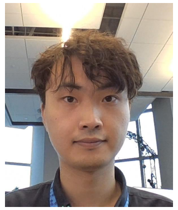
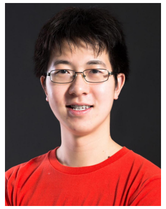
SHANGHAI JIAO TONG UNIVERSITY Microsoft 上海科技大学 ShanghaiTech University SJTU

EEG2Video: Towards Decoding Dynamic Visual Perception from EEG signals


Xuan-Hao Liu[1]*


Yan-Kai Liu[1]*


Yansen Wang[2#]


Kan Ren[3#]


Hanwen Shi[1]


Zilong Wang[2]


Dongsheng Li[2]


Bao-Liang Lu[1]


Wei-Long Zheng[1#]

[1] Shanghai Jiao Tong University

[2] Microsoft Research Asia

[3] ShanghaiTech University

NEURAL INFORMATION PROCESSING SYSTEMS

https://bcmi.sjtu.edu.cn/home/eeg2video

BCMI
Brain-like Computing & Machine Intelligence

Brain Decoding

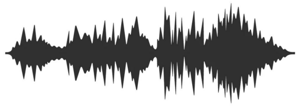
Image

Video

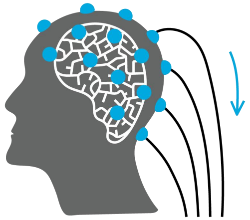
Language

“I love deep learning”

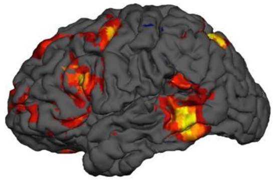
Audio



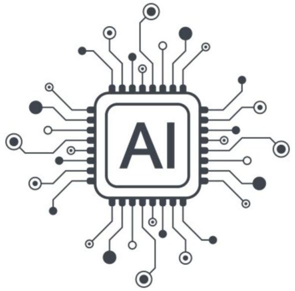
. . .



NeuroImaging Techniques



Brain Activity Recordings

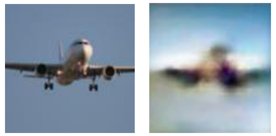


Previous works on Brain Decoding

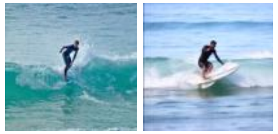
Image Reconstruction

2017
Kamitani
Proof of concept

2019
Beliy used **fMRI**
with self-supervision



2023
Scotti used **fMRI**
with retrieval + SD

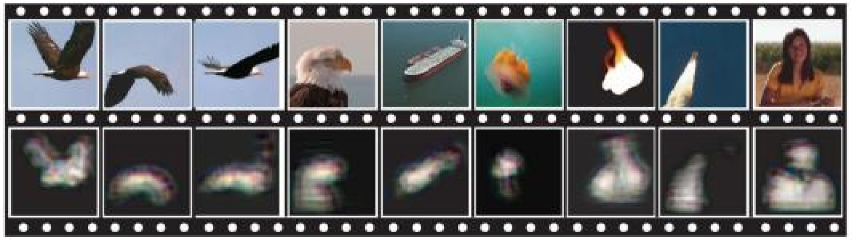


2023
Benchetrit used **MEG**
with SD

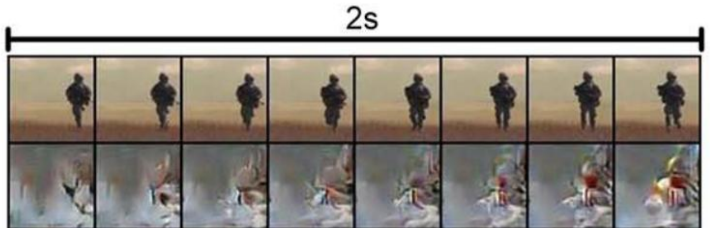
2024
Li used **EEG/MEG**
with retrieval + SD

- fMRI: functional magnetic resonance imaging
- MEG: magnetoencephalogram
- EEG: electroencephalography
- SD: stable diffsuion model
- GAN: generative adversarial network

Video Reconstruction



Wen et.al. 2018

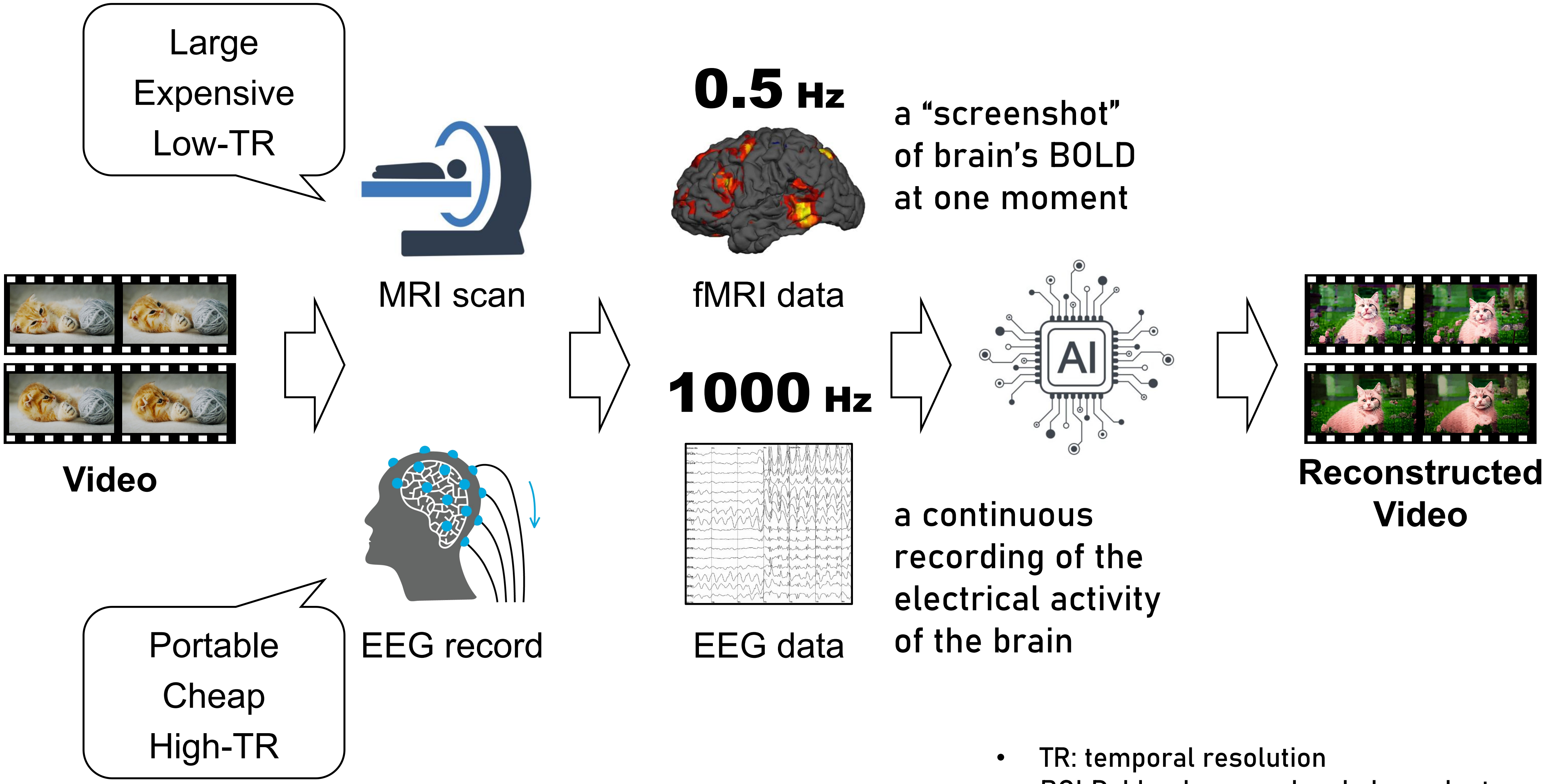


Wang et.al. 2022, GAN

Chen et.al. 2023, MinD-Video

[1] Horikawa T, Kamitani Y. Generic decoding of seen and imagined objects using hierarchical visual features[J]. Nature communications, 2017, 8(1): 15037.
[2] Beliy R, Gaziv G, Hoogi A, et al. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri[J]. Advances in Neural Information Processing Systems, 2019, 32.
[3] Scotti P, Banerjee A, Goode J, et al. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors[J]. Advances in Neural Information Processing Systems, 2024, 36.
[4] Benchetrit Y, Banville H, King J R. Brain decoding: toward real-time reconstruction of visual perception[C]//The Twelfth International Conference on Learning Representations.
[5] Li D, Wei C, Li S, et al. Visual Decoding and Reconstruction via EEG Embeddings with Guided Diffusion[J]. arXiv preprint arXiv:2403.07721, 2024.
[6] Wen H, Shi J, Zhang Y, et al. Neural encoding and decoding with deep learning for dynamic natural vision[J]. Cerebral cortex, 2018, 28(12): 4136-4160.
[7] Wang C, Yan H, Huang W, et al. Reconstructing rapid natural vision with fMRI-conditional video generative adversarial network[J]. Cerebral Cortex, 2022, 32(20): 4502-4511.
[8] Chen Z, Qing J, Zhou J H. Cinematic mindscapes: High-quality video reconstruction from brain activity[J]. Advances in Neural Information Processing Systems, 2023, 36.

Video Reconstruction from Brain Signals



- TR: temporal resolution
- BOLD: blood oxygen level-dependent

Challenges

Challenges of video reconstruction from EEG

- No suitable EEG dataset. How to build a such dataset?
- The EEG's decoding capability remains unclear. How to determine the decoding capability?
- EEG has low spatial resolution and signal-to-noise ratio. How to reconstruct videos from EEG?

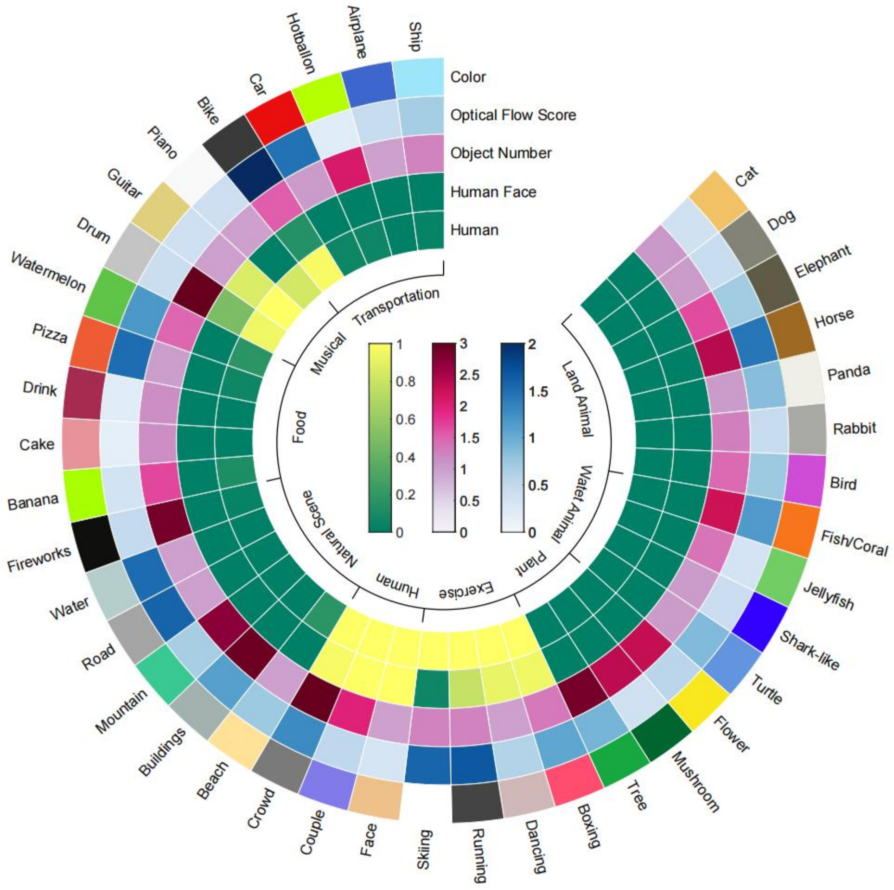
Video Stimuli Selection

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

To build the **S**JTU **EE**G **D**ataset for **D**ynamic **V**ision (SEED-DV), we eleborately select stimuli from 40 concepts across 9 coarser classes following the blow pinciples: *Land Animal, Water Animal, Plant, Exercise, Human, Nutural Scene, Food, Musical Instrument, Transportation.*

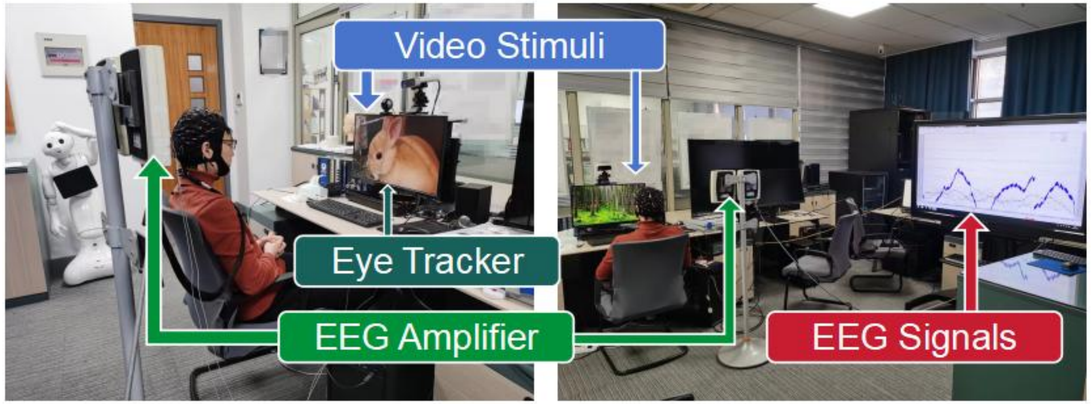
- We choose natural videos rather than artificial ones (like anime).
- We try to cover as diverse natural classes as posible.
- We would like to balance the numbers of the main colors.

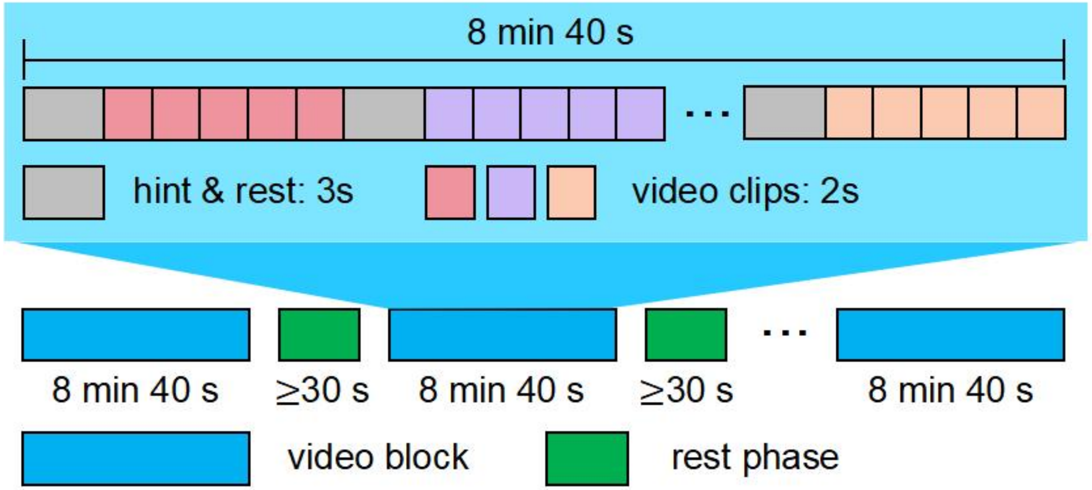


Expeirment Protocol

We recorded 20 subjects' EEG data while they were veiwing video stimuli. For each of 40 concept, 35 two-second video clips are collected from Internet.

- Subjects watched 7 video blocks in total. There is a rest phase between each two blocks.
- Each block includes 40 concepts, the order of these concepts is random across blocks.
- Subjects were first informed of the next concept, then watched 5 video clips of the informed concept.

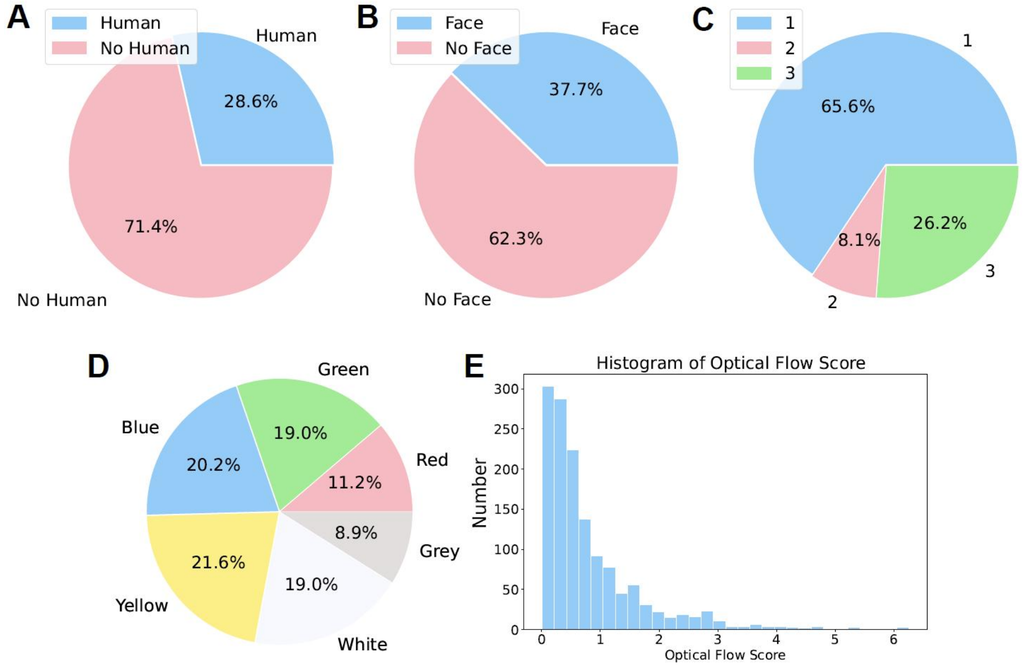




EEG-VP Benchmark

To investigate the EEG's decoding capability. We manually annotated some meta information to conduct the EEG-visual perception (EEG-VP) benchmark.

- Human: the appearance of humans: {*Yes, No* }.
- Face: the appearance of human faces: {*Yes, No* }.
- Number: the number of the main objects: {*One, Two, Many* }.
- Color: the color of the main objects: {*Blue, Green, Red, Grey, White, Yellow, Colorful* }.
- Opitical Flow Score: the optical flow score of the video.



EEG-VP Results

We evaluate a bunch of EEG models on the EEG-VP benchmark and conclude some findings:

- We **can** decode ***Categories*** information from EEG signals.
- We **can** decode ***Color*** information from EEG signals.
- We **can** decode ***Dynamic*** information from EEG signals.
- We **cannot** decode ***numbers, appearance of humans or faces*** from EEG signals.

Table 1: Average classification accuracy (%) and std across all subjects with different EEG classifiers on different tasks. Chance level is the percentage of the largest class. The star symbol (*) represents the result is above chance level with statistical significance (two-sample t-test: $p < 0.05$).

Methods	40-c top-1	40-c top-5	9-c top-1	9-c top-3	Color	Fast/Slow	Numbers	Human Face	Human
Chance level	2.50	12.50	11.11	33.33	20.57	50.00	65.64	62.25	71.43
Raw EEG Signals									
ShallowNet[62]	5.59/2.27*	16.93/4.66*	21.40/1.96*	49.62/2.34*	27.00/2.09*	56.62/1.77*	66.15/0.89	64.87/1.54	73.21/1.52
DeepNet[62]	4.56/1.52*	14.30/3.25*	20.27/1.25*	48.06/1.59*	26.37/1.95*	55.42/0.59*	65.71/0.24	61.58/3.93	72.86/0.40
EEGNet[58]	4.64/0.86*	14.25/1.87*	19.63/0.81*	47.04/1.45*	25.46/1.31*	51.99/2.00	64.67/0.60	61.37/1.31	72.38/0.98
Conformer[59]	4.93/1.57*	15.36/4.44*	20.92/0.98*	49.25/1.49*	27.53/1.37*	55.02/0.83*	65.73/0.26	64.96/1.14	73.00/0.85
TSConv[19]	4.92/0.99*	15.05/2.31*	20.00/1.01*	47.76/1.51*	26.89/1.83*	55.32/0.99*	65.39/0.41	64.39/1.47	72.68/0.67
GLMNet (Ours)	**6.20/3.02***	**17.75/4.24***	**21.93/1.87***	**50.01/2.52***	**27.33/1.45***	**57.35/1.98***	**66.21/0.91**	**65.10/1.45**	**73.34/1.31**
PSD Features									
SVM[63]	5.19/2.81*	-	19.02/3.27*	-	21.31/2.97	53.56/1.11*	64.15/1.22	58.94/2.21	70.91/1.84
MLP	6.20/3.02*	18.91/5.94*	21.59/3.00*	49.86/3.78*	22.02/3.27	55.15/1.20*	64.48/0.92	63.94/1.13	71.74/1.76
GLMNet (Ours)	**6.23/2.91***	**18.98/5.62***	**21.69/3.20***	**50.03/4.10***	**26.40/2.99***	**55.42/1.32***	**64.68/0.92**	**64.22/1.43**	**72.27/1.57**
DE Features									
SVM[63]	4.82/2.80*	-	19.05/3.39*	-	21.07/2.88	53.34/1.25*	63.62/1.73	57.82/3.50	70.25/1.94
MLP	6.12/3.08*	19.02/5.71*	21.17/3.24*	49.40/4.94*	25.91/3.27*	54.76/1.25*	64.10/0.70	63.41/1.57	71.74/1.76
GLMNet (Ours)	**6.16/3.18***	**19.12/6.07***	**21.34/3.34***	**49.55/4.57***	**26.15/3.24***	**55.06/1.20***	**64.25/0.74**	**63.63/1.80**	**72.27/1.58**

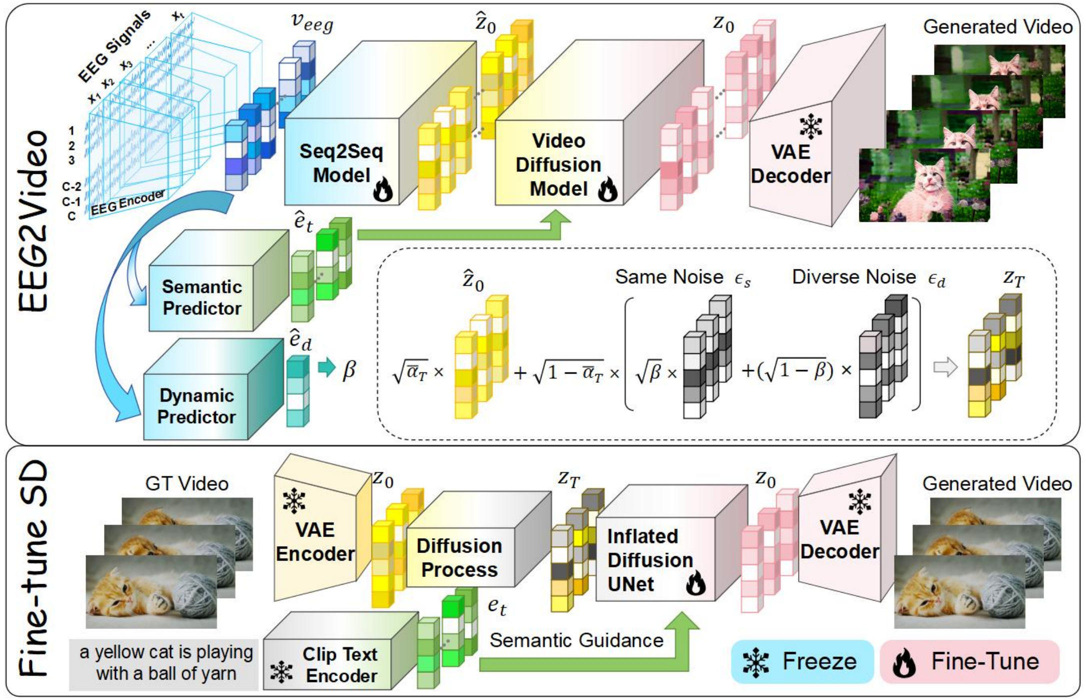
EEG2Video Framework

SHANGHAI JIAO TONG UNIVERSITY

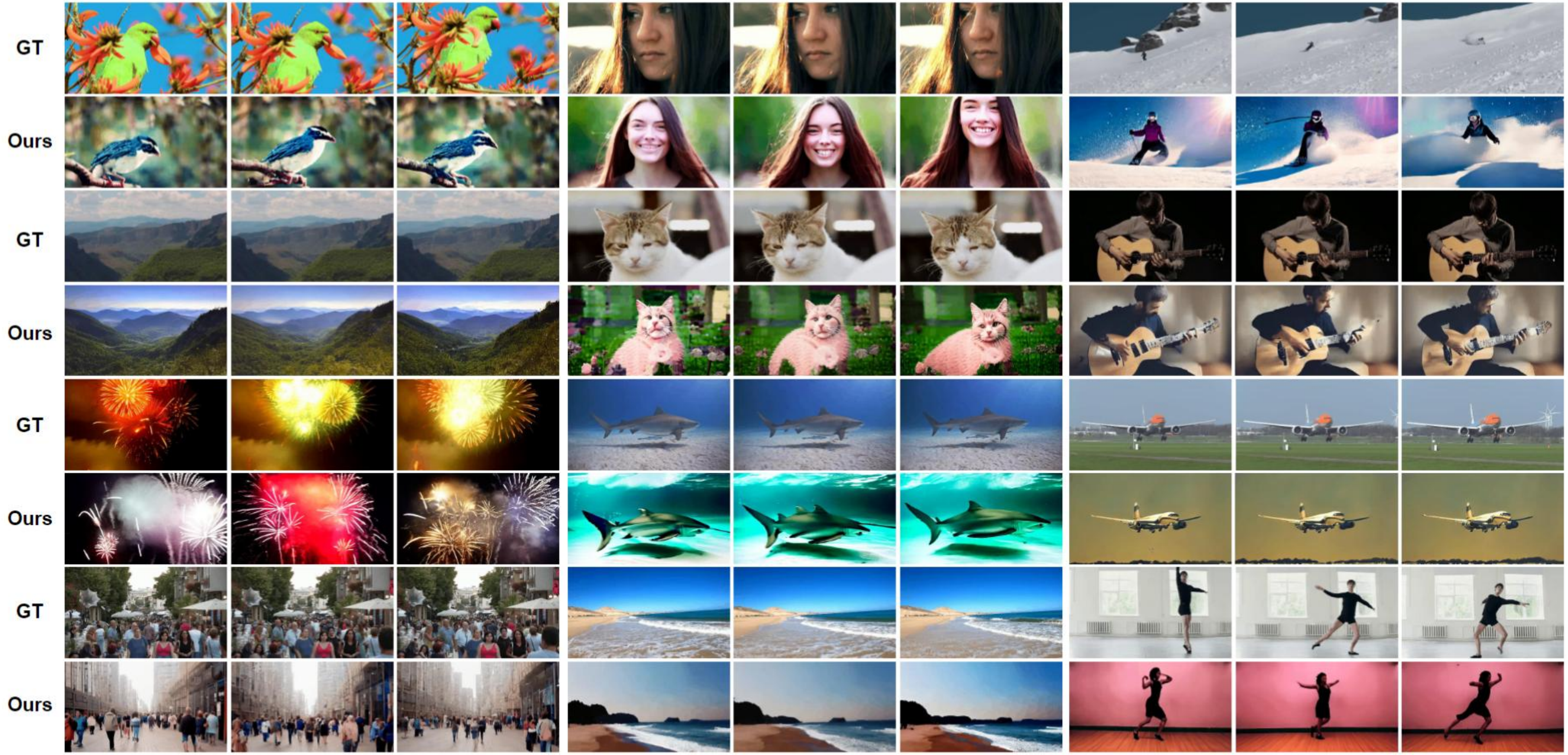
NEURAL INFORMATION PROCESSING SYSTEMS

In this paper, we propose EEG2Video, a pipeline for reconstructing videos from EEG signals. We design several modules based on the results on the EEG-VP benchmark to better decode videos.

- We use a **Seq2Seq model** for densly aligning EEG embeddings with low-level visual information.
- We use a **Semantic predictor** for aligning EEG embeddings with semantic information in the CLIP space.
- We design the dynamic-aware noise-adding (**DANA) modules** to introduce the fast/slow information into the diffusion process.
- We leverage the inflated diffusion models for decoding vivid videos.



Reconstruction Samples



Reconstruction Quantitative Results

- Several metrics across ***semantic-level*** and ***pixel-level*** are used to validate the effectiveness of our EEG2Video framework.
- We conduct the ablation study by removing the Seq2Seq module and the DANA process respectively, and we can see huge performance drop without either module.
- When dealing with smaller subset with less categories, the performance increases.

Table 2: Quantitative results of each methods on different size of subsets. Standard deviation is calculated across random seeds.

# Classes	Metrics	Video-based		Frame-based		
		Semantic-level		Semantic-level		Pixel-level
	Models	2-way	40-way	2-way	40-way	SSIM
10	Full Model	0.852±0.02	0.340±0.01	0.798±0.03	0.232±0.02	0.300±0.03
	w/o Seq2Seq	0.772±0.02	0.117±0.01	0.696±0.02	0.155±0.03	0.187±0.03
	w/o DANA	0.803±0.02	0.183±0.01	0.679±0.02	0.092±0.01	0.292±0.03
40	Full Model	0.798±0.03	0.159±0.01	0.774±0.02	0.138±0.01	0.256±0.03
	w/o Seq2Seq	0.786±0.03	0.113±0.01	0.734±0.02	0.112±0.01	0.189±0.03
	w/o DANA	0.770±0.02	0.128±0.01	0.732±0.03	0.109±0.03	0.217±0.02

NeuroScience Findings

To find electrodes or brain areas most associated with dynamic visual perception, we conduct a one-channel classification task to test the classification quality of each electrode.

- Figure 4(A) shows that the electrodes in the occipital area have higher accuracy on Human/Animal tasks.
- Figure 4(B) reveals that the brain area associated to movements are around the temporal region where the sensory and motor cortex lies.
- Removing occipital region significantly damages the performance ($p < 0.01$).

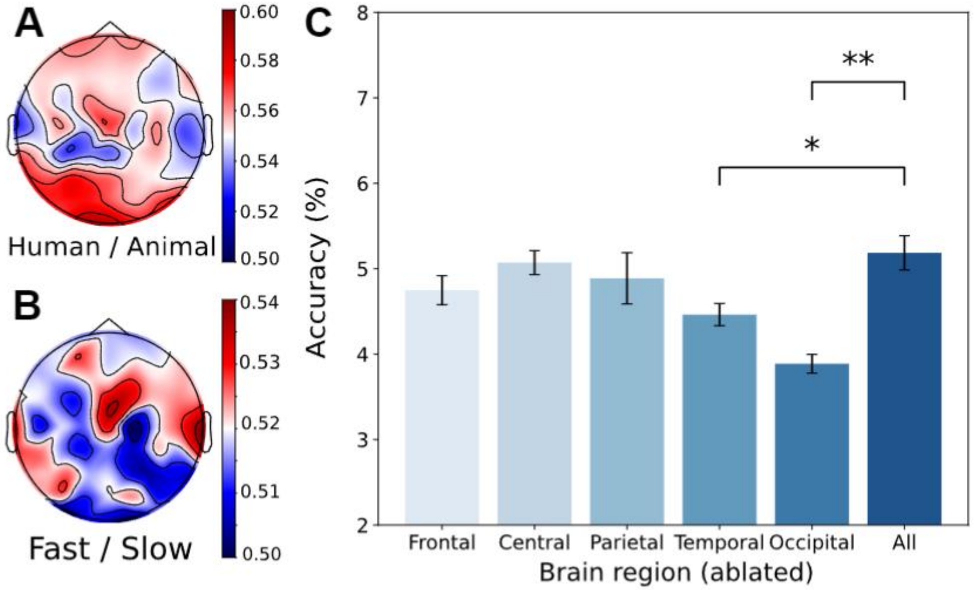
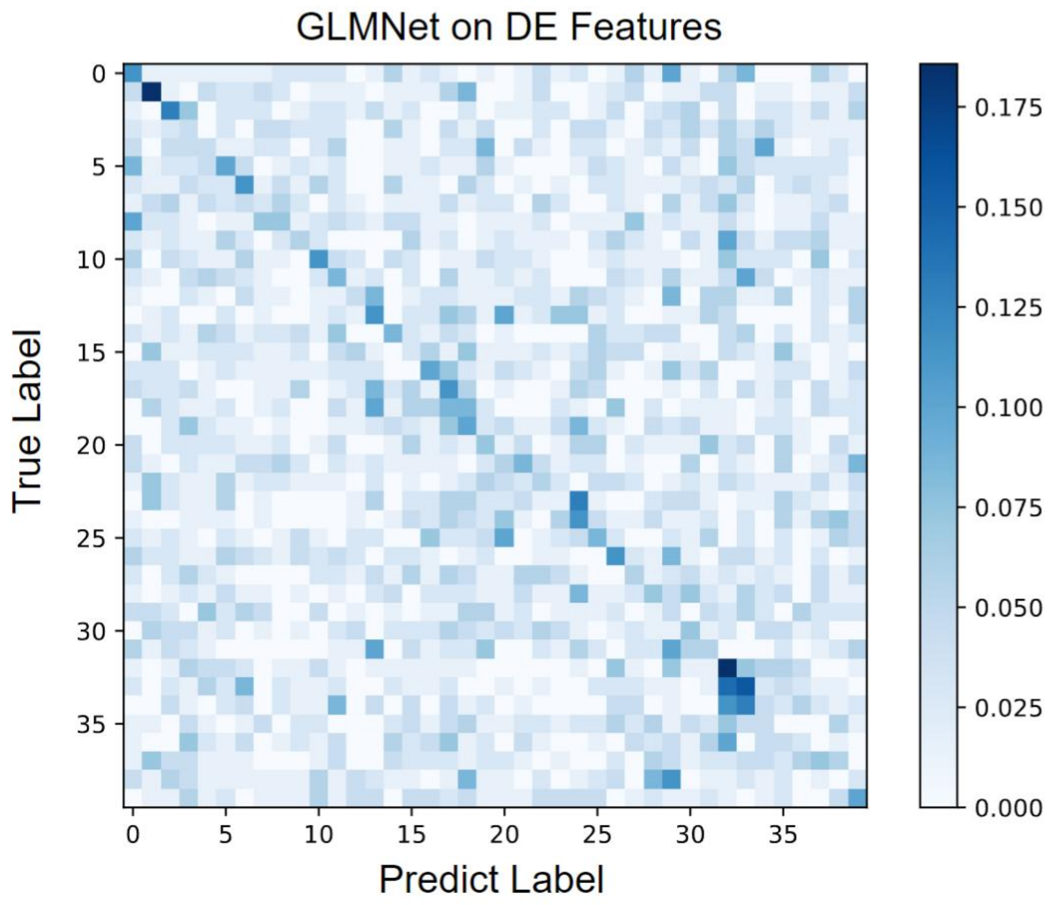


Figure 4: Spatial Analysis. (A-B). Topographies of each electrode's accuracy for Human/Animal and Fast/Slow tasks. (C). Ablate electrodes of different brain regions.

NeuroScience Findings

We plot the confusion matrices of GLMNet on the 40-class task.

- It can be seen that there is a faint diagonal lines.
- Moreover, a small square in the right bottom corner is being observed, of which categories are {Drum, Guitar, and Piano} (32 - 34 class). The musical instruments stimulate the auditory cortex in our brains with these visual cues.



Thanks

Contact us:

haogram_sjtu@sjtu.edu.cn
yansenwang@microsoft.com
renkan@shanghaitech.edu.cn
weilong@sjtu.edu.cn

https://bcmi.sjtu.edu.cn/home/eeg2video