10th International Workshop on Artificial Intelligence and Cognition,
Bologna, 25-26 october 2025

Network-Specific Models for Multimodal Brain Response Prediction

Andrea Corsico, Giorgia Rigamonti, Simone Zini, Luigi Celona, Paolo Napoletano
University of Milano-Bicocca, viale Sarca 336, 69121 Milano, Italy
Intelligent Sensing Lab

Financial support from:

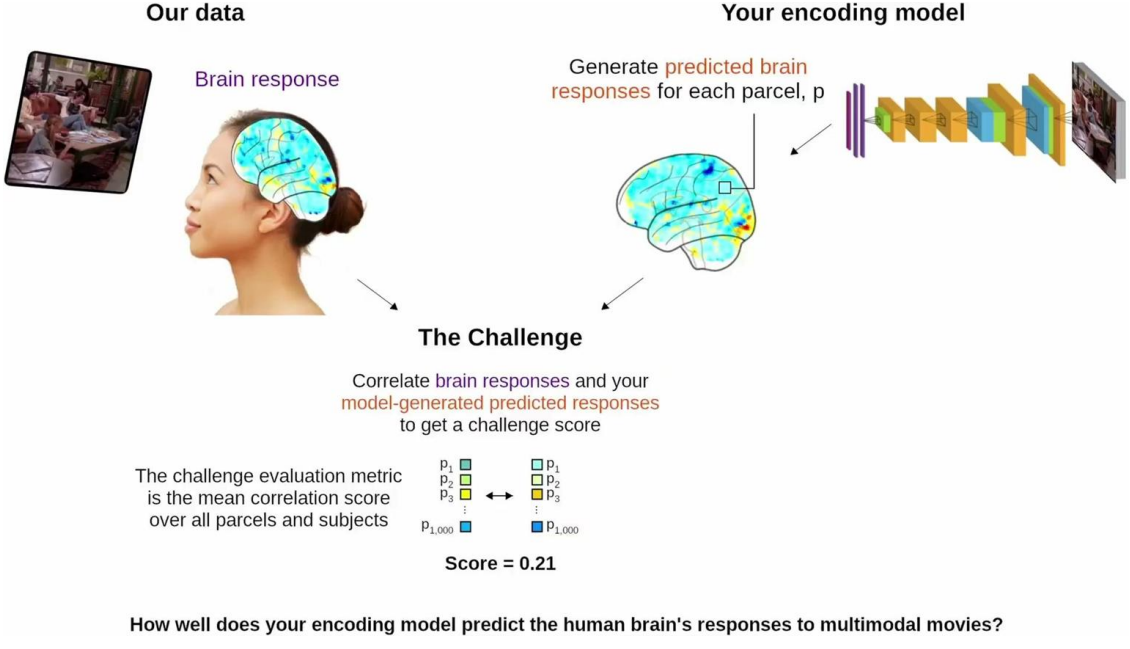
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

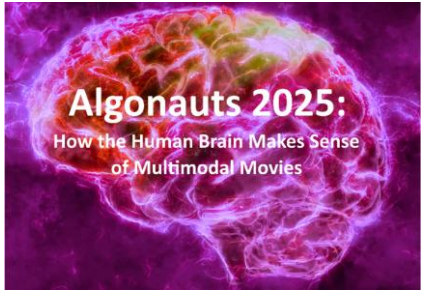
Anthem
AdvaNced Technologies for Human-centEred Medicine

Motivation

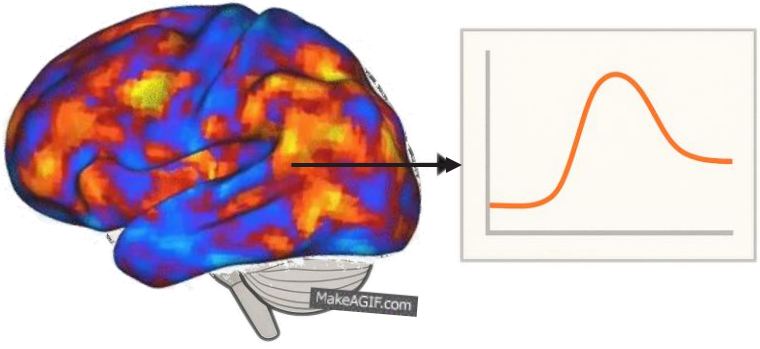
Understanding **how the brain integrates vision, sound, and language** during naturalistic experiences remains a core challenge in cognitive neuroscience

We model **brain activity evoked by movies** using multimodal deep learning and functional brain network priors





What is fMRI?



- **fMRI** (functional Magnetic Resonance Imaging) measures **blood-oxygen-level-dependent (BOLD)** signals – an ***indirect marker*** of neural activity.
- Each scan captures **3D brain volumes over time** (~1.49 s per sample).
- Signals are **parcellated into 1,000 cortical regions** using the Schaefer atlas.
- The BOLD signal **lags neural activity by ~4–6 seconds** due to vascular dynamics.

fMRI captures slow hemodynamic responses,
whereas **EEG** records **fast electrical signals** from neurons

Dataset

Courtois NeuroMod Dataset

65 hours of movie stimuli and corresponding fMRI responses:

55 hours of seasons 1 to 6 of the sitcom Friends

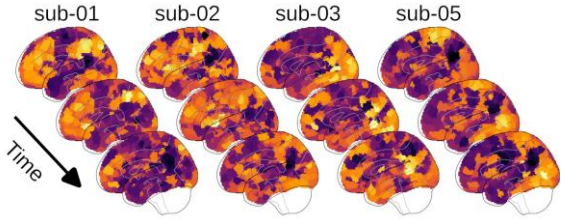
10 hours for the following four movies: ***The Bourne Supremacy***, ***Hidden Figures***, ***Life*** (a BBC nature documentary), and ***The Wolf of Wall Street***

~70h	4	1000
Duration	*Participants*	*Brain Parcels*

Multimodal Stimuli

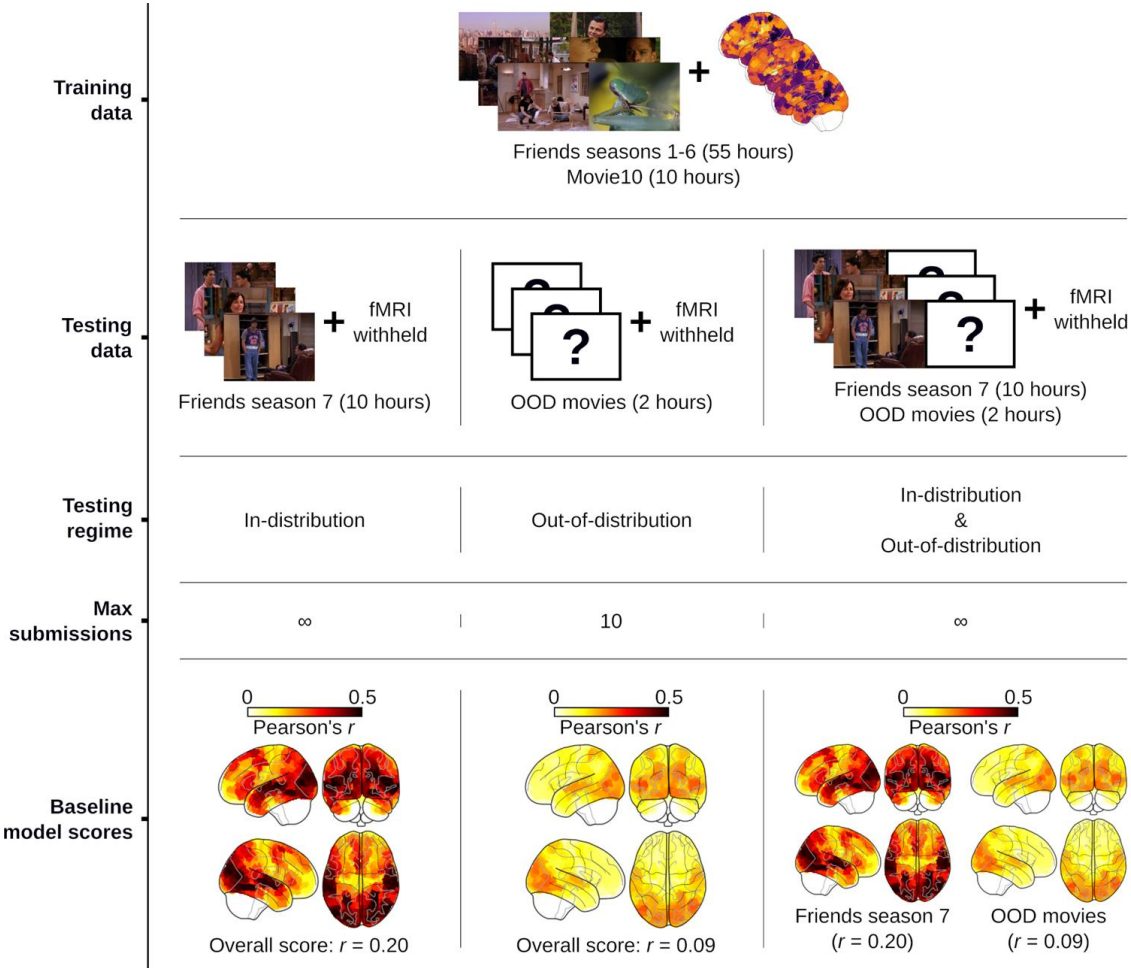
- Frame rate: ~ 30 FPS
- Resolution: 720×480 px
- Audio: 44.1 kHz sampling rate
- Timestamped language transcripts

fMRI Data



- Recording sessions: 12-15 min (~ half episode each)
- Temporal resolution (TR) = 1.49 seconds
- MNI template projection for standardization
- Transformed to Schaefer-1000 functional atlas

The Algonauts 2025 challenge



Task: predict brain responses to movie clips.

Evaluation: Pearson correlation per voxel.

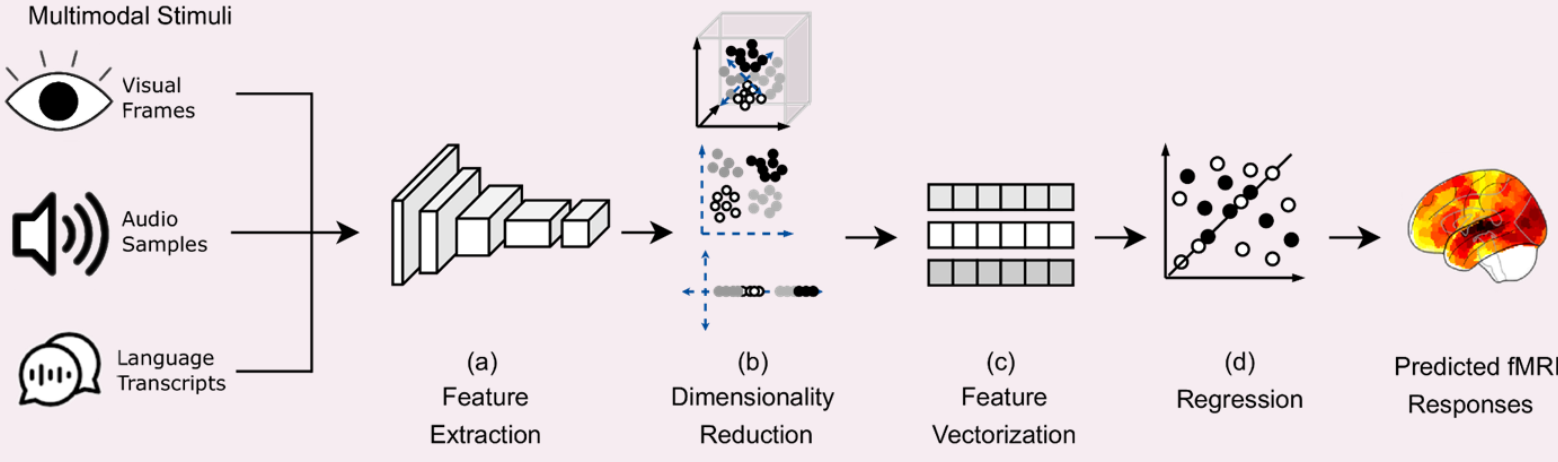
Goal: Achieve high accuracy on a new Friends season (in-domain, ID) and strong generalization to unseen movies (out-of-domain, OOD).

Proposed Model

1.49 s

[0.42, -0.87, 1.15, ..., 0.08, -1.02, 0.67]

BOLD measurements for the
1000 cortical regions



Proposed Model - Feature Extraction

ViNET captures attentional saliency, highlighting ***where humans focus visually*** in each frame.

VideoMAE2 encodes ***spatiotemporal dynamics***, modeling ***object motion, scene changes, and visual context over time***.

Together, they represent both **low-level perception** (salient regions) and **high-level visual semantics** (actions and scenes)

Wav2Vec2.0 captures speech content and phonetic structure, reflecting ***linguistic and prosodic information***.

AudioPANNs extract environmental and musical cues, encoding ***non-speech sounds*** like ambient noise or music.

openSMILE models low-level acoustic and emotional prosody, capturing ***intonation, rhythm, and affective tone***.

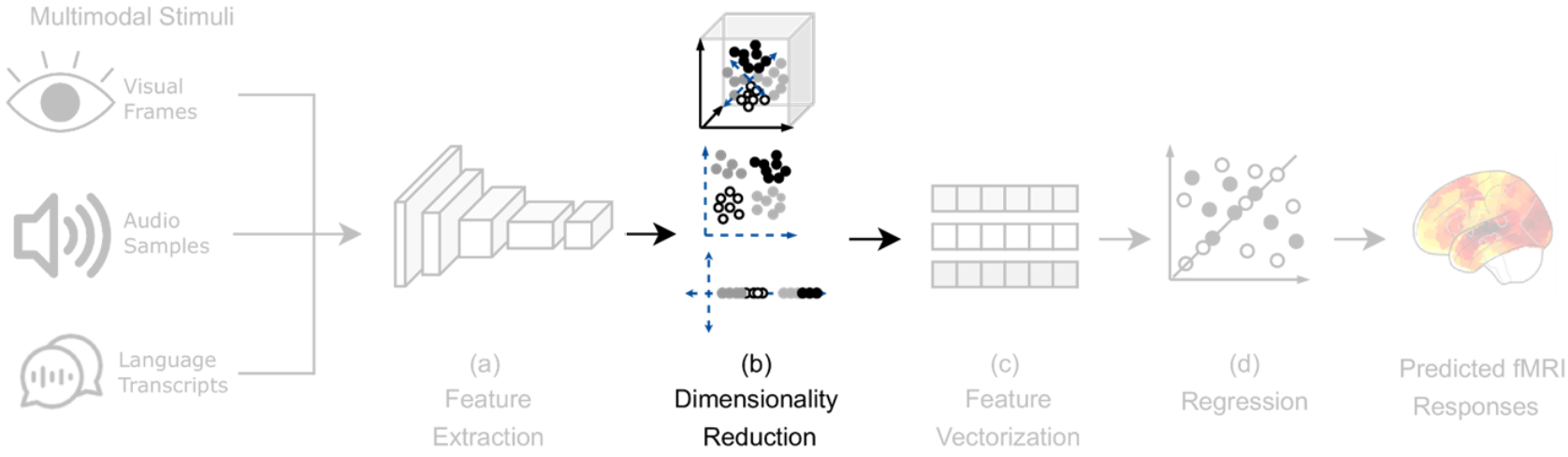
Combined, they encode the **semantic, environmental, and emotional** dimensions of the auditory stream.

RoBERTa represents semantic and contextual meaning from subtitles aligned with the fMRI timeline.

Intermediate layer embeddings (layer 8) capture ***conceptual and syntactic structure*** most correlated with brain activity.

Reflects **high-level linguistic processing** in cortical language networks

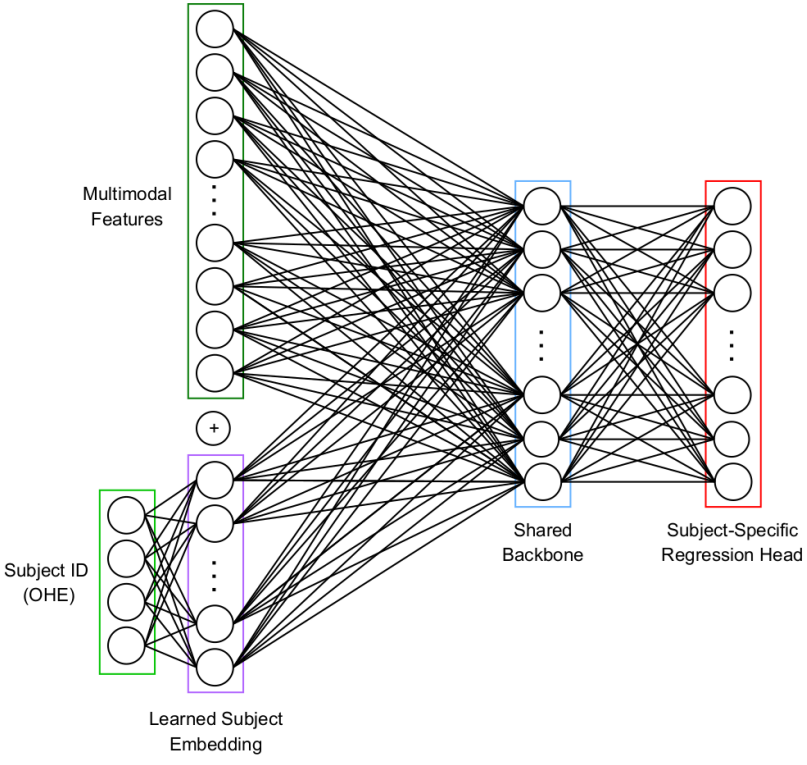
Proposed Model - Dimensionality Reduction



Principal Component Analysis
applied uniformly across modalities, compressing each to a 250-dimensional feature vector

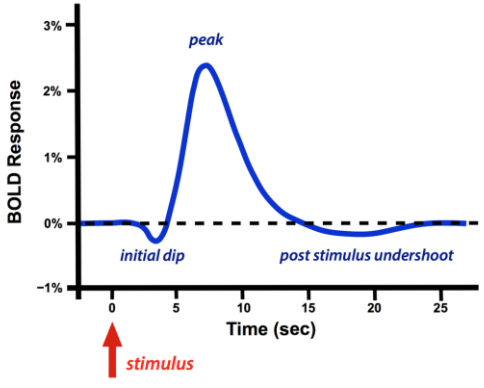
Proposed Model - Feature Vectorization & Regression

- **Multimodal features** are concatenated with a **learned subject embedding** derived from one-hot encoded subject IDs.
- A **shared backbone network** learns **common representations** across all participants.
- **Subject-specific regression heads** model **individual variability** in brain responses.



Proposed Model - Yeo Network–Specific Memory Dynamics

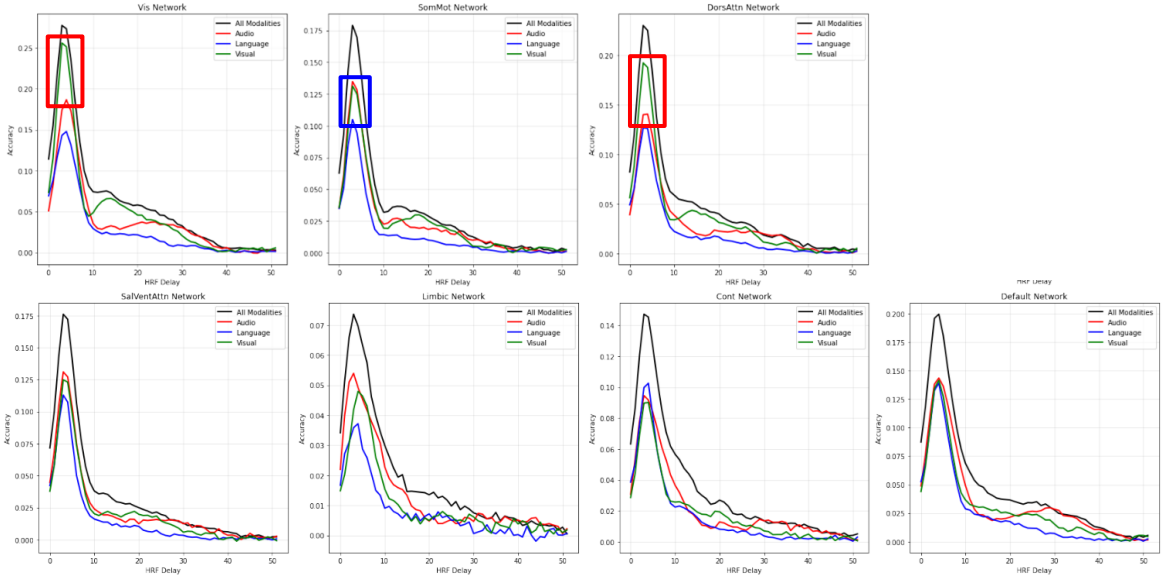
Hemodynamic Response Function



Models the **delayed BOLD response** to neural activity.

Peak ≈ 4–6 s, return to baseline ≈ 12–20 s. We found that **2 s works better**.

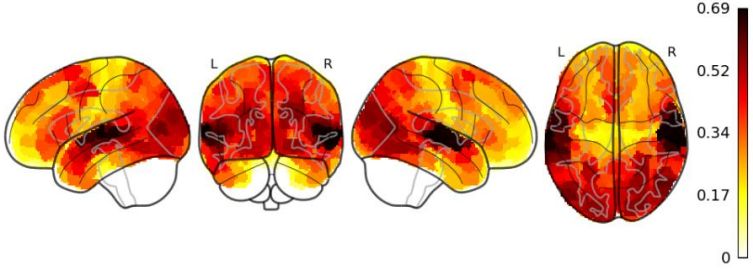
Essential for accurate **temporal encoding** in fMRI models.



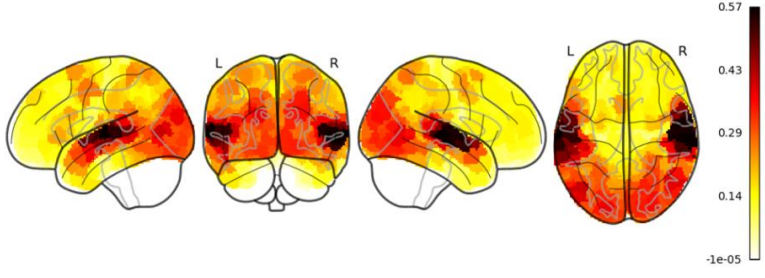
- **Unique temporal profiles for each modality–network pair**, highlighting memory-related effects.
- Visual and Dorsal Attention benefited from visual memory; Somatomotor from both visual and auditory.
- Four MLPs—one per network with distinct modality behavior, plus a fourth for the remaining four networks.

Results Overview

In-distribution



Out-of-distribution



- Robust predictions in **auditory and language regions**.
- **Stable patterns** across subjects and stimulus domains.
- **Performance drop** in visual regions → domain gap (animated vs. live-action).
- Maintained **functional consistency** in temporal cortex.

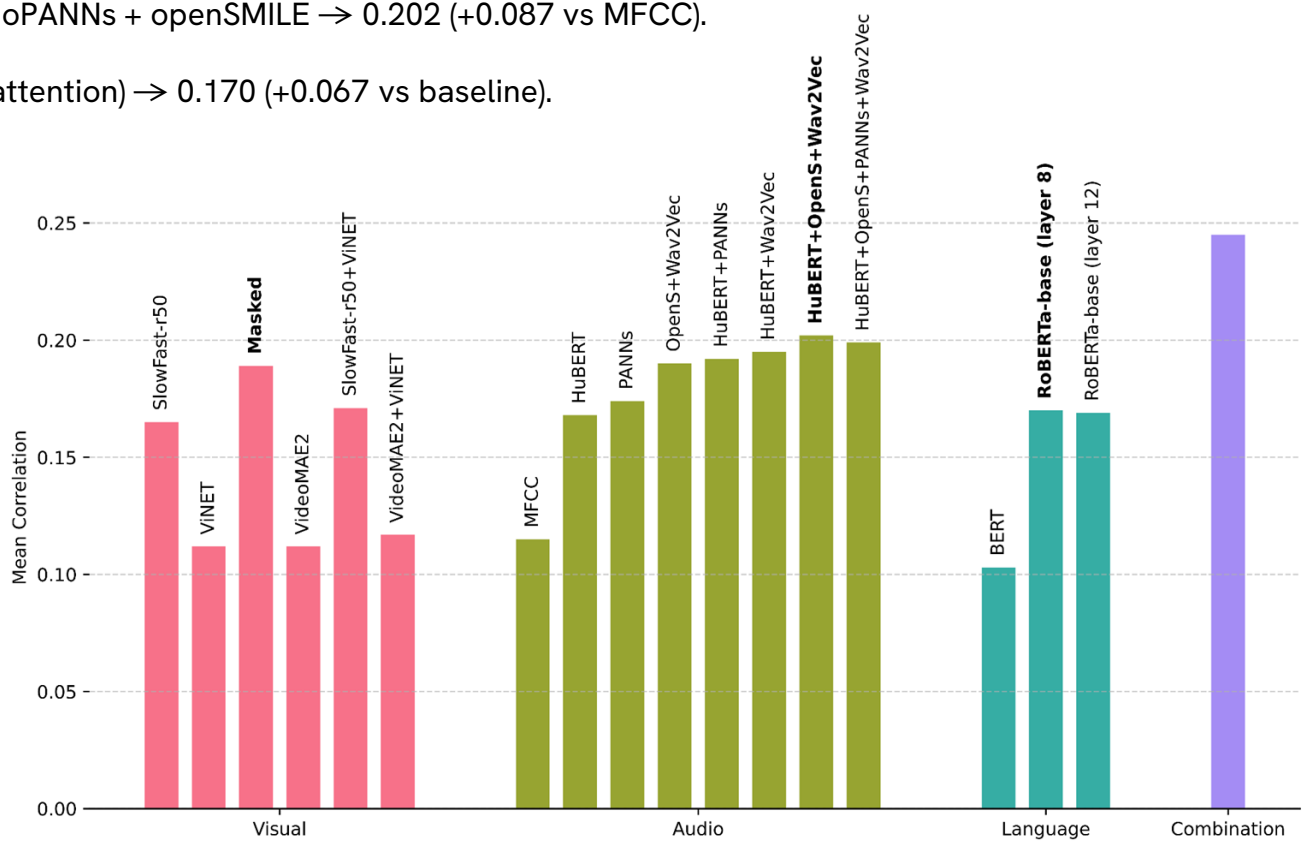
Challenge Leaderboard

Rank	Team	Score
1	NCG	0.320
2	sdascoli	0.319
3	SDA	0.313
4	angelneer926	0.296
5	CVIU-UARK	0.296
6	VIL	0.295
7	MedARC	0.288
8	ckadirt	0.273
9	**corsi01**	**0.266**
10	ICL_SNU	0.263
⋮	⋮	⋮
34	Baseline	0.203

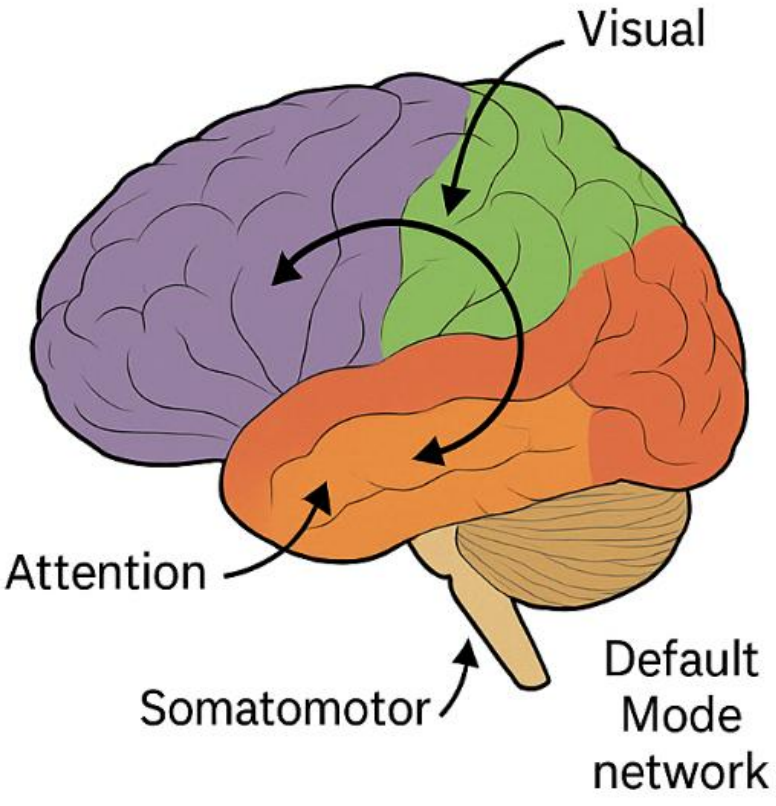
Rank	Team	Score
1	sdascoli	0.215
2	NCG	0.210
3	SDA	0.209
4	ckadirt	0.209
5	CVIU-UARK	0.205
6	angelneer926	0.199
7	ICL_SNU	0.161
8	**corsi01**	**0.158**
9	alit	0.157
10	robertscholz	0.150
⋮	⋮	⋮
21	Baseline	0.090

Ablation Study

- **Visual:** ViNET saliency masking → best performance (↑ focus on relevant regions).
- **Audio:** Wav2Vec2.0 + AudioPANNs + openSMILE → 0.202 (+0.087 vs MFCC).
- **Language:** RoBERTa (L8 + attention) → 0.170 (+0.067 vs baseline).



Cognitive Interpretation



- Distinct networks show specialized multimodal integration.
- Visual & Attention systems benefit from temporal memory → predictive processing.
- Language regions align with semantic Transformer embeddings (RoBERTa L8).
- Consistent spatial patterns across subjects → shared neural coding principles.

Conclusions and Future Work

Main Contributions

- Network-specific multimodal model based on Yeo 7 networks.
- Rich visual, audio, and language features aligned with fMRI timing.
- Multi-subject MLP capturing shared and individual variability.
- Incorporating network-dependent temporal memory enhances sensory response prediction.
- Top-10 in Algonauts 2025, doubling baseline OOD accuracy.

Key Insights

- Brain networks show specialized multimodal integration.
- Supports predictive-processing and hierarchical cognition.

Future Work

- Add recurrent / transformer-based memory.
- Extend to EEG, MEG, and larger datasets.
- Align with multimodal foundation models.

Scan for GitHub repository

THANKS FOR YOUR ATTENTION!

Intelligent Sensing Lab

https://islab.disco.unimib.it/