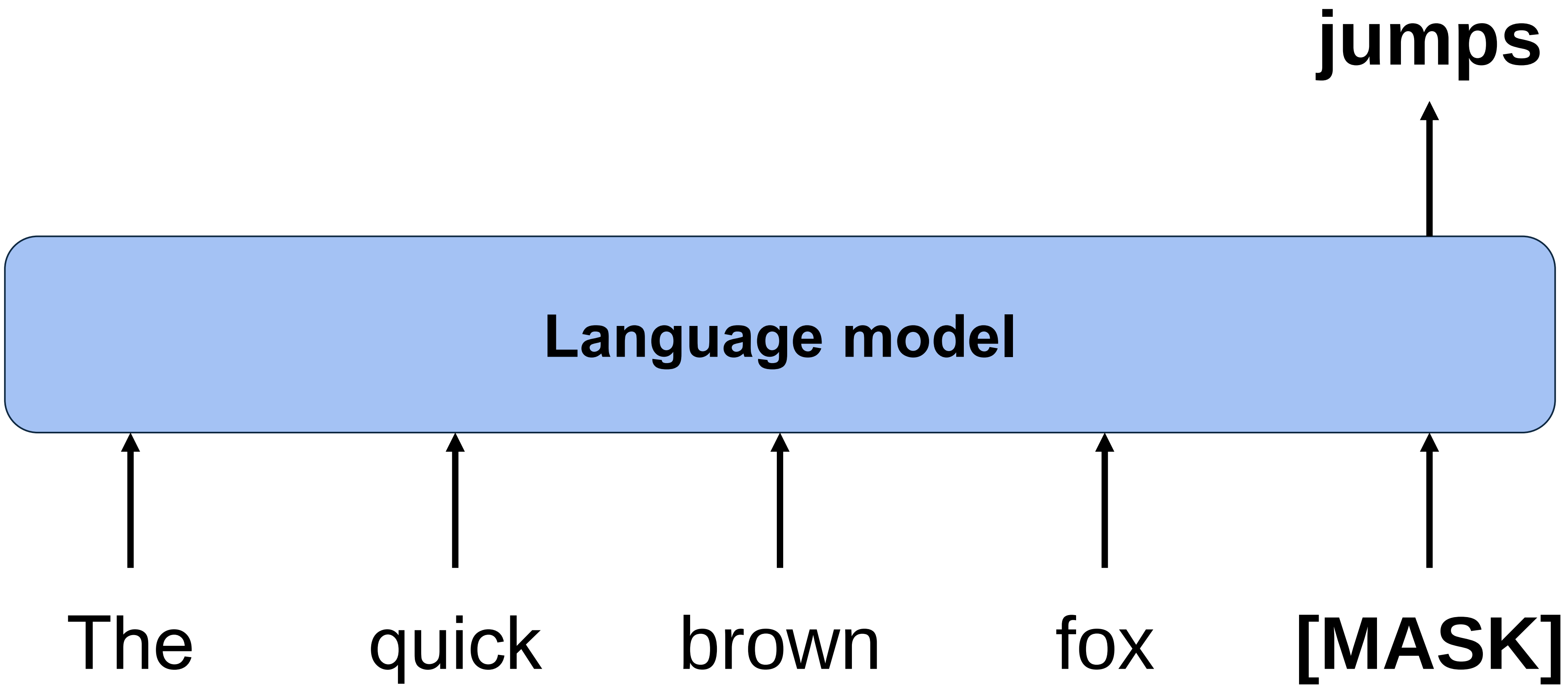
2025

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

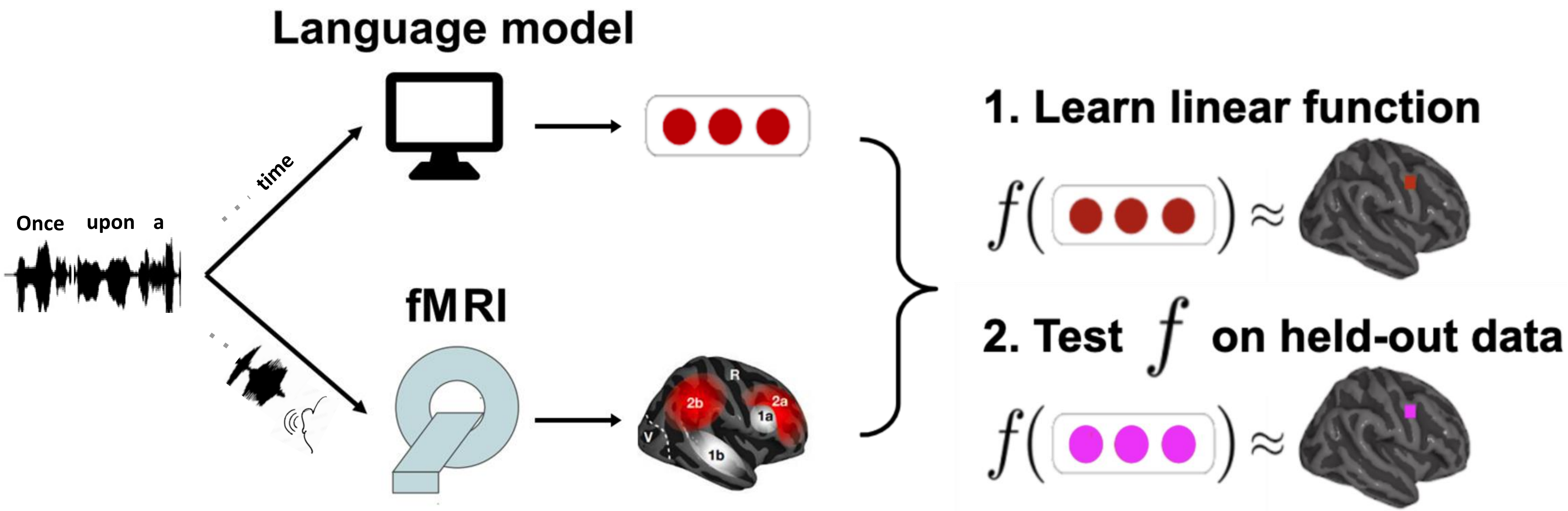
Subba Reddy Oota   Akshett Jindal   Ishani Mondal   Khushbu Pahwa
Satya Sai Srinath   Manish Shrivastava   Maneesh Singh   Bapi S. Raju   Manish Gupta

subba.reddy.oota@tu-berlin.de

Language models (LMs) are trained to predict missing words



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



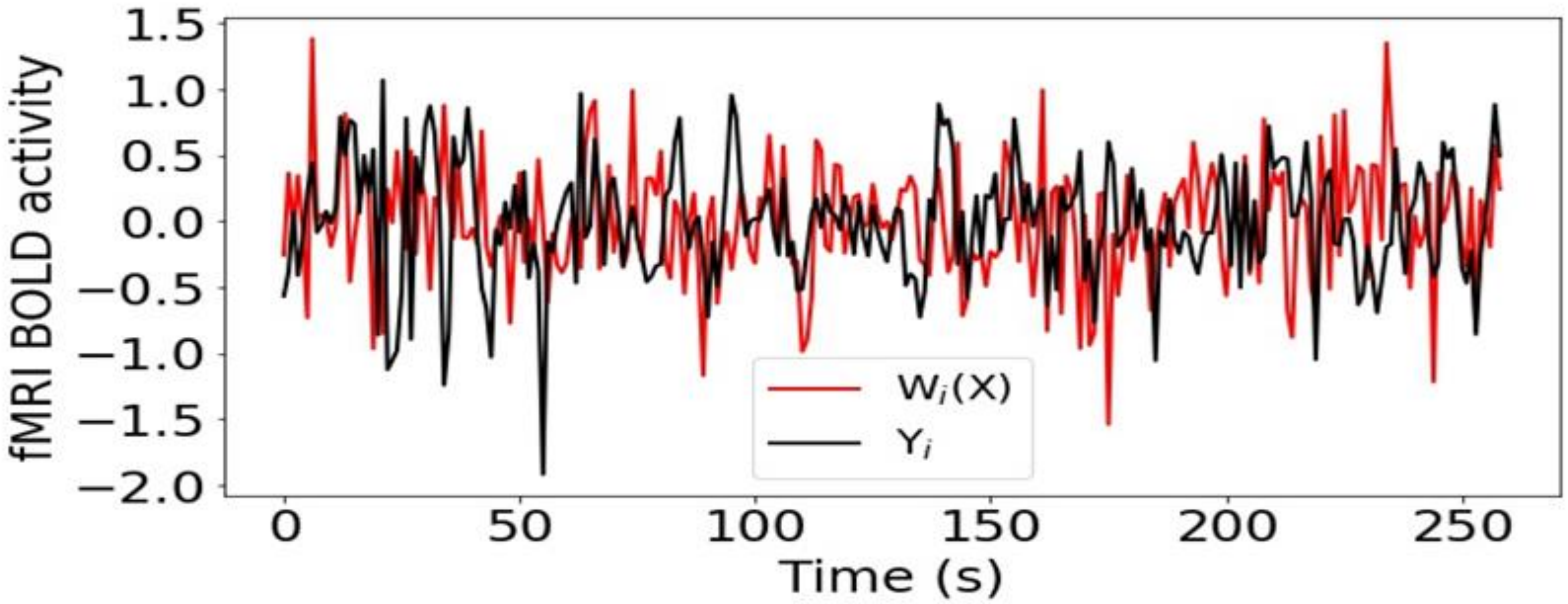
Brain alignment of an LM ⇒ how similar its representations are to a human brain

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

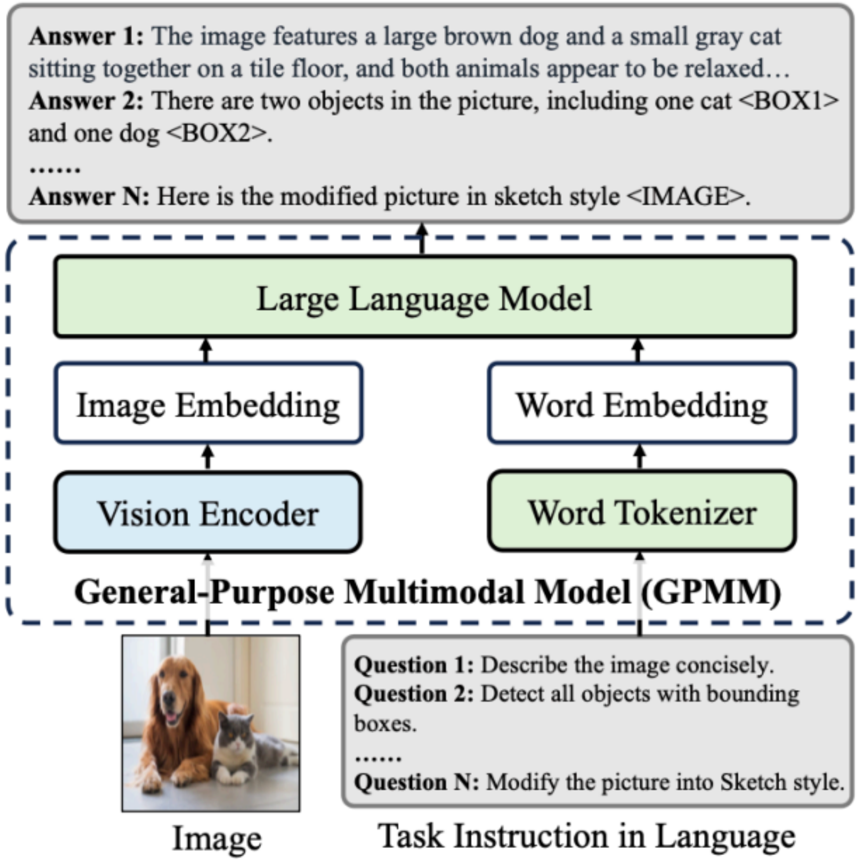


brain alignment$_i$ = Pearson corr(true v_i, pred v_i)

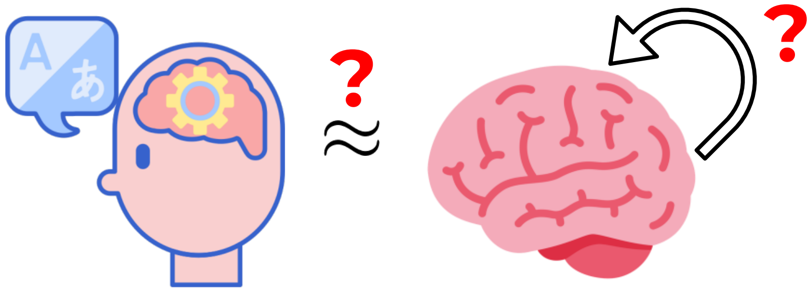
Brain alignment of a LM ⇒ Advances in model size, instruction-tuning, and multimodality have improved alignment with neural data.

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)

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

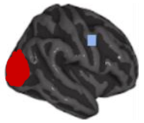
NSD dataset naturalistic Image stimulus

+

Image Captioning:
What is the caption of the image?
Image Understanding:
Describe the most dominant color in the image.
Visual Relationship:
What objects are being used by the largest animal in this image?

Task-specific instructions

→ ***Multimodal instruction-tuned model (MLLM)*** → **Instruction Embedding** → **estimate alignment** → $f(MLLM) \approx$



- 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.

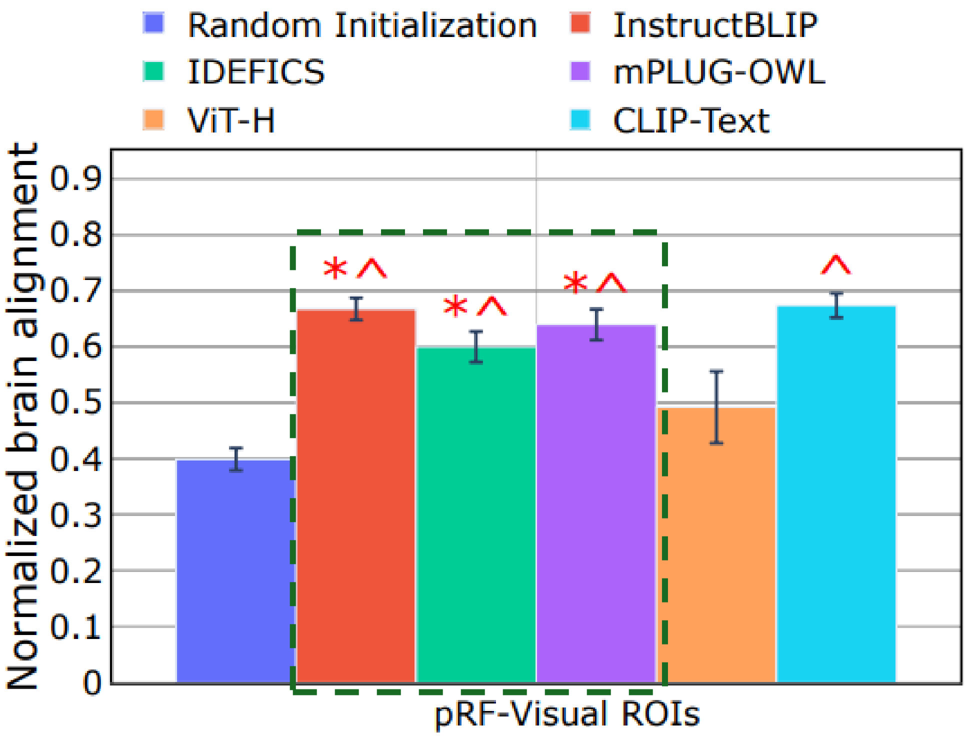
Task-specific natural instructions

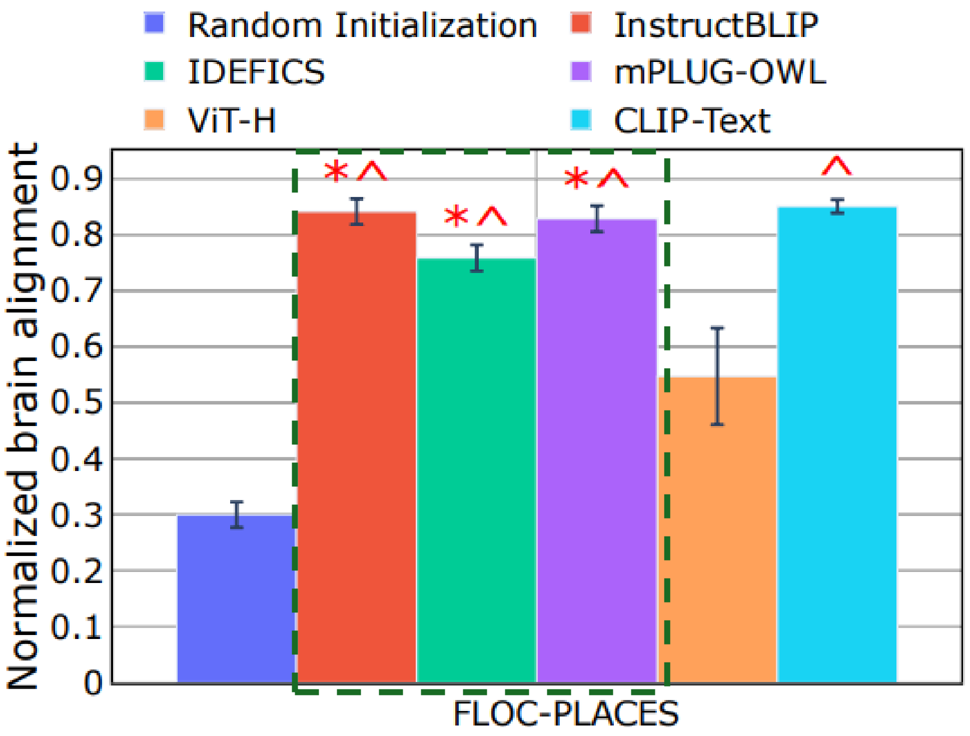
Task	Description
Image Understanding	IU1: Describe the most dominant color in the image IU2: List any food items visible. IU3: How many animals are there in the image?
Visual Question Answering	VQ1: What is in this image? VQ2: Are there any people in this image? If yes, describe them. VQ3: What is the 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?

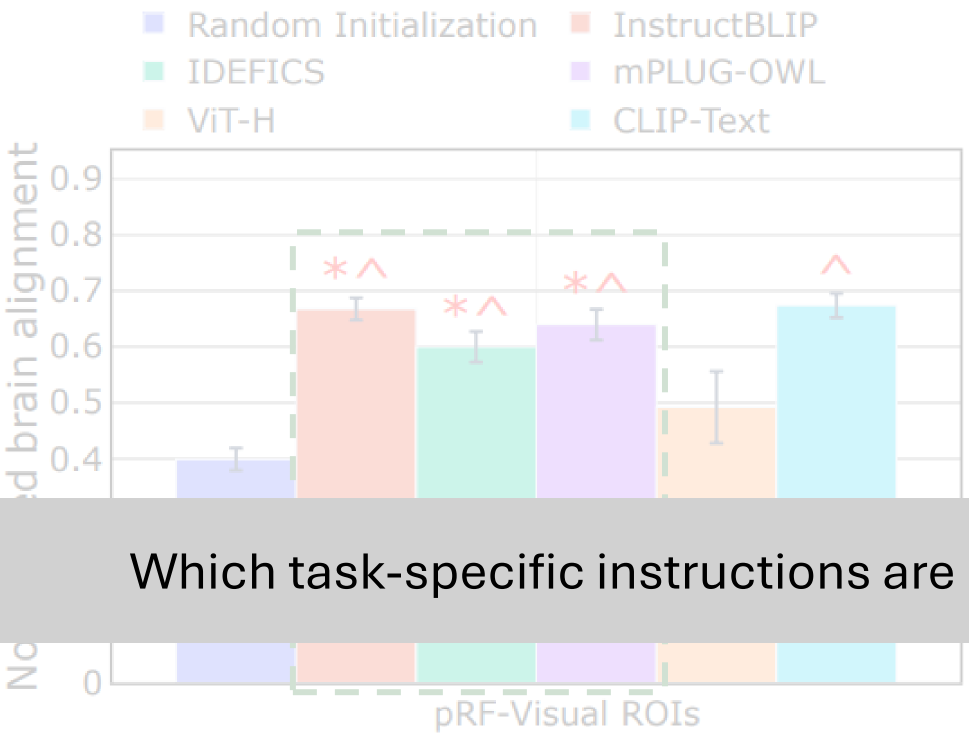
Result-1: MLLMs vs. Unimodal vs. Multi-modal models and brain alignment

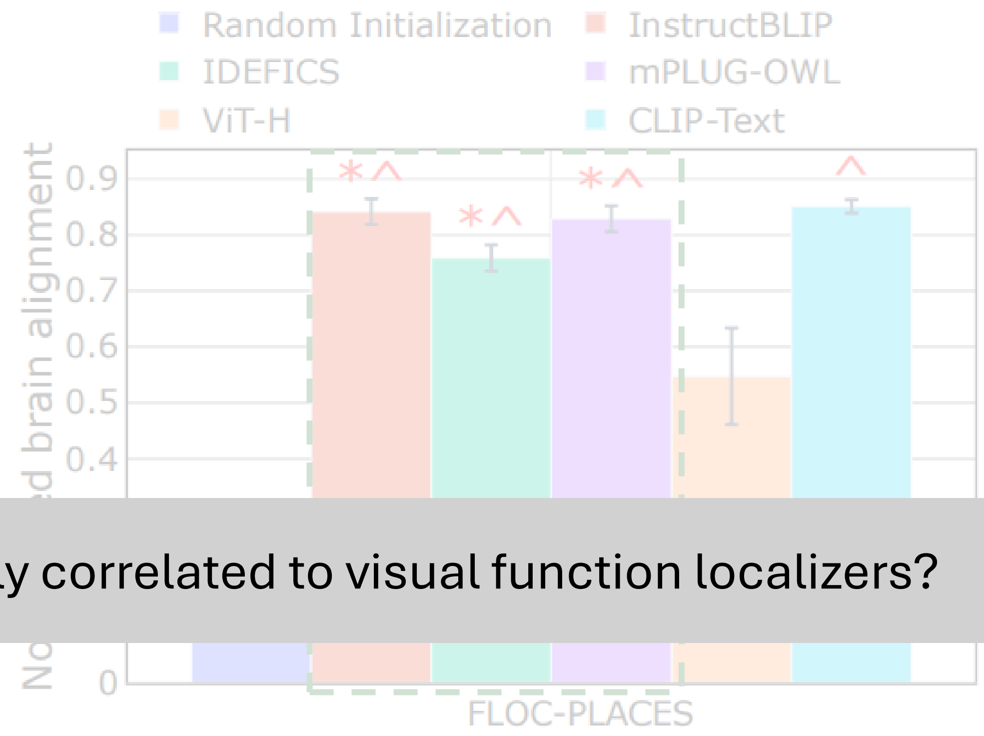




- **Early-visual regions**
  - Both **MLLMs** and **multi-modal** models show significantly high brain alignment than baseline and unimodal video models
  - Surprisingly, brain alignment of **random initialization of MLLMs** is closer to that of **unimodal video models**
- **Higher-visual regions**
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

Result-1: MLLMs vs. Unimodal vs. Multi-modal models and brain alignment

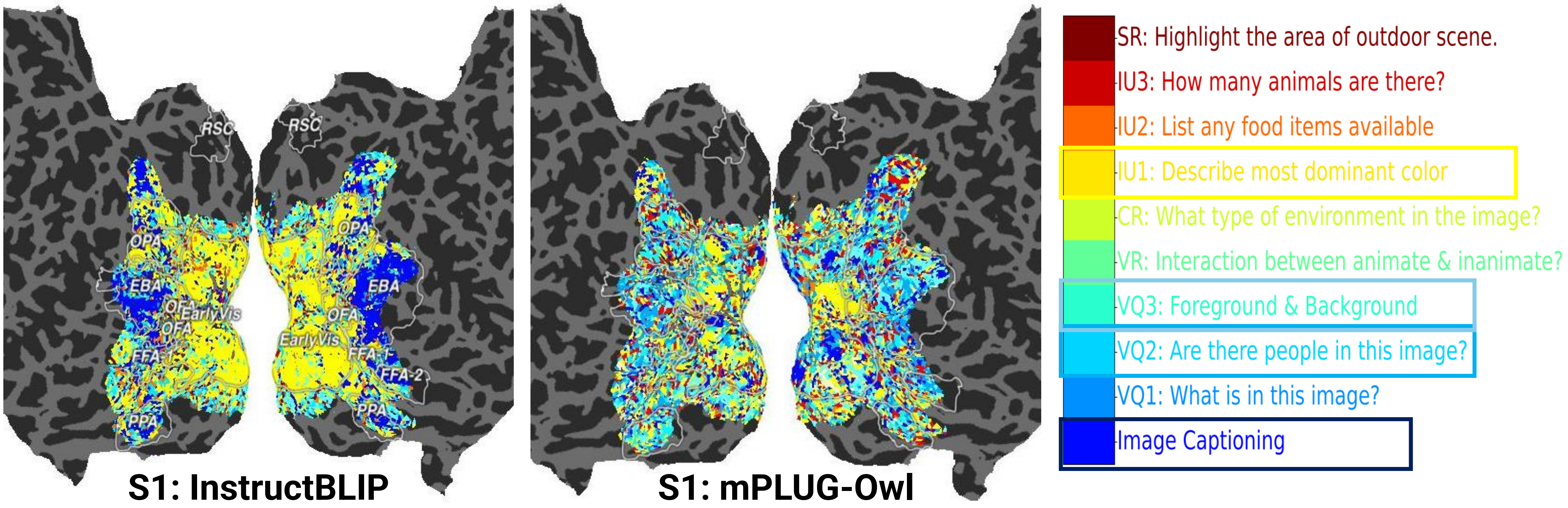




Which task-specific instructions are highly correlated to visual function localizers?

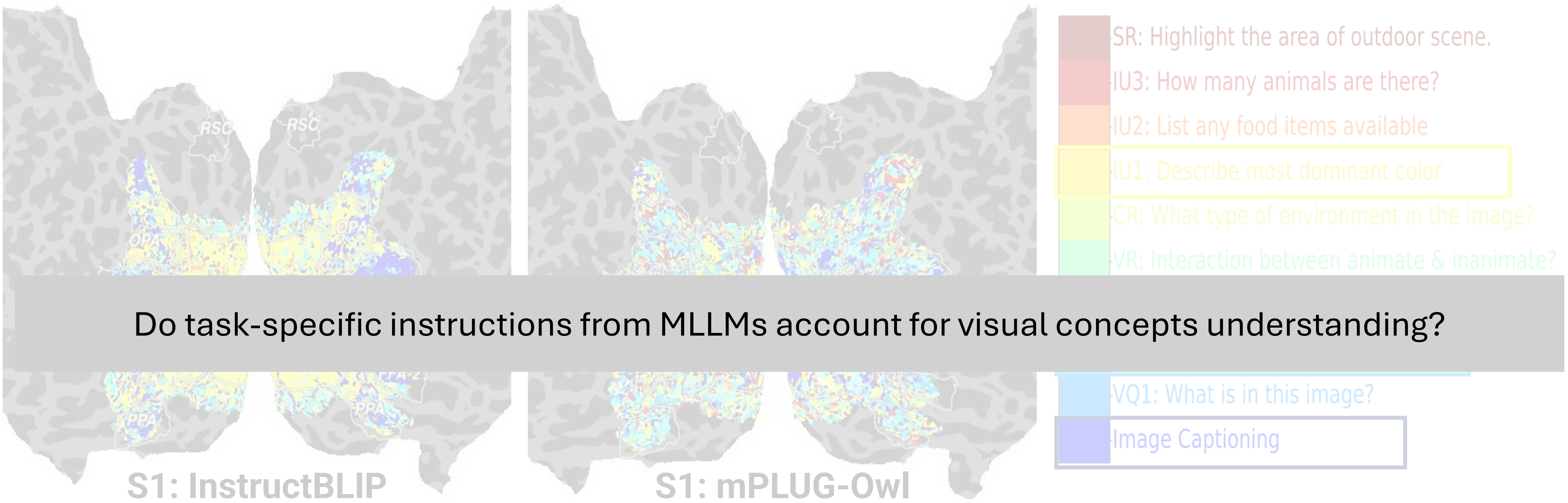
- **Early-visual regions**
  - Both **MLLMs** and **multi-modal** models show significantly high brain alignment than baseline and unimodal video models
  - Surprisingly, brain alignment of **random initialization of MLLMs** is closer to that of **unimodal video models**
- **Higher-visual regions**
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

Result-2: Which task-specific instructions are highly correlated to visual function localizers?



- **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**

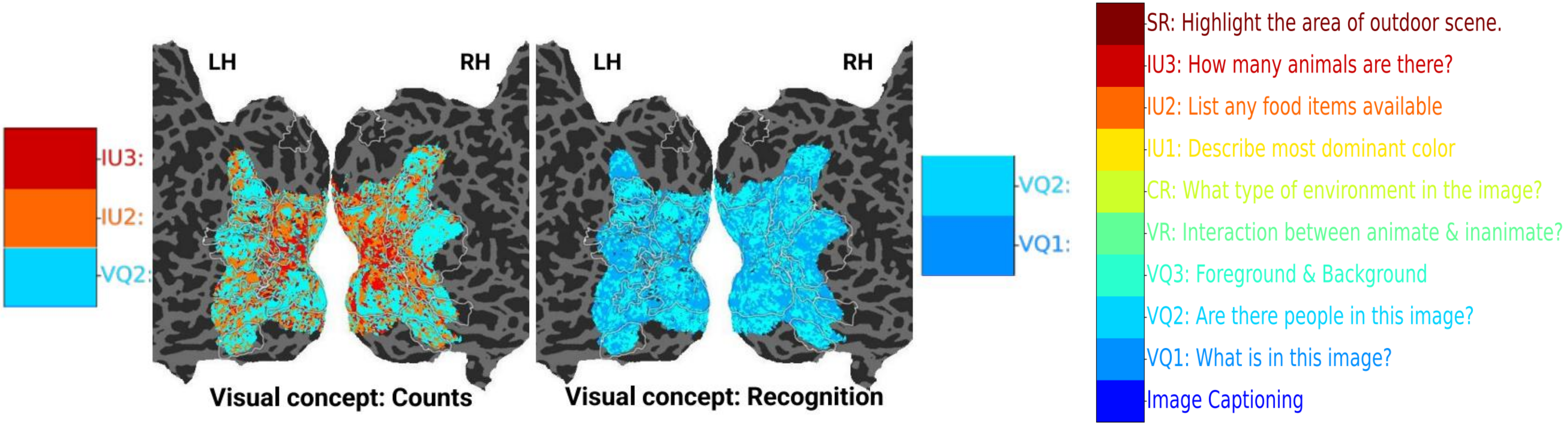
Result-2: Which task-specific instructions are highly correlated to visual function localizers?



Do task-specific instructions from MLLMs account for visual concepts understanding?

- **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**

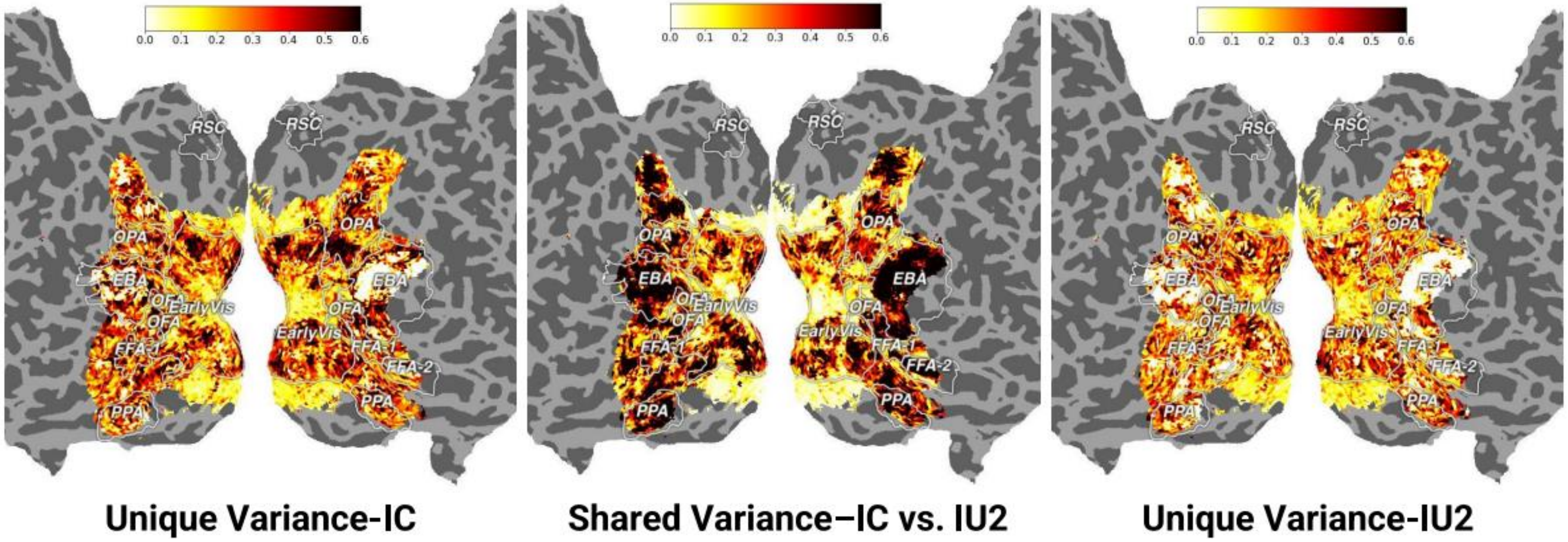
Result-3: MLLMs capture count- and recognition-related visual concepts effectively across instructions

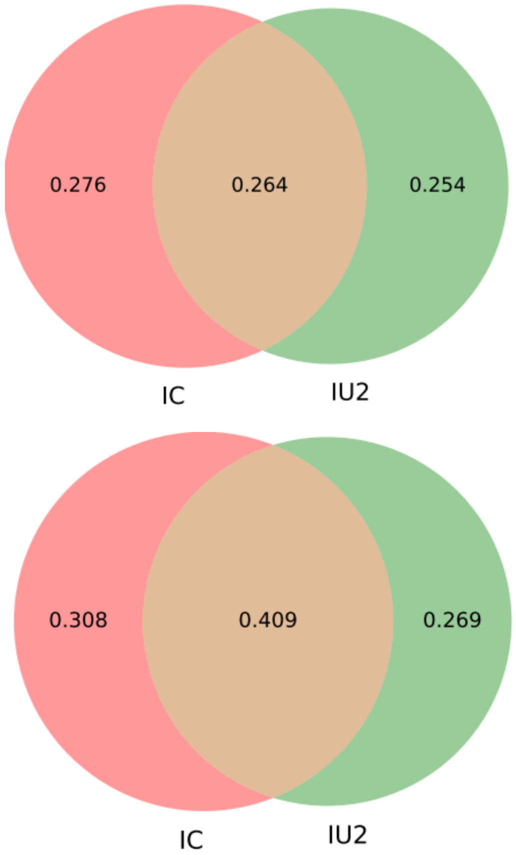


- **Visual concept-Count**
  - **VQ2** instruction shows significantly high brain alignment in **high-level visual regions**, while **IU2** and **IU3** instructions show higher alignment in **early visual regions**
- **Visual concept-Recognition**
  - Both **VQ1** and **VQ2** instruction show significantly high brain alignment across **high-level** and **early-visual 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





- 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**

Conclusions

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

Correlating instruction-tuning (in multimodal models) with vision-language processing (in the brain) (ICLR-2025)


Subba Reddy Oota


Akshett Jindal


Ishani Mondal


Khushbu Pahwa


Satya Sai Srinath


Manish Shrivastava


Maneesh Singh


Manish Gupta


Bapi S. Raju

Technische Universität Berlin

University of Maryland

Rice University

Microsoft

International Institute of Information Technology Hyderabad