Revealing Vision-Language Integration in the Brain with Multimodal Networks

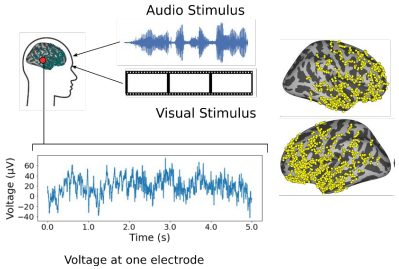
Vighnesh Subramaniam*, Colin Conwell, Christopher Wang, Gabriel Kreiman, Boris Katz, Ignacio Cases, Andrei Barbu

Summary: We use multimodal networks of vision and language to identify areas of vision-language integration in the brain.

Motivation: Little is known about multimodal processing in the brain, particularly vision and language.

We apply ridge regression to fit activity in the brain using representations. We find areas where multimodal networks are better than unimodal networks.

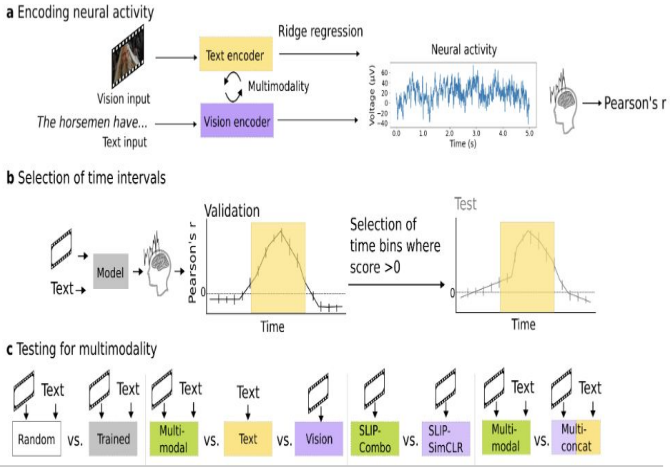
Data: 7 subjects, 7 movies, 1006 sEEG electrodes (yellow)



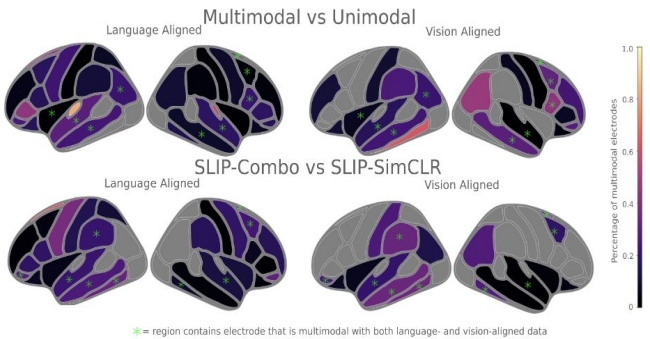
Networks:

- Multimodal: CLIP, SLIP, ALBEF, BLIP, Flava
- Language: SBERT, SimCSE
- Vision: ConvNeXt, SimCLR, BEIT

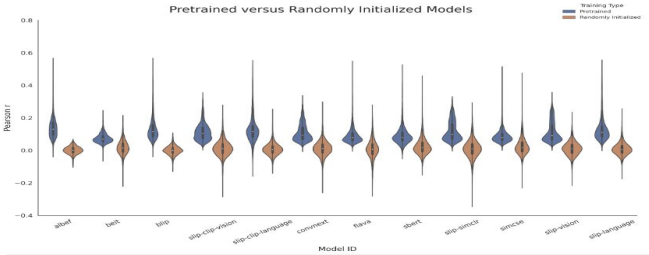
Overview



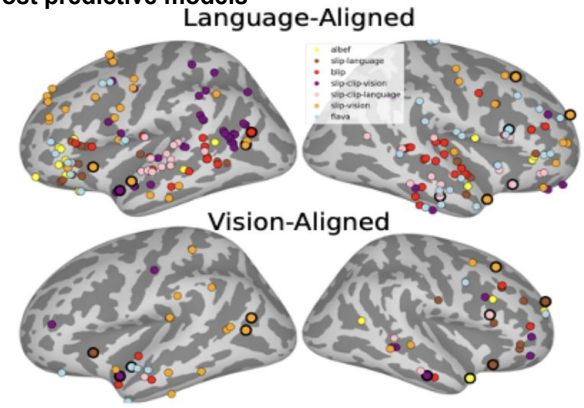
Multimodal Regions



Random vs Trained



Most predictive models



Future work:

More modalities: Audio? Touch?
Deeper explanation for CLIP/SLIP success?
Temporal Processing -- exploration of multimodal integration across time?

* *Correspondence: vsub851@mit.edu*