The Algonauts Project 2025 Challenge:

Andrea Corsico, Giorgia Rigamonti, Simone Zini, Luigi Celona and Paolo Napoletano

Department of Informatics, Systems and Communication, University of Milano-Bicocca, viale Sarca 336 – 20126, Milano, Italy

Abstract

In this work, we present our approach for predicting brain responses to complex multimodal movies by leveraging network-specific modelling based on the Yeo 7-network parcellation of the Schaefer atlas.Instead of treating the brain as a homogeneous system, we grouped the seven Yeo functional networks into four distinct clusters. We implemented separate multi-layer perceptron (MLP) multisubject models for each cluster, enabling network-specific optimization and memory modelling. This architecture allowed us to incorporate variable memory components that adapt both temporal dynamics and modality usage based on the functional properties of each network cluster. Our results demonstrate that this network-clustered approach, combined with multi-subject modeling, significantly improves prediction accuracy across the 1000 regions of the Schaefer atlas, achieving an eighth-place ranking in the challenge with correlation scores nearly double those of the baseline in the model selection phase (OOD testing). Code is available at https://github.com/Corsi01/algo2025

Keywords

Brain encoding model, Deep Learning, fMRI, Neuroimaging

1. Introduction

A major goal of computational neuroscience is to model how the human brain responds to naturalistic stimuli. Traditional brain encoding models have focused on single sensory modalities using controlled laboratory stimuli, achieving success in predicting neural responses within specific cortical regions. However, real-world perception involves simultaneous processing of multiple sensory modalities—visual, auditory, and linguistic information—across distributed brain networks.

Recent advances in deep learning have provided powerful computational tools for modeling the brain. Linear encoding models initially demonstrated the feasibility of predicting neural responses using features extracted from artificial neural networks. More sophisticated approaches have since emerged, including the use of transformer architectures and multimodal models that can process diverse types of sensory input simultaneously. The development of large-scale neuroimaging datasets has enabled more comprehensive brain modeling approaches. While earlier datasets were limited in scope or modality, the CNeuroMod dataset provides extensive fMRI recordings of brain responses to naturalistic movie stimuli, offering an unprecedented opportunity for developing and testing whole-brain encoding models.

The Algonauts Project 2025 Challenge [1] leverages this dataset to evaluate computational models on their ability to predict brain responses to multimodal movie content while generalizing to new stimulus distributions.

2. Background and Related Work

Recent work in the Algonauts Project has highlighted two key insights that, while developed for previous challenge scales, provide valuable guidance for the current multimodal whole-brain approach.

First, [2] demonstrated the importance of modeling temporal dynamics and memory components in brain encoding models, showing that incorporating information from past stimuli significantly improves prediction accuracy. Second, [3] showed the benefits of training encoding models across multiple

a.corsico@campus.unimib. (A. Corsico); giorgia.rigamonti@unimib.it (G. Rigamonti); simone.zini@unimib.it (S. Zini); luigi.celona@unimib.it (L. Celona); paolo.napoletano@unimib.it (P. Napoletano)
0009-0006-4253-1020 (G. Rigamonti); 0000-0002-8505-1581 (S. Zini); 0000-0002-5925-2646 (L. Celona); 0000-0001-9112-0574 (P. Napoletano)

subjects, followed by subject-specific fine-tuning, leveraging shared patterns of neural organization while accounting for individual differences.

The fMRI data to be predicted in this challenge are organized according to the Schaefer 1000-region atlas, which provides a comprehensive parcellation of the human cerebral cortex into 1000 functionally-defined regions that offer whole-brain coverage while maintaining sufficient spatial granularity [4]. The Yeo 7-network parcellation organizes these regions into seven large-scale functional networks [5]. Each network exhibits distinct functional properties and connectivity patterns, with each of the 1000 Schaefer regions assigned to one of these seven networks based on their predominant functional connectivity profile (Figure 1).

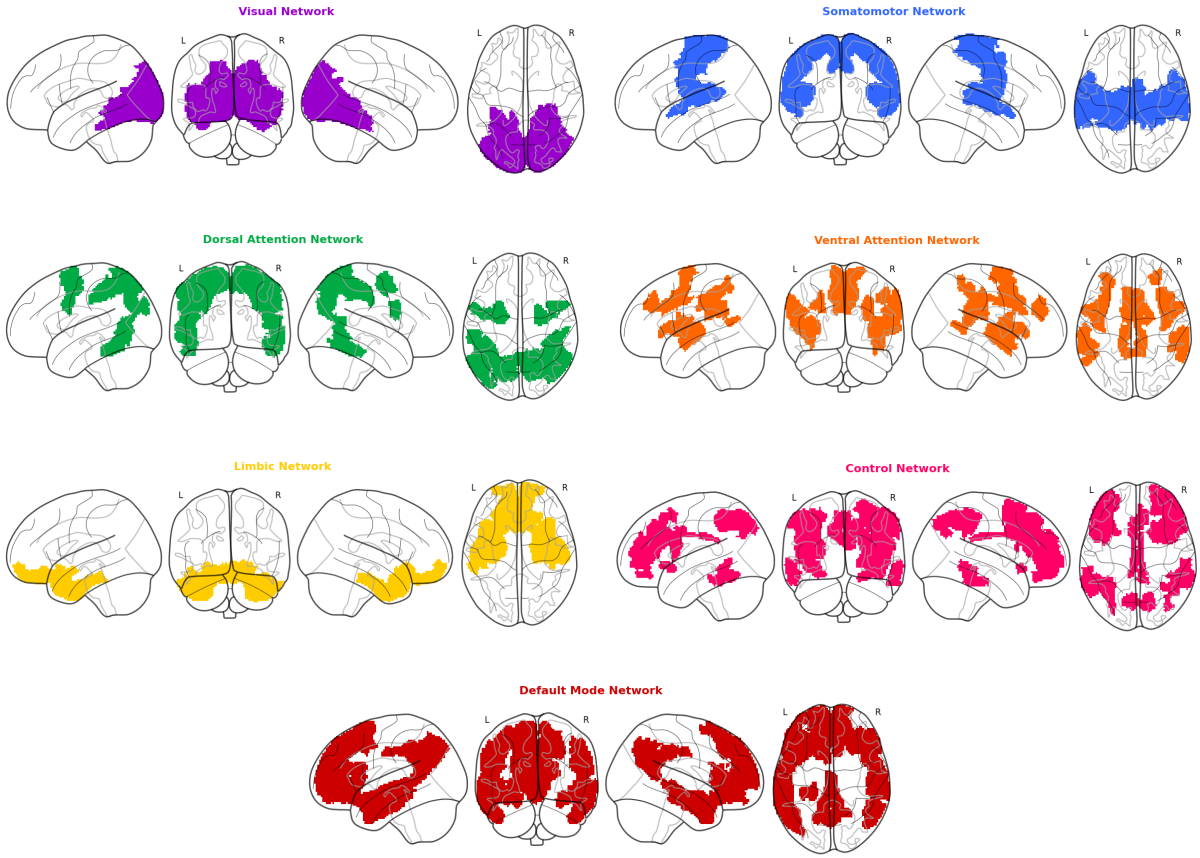


Figure 1: *Yeo 7-network parcellation of Schaefer atlas*

3. The Algonauts Challenge

3.1. Problem Definition

The Algonauts Project 2025 Challenge introduces multimodal movie stimuli containing simultaneous visual, auditory, and linguistic information, representing a significant increase in complexity compared to previous editions that focused on static images or visual scenes [6, 7]. This multimodal nature fundamentally changes the encoding problem, requiring models to integrate information across sensory modalities while predicting responses across the complete Schaefer 1000-region atlas.

Traditional layer-wise feature selection approaches, where different computational model layers are mapped to hierarchically organized brain regions, become computationally intractable at this scale. The challenge requires prediction across functionally heterogeneous brain systems, including sensory, motor, attention, language, and default mode networks. Each system exhibits distinct temporal

dynamics, modality preferences, and information processing characteristics, necessitating flexible modeling approaches that can adapt to this functional diversity. A critical aspect of the 2025 challenge is the test of out-of-distribution generalization. Models trained on Friends episodes must generalize to entirely different movie content with distinct visual styles, narrative structures, and acoustic properties. This requirement ensures that successful models capture fundamental principles of brain organization rather than stimulus-specific patterns.

The temporal dimension adds further complexity, as movie stimuli unfold over time with different brain networks exhibiting varying temporal receptive windows. Primary sensory areas show rapid responses while association cortex demonstrates extended integration periods, requiring adaptive temporal modeling strategies.

Finally, the challenge provides fMRI data from four subjects experiencing identical stimulation, creating constraints on data diversity while emphasizing the importance of models that can capture both shared neural principles and individual differences in brain response patterns.

3.2. Data

This challenge is based on the CNeuroMod dataset [8], one of the largest available collections of human brain responses to naturalistic movie stimuli. This dataset provides over 80 hours of fMRI data across four subjects, all experiencing identical stimulation protocols. The training data consists of two main components: the complete Friends series (seasons 1 - 6) and the Movie10 dataset, which includes four feature films: Life, The Bourne Identity, The Wolf of Wall Street, and Hidden Figures.

The fMRI data are provided in preprocessed format with a temporal resolution of TR = 1.49 seconds. All brain data have been normalized and transformed to the Schaefer 1000-region atlas, a comprehensive parcellation of the human cortex that divides the brain into 1000 functionally defined regions of interest. This atlas provides whole-brain coverage while maintaining sufficient spatial granularity to capture regional response differences across diverse functional systems.

The challenge evaluation follows a two-phase structure that tests both in-distribution and out-of-distribution generalization capabilities. During the model-building phase, encoding models are trained on the available data and evaluated on Friends season 7, maintaining stimulus similarity while testing temporal generalization. The out-of-distribution evaluation phase employs an entirely different set of movies: Chaplin, Princess Mononoke, Planet Earth, Around the World in 80 Days, World of Tomorrow, and Pulp Fiction. This OOD test set was specifically designed to challenge model generalization across multiple dimensions. The films vary in language content (English-speaking, French-speaking, non-verbal), visual styles (realistic cinematography, nature documentary, black-and-white silent film, and stick figure animation), and scene characteristics (indoor dialogue scenes, outdoor action sequences, and natural environments). This diversity ensures that successful models must capture fundamental principles of brain organization rather than stimulus-specific patterns, providing a rigorous test of encoding model robustness and biological validity.

4. Proposed Model

4.1. Feature extraction

A consistent approach was adopted across all modalities for feature extraction. Features were extracted from 1.49-second stimulus windows corresponding to the fMRI TR, ensuring precise temporal alignment between stimulus features and neural responses.

The extraction process typically yielded high-dimensional embeddings for video and audio or variable-length representations for text, creating dimensionality challenges for subsequent modeling. These inconsistencies were addressed through statistical pooling operations (mean, maximum, and standard deviation) to generate fixed-size feature vectors for each temporal window. Following statistical pooling, features from each modality were subjected to Principal Component Analysis (PCA) for dimensionality reduction.

4.1.1. Visual Features

Two complementary approaches were employed for visual feature extraction. First, we utilized the ViNET model to extract saliency maps from video frames, which were then used to mask the original video content [9]. The masked videos were processed through the same model to extract feature vectors from the backbone network. This approach aimed to filter visual information by maintaining only the most salient regions for human perception, potentially improving the biological relevance of the extracted features.

Second, we employed VideoMAE V2, a transformer-based model for video understanding that uses masked autoencoder pre-training on video sequences [10]. Features were extracted from the final layer of this pre-trained model to capture high-level temporal and spatial video representations.

4.1.2. Audio features

Three complementary approaches were employed for audio feature extraction to capture different aspects of the auditory signal. First, Wav2Vec2.0 was used to extract features related to speech content, leveraging its self-supervised training on speech data [11]. Second, openSMILE was utilized to capture low-level acoustic features including spectral, prosodic, and temporal characteristics of the audio signal [12]. Third, AudioPANNs (Pre-trained Audio Neural Networks) were employed to extract features related to environmental sounds, music, and other non-speech audio events [13]. This multi-faceted approach aimed to comprehensively represent the diverse auditory information present in naturalistic movie stimuli.

4.1.3. Language Features

Language features were extracted using RoBERTa-base, a transformer-based language model pre-trained on large-scale text corpora [14]. Contextualized word embeddings were extracted from the 8th hidden layer, which has been shown to produce optimal predictions of brain responses [15]. To obtain coherent sentence-level representations, statistical pooling operations were applied to the word embeddings within each temporal window. Additionally, attention weights from all transformer layers were extracted and incorporated as features. This approach was motivated again by findings from Lamarre et al. (2023), which demonstrated that attention weights accurately predict language representations in the brain and capture information about contextual integration processes that are not fully contained in the hidden state representations [15]. The attention patterns provide complementary information to the contextual embeddings, reflecting the mechanism by which the transformer integrates information across words.

4.2. Multi subjects model

To leverage data from all four subjects while accounting for individual differences in brain organization, we implemented a multi-subject MLP architecture. The model consists of a shared backbone network that learns common feature representations across subjects, combined with subject-specific prediction heads that account for individual neural response patterns.

The architecture incorporates trainable subject embeddings that transform subject identity from one-hot encoding to dense representations. These subject embeddings are concatenated with the input features and processed through a shared backbone that extracts subject-agnostic feature representations that capture shared patterns of neural encoding across individuals.

Individual differences are modeled through separate linear prediction heads for each subject, which map from the shared backbone representations to the region brain responses. This approach enables the model to learn both common neural encoding principles and subject-specific response characteristics, effectively increasing the diversity and size of the training data while maintaining the ability to capture individual variations in brain organization.

4.3. Network memory modeling

To explore the temporal dynamics of neural responses across functional networks, we conducted a systematic analysis of memory effects by fitting models with different lag windows (Figure 2). Each modality was tested across all seven Yeo networks, with results averaged across the four subjects. This revealed distinct temporal response profiles for each network-modality combination.

Based on these network-specific temporal characteristics, we implemented a data-driven approach to incorporate additional memory features by concatenating them to the input feature vector. The exploration revealed that three networks benefited significantly from memory components: Visual and Dorsal Attention networks showed improved performance with visual memory features, while the Somatomotor network benefited from both visual and audio memory features.

Consequently, we created four separate multi-subject MLP models: individual models for Visual, Somatomotor, and Dorsal Attention networks (each with their respective memory augmentations), and a combined model for the remaining four networks (Ventral Attention, Limbic, Frontoparietal, and Default Mode) that did not show substantial memory benefits. This architecture allows each model to optimize for the specific temporal dynamics and modality preferences observed in the corresponding brain networks.

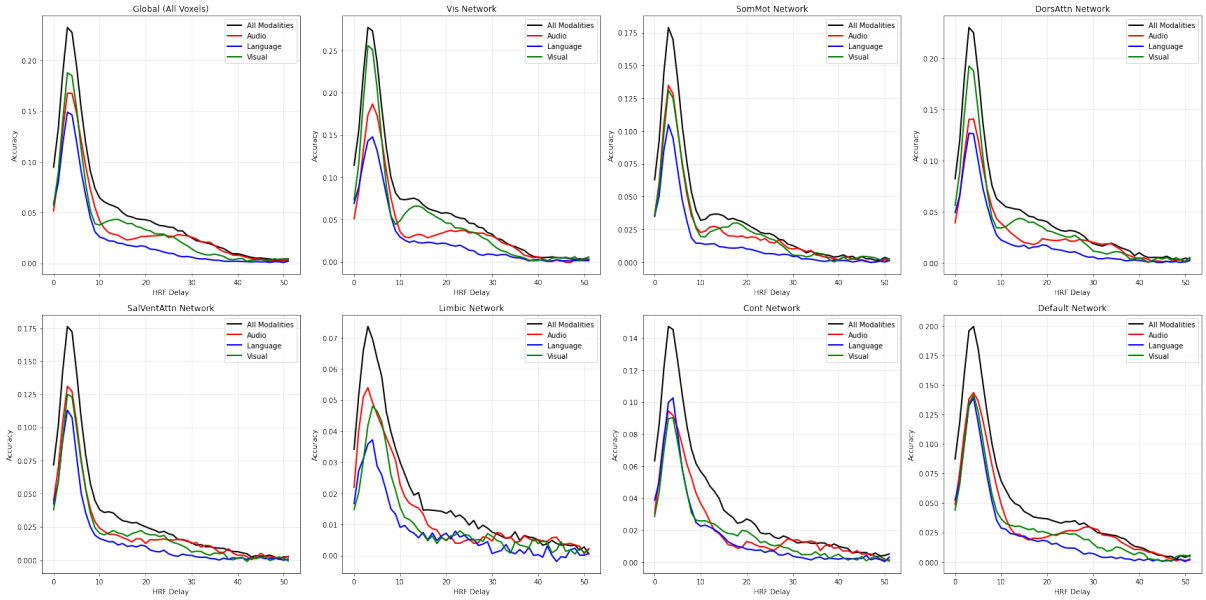


Figure 2: *Temporal response patterns across Yeo networks and modalities.* Correlation performance as a function of HRF delay (0-50 time points) for individual modalities and combined features across the seven functional networks.

5. Experiments

The model predicts neural responses at each time point independently, using a temporal window of past stimuli to account for the hemodynamic response function delay that characterizes the relationship between neural activity and the BOLD signal measured by fMRI. After systematic exploration of temporal parameters, we selected an HRF delay of 2 time points and a stimulus window of 7 time points for all modalities, obtained through grid search over various combinations.

Training was performed using the Adam optimizer with subject-weighted MSE loss to handle imbalanced subject representation in batches. All hyperparameters were optimized using Optuna [16] for each of the four network models separately, including the dimensions of the subject embedding and shared hidden layer, as well as heavy dropout and weight decay parameters. Given that the two visual feature types (ViNET saliency-masked and VideoMAE) did not perform well when used together,

we trained separate models for each visual feature approach.

The MLP architecture proved superior to independent ridge regression for each brain region but required strong regularization to prevent overfitting. All hyperparameter optimization and training procedures used Friends season 6 as the validation set, with final models retrained on the complete dataset using the optimized parameters before submission.

5.1. Brain responses to in-distribution (ID) movies

The model was evaluated on Friends season 7 as the in-distribution test set. Among the two visual feature approaches, the ViNET saliency-masked features consistently outperformed VideoMAE features across all network models, leading to the selection of ViNET-based models for final submission. Validation results on Friends season 6 revealed distinct performance patterns across the functional networks, confirming the effectiveness of the network-clustered approach. The network-specific correlations demonstrated that different brain networks exhibited varying degrees of predictability, with some networks showing substantial benefits from the memory-augmented modeling while others performed well with the standard approach (Table 1).

Table 1
Model performance across brain networks on validation set. (1) Memory models, (2) No-memory models, (a) ViNet, (b) VideoMAE2. Network sizes in parentheses: Visual(162), SomatoMotor(194), DorsalAttention(122), VentralAttention(121), Limbic(60), Default(212), Control(129), Mean(1000).

Model	Vis	Som	Dors	Vent	Limb	Def	Ctrl	Mean
MLP (1a)	0.3901	0.2286	0.3095	0.2196	0.1131	0.2747	0.2309	0.2667
MLP (1b)	0.3873	0.2232	0.2968	0.2135	0.1125	0.2735	0.2298	0.2625
MLP (2a)	0.3633	0.2178	0.3014	0.2174	0.1092	0.2713	0.2291	0.2586
Ridge (2a)	0.3613	0.2098	0.2921	0.2091	0.1015	0.2626	0.2180	0.2501

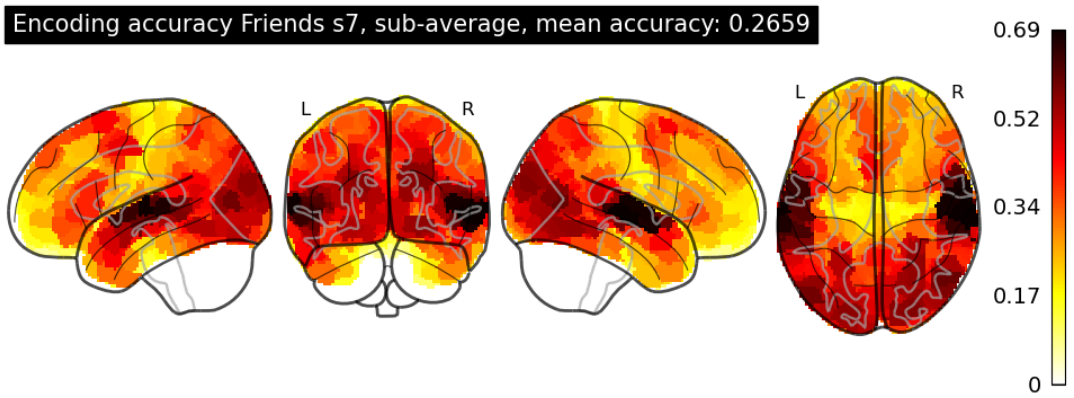
5.2. Brain responses to out-of-distribution (OOD) movies

The OOD films presented diverse visual and content characteristics. Chaplin, being a silent film, lacked speech content entirely. We addressed this by using the typical language feature representing the absence of speech for the entire duration of the film. Given the distinct visual characteristics of the OOD films, we adapted our visual feature extraction strategy. Since the ViNET backbone was trained on Kinetics-400 (primarily featuring common human actions), we employed VideoMAE features for Planet Earth (natural scenes) and Chaplin (black-and-white cinematography). This adaptation proved effective during the OOD evaluation phase, where VideoMAE demonstrated superior performance compared to ViNET for these specific film types.

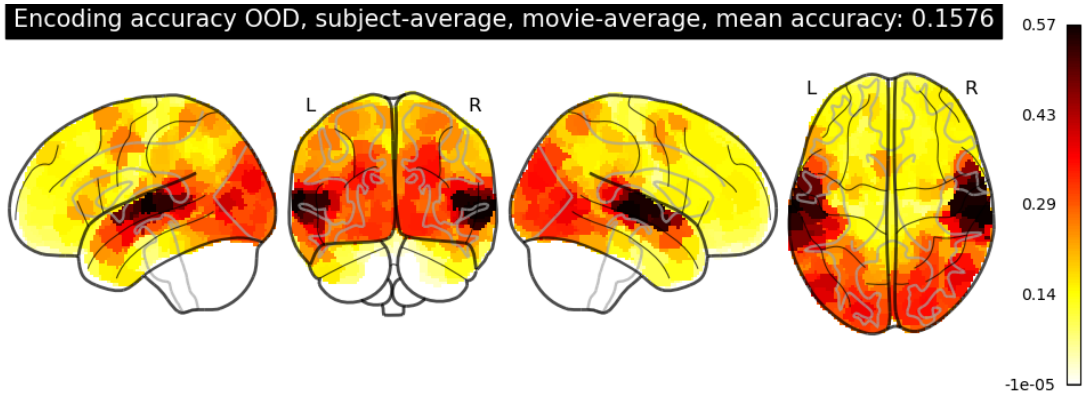
6. Results

Our approach achieved competitive performance in both evaluation phases and out-of-distribution evaluation. The model demonstrated substantial improvement over the baseline, validating the effectiveness of the network-clustered, multi-subject approach for multimodal brain encoding. The results revealed that language and auditory processing areas showed the highest prediction accuracy, with superior temporal regions and temporal cortex exhibiting the strongest correlations.

The transition from ID to OOD evaluation showed an expected decrease in overall performance due to the strong shift in stimulus characteristics. However, the spatial pattern of predictable regions remained relatively consistent, with language and auditory areas maintaining higher accuracy compared to visual regions (Figure 3). This suggests that our feature extraction approach successfully captured generalizable representations for audio-linguistic processing, while visual features may benefit from a broader range of characteristics represented.



(a) In-distribution



(b) Out-of-distribution

Figure 3: *Brain encoding performance comparison between in-distribution and out-of-distribution evaluation.* Correlation maps showing prediction accuracy across cortical regions, averaged across subjects(ID) and across cortical regions, movies and subjects(OOD).

The results indicate that higher-order processing areas involved in complex cognitive functions proved more challenging to predict than primary sensory regions, highlighting the inherent difficulty in modeling abstract neural computations across different stimulus distributions.

7. Conclusion

This work presented a novel approach for predicting brain responses to multimodal stimuli by leveraging functional brain organization. The network-clustered architecture, based on Yeo's 7-network parcellation, enabled specialized modeling approaches for different brain systems while incorporating adaptive memory components where beneficial.

Key contributions include the development of custom multimodal feature extraction pipelines and the effective scaling of encoding models to whole-brain prediction across 1000 regions. The multi-subject approach successfully captured both shared neural principles and individual differences, achieving competitive performance in the Algonauts 2025 Challenge.

Future work could explore more sophisticated memory mechanisms, more extensive visual feature extraction, and extend the network-clustered approach to other neuroimaging datasets and stimulus types.

Table 2
The Algonauts Project 2025 Challenge Model Selection and Evaluation Phases leaderboard. The Challenge Score is the Pearson's correlation between predicted and withheld fMRI responses averaged (a) across all brain parcels and subjects for ID data, and (b) across all brain parcels, movies, and subjects for OOD data.

Rank	Team Name	Challenge Score
1	NCG	0.3198
2	sdascoli	0.3195
3	SDA	0.3130
4	angelneer926	0.2961
5	CVIU-UARK	0.2958
6	VIL	0.2947
7	MedARC	0.2878
8	ckadirt	0.2730
9	**corsi01**	**0.2659**
10	ICL_SNU	0.2626
⋮	⋮	⋮
34	Baseline	0.2033

(a)

Rank	Team Name	Challenge Score
1	sdascoli	0.2146
2	NCG	0.2096
3	SDA	0.2094
4	ckadirt	0.2085
5	CVIU-UARK	0.2054
6	angelneer926	0.1986
7	ICL_SNU	0.1612
8	**corsi01**	**0.1576**
9	alit	0.1574
10	robertscholz	0.1496
⋮	⋮	⋮
21	Baseline	0.0895

(b)

Acknowledgment

Financial support from ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union – NextGenerationEU.

This work was partially funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM)[1]. This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

References

[1] A. T. Gifford, D. Bersch, M. St-Laurent, B. Pinsard, J. Boyle, L. Bellec, A. Oliva, G. Roig, R. M. Cichy, The algonauts project 2025 challenge: How the human brain makes sense of multimodal movies, 2025. URL: https://arxiv.org/abs/2501.00504. arXiv:2501.00504.

[2] H. Yang, J. Gee, J. Shi, Memory encoding model, 2023. URL: https://arxiv.org/abs/2308.01175. arXiv:2308.01175.

[3] X.-B. Nguyen, X. Liu, X. Li, K. Luu, The algonauts project 2023 challenge: Uark-ualbany team solution, 2023. URL: https://arxiv.org/abs/2308.00262. arXiv:2308.00262.

[4] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, B. T. T. Yeo, Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri, Cerebral Cortex 28 (2017) 3095–3114.

[5] B. T. Yeo, F. M. Krienen, J. Sepulcre, et al., The organization of the human cerebral cortex estimated by intrinsic functional connectivity, Journal of Neurophysiology 106 (2011) 1125–1165.

[6] A. T. Gifford, B. Lahner, S. Saba-Sadiya, M. G. Vilas, A. Lascelles, A. Oliva, K. Kay, G. Roig, R. M. Cichy, The algonauts project 2023 challenge: How the human brain makes sense of natural scenes, arXiv preprint arXiv:2301.03198 (2023).

[1] https://fondazioneanthem.it/

[7] R. M. Cichy, K. Dwivedi, B. Lahner, A. Lascelles, P. Iamshchinina, M. Graumann, A. Andonian, N. Murty, K. Kay, G. Roig, et al., The algonauts project 2021 challenge: How the human brain makes sense of a world in motion, arXiv preprint arXiv:2104.13714 (2021).

[8] J. Boyle, B. Pinsard, V. Borghesani, F. Paugam, E. DuPre, P. Bellec, The courtois neuromod project: quality assessment of the initial data release (2020), in: 2023 Conference on Cognitive Computational Neuroscience, 2023, pp. 2023–1602.

[9] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, V. Gandhi, Vinet: Pushing the limits of visual modality for audio-visual saliency prediction, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2021, pp. 3520–3527.

[10] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, Y. Qiao, Videomae v2: Scaling video masked autoencoders with dual masking, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 14549–14560.

[11] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[12] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.

[13] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, Panns: Large-scale pretrained audio neural networks for audio pattern recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 2880–2894.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[15] M. Lamarre, C. Chen, F. Deniz, Attention weights accurately predict language representations in the brain, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 4513–4529.

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.