NEURAL INFORMATION PROCESSING SYSTEMS

2026 Oral Presentation

MindAlign: Bridging EEG, Vision, and Language for Zero-Shot Visual Decoding

Zexuan Chen[1], Sichao Liu[1,2,3*], Runhao Iu[2,4], Huichao Qi[5], Alexandra Woolgar[2], Xi Vincent Wang[1], Lihui Wang[1]

[1] KTH [2] University of Cambridge [3] EPFL [4] McGill University [5] Karolinska Institutet

- Presented by: Zexuan Chen

December 08, 2026

Motivation and Contributions

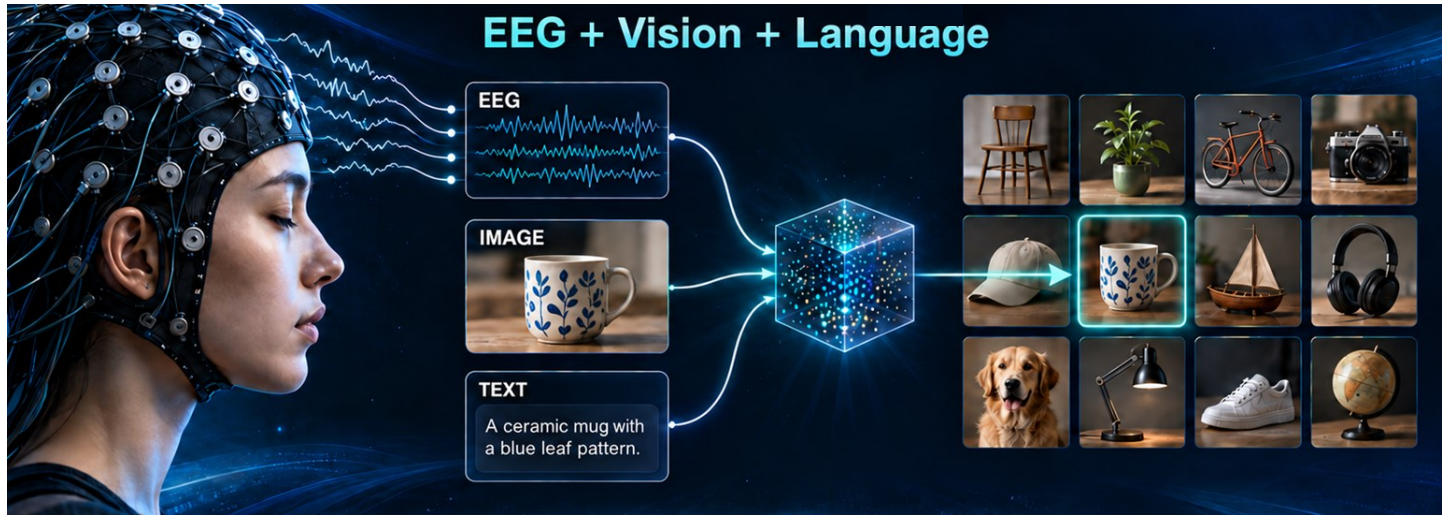
EEG-based visual decoding aims to identify visual stimuli from non-invasive brain signals, with potential applications in neuroscience and BCIs.

➢ **Challenges & Research Gap:**

1. Low **signal-to-noise ratio (SNR)** and **inter-subject variability** hinder robust learning of EEG representations.
2. Rich **semantic information** remains underexplored beyond **pairwise EEG–image alignment**.

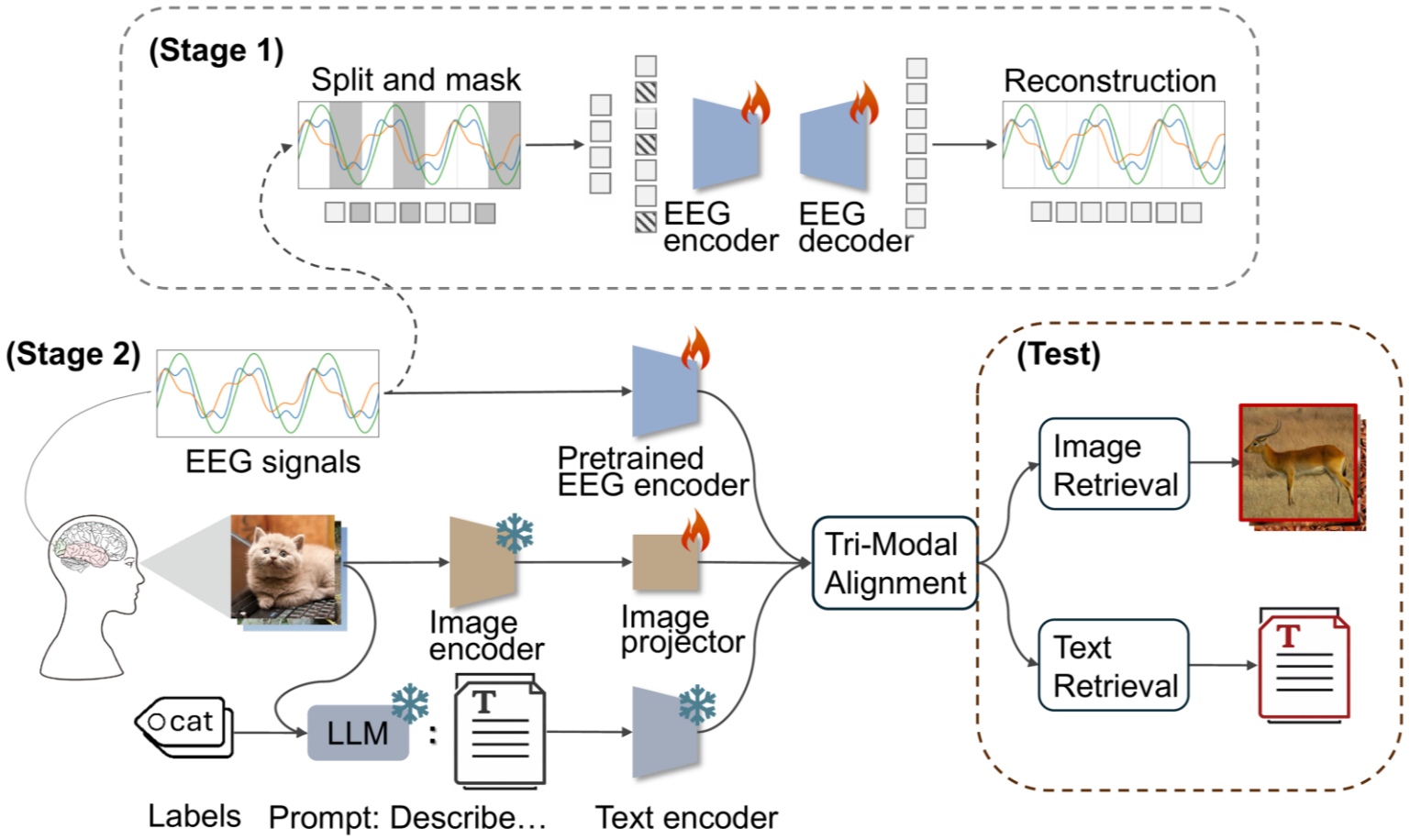
➢ **Contributions:**

1. Self-supervised EEG Pretraining: Learn robust neuralrepresentations throughmasked signal reconstruction.
2. Tri-modal alignment: Incorporates LLM-generated image descriptions to enrich semantic supervision.
3. High-performance EEG encoder design.



Achieves 54.1% Top-1 / 83.4% Top-5 accuracy on 200-way zero-shot image retrieval in Things-EEG2.

Two stage training strategy



Stage 1

Masked EEG pre-training

The EEG encoder is pretrained by masked reconstruction. Discards the decoder and transfers encoder weights into Stage 2.

Stage 2

Tri-Modal alignment

The pretrained EEG encoder is jointly trained with image projector. Contrastive losses align all three modalities in a shared space.

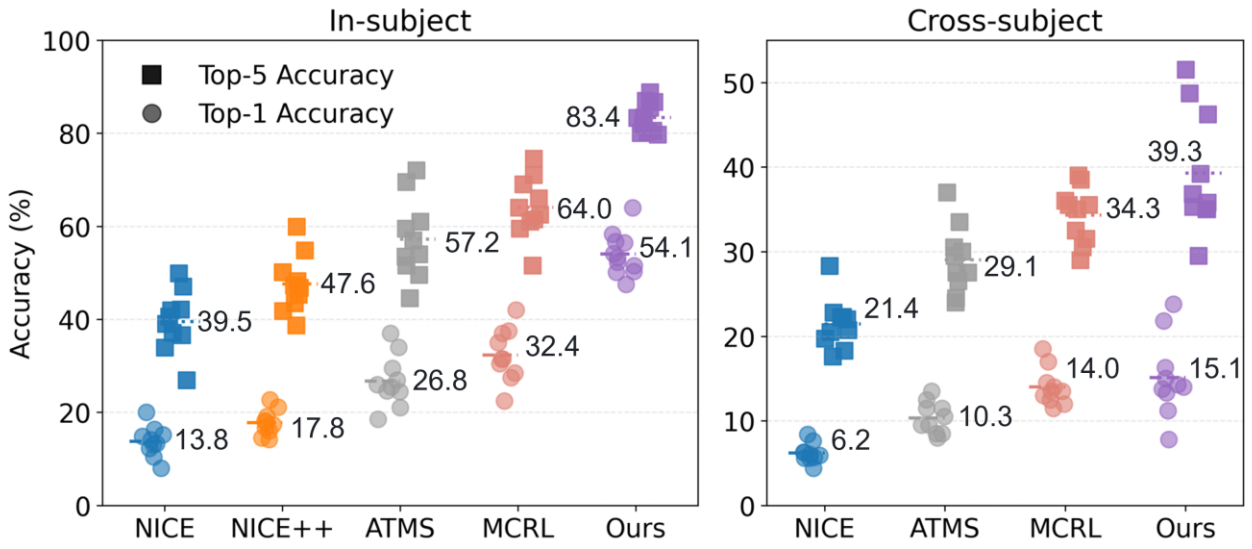
Results (Performance against SOTA methods)

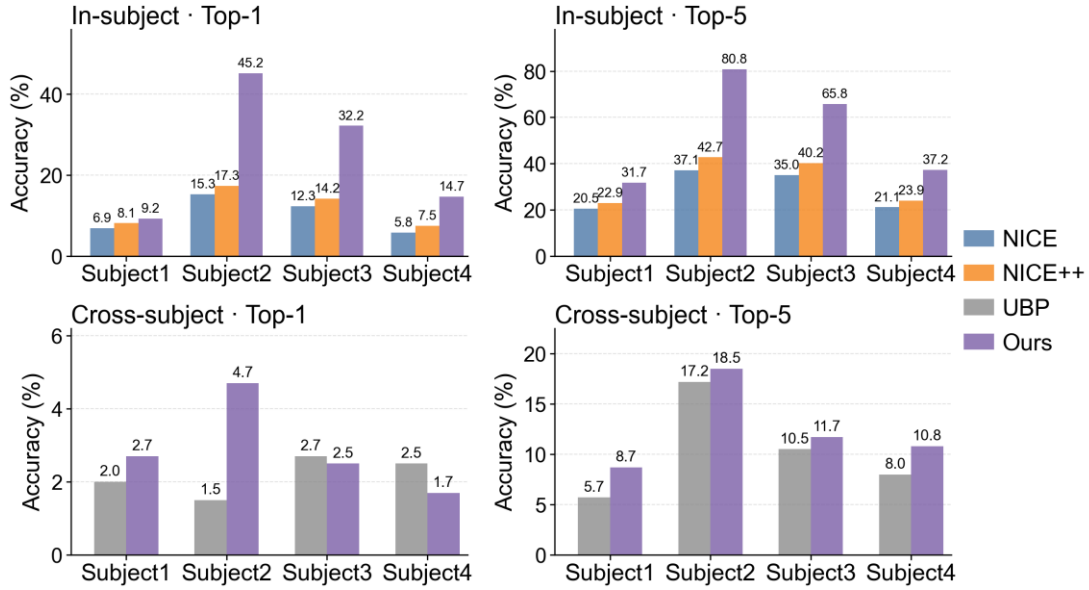
Things-EEG2 benchmark

63-channel EEG from 10 participants. Train on 1,654 concepts and test on 200 disjoint concepts under 200-way zero-shot retrieval.

Things-MEG validation

271-channel MEG from 4 participants across 1,854 concepts. Used as supporting evidence that gains generalize beyond EEG.

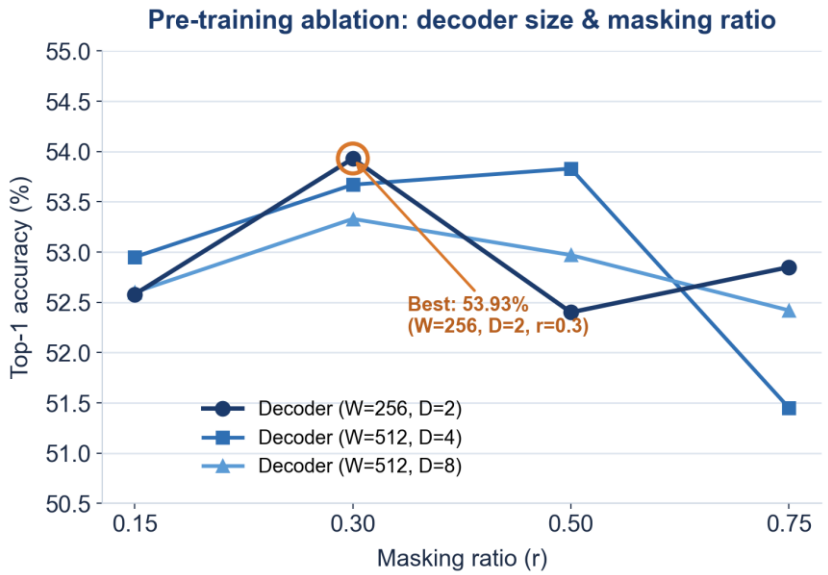




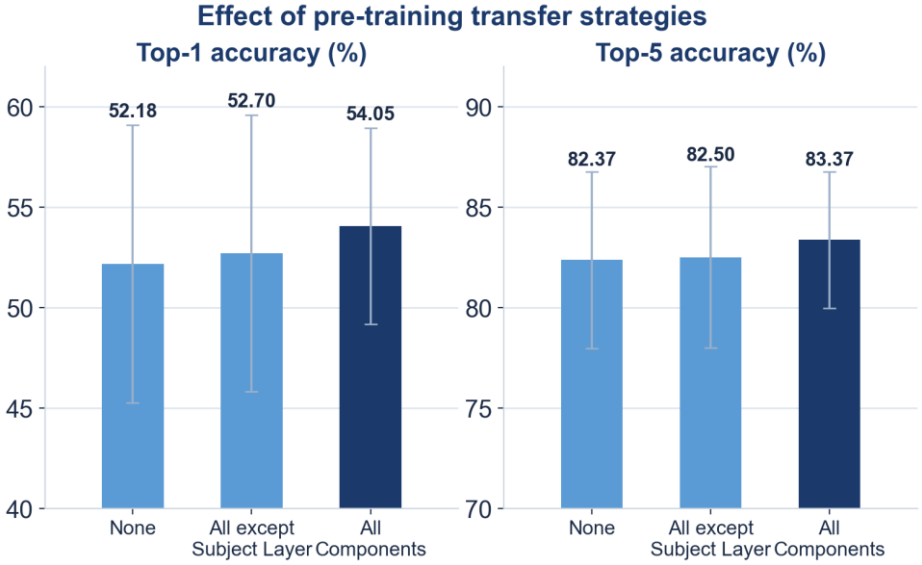
In-subject: trained and tested on data from the same participant. **Cross-subject**: LOSO (leave one subject out).

Results (Ablation study)

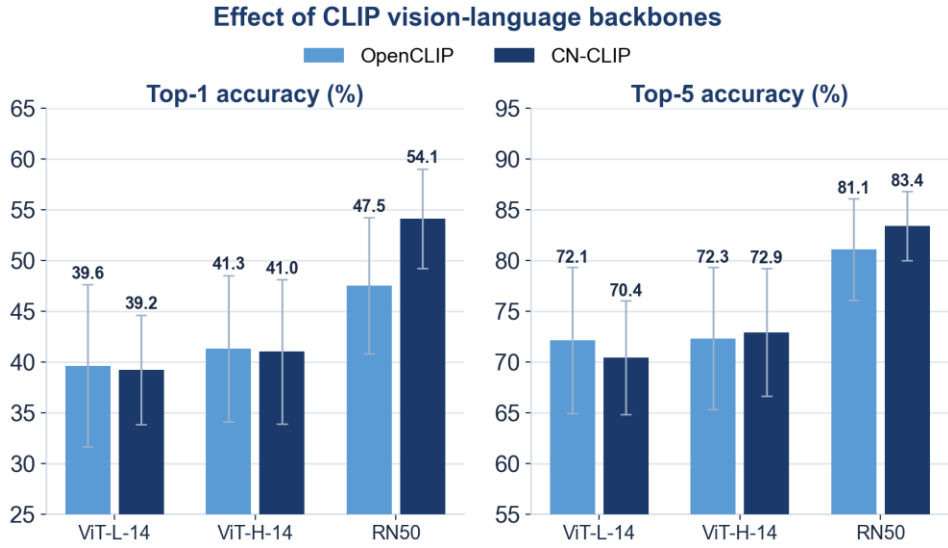
(1a)



(1b)



(2a)



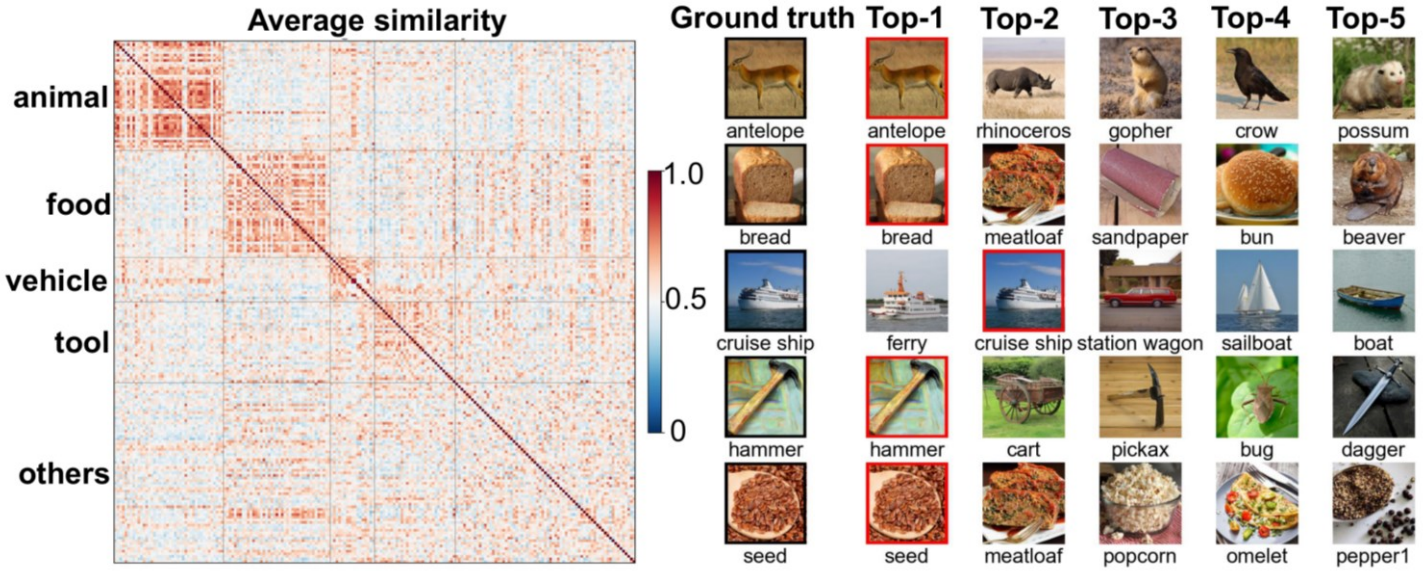
(2b)

Effect of the LLM chosen and alignment weight α

$$L_{\text{total}} = (1-\alpha)\,L_{\text{EI}} + \alpha\,L_{\text{IT}}$$

Method / α	Top-1	Top-3	Top-5
α = 0	52.80 ± 7.85	74.02 ± 6.31	81.87 ± 5.79
Label-only, α = 0.1	53.00 ± 7.44	73.92 ± 5.78	82.42 ± 4.90
LLaVA, α = 0.1	53.02 ± 7.31	74.03 ± 5.66	82.05 ± 5.29
Qwen, α = 0.1	**54.05 ± 4.87**	**75.73 ± 4.37**	**83.37 ± 3.39**
Qwen, α = 0.1 (Label-free)	53.02 ± 6.62	74.17 ± 5.00	82.45 ± 4.44
Qwen, α = 0.3	53.18 ± 6.53	74.60 ± 5.53	83.23 ± 4.47
Qwen, α = 0.5	49.93 ± 6.64	71.92 ± 5.62	80.68 ± 4.79

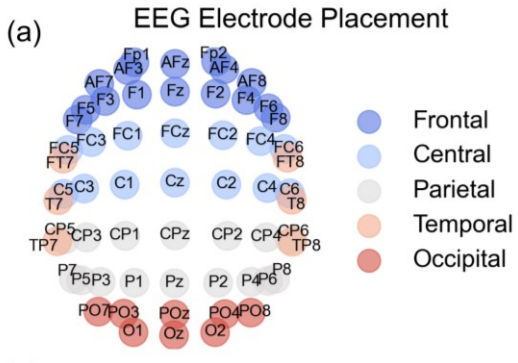
Neural Dynamics & Semantic Analysis

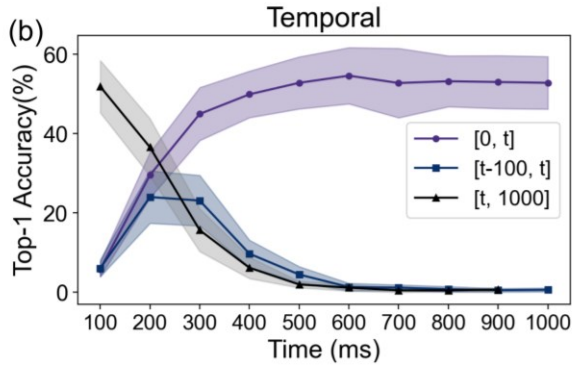


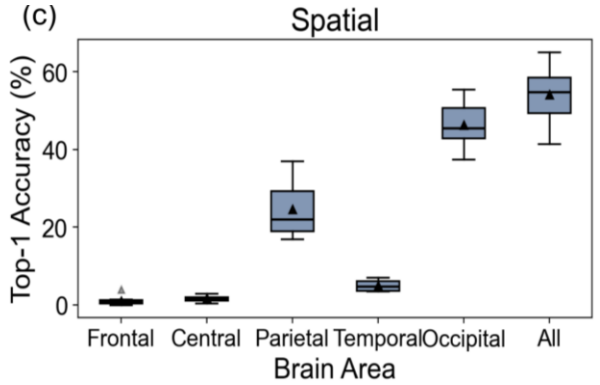
Left: Representational similarity analysis (RSA) on the learned EEG features.

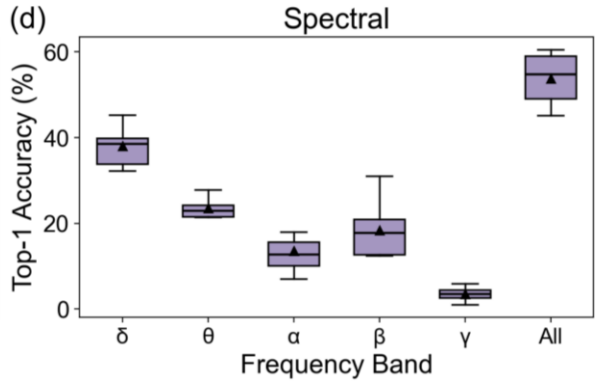
Right: Top-1 to Top-5 image retrieval results per category.

Temporal, spatial, and spectral analyses on EEG.









Takeaways & Future work

Self-supervised pretraining

Pretrain the EEG encoder with a masked reconstruction objective. Encourage the encoder to learn intrinsic spatio-temporal regularities of EEG signals

Multimodal Alignment

Jointly align EEG, image, and text representations within the shared embedding space. Enrich the supervisory signal itself with LLM-generated text.

Performance Improvements

Achieves substantial gains in decoding accuracy and cross-subject robustness Things-MEG results also suggest broader modality potential.

Limitations and next research directions

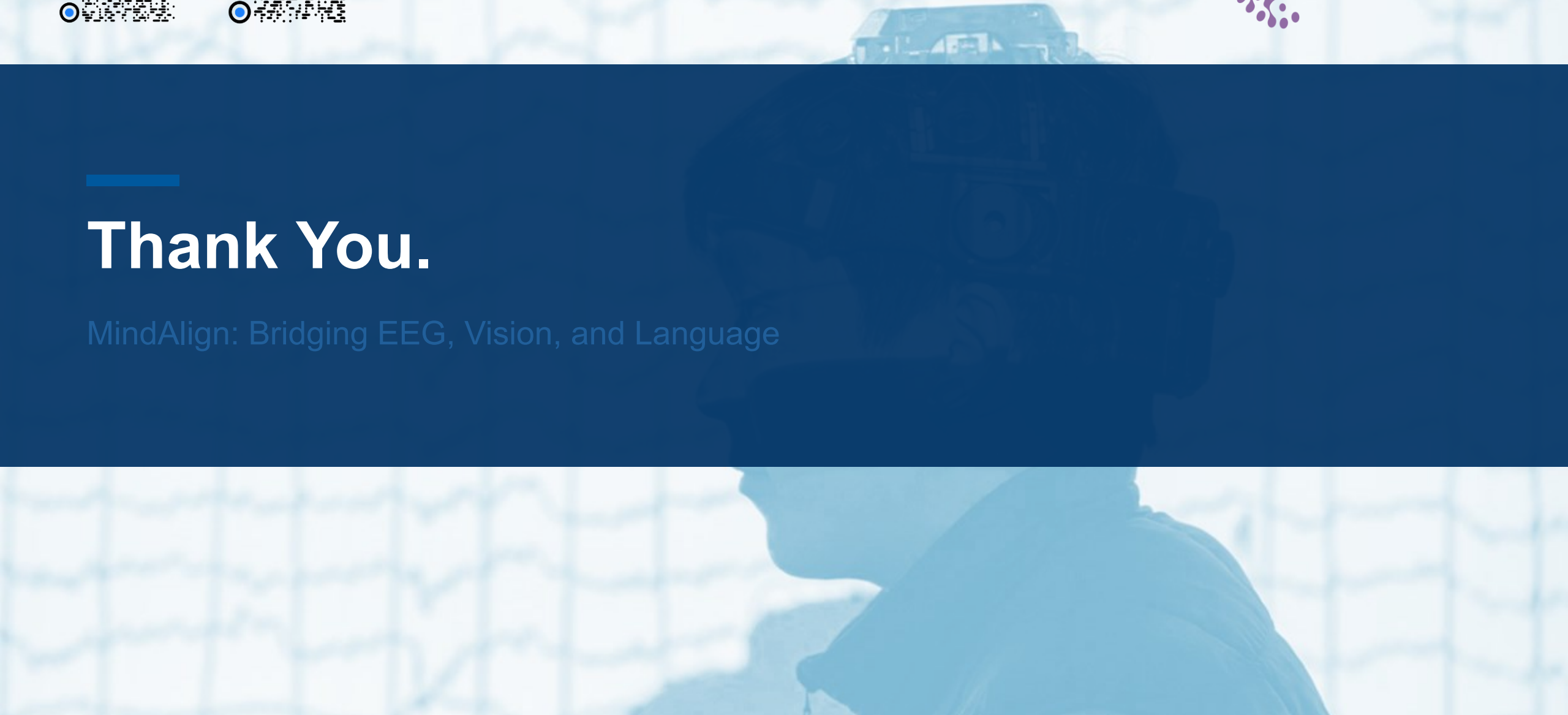
The main limitation is that cross-subject Top-1 accuracy remains substantially lower than in-subject performance, highlighting the need to further investigate inter-subject variability.

Future work will focus on extending the framework to other brain signals beyond EEG and exploring generative reconstruction tasks, such as diffusion-based image synthesis.

Code:

Paper:

NEURAL INFORMATION PROCESSING SYSTEMS



Thank You.

MindAlign: Bridging EEG, Vision, and Language