Brain Alignment:
Insights from Multi-Modal and Instruction-Tuned LLMs

Raju Bapi & Subba Reddy Oota
IIIT-Hyderabad, India & TU Berlin, Germany

raju.bapi@iiit.ac.in
subba.reddy.oota@tu-berlin.de

Technische Universität Berlin

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Neuro-AI Pipeline

DNN Model Representations

Stimulus

Language | Audio | Visual | Multimodal

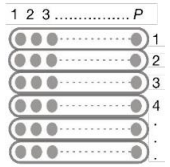
DNN Models

Language Model | Audio Model | Vision Model | Multimodal Model

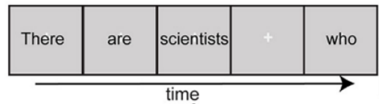
Stimulus Feature Space

Down sampling: Interpolation of the feature matrix

HRF estimation: use of FIR model, different delays



Human Brain Recordings



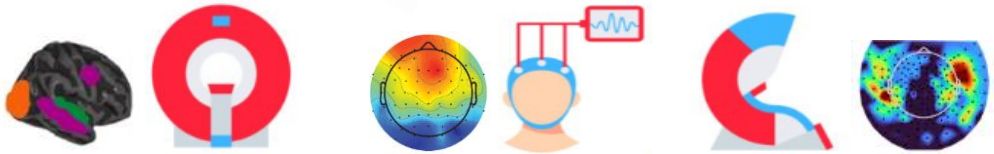
Reading

Listening

Static Image

Video clip (with or without audio)

Visual | Auditory



fMRI | EEG | MEG

Evaluation Metrics

PCC, R^2, 2V2 Accuracy, RDM, CKA, Noise Ceiling, Normalized brain alignment

fMRI: Whole brain, ROI level, Sub-ROI level, task-specific voxels

MEG: Sensor recordings over time points

EEG: Electrode signals recorded over time

Linear

Encoding Models →

Ridge | Bootstrap Ridge | Banded Ridge

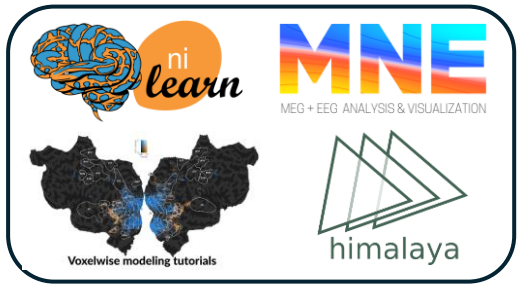
Lasso | PLS | Kernel Ridge

← **Decoding Models**

Non-Linear

Multi-Layer Perceptron | DNN Models

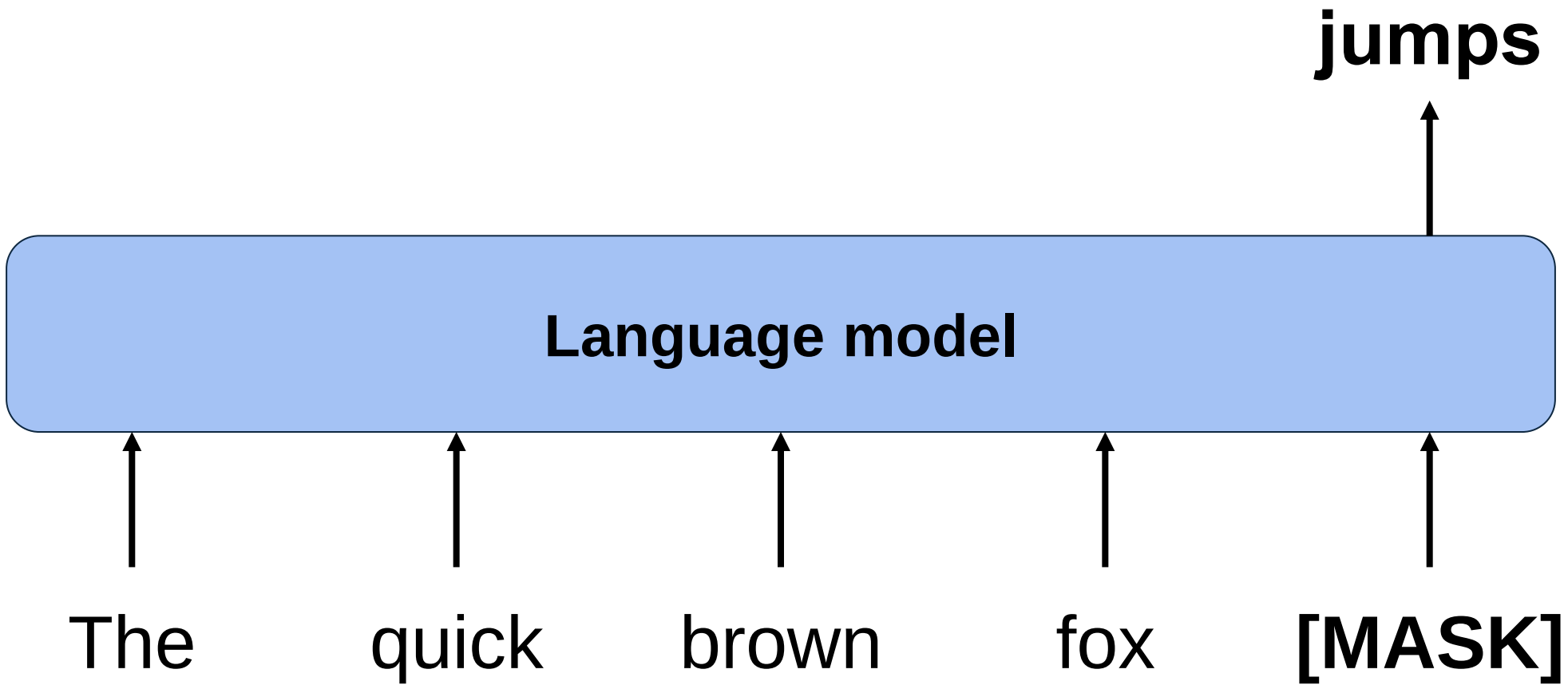
Visualization Tools



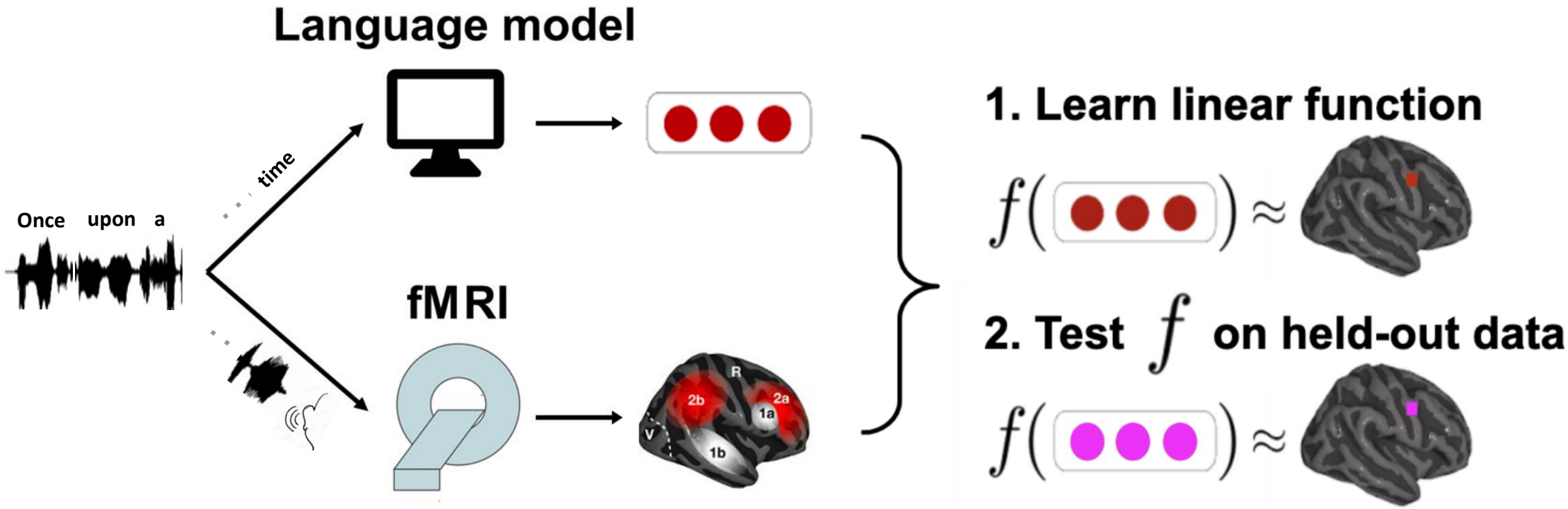
Oota et al (2025). Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey) [TMLR]

https://openreview.net/pdf?id=YxKJihRcby

Language models (LMs) are trained to predict missing words



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



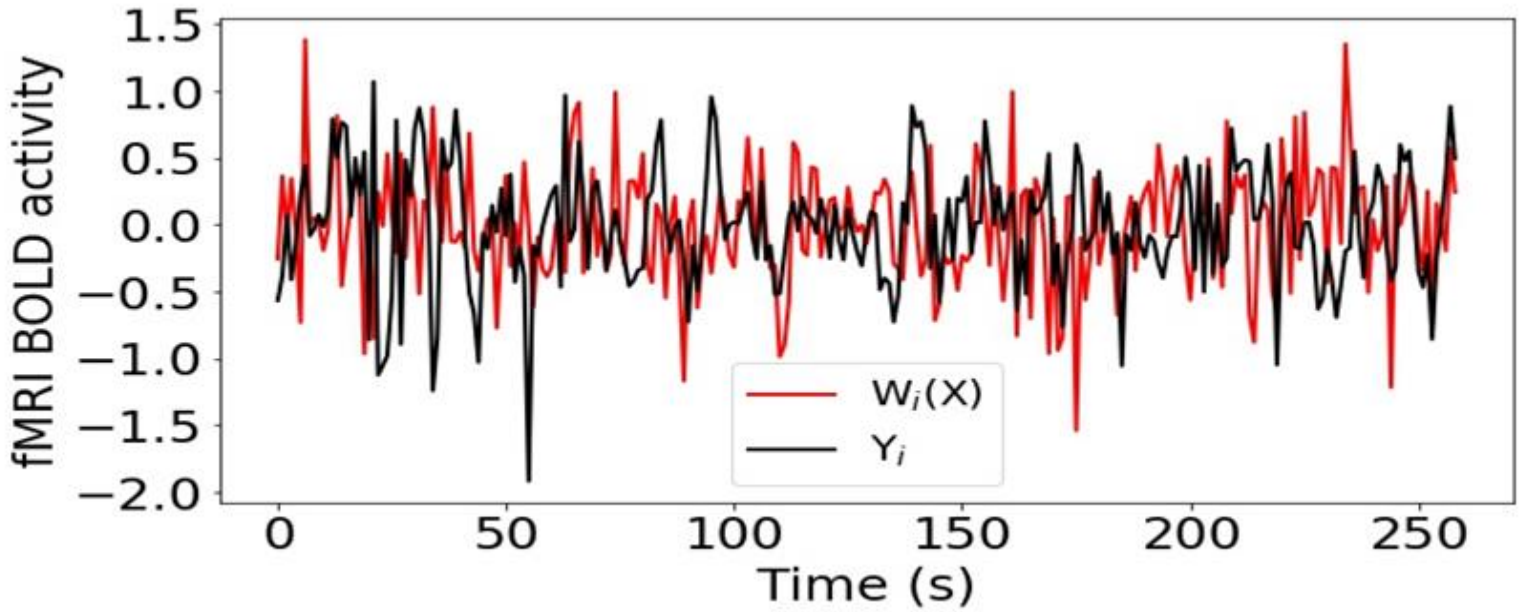
Brain alignment of an LM ⇒ how similar its representations are to a human brain

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 tasks (e.g. listening to a story) to an impressive degree



brain alignment$_i$ = Pearson corr(true v_i, pred v_i)

Jain and Huth. Incorporating context into language encoding models for fMRI. (NeurIPS 2018)
Toneva and Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). (NeurIPS 2019)

Neuro-AI: Questions

1. Although participants experience stimuli in one modality (ex., image), evoked brain response is rich!
   - Is it possible to align representations from unseen modalities (ex., image captions)? [Ooota et al., COLING 2022a; Oota et al. COLING 2022b]
2. Engagement of the brain with the stimulus material is rich! While viewing an image, there may be evoked activity corresponding to reflection on the content of the image (summarization, etc).
   - Do task-based representations from DL models have better alignment than representations from pre-trained models? [Oota et al., NAACL 2022 (Language); Oota et al., Interspeech 2023 (Speech); Oota et al., NeurIPS 2023 (Language)]
3. More work is needed to explore true multi-modal models that integrate both modalities ( and ) with balanced knowledge transfer and deeper brain-like understanding  .
   - Do recent advances in multimodality and instruction-tuning lead to improved brain alignment of a LM? [Focus of this talk!]

ICLR
International Conference On Learning Representations
2025

Multi-modal brain encoding models for multi-modal stimuli

Subba Reddy Oota Khushbu Pahwa Mounika Marreddy
Maneesh Singh Manish Gupta Raju S. Bapi

subba.reddy.oota@tu-berlin.de

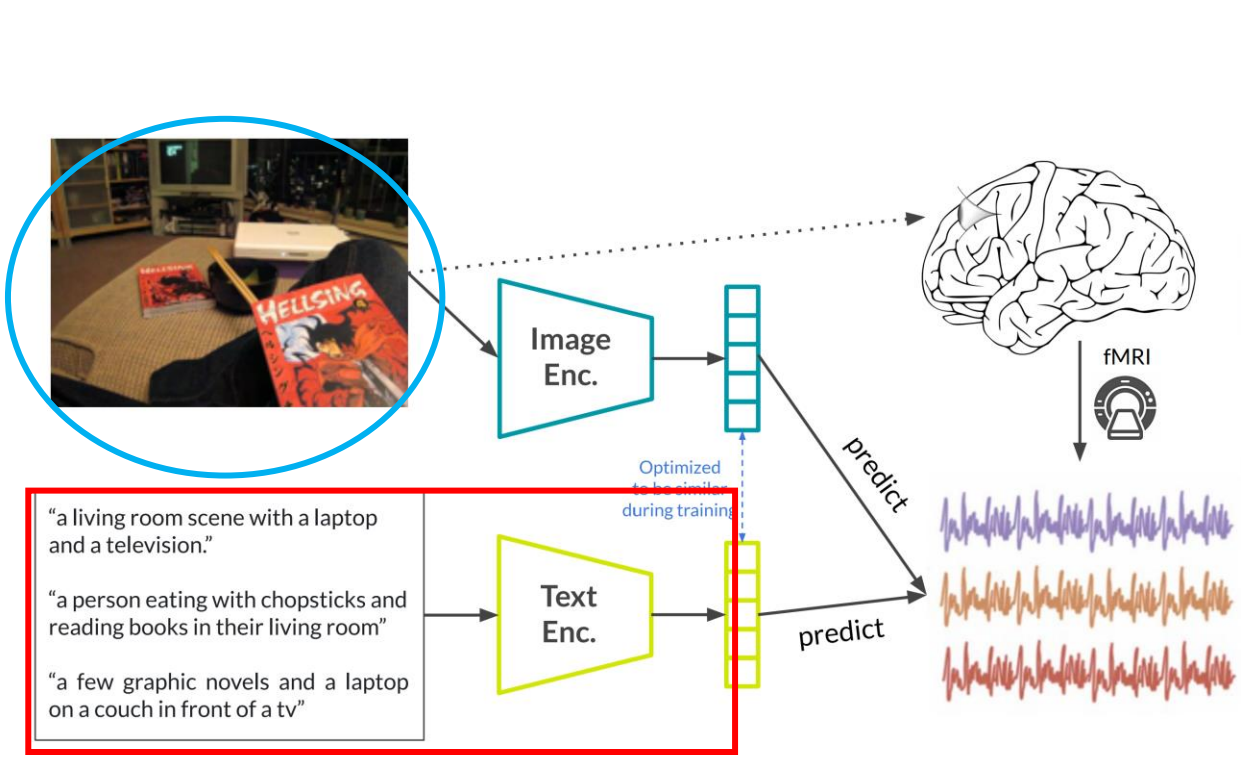
Technische Universität Berlin

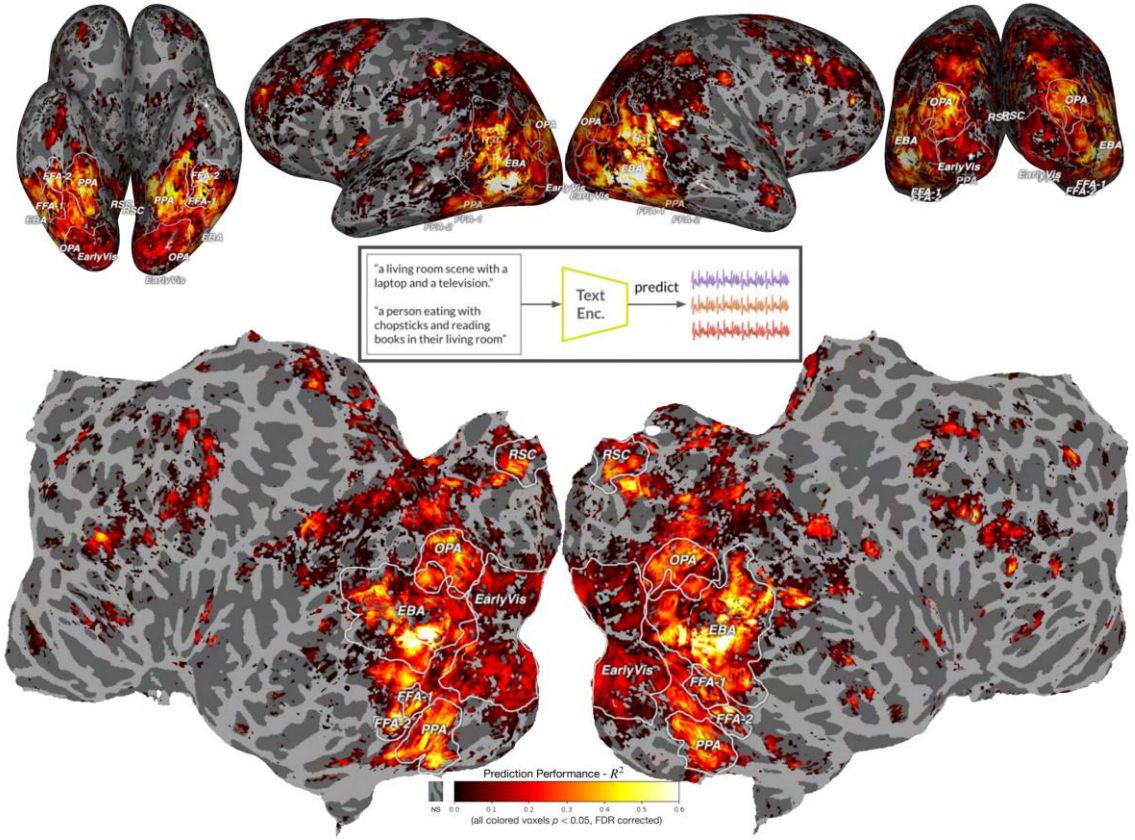
RICE UNIVERSITY

Microsoft

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

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

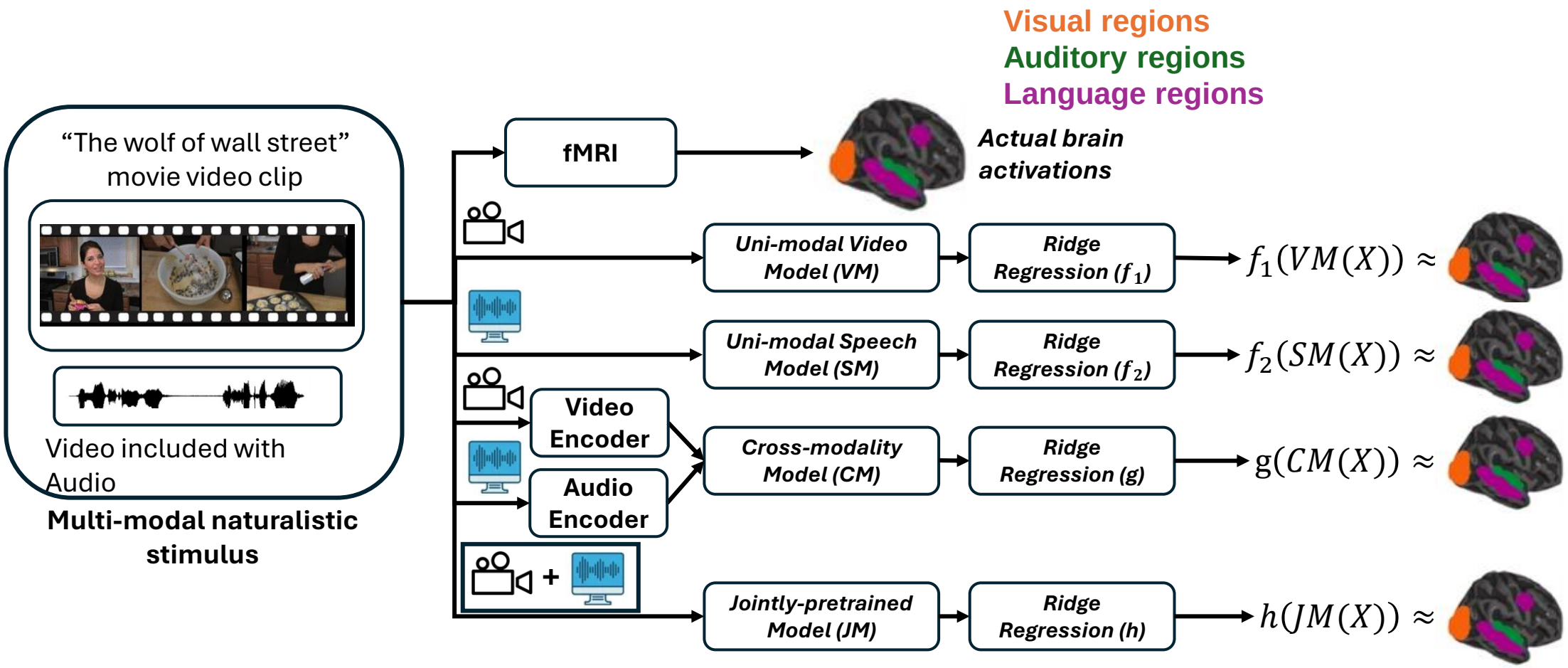




How accurately do multi-modal models predict brain activity evoked by multi-modal stimuli?

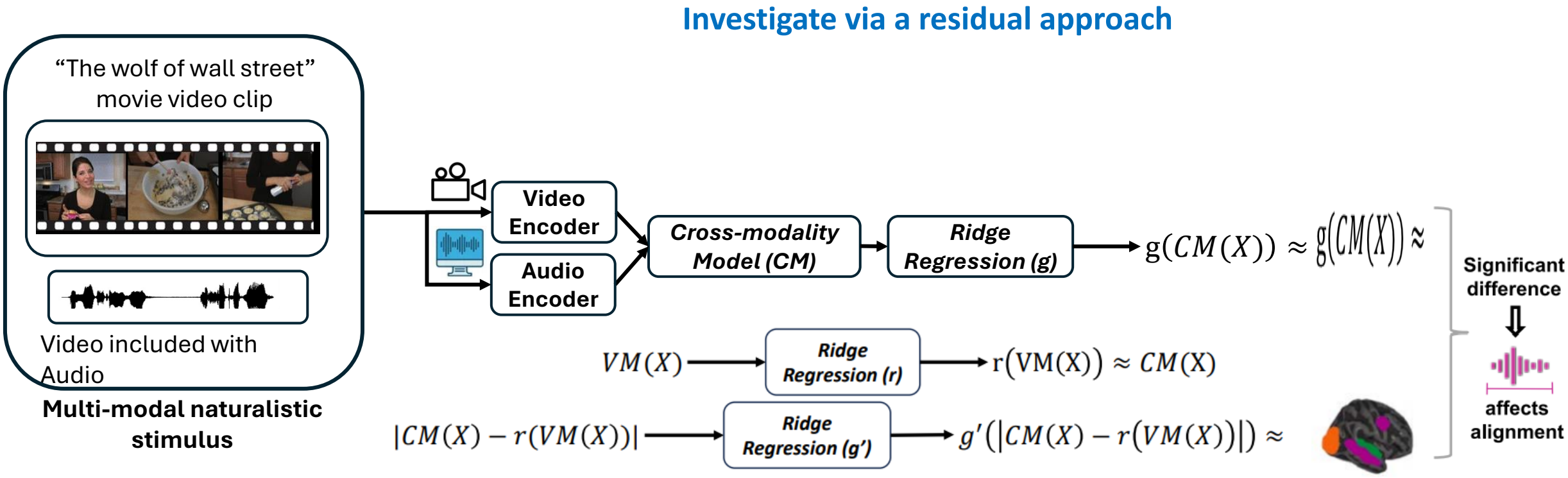
Aria Y. Wang, Kendrick Kay. "Incorporating natural language into vision models." Bioarxiv 2023.

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?



Toneva et al. 2022 Nature Computational Science, Oota et al. 2023 NeurIPS

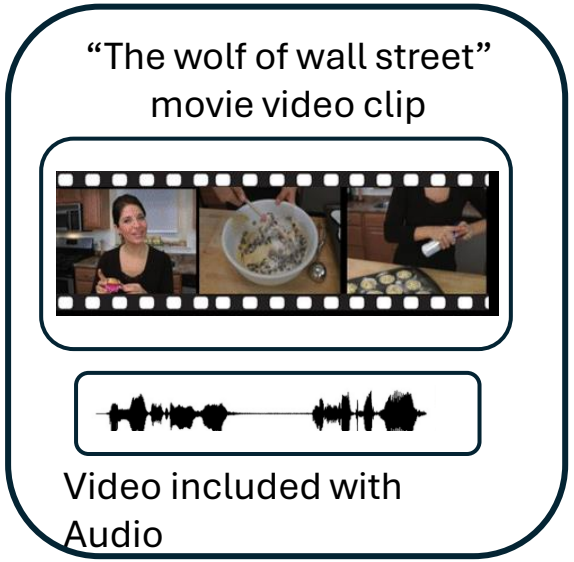
Datasets & Models

- 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)

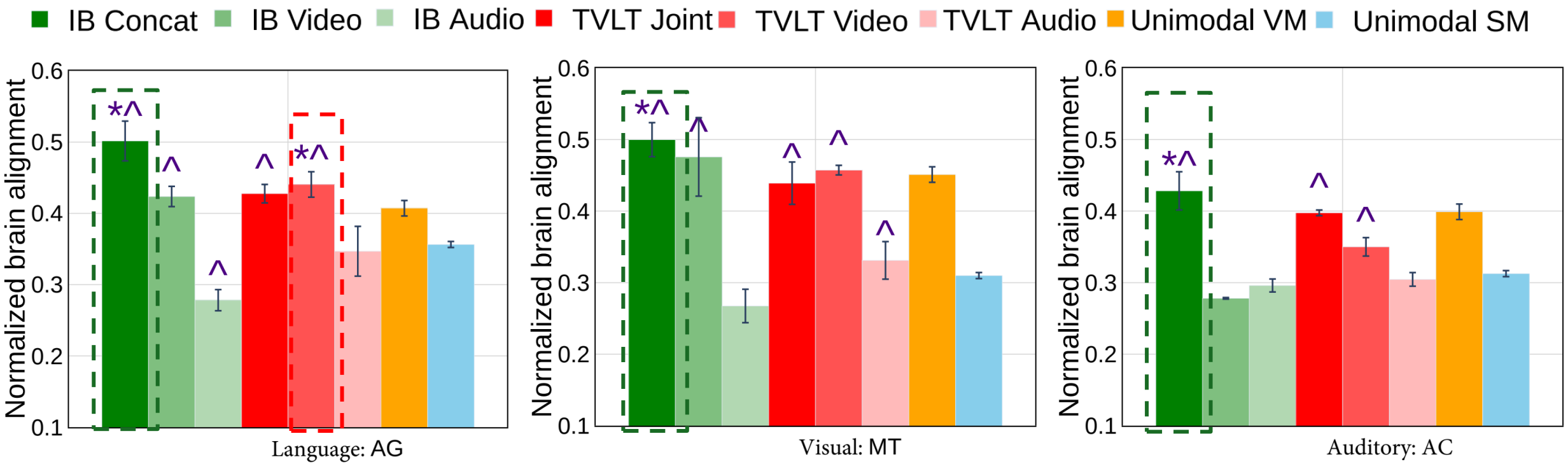


Multi-modal naturalistic stimulus

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

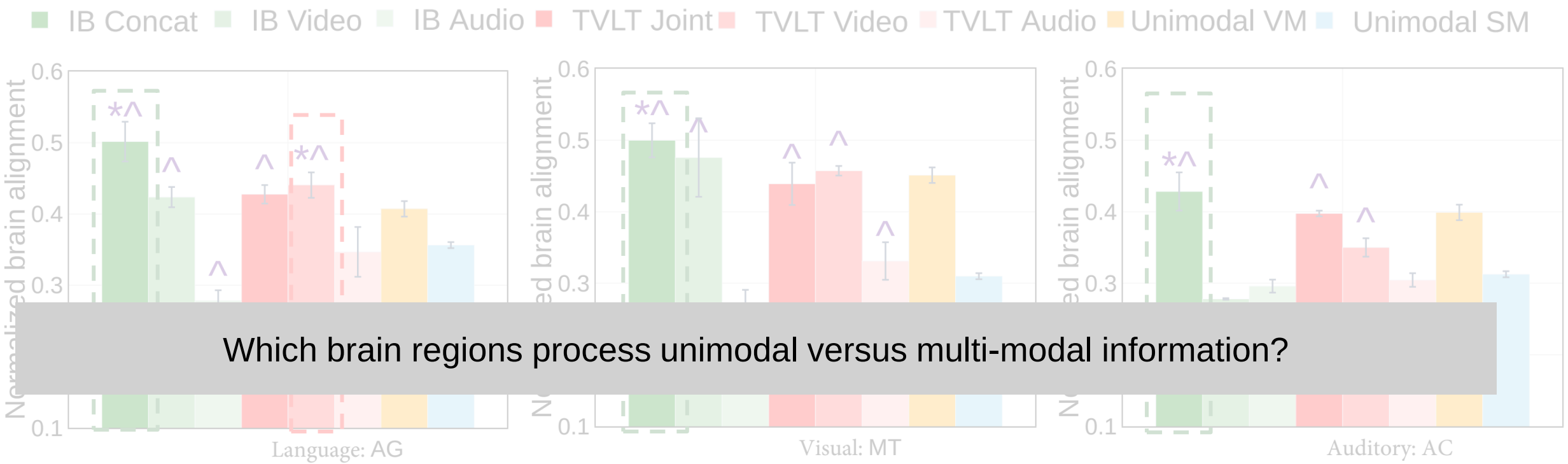
Multi-modal stimulus: How do multi-modal and unimodal models differ in their ability to predict brain activity in late language regions, higher visual regions and early sensory regions?

Result: 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.

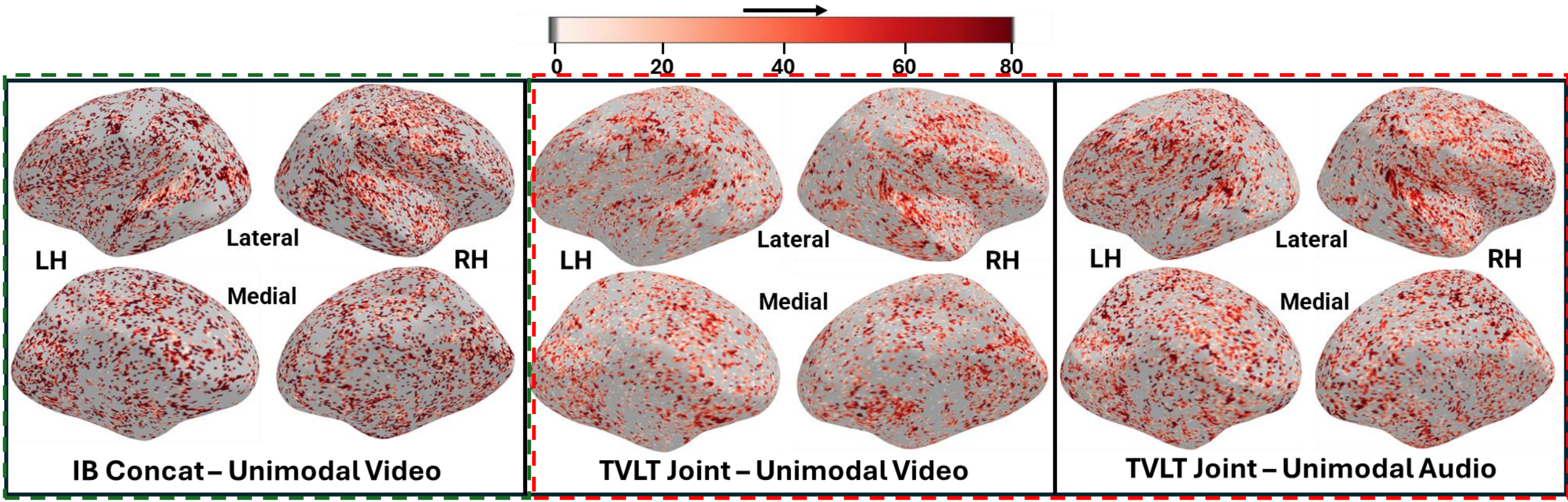
Result: Multi-modal vs. Unimodal models & brain alignment



Which brain regions process unimodal versus multi-modal information?

- **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.

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

Interim Conclusions: Multimodal Alignment

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 🧠**.**

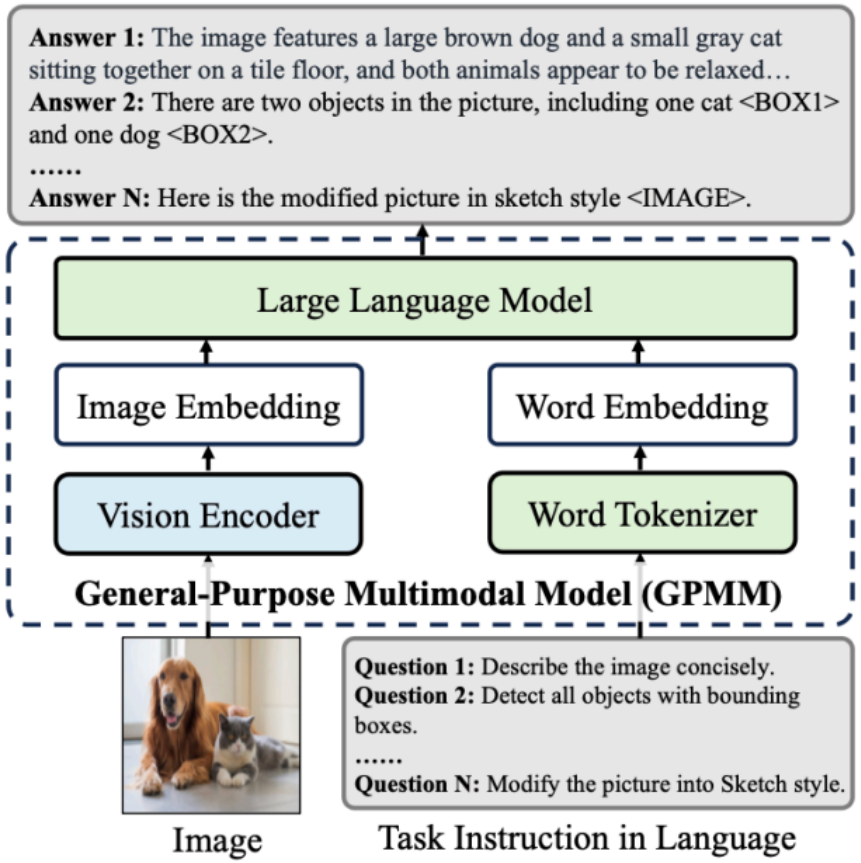
ICLR

Correlating instruction-tuning (in multimodal models) with vision-language processing (in the brain)

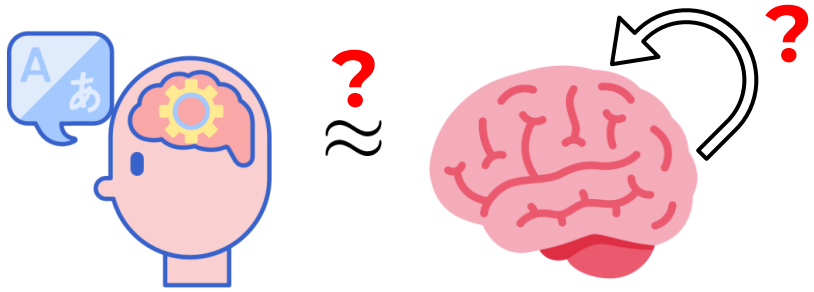
Subba Reddy Oota Akshett Jindal Ishani Mondal Khushbu Pahwa

Satya Sai Srinath Manish Shrivastava Maneesh Singh Raju S.Bapi Manish Gupta

Multimodal instruction tuning enables models to generalize to new tasks by following unseen instructions



INPUT: <image>Describe this image in detail.
OUTPUT: <long descriptions>



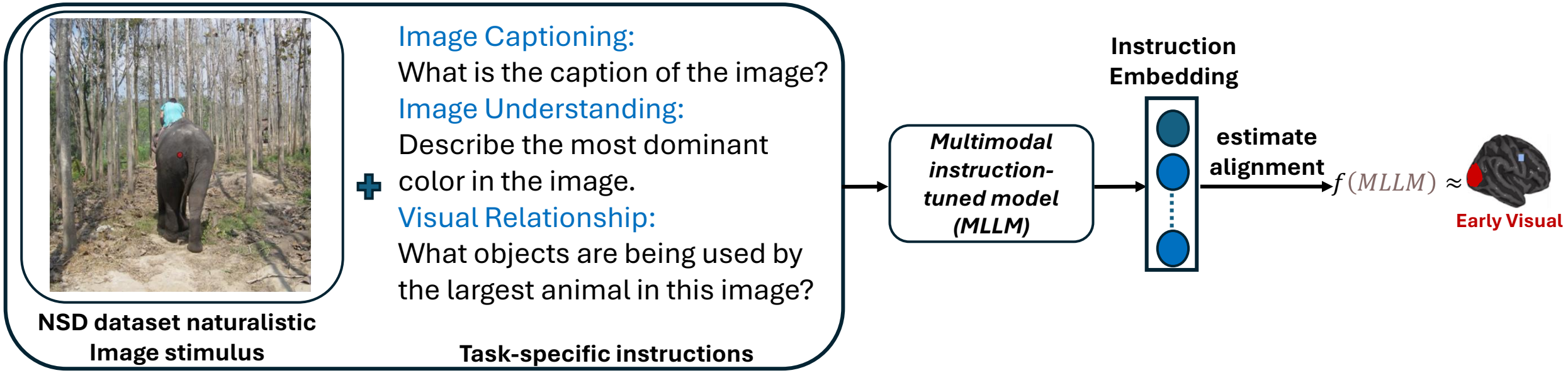
How do multimodal instruction-tuned LLMs process visual images when guided by natural language task instructions?

How does the brain integrate information during the processing of visual images?

Do multimodal instruction-tuned models prompted with natural language improve brain alignment and capture instruction-specific representations?

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, Enhong Chen. "A Survey on Multimodal Large Language Models." Arxiv 2024.

Multi-modal Instruction-tuned LLMs (MLLMs): brain alignment



- How well do MLLMs predict brain activity evoked by visual stimuli under task-specific instructions compared to unimodal and multimodal models?
- Do instruction-specific representations in MLLMs differentiate visual brain regions involved in processing, thereby aligning with the mechanisms of human visual cognition?

Datasets & Models

- Brain: fMRI recordings from NSD dataset [St-Laurent et al. 2023]
  - Passively watching natural scene images
  - N=4

- 3 multimodal instruction-tuned large language models
  - InstructBLIP
  - mPLUG-Owl
  - IDEFICS

- unimodal and multi-modal models
  - ViT-H
  - CLIP

NSD dataset naturalistic Image stimulus

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

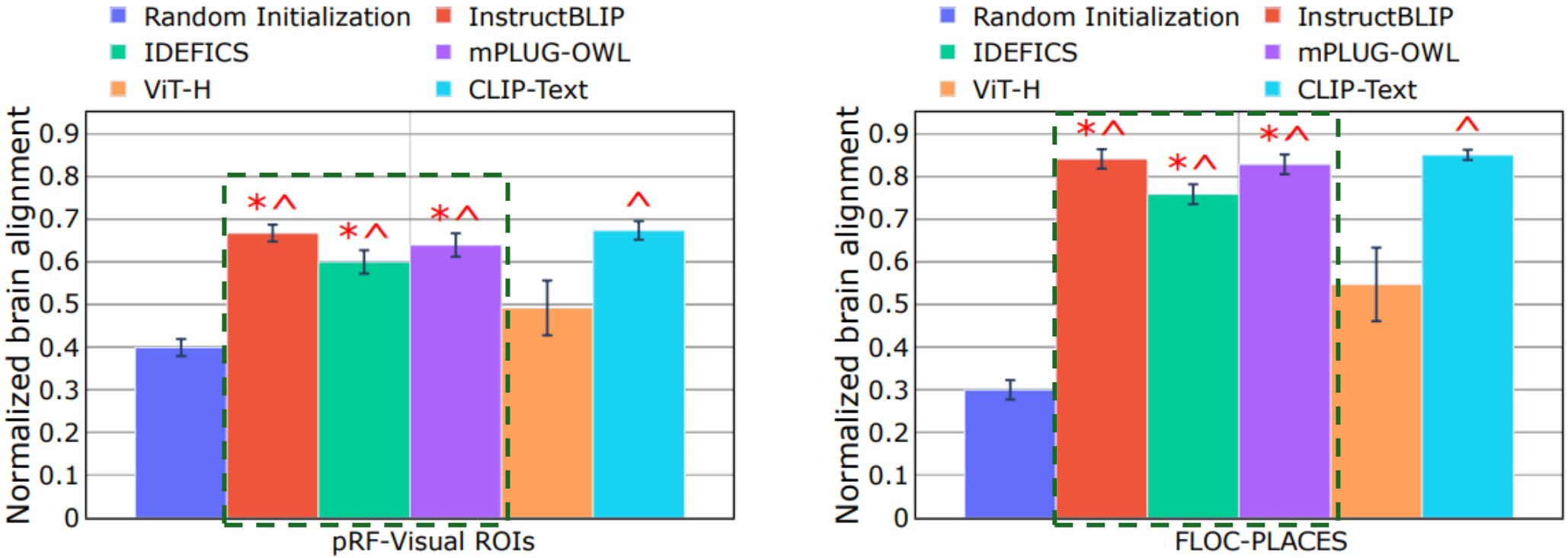
Task-specific natural instructions

Task	Description
Image Understanding	IU1: Describe the most dominant color in the image IU2: List any food items visible. IU3: How many animals are there in the image?
Visual Question Answering	VQ1: What is in this image? VQ2: Are there any people in this image? If yes, describe them. VQ3: What is the foreground of the image? What is in the background?
Image Captioning	IC: Generate some text to describe the image
Scene Recognition	SR: Highlight the area that shows a natural outdoor scene.
Commonsense Reasoning	CR: What type of environment is shown in the image?
Visual Relationship	VR: What kind of interaction is happening between the animate and inanimate objects here?

These tasks which are generally applicable to any image regardless of the contents in the image

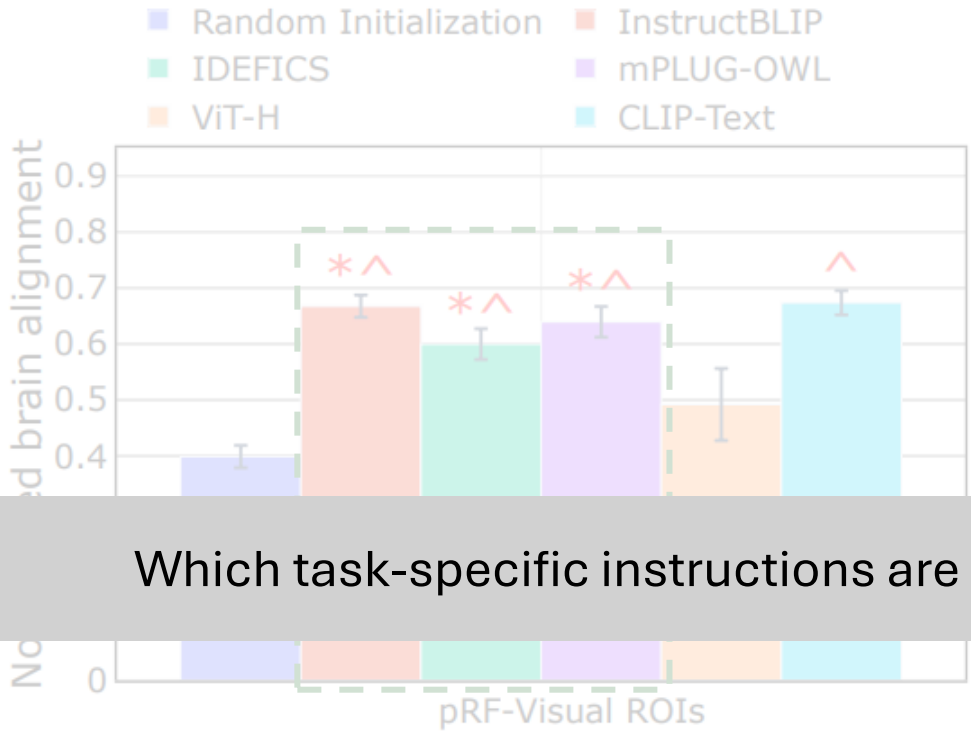
How do MLLMs, unimodal and multi-modal models differ in their ability to predict brain activity in higher visual and early visual regions?

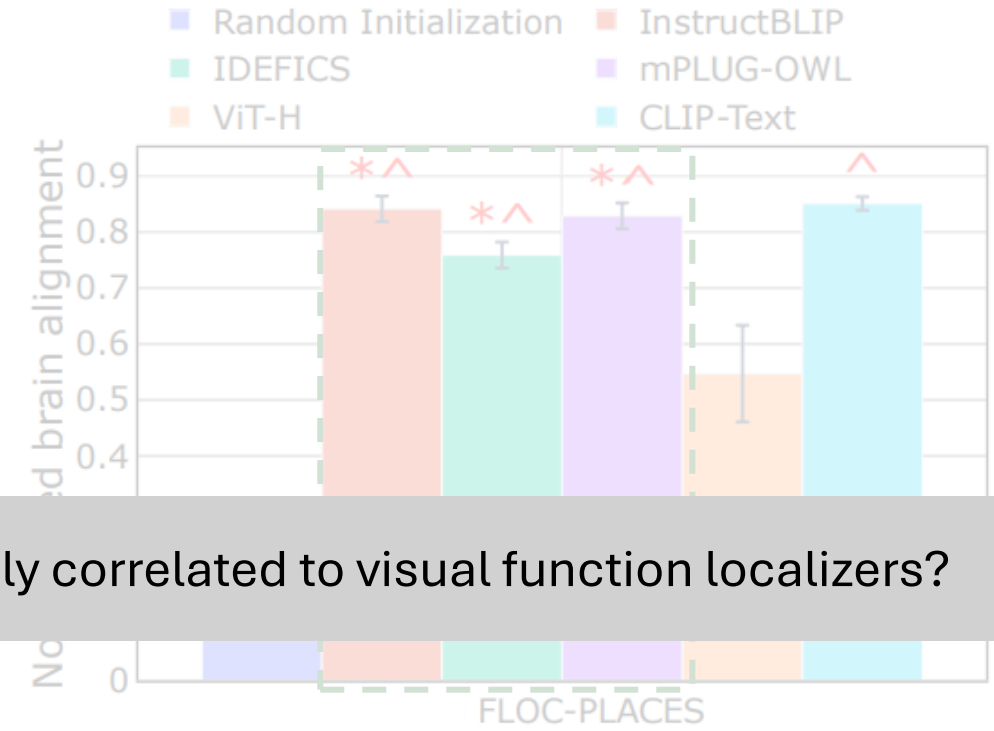
Result: MLLMs vs. Unimodal vs. Multi-modal models & brain alignment



- **Early-visual regions**:
  - Both **MLLMs** and **multi-modal** models show significantly high brain alignment than baseline and unimodal video models
  - Surprisingly, brain alignment of **random initialization of MLLMs** is closer to that of **unimodal video models**
- **Higher-visual regions:**
  - Both **MLLMs** and **multi-modal** models show better brain relevant representations (~0.8) than early visual areas (~0.6).

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

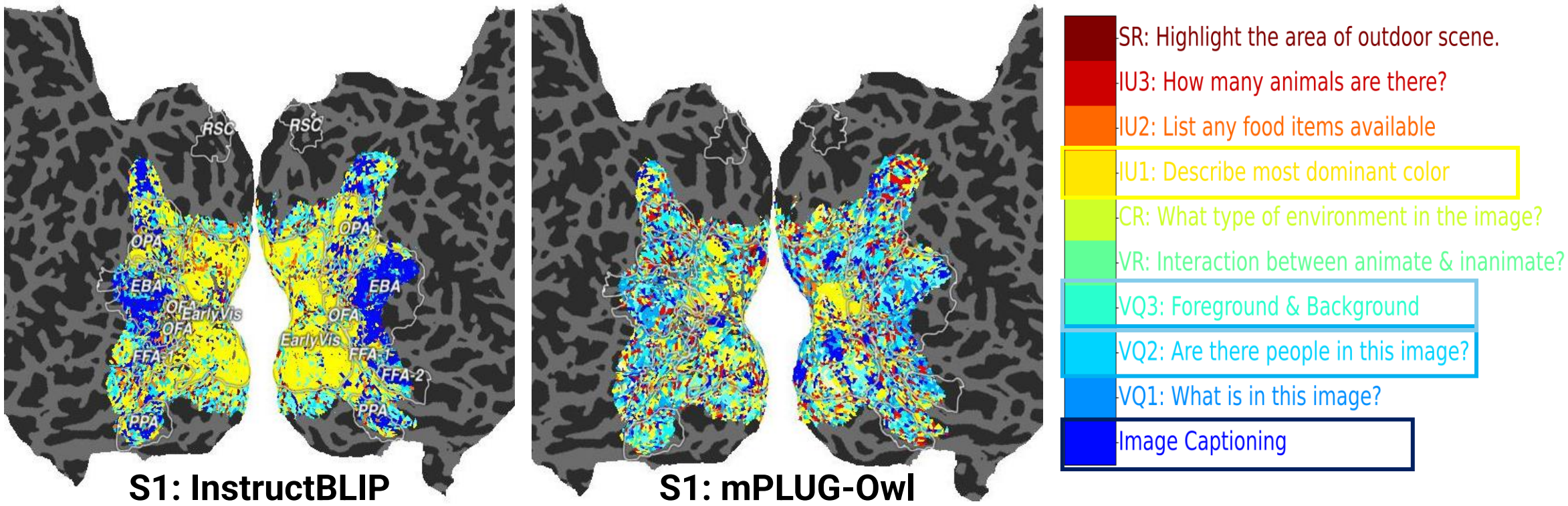




Which task-specific instructions are highly correlated to visual function localizers?

- **Early-visual regions**:
  - Both **MLLMs** and **multi-modal** models show significantly high brain alignment than baseline and unimodal video models
  - Surprisingly, brain alignment of **random initialization of MLLMs** is closer to that of **unimodal video models**
- **Higher-visual regions:**
  - Both **MLLMs** and **multi-modal** models show better brain relevant representations (~0.8) than early visual areas (~0.6).

Result: Which task-specific instructions are highly correlated to visual function localizers?

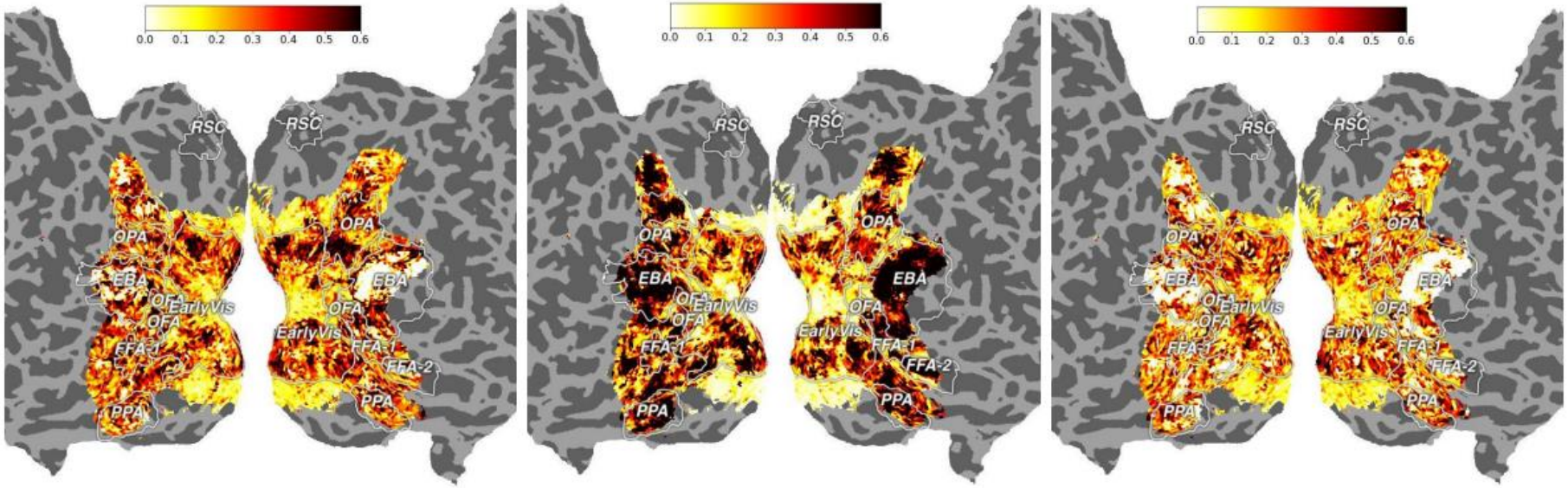


S1: InstructBLIP **S1: mPLUG-Owl**

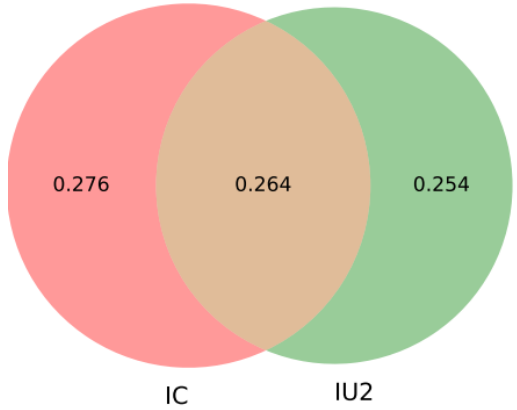
- **Early-visual regions**:
  - **Image understanding** instruction shows significantly high brain alignment across MLLMs
- **Higher-visual regions:**
  - **Image captioning** instruction shows significantly high brain alignment in the EBA, PPA, and FFA regions
  - **Visual question answering** instructions shows significantly high brain alignment in the PPA, and FFA regions
- **Not all instructions lead to increased brain alignment across all regions**

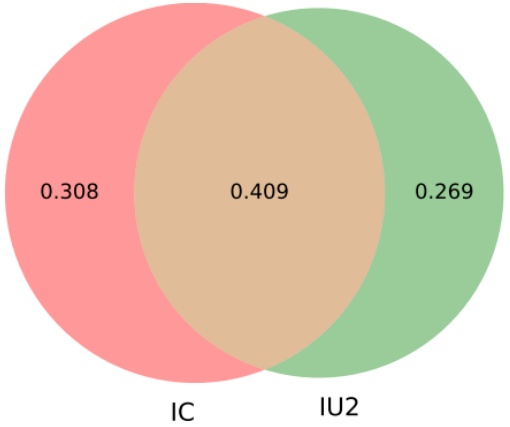
What is the unique and shared variance of each task-specific instruction to brain responses?

Result-4: Partitioning explained variance between task-specific instructions



Unique Variance-IC | Shared Variance–IC vs. IU2 | Unique Variance-IU2





- Between **Image Captioning (IC)** and **Image Understanding (IU2):** there is no unique variance for **IU2** in the **EBA region (higher-visual)**, while **IC** retains some unique variance.
- Task-specific instructions exhibit **moderate shared variance** in the **early visual cortex**, while **shared variance is significantly higher** in **higher visual ROIs**

Interim Conclusions: Instruction-tuned Multimodal model Alignment

1. 🗣️ **MLLMs** generate task-specific output tokens based on instructions, but **not all instructions lead to better brain alignment**

2. 👁️ They capture multiple **visual concepts**, yet exhibit **similar brain alignment** across different types of visual stimuli

3. The **variance** in brain alignment is shared across task-specific instructions:
   - Moderate in 🧠 *early visual areas*
   - Higher in 🧠 *high-level visual regions*

4. **But,** more work to do - especially in enhancing MLLMs' ability to **differentiate between instruction types** in terms of neural alignment

Summary & Concluding Thoughts

1. Cross-modal encoding models are feasible from uni-modal brain responses.

2. Task-based representations give raise to better encoding models.

3. Representations from multimodal and instruction-tuned MLLMs exhibit impressive alignment performance.

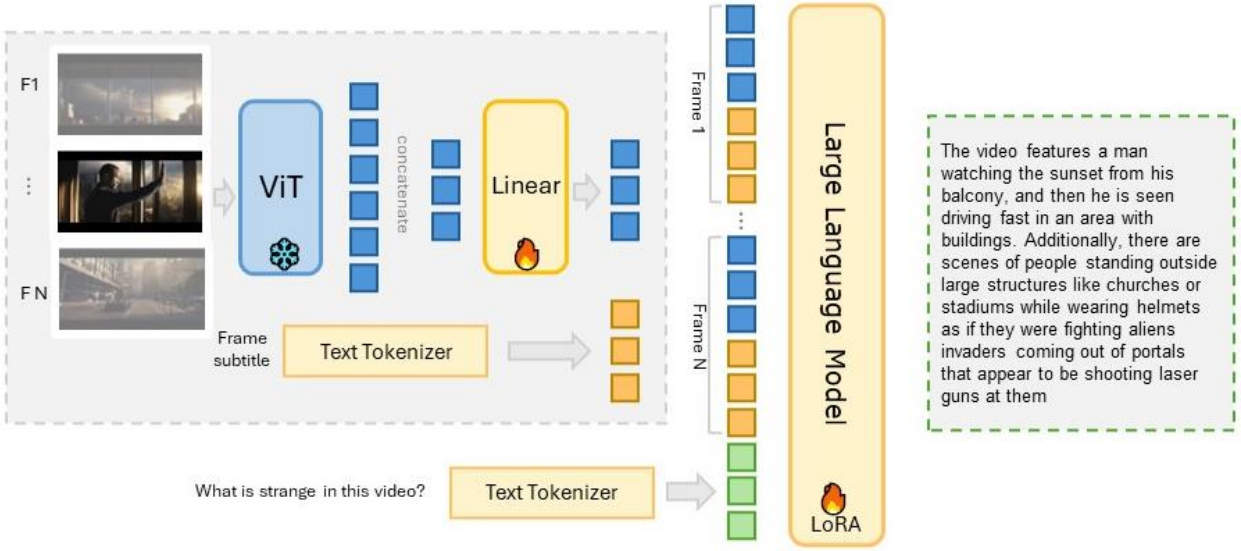
4. Is it possible to look at multilingual representation of language using such naturalistic tasks?

5. Can we characterize how neurodegenerative disorders affect degradation of language function using DL model alignment?

6. EEG Brain decoding for BCI applications.

Instruction-tuned MLLMs for Video and Audio

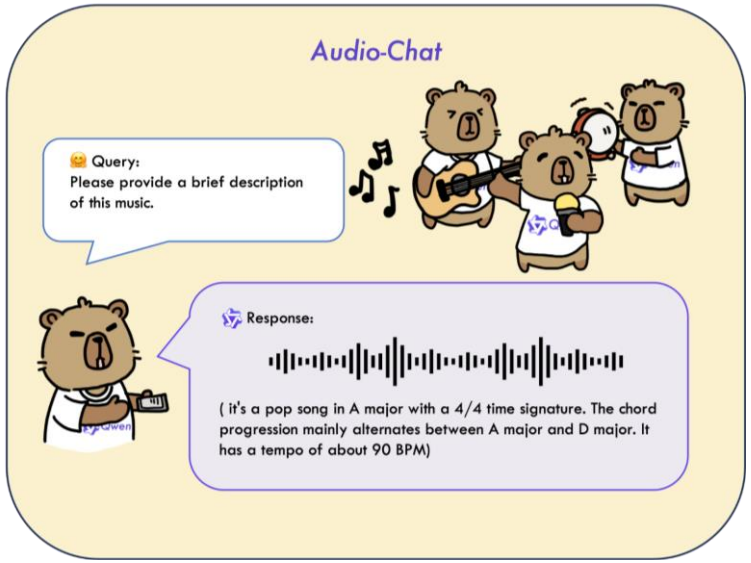
Instruction-tuned video MLLM



INPUT: <video> Describe this video in detail

OUTPT: <long descriptions>

Instruction-tuned audio MLLM



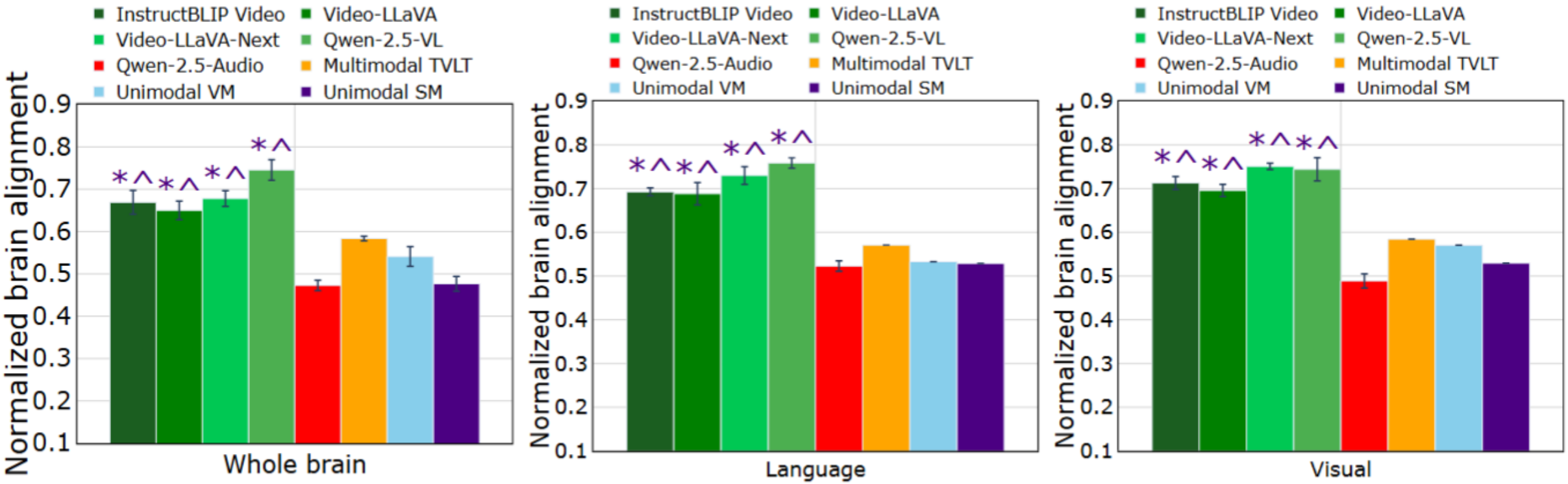
INPUT: <audio> Describe this audio in detail

OUTPT: <long descriptions>

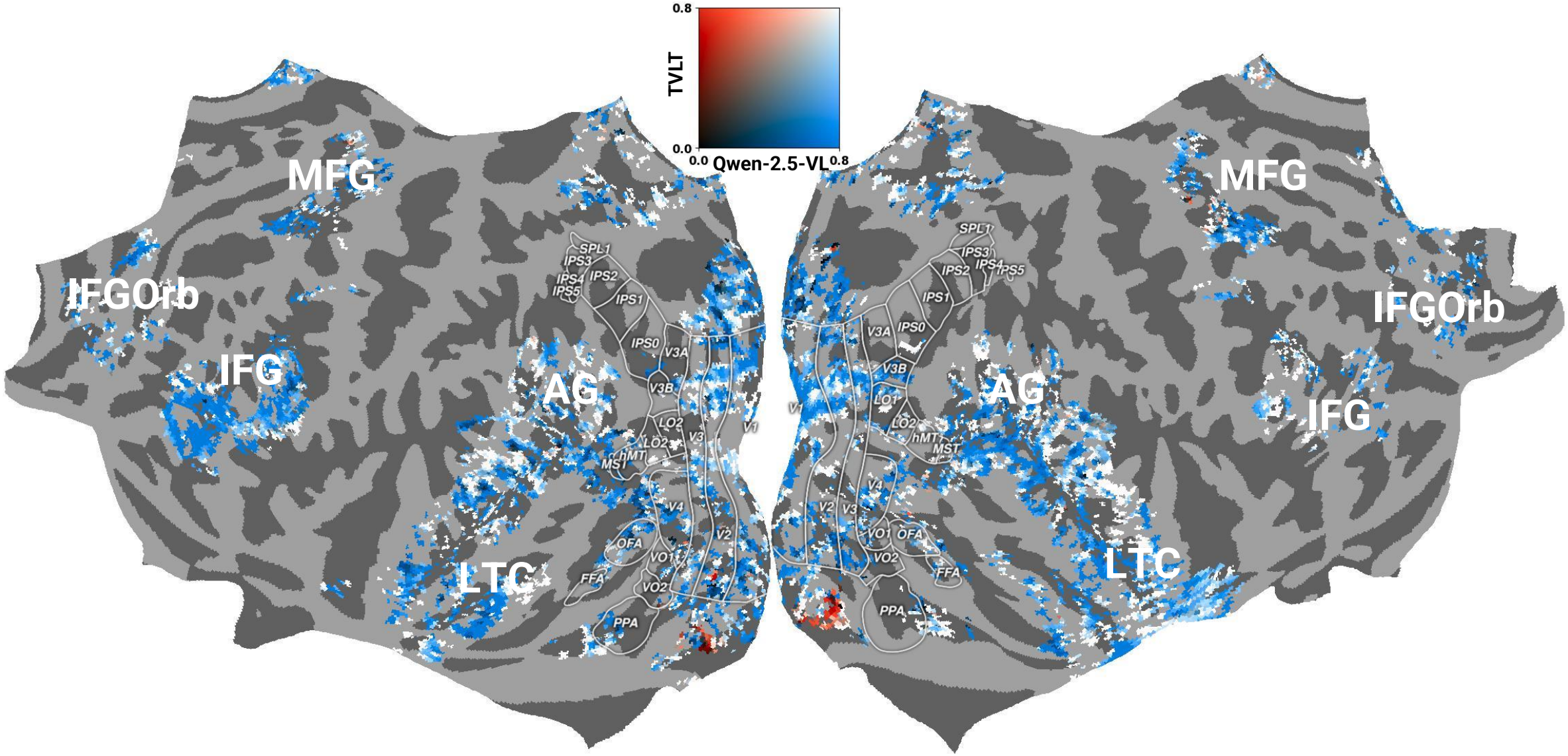
Video: Task-specific natural instructions

Task	Description
Action Recognition	What are the main events captured in the video? Describe the activity in the video
Video Understanding	Can you describe the video content in detail? Can you provide a summary of video?
Visual Question Answering	How many people are in the video, and what are they doing?
Video Captioning	Caption the video in a short sentence. What is this video about?
Object and Scene Recognition	What are the main objects and people visible in the video? Describe each one briefly.
Commonsense Reasoning	Why did the character take this action? What could have motivated them to do this?
Spatial Understanding	Where is this video taken from? What place/landmark is shown in the video?
Temporal Ordering	Step-by-step describe the activity shown in the video.
Video reasoning	Why is this video funny? What is unusual about this video?
Narrative Understanding	Summarize the main storyline of the movie. What is the central conflict, and how is it resolved?
Emotion and Sentiment Analysis	What emotions do the characters express during the video? How does the video make you feel overall?
Cultural Context Understanding	What cultural references or norms are represented in the video? How does this shape the characters' actions?
Global Appearance	Describe the changes in characters' appearances throughout the video, including any noticeable outfit changes.
Linking Events	Explain how an early event in the video influences later developments.

Instruction-tuned Video MLLMs have improved brain alignment across language, visual and auditory regions



Instruction-tuned Video MLLMs have improved brain alignment across language, visual and auditory regions



Mariya Toneva

Xavier Hinaut

Manish Gupta

Mounika Marreddy

Satya Sai Srinath

Fatma Deniz

Frederic Alexandre

Maneesh Singh

Khushbu Pahwa

Ishani Mondal

Manish Shrivastava

Technische Universität Berlin

RICE UNIVERSITY

UNIVERSITY OF MARYLAND

Microsoft

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD