Animate Your Thoughts: Reconstruction of Dynamic Natural Vision from Human Brain Activity

•Yizhuo Lu[1,2], Changde Du*[1], Chong Wang[4], Xuanliu Zhu[5], Liuyun Jiang[1,2], Xujin Li[1,2], Huiguang He[∓1,2,3],

•[1] State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, CASIA,

•[2] School of Future Technology, University of Chinese Academy of Sciences, [3] School of Artificial Intelligence, UCAS,

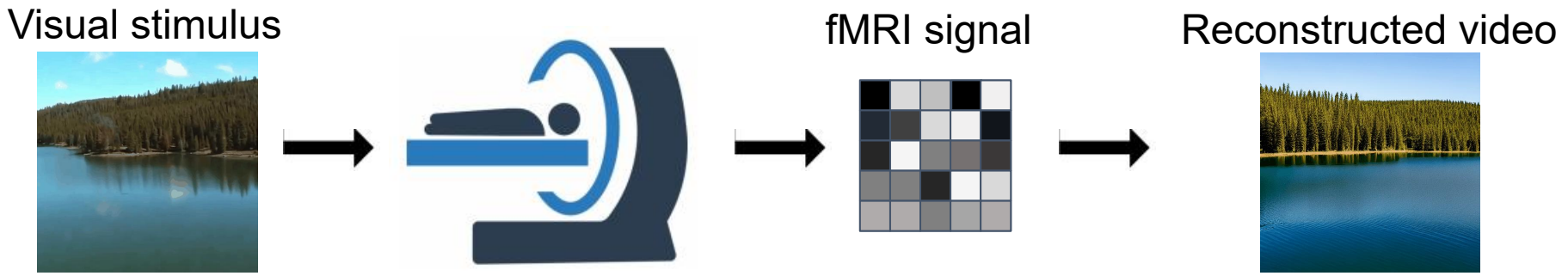
•[4] School of Computer and Artificial Intelligence, Zhengzhou University, [5] Beijing University of Posts and Telecommunications

•*Email: luyizhuo2023@ia.ac.cn, huiguang.he@ia.ac.cn*

1 中国科学院自动化研究所 INSTITUTE OF AUTOMATION CHINESE ACADEMY OF SCIENCES

2 中国科学院大学 University of Chinese Academy of Sciences

Introduction

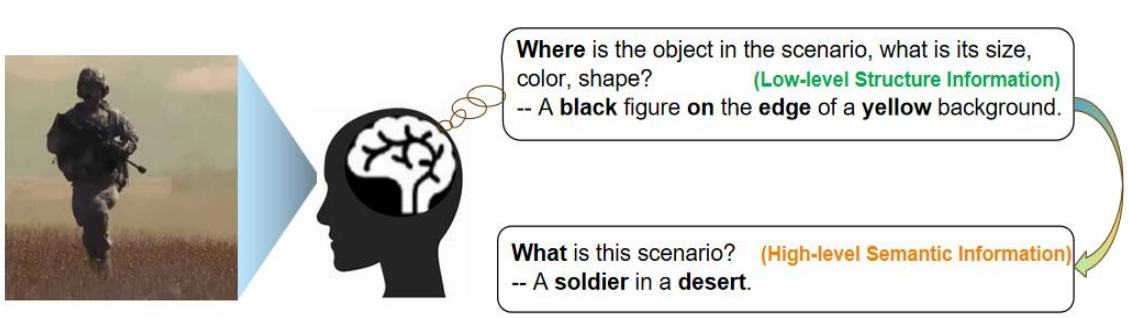


(1) Method: We propose Mind-Animator, which enables video reconstruction by decoupling semantic, structural, and motion information from fMRI data for the first time.

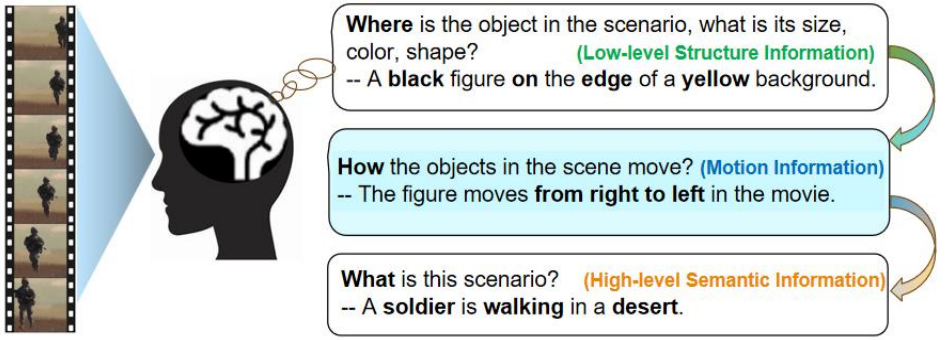
(2) Interpretability: We use voxel-wise and ROI-wise visualization techniques to elucidate the interpretability of our proposed model from a neurobiological perspective.

(3) Comprehensive evaluation: We introduce eight evaluation metrics that comprehensively assess the reconstruction results of our model and all previous models across three dimensions—semantic, structure, and spatiotemporal consistency—on three publicly available video-fMRI datasets. This establishes our work as the first unified benchmark for subsequent researchers. We will release all data and code to facilitate future research.

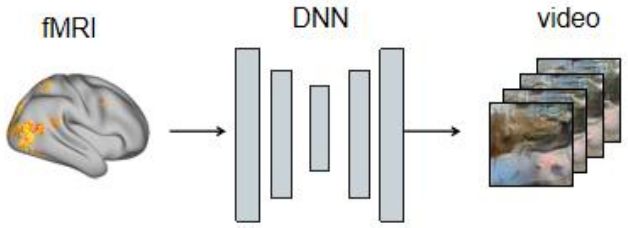
Motivation and Related works



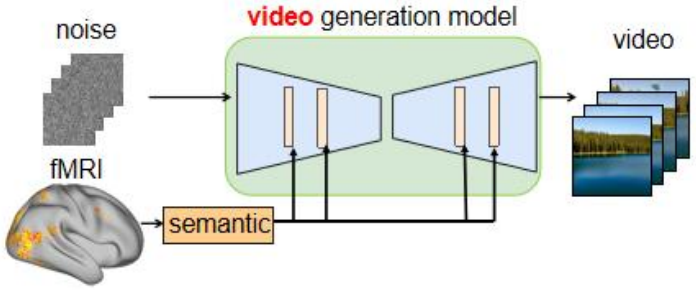
(a) Static image stumilus



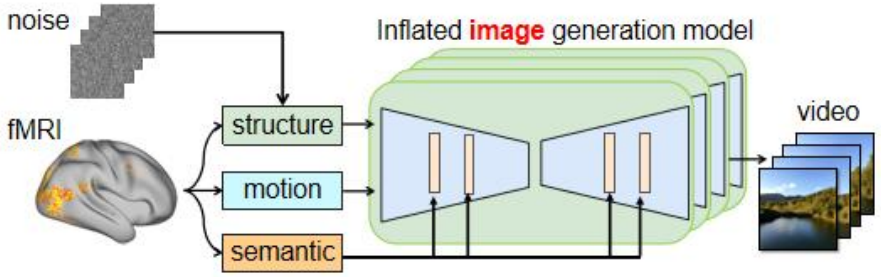
(b) Dynamic video stumilus



(c) End-to-end models



(d) Video generation model based



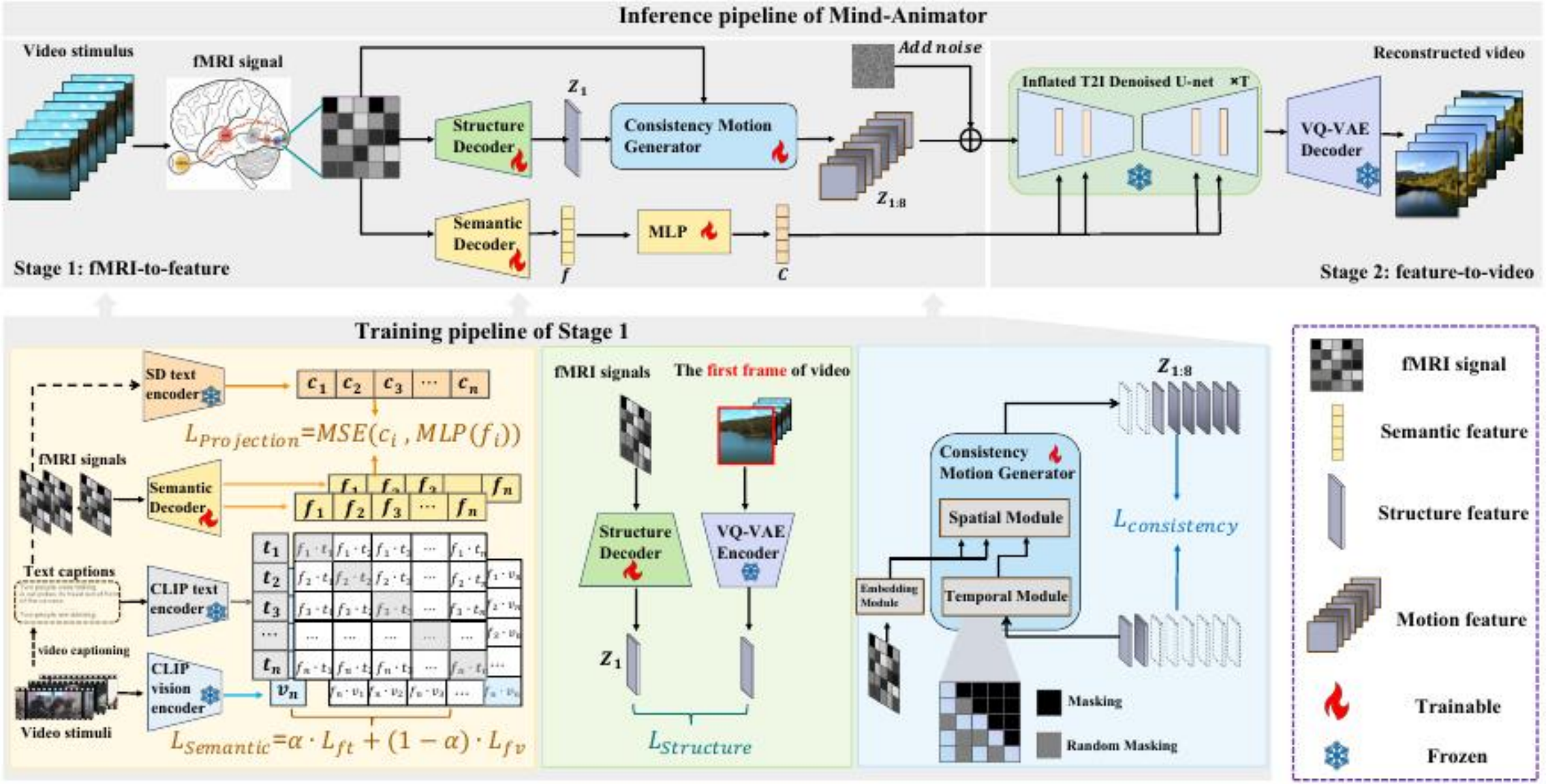
(e) Our model

Existing Issues:

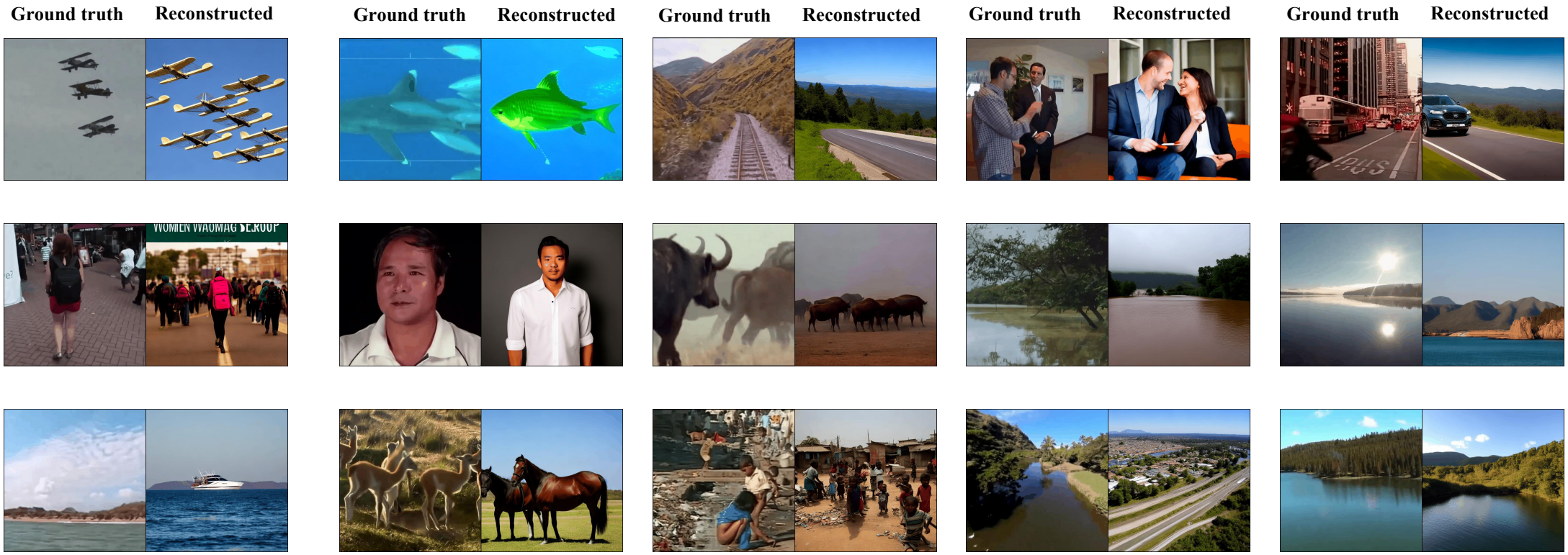
(1) Unable to accurately model semantic, structural, and motion features.

(2) External video data (e.g., motion priors learned by video generation models) interfere with the motion features in the reconstructed video.

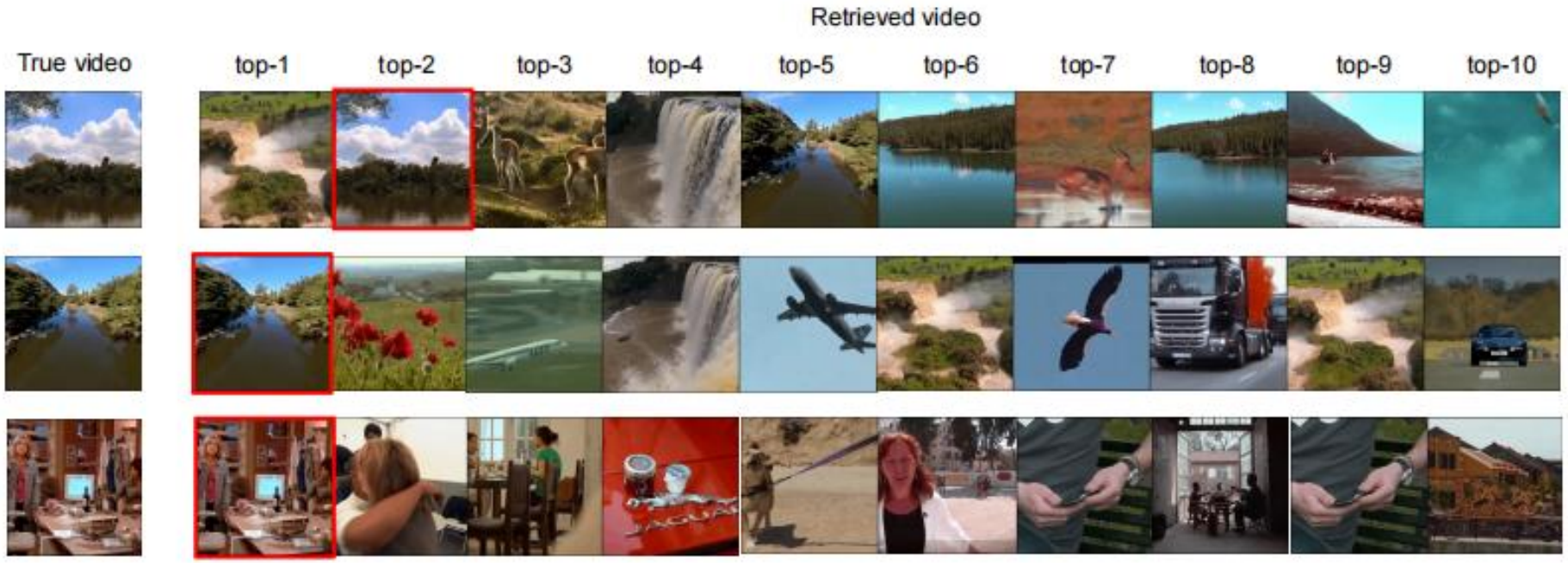
Methodology



Results (reconstruction task)



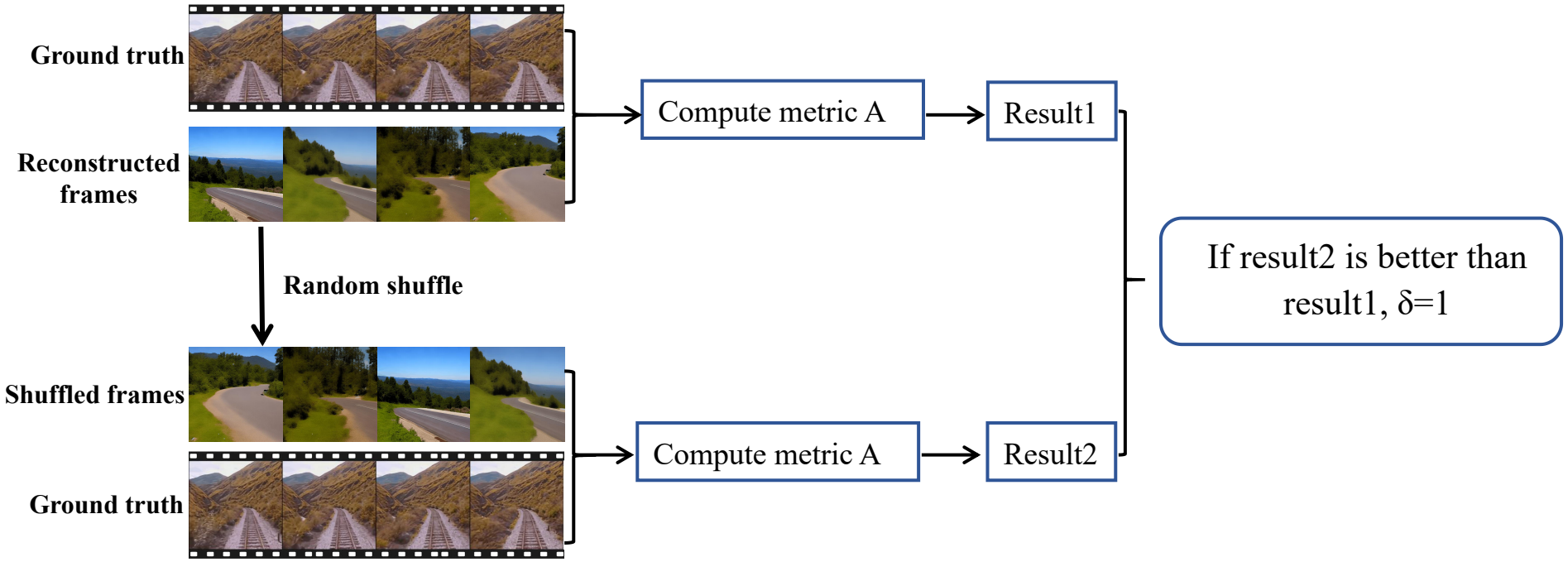
Results (retrieval task)



Dataset		CC2017							
		Subjet1		Subjet2		Subjet3		Average	
Model	Test set	top-10	top-100	top-10	top-100	top-10	top-100	top-10	top-100
Wen (Wen et al. (2018))	Small	2.17*	19.50*	3.33*	19.17*	——	——	2.75*	19.33*
Kupershmidt (Kupershmidt et al. (2022))	Small	1.09*	8.57*	0.92*	8.24*	0.84*	8.24*	0.95*	8.35*
Mind-video (Chen et al. (2024))	Small	**3.22***	19.08*	2.75*	16.83*	3.58*	22.08*	3.18*	19.33*
Ours	Small	3.08	**22.58**	**4.75**	**26.90**	**4.50**	**24.67**	**4.11**	**24.72**
Wen (Wen et al. (2018))	Large	1.41*	11.58*	2.08*	9.58*	——	——	1.75*	10.58*
Kupershmidt (Kupershmidt et al. (2022))	Large	0.17*	2.94*	0.17*	2.77*	0.25*	2.18*	0.19*	2.63*
Mind-video (Chen et al. (2024))	Large	1.75*	7.17*	0.83*	5.17*	1.25*	9.00*	1.28*	7.11*
Ours	Large	**2.17**	**12.50**	**2.25**	**17.00**	**2.75**	**16.42**	**2.39**	**15.31**

Interpretability

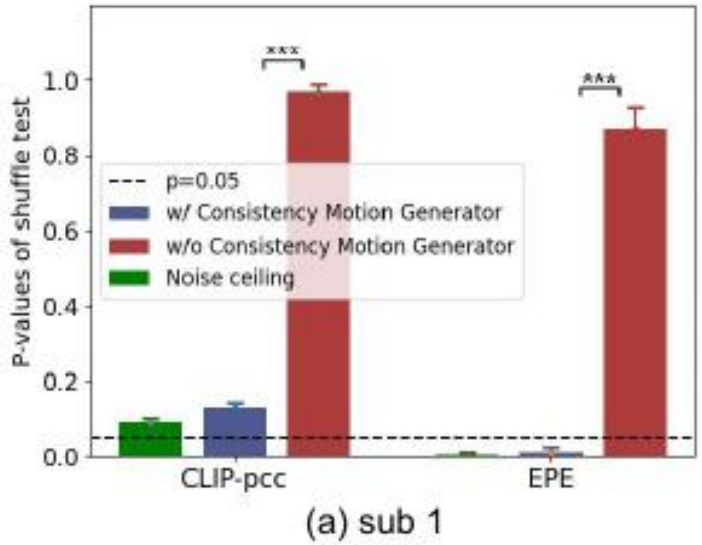
· **Have we truly decoded motion information from fMRI?**



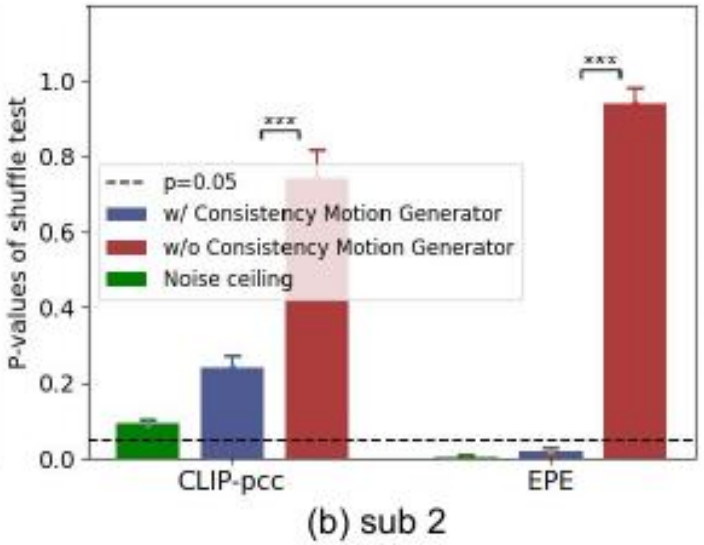
Repeating the aforementioned procedure 100 times, the P-value of the shuffle test can be estimated as $P = \sum_{i=1}^{100} \delta_i / 100$

A lower P-value indicates a higher consistency between the reconstructed video frames and the ground truth before shuffling.

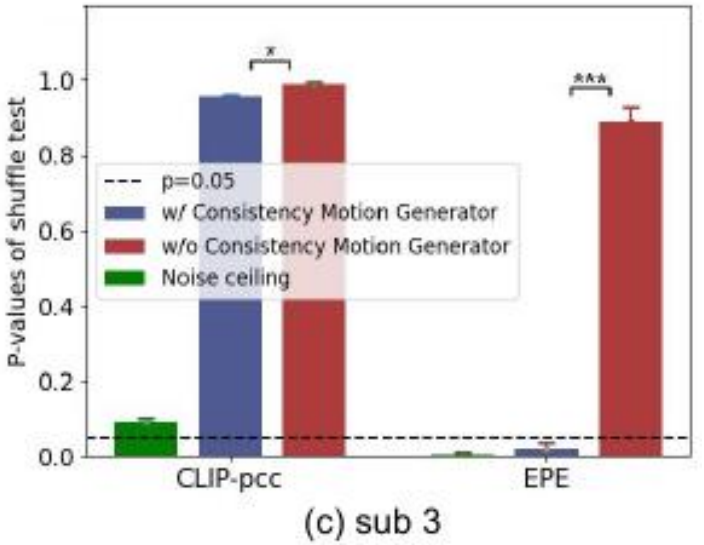
Interpretability



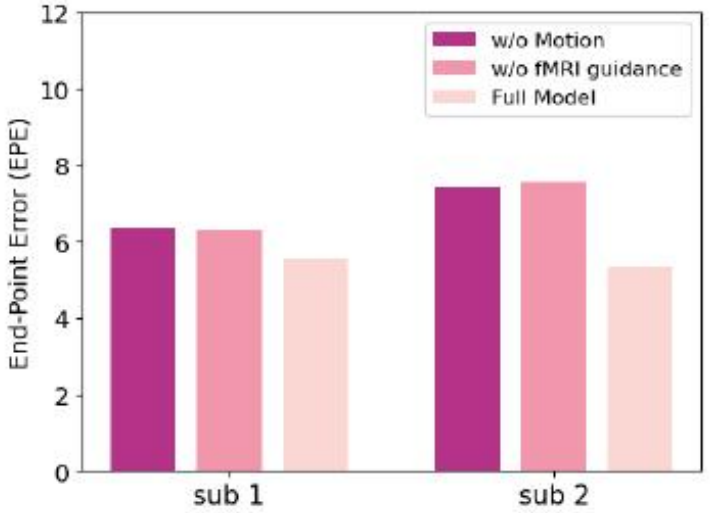
(a) sub 1



(b) sub 2

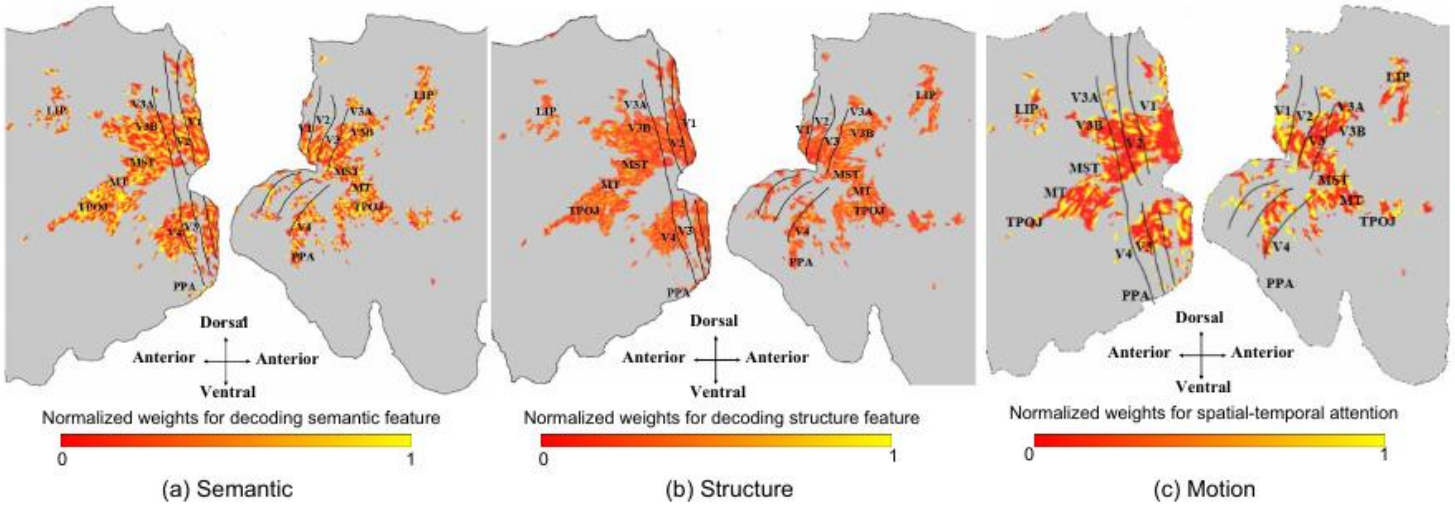


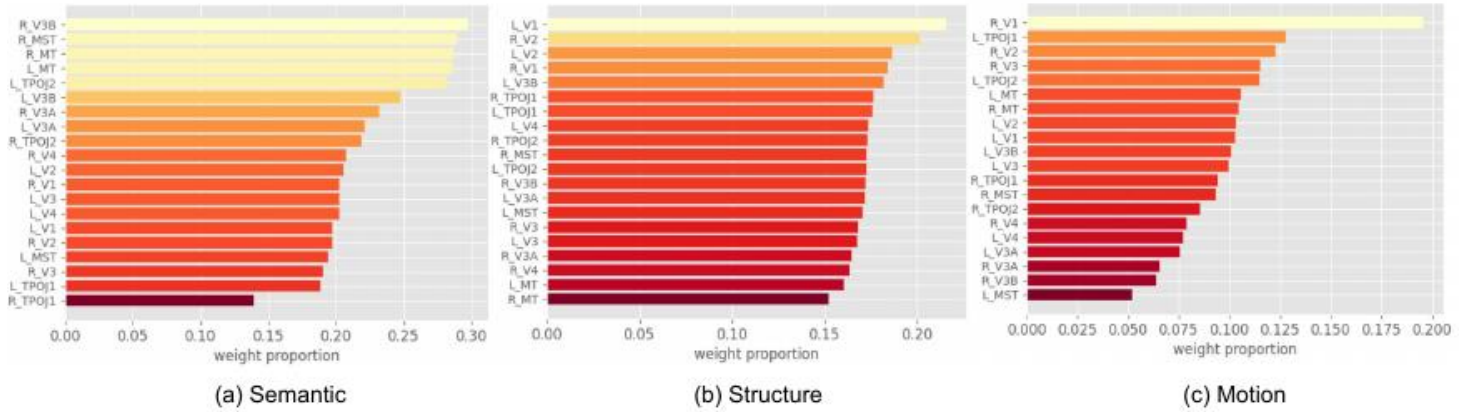
(c) sub 3



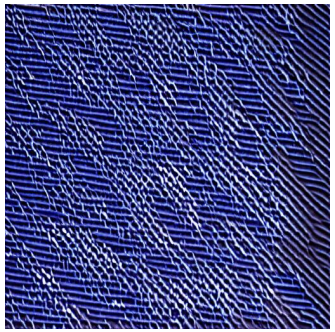
Interpretability

· **Which brain regions are responsible for decoding different features, respectively?**





Ablation Study

Ground truth	Reconstructed	Without Semantic	Without Structual	Without Motion
				
				

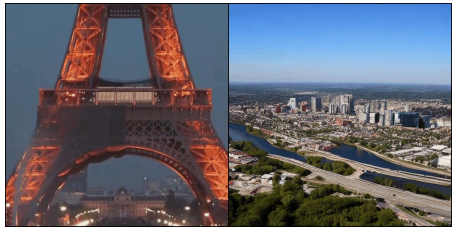
Fail Cases

Case 1: Decoding errors in semantic, structural, or motion features due to low decoding accuracy.

Ground truth Reconstructed

Ground truth Reconstructed

Ground truth Reconstructed



Ground truth Reconstructed

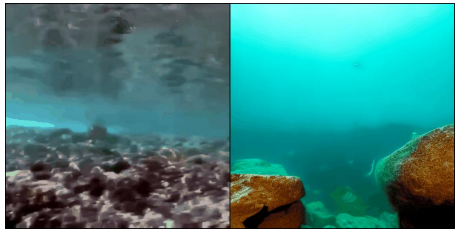
Case 2: The data acquisition paradigm causes abrupt content transitions at the boundaries of video clips, which are uniformly segmented from the complete videos viewed by the subjects during data collection.

Ground truth Reconstructed

Ground truth Reconstructed

Ground truth Reconstructed

Ground truth Reconstructed



Thanks!