



**ICLR**  
International Conference On  
Learning Representations

2025

# Multi-modal brain encoding models for multi-modal stimuli

Subba Reddy Oota

Maneesh Singh

Khushbu Pahwa

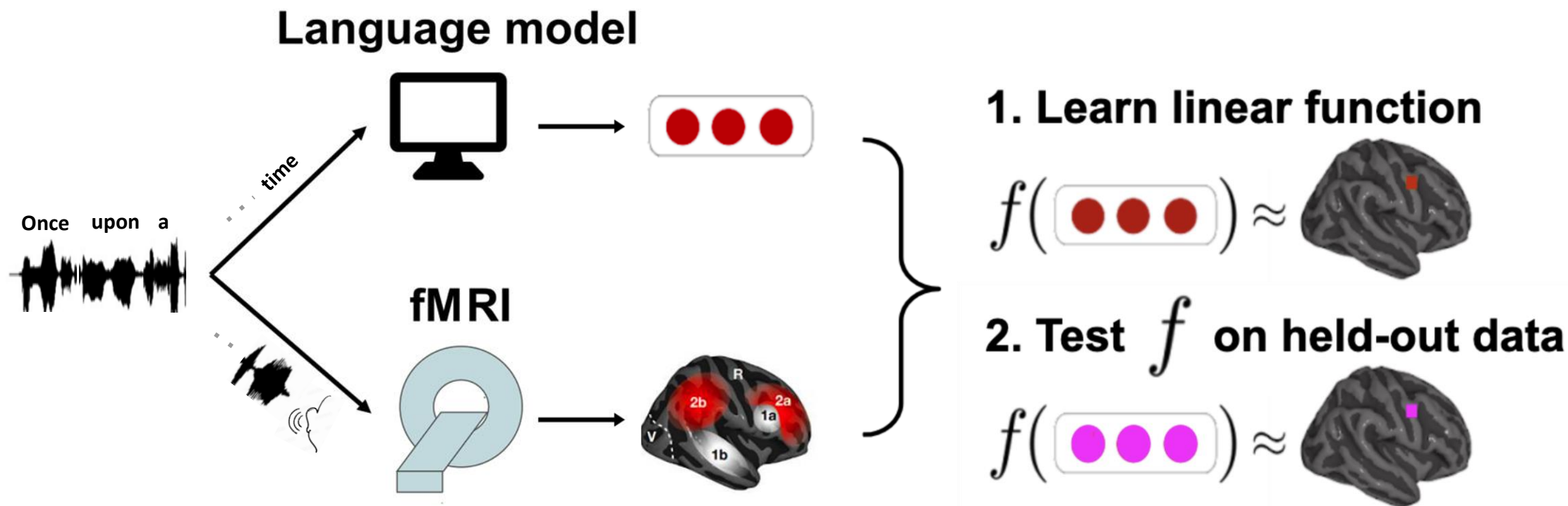
Manish Gupta

Mounika Marreddy

Bapi S. Raju

[subba.reddy.oota@tu-berlin.de](mailto:subba.reddy.oota@tu-berlin.de)

# Language models (LMs) predict brain activity evoked by complex language (e.g. listening a story) to an impressive degree



Brain alignment of an LM  $\Rightarrow$  how similar its representations are to a human brain

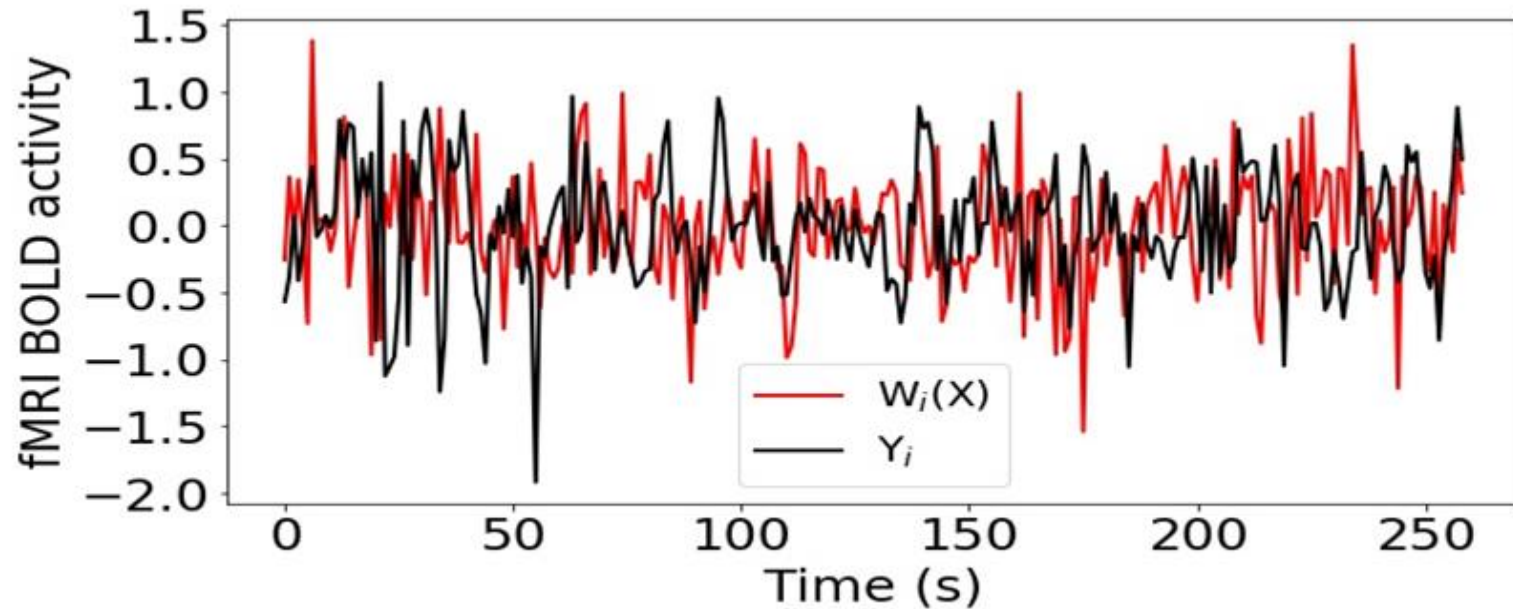
Wehbe et al. 2014,  
Jain and Huth 2018,  
Gauthier and Levy 2019

Toneva and Wehbe 2019,  
Caucheteux et al. 2020,  
Toneva et al. 2020

Jain et al. 2020,  
Schrimpf et al. 2021,  
Goldstein et al. 2022

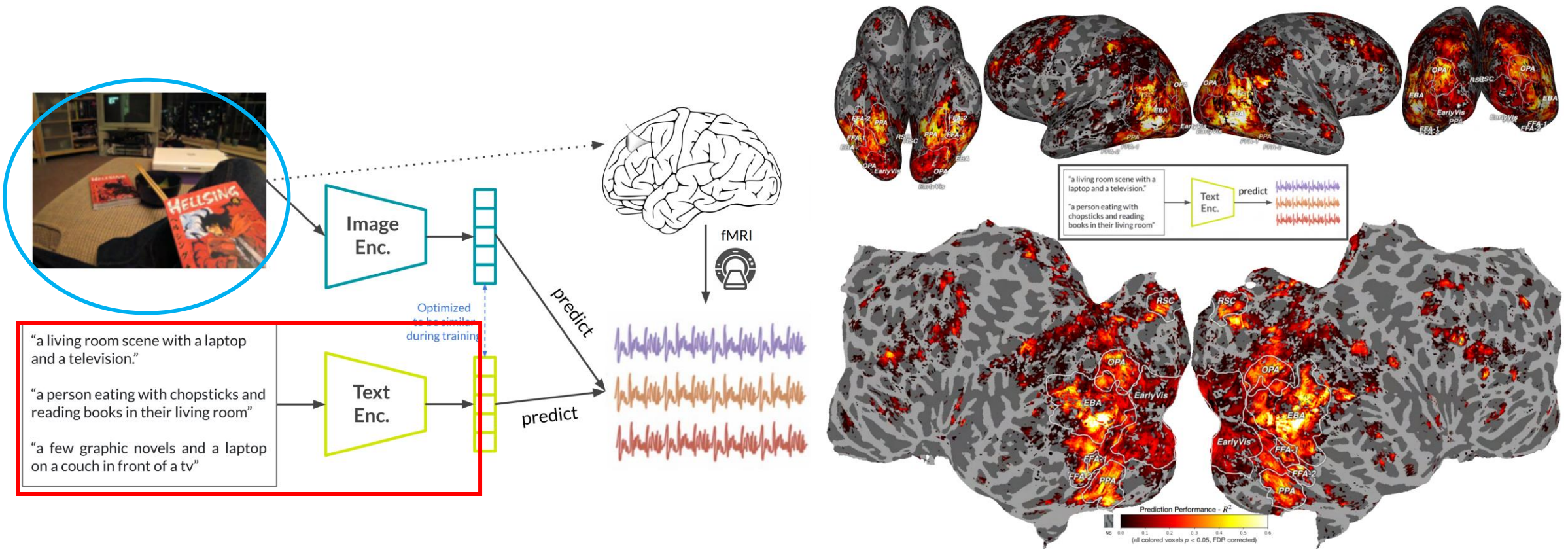
...

# Language models (LMs) predict brain activity evoked by complex language (e.g. listening a story) to an impressive degree



$$\text{brain alignment}_i = \text{Pearson corr}(\text{true } v_i, \text{pred } v_i)$$

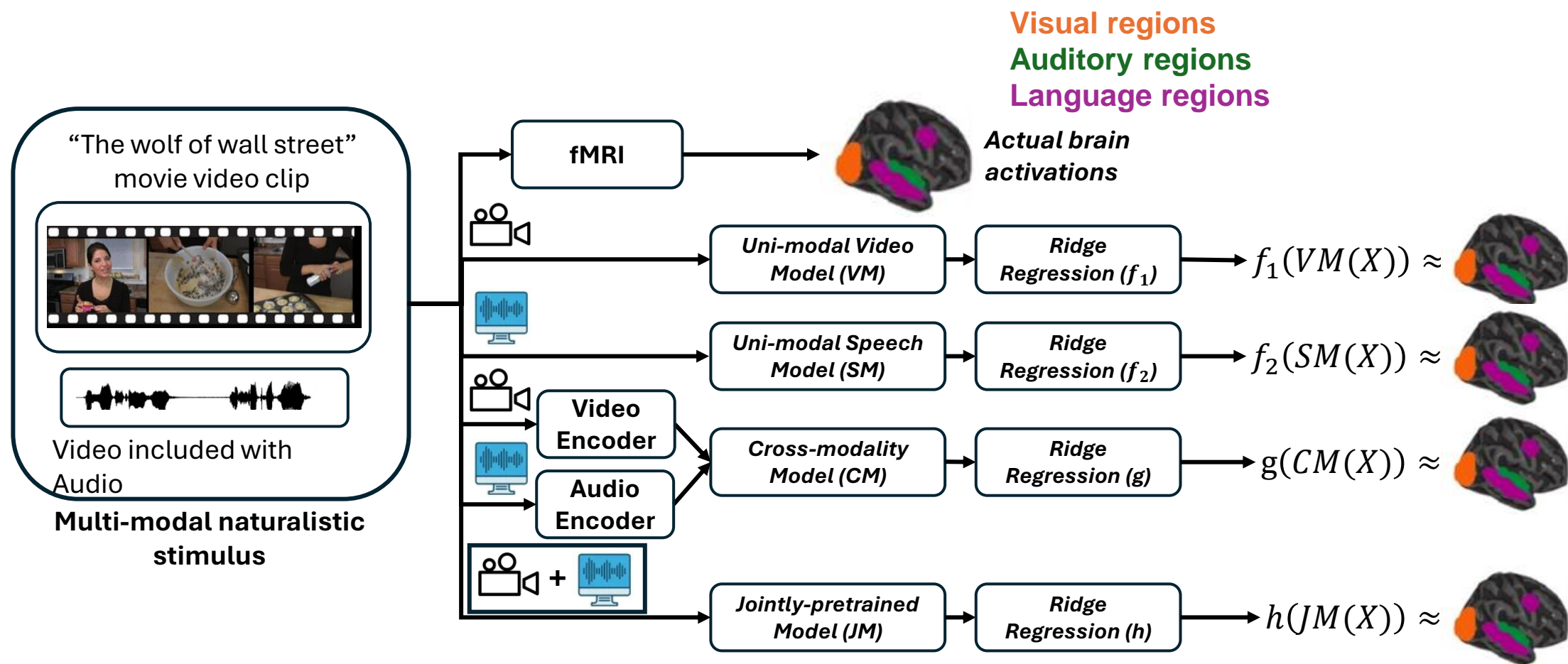
# Multi-modal Transformer models can predict visual brain activity impressively well, even with text modality representations



How accurately do multi-modal models predict brain activity evoked by multi-modal stimuli?



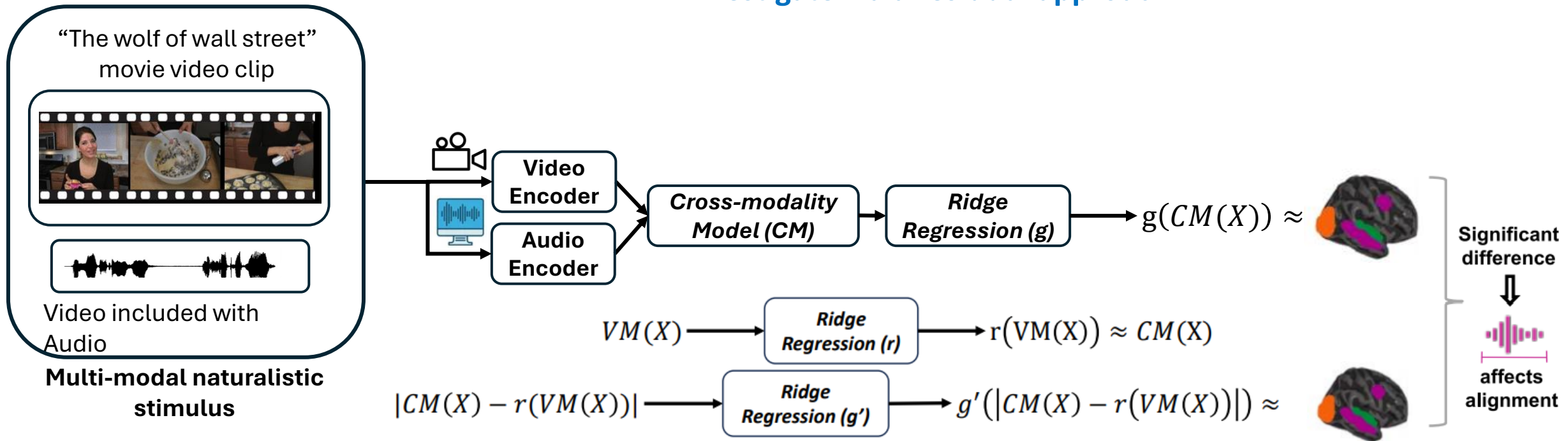
# Multi-modal vs. Unimodal models: Brain alignment



- How well do multi-modal models predict multi-modal stimulus-evoked brain activity over unimodal models?
- How our brains separates and integrates information across modalities through a hierarchy of early sensory regions to higher cognition (language regions)?

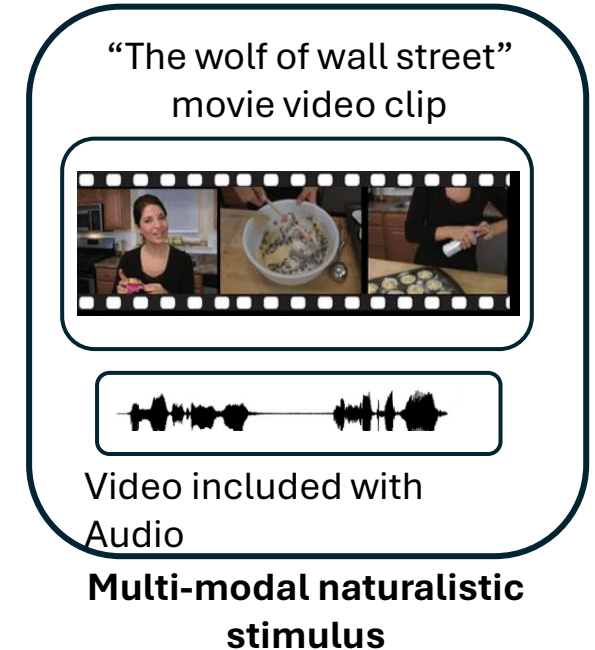
# Which modality of representations in multi-modal models leads to high brain alignment?

Investigate via a residual approach



# Datasets & Models

- Brain: fMRI recordings from NeuroMod Movie10 [St-Laurent et al. 2023]
  - Passively watching 4 movies
  - N=6
- 3 unimodal video-based Transformer models
  - VideoMAE
  - ViViT
  - ViT-H
- 2 unimodal Audio-based Transformer models
  - Wav2Vec2.0
  - AST
- 2 multi-modal Transformer models
  - Cross-modal model (ImageBind)
  - Jointly-pretrained model (TVLT)

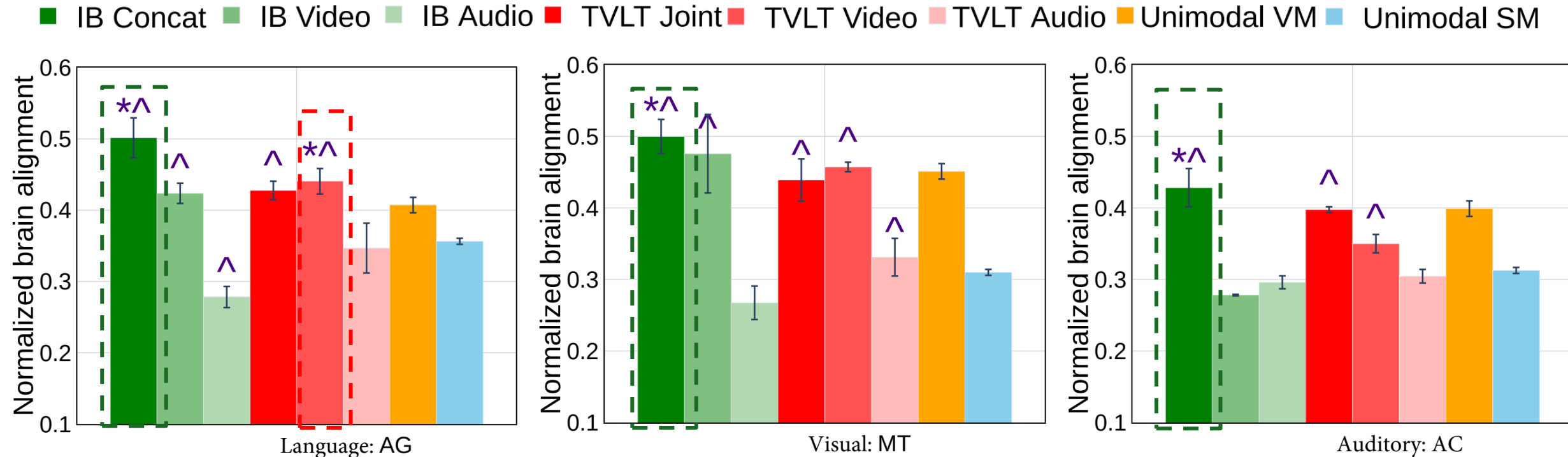


To quantify model predictions, we have an estimate of the explainable variance and use that to measure normalize brain alignment.

**Multi-modal stimulus: How do multi-modal and unimodal models differ in their ability to predict brain activity in late language regions, higher visual regions and early sensory regions?**

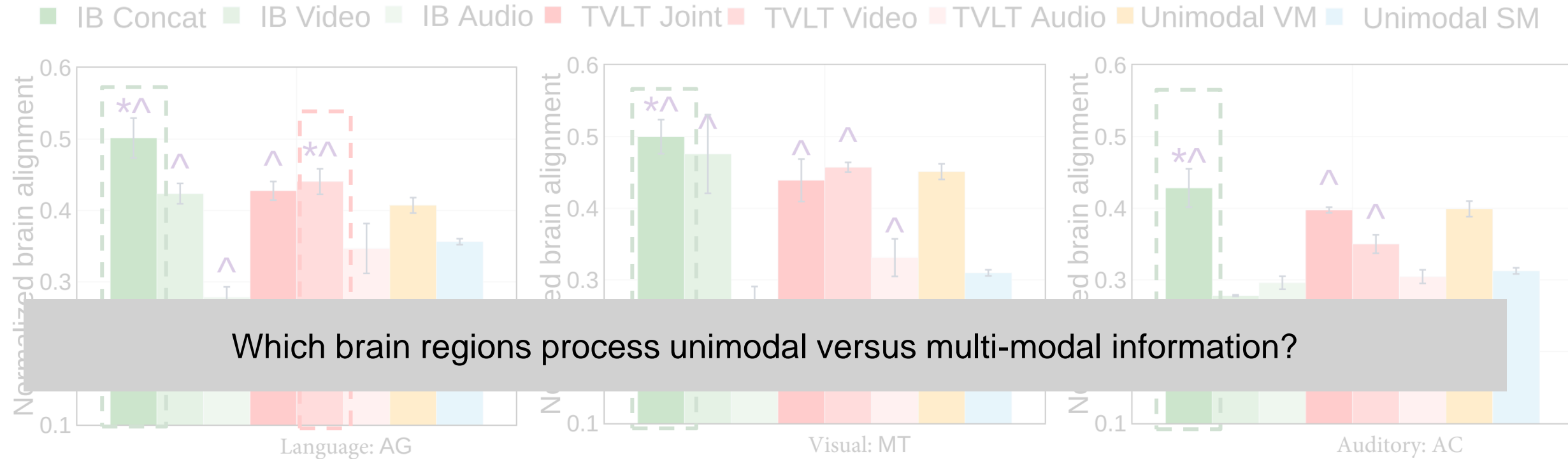


# Result-1: Multi-modal vs. Unimodal models & brain alignment



- **Language region (AG)**
  - Both types of multi-modal models show higher brain alignment than unimodal video and speech models with **language regions (PTL, IFG)**, but **audio models** trail behind **video models**.
- **Higher-visual (MT) and Early-sensory (AC)**
  - **Cross-Modal Models**: Concat embeddings improve alignment, while **jointly-pretrained models** perform similar to unimodal video models.
  - In AC, surprisingly, **unimodal video models** show improved brain alignment over **unimodal speech models**.

# Result-1: Multi-modal vs. Unimodal models & brain alignment



- **Language region (AG)**
  - Both types of multi-modal models show higher brain alignment than unimodal video and speech models with language regions (PTL, IFG), but **audio models** trail behind **video models**.
- **Higher-visual (MT) and Early-sensory (AC)**
  - **Cross-Modal Models:** Concat embeddings improve alignment, while **jointly-pretrained models** perform similar to unimodal video models.
  - In AC, surprisingly, **unimodal video models** show improved brain alignment over **unimodal speech models**.

## Result-2: Which brain regions process unimodal versus multi-modal information?

- **Multi-modal models**

- Improved alignment in **semantic language regions** such as **AG, PCC and dmPFC**
- Improved alignment in **syntactic language regions** such as **PTL and IFG**

- **Multi-modal and Unimodal models**

- Both models exhibit similar brain alignment in several language regions such as ATL, IFGOrb and MFG, visual regions such as PPA (scene visual area), and EVC (early visual cortex).

- **Multi-modal effects:** In general, multimodal models have better predictivity in the language regions
- **Unimodal effects:** Unimodal models have higher predictivity in the early sensory regions (visual and auditory)

## Result-2: Which brain regions process unimodal versus multi-modal information?

- **Multi-modal models**

- Improved alignment in **semantic language regions** such as **AG, PCC** and **dmPFC**
- Improved alignment in **syntactic language regions** such as **PTL** and **IFG**

- **Multi-modal & Unimodal models**

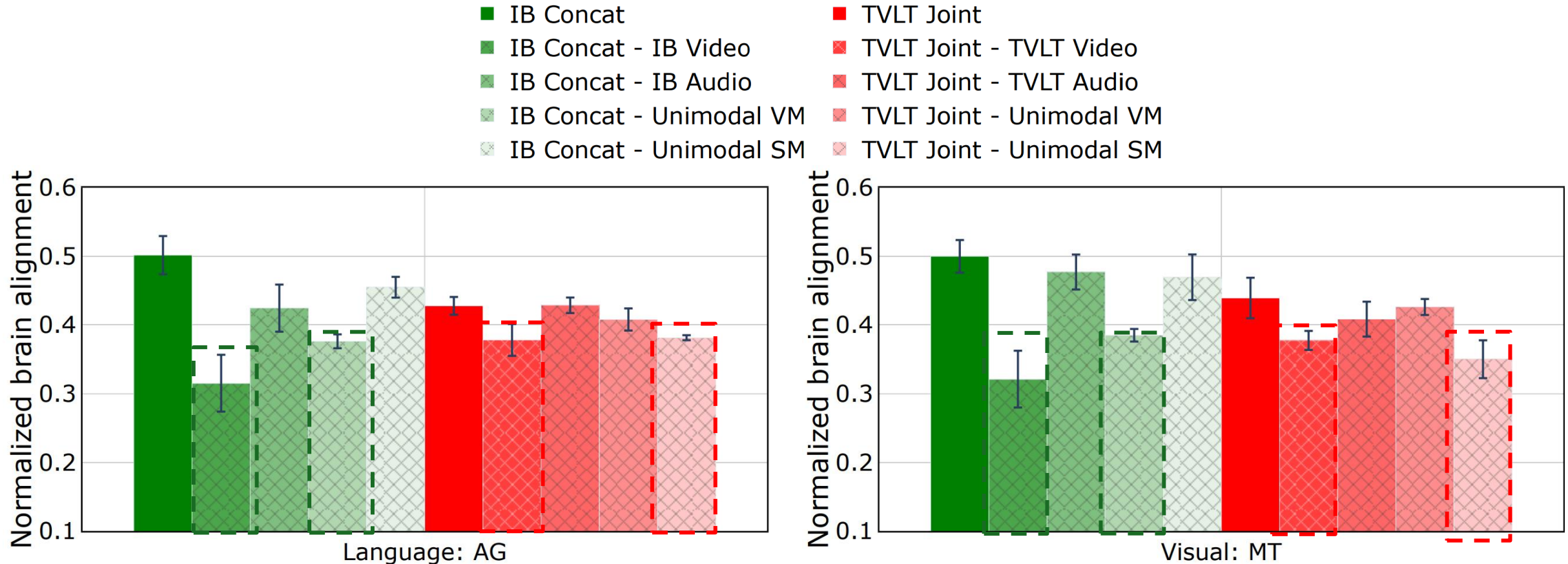
- Both models exhibit similar brain alignment in several language regions such as **ATL, IFGOrb**

Which modality of representations in multi-modal models lead to high brain alignment?

- **Multi-modal effects:** In general, multimodal models have better predictivity in the language regions
- **Unimodal effects:** Unimodal models have higher predictivity in the early sensory regions (visual and auditory)

**How is the alignment between brain recordings and multi-modal model representations affected by the elimination of modality-specific features?**

# Result-3: Modality-specific contribution in language and visual regions



- **Cross-modal models**

- Brain alignment is **partially explained** by video features, but removal of speech features does not lead to drop in brain alignment.

- **Jointly-pretrained models**

- Brain alignment is **partially explained** by both video and audio features.



## Result-3: Modality-specific contribution in language and visual regions

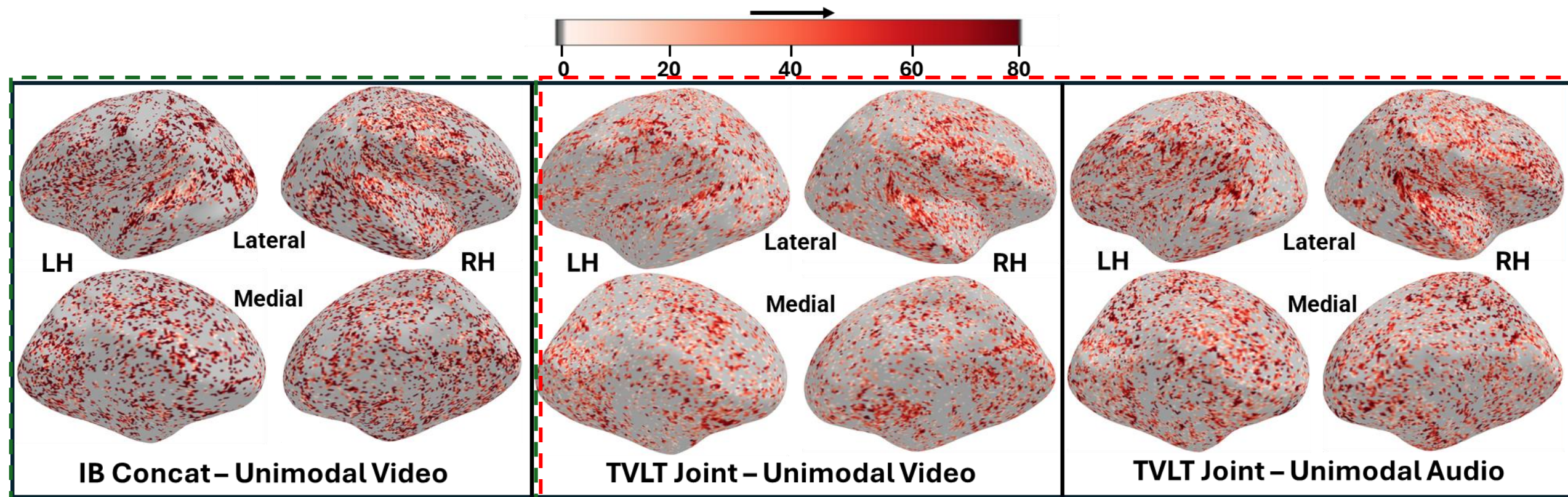
- **Cross-modal models**

- Alignment with language and visual regions is only **partially explained** by **video modality features**.
- Model relies heavily on visual information for improved brain alignment.

- **Jointly-pretrained models**








- Alignment with language and visual regions is only **partially explained** by both **video modality** and **audio modality features**.
- Model is capturing more high-level abstract and semantic information that goes beyond specific features of just one modality.

# Qualitative Analysis: Effect of removal of modality-specific features



- **Cross-modal model**: removal of unimodal video features leads to a significant drop in visual regions
- **Jointly-pretrained model**: removal of unimodal video and audio features leads to a significant drop in language regions

# Conclusions

1. Improved alignment in **cross-modal models** is mainly driven by the removal of  features, not  features:
2. **Jointly-pretrained model** reflects human-like learning via simultaneous multi-modal experiences  .
3. **But** more work is needed to explore true multi-modal models that integrate both modalities ( and ) with balanced knowledge transfer and deeper brain-like understanding .

## Multi-modal brain encoding models for multi-modal stimuli (ICLR-2025)



**Subba Reddy Oota**



**Khushbu Pahwa**



**Mounika Marreddy**



**Maneesh Singh**



**Manish Gupta**



**Bapi S. Raju**