2). To demonstrate the versatility of our approach, we show its extension to alternative backbones like MindEye and NeuroClips (Sec. C.3). We further validate the scalability of our Mixture-of-Memories (MoM) strategy by expanding the memory pool with external datasets (Sec. C.4). Additionally, we provide qualitative visualizations of reconstructed videos (Sec. C.5) and assess the model's robustness regarding out-of-distribution (OOD) concepts (Sec. C.6), cross-dataset generalization on BOLD-Moments (Sec. C.7), and neural interpretability (Sec. C.8). We further perform sensitivity analysis of the model's ro-

Table S2. **Quantitative comparison results of frame-based metrics on the CineBrain dataset.** Results for the CineBrain dataset are quoted from [20]. "*" denotes methods reimplemented using the same decoder model and fMRI input as CINENEURON.

METHODS	Semantic-level		Pixel-level	
	2-way	50-way	SSIM	PSNR
GLFA [49]	0.847	0.225	0.123	7.526
CineSync [20]	0.926	0.358	0.240	11.92
CineSync*	0.926±0.04	0.293±0.03	0.267±0.04	**16.04**±2.35
CINENEURON (Avg)	**0.949**±0.03	**0.438**±0.04	**0.271**±0.05	16.02±2.38
Subject1	0.946±0.02	0.425±0.05	0.272±0.05	16.11±2.65
Subject5	0.951±0.02	0.451±0.03	0.269±0.04	15.92±2.11

bustness to noise (Sec. C.9). Finally, we report detailed metrics for retrieval accuracy (Sec. C.10) and classification performance (Sec. C.11).

C.1. Results of Each Subject

We report the fMRI-to-video reconstruction performance of our CINENEURON on each subject in the cc2017 and CineBrain datasets in Tab. S1. The cc2017 dataset contains fMRI data from 3 subjects, while the CineBrain dataset currently provides fMRI data only for Subjects 1 and 5. The results in Tab. S1 demonstrate that our method consistently outperforms the baselines in average performance across all subjects, validating its effectiveness and robustness.

C.2. Results of Frame-Based Metrics

We provide frame-based evaluation results in Tab. S2, along with average results for our method across all subjects. The results show that CINENEURON excels in semantic-level frame understanding while maintaining competitive pixel-level performance. In Tab. S2, CINENEURON achieves the highest semantic-level 2-way accuracy, exceeding CineSync and CineSync* by 2.3% and GLFA by 10.2%. For 50-way semantic classification, our method provides an 8% improvement over CineSync, demonstrating its ability to recognize fine-grained semantics. Regarding pixel-level metrics, CINENEURON achieves the best SSIM, surpassing CineSync by 3.1%, and performs second-best in PSNR, competitive with CineSync*. These results demonstrate that our method improves frame-based video comprehension by effectively integrating multimodal semantics.

C.3. Extension to Other Architectures

We highlight that our CINENEURON is architecture-agnostic and can extend to other architectures. In addition to transformer-based architectures such as MindVideo [10], our method is also applicable to the architectures of MindEye [80] and NeuroClips [25].

MindEye and NeuroClips employ an MLP backbone coupled with a diffusion prior [78], alongside the MixCo [43] contrastive learning loss during training. The MLP backbone includes a ridge regression module and a Residual MLP module. The ridge regression module maps fMRI data to a lower dimension, and the Residual MLP module further refines the representation in an enhanced hidden space. Following NeuroClips, we initialize the MLP backbone using a pretrained checkpoint from MindEye2 [81].

Since we focus on learning semantics for fMRI embeddings, we utilize the MLP backbone as the fMRI encoder. In the Bottom-Up Semantic Enrichment stage, our training protocol adheres to the default NeuroClips semantic learning settings, incorporating our designed classification and action alignment tasks. In the Top-Down Memory Integration stage, we train the MLP backbone and our newly introduced components: the routing network and the fusion mechanism, and implement LoRA tuning within the Video DiT model.

The output of the MLP backbone serves as fMRI embeddings for alignment, classification, and memory pool retrieval during training. To match the embedding dimension and token length between the MLP backbone output and the retrieved text embeddings, we employ an additional cross-attention layer to integrate the fMRI embeddings into the retrieved text embeddings. During training, except for hyperparameters specifically noted in NeuroClips, all other settings remain consistent with those outlined in Sec. 4.1 of the main text. During inference, our method remains straightforward and efficient among different architectures. It requires only the input of fMRI data to generate decoded video, eliminating the need for complex steps like key frame reconstruction, ControlNet integration, or additional condition generation.

We present quantitative results of our CINENEURON using the MindEye and NeuroClips architectures on the cc2017 dataset, as shown in Tab. S3. Building upon the MindEye and NeuroClips architectures, our method achieves improved semantic-level metrics and comparable spatiotemporal-level metrics, indicating the scalability of our approach across different architectures.

Table S3. **Quantitative results of our CINENEURON based on MindEye/NeuroClips architecture (denoted as †) on Subject 1 of cc2017 dataset.**

METHODS	Semantic-level		Spatiotemporal-level		
	2-way	50-way	CLIP-pcc	DTC	MS
Wen [110]	-	0.166±0.02	-	-	-
Wang [96]	0.773±0.03	-	0.402±0.41	-	-
Kupershmidt [47]	0.771±0.03	-	0.386±0.47	-	-
MinD-Video [10]	0.839±0.03	0.197±0.02	0.408±0.46	0.884±0.08	0.901±0.05
NeuroClips [25]	0.834±0.03	0.220±0.01	0.738±0.17	0.926±0.05	0.955±0.01
CINENEURON	0.846±0.02	0.237±0.02	0.973±0.01	**0.959**±0.01	**0.967**±0.02
CINENEURON†	**0.860**±0.02	**0.242**±0.02	**0.974**±0.01	0.958±0.01	0.965±0.01

Table S4. **Quantitative ablation study of expanding the memory pool on Subject 1 of the cc2017 dataset.**

POOL SIZE	Semantic-level		Pixel-level		Spatiotemporal-level	
	Acc_2	**Acc_{50}**	**SSIM**	**PSNR**	**CLIP-pcc**	**DTC**
Reduce 50%(2160)	0.845	0.235	0.372	9.429	0.970	0.949
Ours (4320)	0.846	0.237	**0.376**	9.474	0.973	0.959
+AnimalKindom (5185)	0.854	0.243	0.373	9.434	**0.976**	**0.964**
+BOLDMoments (6185)	**0.858**	**0.254**	0.374	**9.486**	0.970	0.959

Table S5. **Comparison on OOD samples of the cc2017 dataset.**

METHOD	**Acc_2**	**Acc_{50}**	**SSIM**	**PSNR**	**DTC**	**CLIP-pcc**
NeuroClips	0.801	0.193	0.210	9.193	0.926	0.959
MindAnimator	0.784	0.162	**0.274**	9.082	0.606	0.829
Ours	**0.821**	**0.197**	0.267	**9.237**	**0.953**	**0.970**

Table S6. **Comparison on BOLDMoments [48] dataset.**

METHOD	**Acc_2**	**Acc_{50}**	**SSIM**	**PSNR**	**DTC**	**CLIP-pcc**
NeuroClips	0.736	0.154	0.181	8.997	0.922	0.973
Ours	**0.791**	**0.192**	**0.230**	**9.078**	**0.968**	**0.986**

C.4. Expanding the Memory Pool

To further validate the effectiveness and robustness of our proposed Mixture-of-Memories, we ablate the memory pool size by both reducing its capacity and expanding it using external data. The expansion strategy aims to simulate a more comprehensive cognitive process akin to the human brain, which utilizes both established memories and broader external knowledge to better understand and learn from current stimuli.

To achieve this, we incorporate additional datasets including Animal Kingdom [69], Dream-1K [100], and BOLDMoments [48]. We process their videos into short clips, randomly selecting 665 animal videos from Animal Kingdom and 200 human videos from Dream-1K to create an expanded pool of 5185 videos. We then further scale the pool to 6185 videos using the BOLDMoments dataset. For all external data, we extract corresponding text, image, and action embeddings for retrieval.

The quantitative results in Tab. S4 demonstrate the stability and scalability of our method across the semantic, pixel, and spatiotemporal levels. Notably, a 50% reduction in memory size (to 2160 videos) results in only minor degradation (e.g., a 0.1% drop in Acc_2), highlighting the robustness of the retrieval mechanism. Conversely, larger pools consistently boost performance across the metrics.

C.5. More Visualization Results

To demonstrate the effectiveness of our CINENEURON, we present additional qualitative results on the cc2017 and CineBrain datasets, displayed in Figs. S3 and S4.

C.6. Generalization to OOD Concepts

Our MoM's multi-modal design ensures robustness for OOD samples: if specific image concepts are missing, retrieved text or action priors still enhance generation. For OOD evaluation, we identify 63 unseen videos based on cc2017 categories and evaluate methods on this subset. Tab. S5 shows our method leads on most metrics, indicating

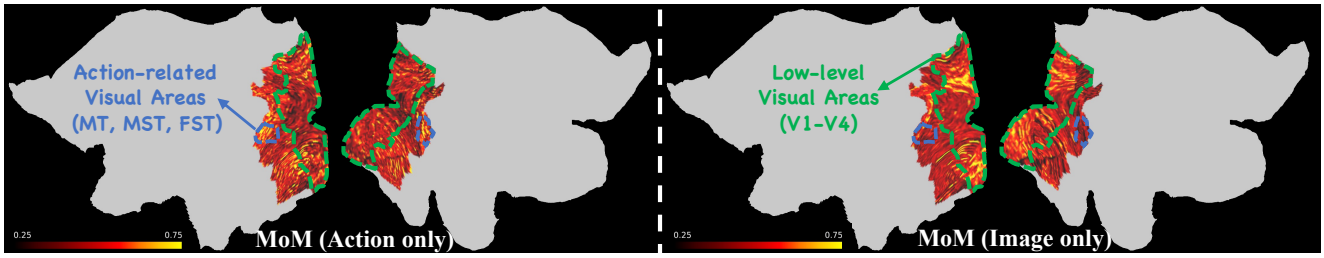

Figure S2. **More Visualization of voxel weights from the first regression layer.** Voxel weights are averaged and normalized to $[0, 1]$, displayed with a 0.25 to 0.75 colorbar. The blue and green dotted lines indicate the action-related and low-level visual areas, respectively.

Table S7. **Noise sensitivity analysis.**

Configuration	Acc_2	Acc_{50}	SSIM	PSNR	DTC	CLIP-pcc
No Noise	0.846	0.237	0.376	9.474	0.959	0.973
add 10% Noise	0.845	0.229	0.371	9.303	0.928	0.965
add 25% Noise	0.843	0.225	0.365	9.228	0.923	0.963

Table S8. **The averaged top-1 retrieval accuracy on the cc2017 dataset.**

METHOD	fMRI-to-image Retrieval (↑)	image-to-fMRI Retrieval (↑)
NeuroClips [25]	22.2%	18.8%
CINENEURON	**28.3%**	**26.2%**

strong OOD generalization. Tab. S4 also shows consistent gains with larger pools. This validates our scalability and offers a unique solution to OOD issues by simply increasing memory diversity, a capability absent in other methods.

C.7. Results on BOLDMoments Dataset

We report new results on the BOLDMoments [48] dataset (Version B, MNI152). Tab. S6 shows that our method consistently outperforms NeuroClips on BOLDMoments, indicating strong cross-dataset generalization. Tab. S4 also shows that adding BOLDMoments data into the memory pool further boosts performance, demonstrating our scalability with increased memory diversity.

C.8. More Interpretation Results

We conduct more interpretability analysis in Fig. S2, which reveals that: (**i**) incorporating action memory selectively activates dorsal motion regions (*e.g.*, MT); and (**ii**) incorporating image memory emphasizes low-level visual areas (*e.g.*, V1-V4). These findings show that MoM aligns with genuine neural processing and our method effectively enhances both motion perception and low-level visual details.

C.9. Noise Sensitivity Analysis

Tab. S7 shows that our method remains robust under up to 25% noise, verifying that MoM provides robust semantic guidance.

C.10. Retrieval Accuracy

Following NeuroClips [25], we evaluate the retrieval performance of our CINENEURON using Top-1 fMRI-to-image retrieval accuracy (forward retrieval accuracy) and Top-1 image-to-fMRI retrieval accuracy (backward retrieval accuracy). For fMRI-to-image retrieval, each test fMRI data is converted into an fMRI embedding. The fMRI embedding is used to query the target embedding based on the CLIP cosine similarity from a set that includes the target embedding along with 299 other randomly selected test embeddings. Retrieval is successful if the cosine similarity is highest between the fMRI embedding and its corresponding image embedding. The test set comprises 1,200 fMRI-video pairs, divided into 4 subsets of 300 pairs each for evaluation. We report the average retrieval accuracy across these subsets. Image-to-fMRI retrieval follows the same protocol, with fMRI and image roles reversed.

Tab. S8 presents the retrieval accuracy comparison with NeuroClips, demonstrating improvements of 6.1% in fMRI-to-image retrieval and 7.4% in image-to-fMRI retrieval. These results indicate that our proposed method effectively enhances retrieval accuracy, further validating the efficacy of our Mixture-of-Memories strategy.

Despite recent advancements in fMRI-to-image reconstruction achieving retrieval accuracies exceeding 90%, both NeuroClips and our retrieval accuracy in the cc2017 dataset are lower. This discrepancy is attributed to three factors: (1) Task difficulty: Visual stimuli from videos are inherently more complex than images, making the retrieval

more challenging. (2) Dataset distribution: The cc2017 test set includes numerous object categories absent in the training set, increasing generalization difficulty. (3) Lack of large-scale pretrained models: Unlike the fMRI-to-image reconstruction task, which benefits from robust models pretrained on large-scale datasets, the fMRI-to-video reconstruction suffers from smaller datasets and lacks pretrained models, complicating the retrieval of unseen samples.

Notably, our approach does not solely depend on retrieval accuracy. Instead, it employs a dynamic and end-to-end retrieval and fusion process, allowing the model to continuously refine the fMRI embeddings and their associated semantics during training. Because the retrieved samples are not fixed but updated based on the current representation, the model can iteratively adjust and improve its alignment between fMRI signals and semantic space.

Future research should focus on constructing large-scale fMRI-video datasets and developing pretrained models to address these issues and further enhance retrieval accuracy.

C.11. Classification Accuracy

We present our classification accuracy results for each category on the cc2017 dataset, as shown in Tab. S13. The results demonstrate that utilizing superclasses and reducing the number of classes, along with the introduction of Focal Loss, effectively enhance classification performance.

D. More Ablation Studies

In this section, we provide more ablation studies and further analysis, investigating the impact of EEG input (Sec. D.1), superclass pre-processing (Sec. D.2), and different semantic alignments (Sec. D.3).

D.1. Ablation of EEG Input

While fMRI can probe deep-brain neural activities, it is limited by relatively low temporal resolution. In contrast, EEG provides superior temporal resolution that is well-suited for capturing rapid neural oscillations. Combining these two modalities is a promising way to leverage the high temporal resolution of EEG to compensate for the temporal limitations of fMRI.

We conduct preliminary experiments on the CineBrain dataset using both fMRI and EEG data as inputs. As shown in Tab. S9, incorporating EEG data further improves both semantic and spatiotemporal decoding performance compared with using fMRI alone. These results demonstrate the benefit of integrating EEG with fMRI and suggest that jointly modeling multimodal neural signals is a promising direction for future work.

D.2. Ablation on Superclass Pre-processing

Tab. S10 shows that finer-grained categories (w/ Subclasses) hurt both semantic and pixel-level metrics, confirming the necessity of superclass preprocessing for noisy fMRI to reduce learning difficulty.

D.3. Ablation on Different Semantics

Tab. S11 shows that, in the first stage (bottom-up semantic enrichment), action alignment yields better Temporal Consistency and EPE than category alignment, highlighting its importance for motion decoding, whereas category alignment primarily benefits semantic accuracy.

Furthermore, Tab. S12 shows that, in the second stage (top-down memory integration), image and action memories play complementary roles in MoM: images improve low-level details (SSIM) while actions enhance spatiotemporal consistency (DTC).

E. Limitations

While CINENEURON has achieved semantically-enhanced and high-quality fMRI-to-video reconstruction, certain limitations remain. Our method struggles with accurately reconstructing cross-scene fMRI, *i.e.*, fMRI recorded during transitions between video clips, a challenge also noted in NeuroClips [25]. Although such fMRI instances are infrequent, addressing this issue is a potential avenue for future research. Additionally, due to the limited dataset size, we employed parameter-efficient LoRA for fine-tuning. With sufficient data, fully fine-tuning the model might enhance performance further. Constructing large-scale fMRI-to-video datasets will require collaboration across various fields, including neuroscience and artificial intelligence, to promote the future research in the community.

F. Ethical Considerations and Social Impacts

This work investigates the potential of video generation models for decoding human brain activity, specifically focusing on fMRI data. This approach aims to enhance our understanding of brain function and contribute to advancements in neuroscience, such as the field of brain-computer interfaces. While the research holds practical significance, addressing concerns regarding participant privacy and data security is also essential. In this work, we utilize two public, de-identified datasets as our training data, thereby strictly adhering to ethical standards. To further reduce privacy risks, data collection agencies must adhere to stringent protocols and ethical guidelines. Additionally, the community and government should implement measures to safeguard private data and prevent misuse.

Table S9. **Quantitative ablation of adding EEG input on Subject 5 of CineBrain dataset.**

METHODS	Semantic-level		Spatiotemporal-level		
	2-way (↑)	50-way (↑)	CLIP-pcc (↑)	DTC (↑)	MS (↑)
only fMRI data	0.938±0.01	0.401±0.02	0.990±0.01	0.978±0.01	0.976±0.01
fMRI data + EEG data	**0.949**±**0.01**	**0.471**±**0.04**	**0.995**±**0.01**	**0.984**±**0.01**	**0.979**±**0.01**

Table S10. **Ablation on use of superclasses.**

Configuration	**Acc$_2$**	**Acc$_{50}$**	**SSIM**	**PSNR**	**DTC**	**CLIP-pcc**
w/ Subclasses	0.841	0.217	0.342	8.945	0.957	0.972
w/ Superclasses (**Ours**)	**0.846**	**0.237**	**0.376**	**9.474**	**0.959**	**0.973**

Table S11. **Ablation on category and action semantics.**

Method	**Temp. Cons. ↑**	**EPE ↓**	**Acc$_2$ ↑**	**Acc$_{50}$ ↑**
Category Only	0.965	1.861	<u>0.840</u>	<u>0.229</u>
Action Only	<u>0.972</u>	**1.620**	0.836	0.212
Full Model	**0.973**	<u>1.628</u>	**0.846**	**0.237**

Table S12. **Ablation on MoM.**

Configuration	**Acc$_2$**	**Acc$_{50}$**	**SSIM**	**PSNR**	**DTC**	**CLIP-pcc**
MoM w/o Image	0.842	0.225	0.344	9.274	0.950	0.970
MoM w/o Action	0.840	0.229	0.369	9.339	0.933	0.965
Ours	**0.846**	**0.237**	**0.376**	**9.474**	**0.959**	**0.973**

Figure S3. **More qualitative results of our CINENEURON on the cc2017 dataset.**

Figure S4. **More qualitative results of our CINENEURON on the CineBrain dataset.**

Table S13. **Classification accuracy of all categories on the cc2017 dataset.** "-" denotes no such category in the test set.

Index	Class Name	Accuracy
0	accessory	0.975
1	animal	0.720
2	appliance	-
3	electronic	0.967
4	food	0.983
5	furniture	0.873
6	indoor	0.749
7	kitchen	0.972
8	man	0.718
9	others	0.972
10	outdoor	0.616
11	crowd	0.673
12	sports	0.967
13	vehicle	0.787
14	woman	0.820