A multimodal encoding model for predicting human brain responses to complex naturalistic movies

Viacheslav Fokin (glorianfox@gmail.com)
Lumiere Education, 919 North Market Street, Wilmington, Delaware, 19801

Arefeh Sherafati (arefeh@wustl.edu)
Lumiere Education, 919 North Market Street, Wilmington, Delaware, 19801

Abstract

Accurately modeling human brain responses to complex, multimodal sensory inputs remains a core challenge in computational neuroscience. The Algonauts Project 2025 provides a large whole-brain fMRI dataset collected during naturalistic movie viewing, presenting the challenge of improving multimodal encoding models. Here, we propose a biologically informed encoding model that integrates visual, auditory, and language features, extracted using established deep learning and signal processing approaches, within predefined functional connectivity (FC) networks. By applying predictive modeling within an FC-based cortical mask, we achieve a 45.39% performance gain over the full cortex baseline. Our findings demonstrate the value of incorporating functional brain organization into encoding models and lays a foundation for future biologically grounded AI systems that integrate sensory information across domains.

Keywords: naturalistic viewing, encoding model; fMRI; multimodal processing; functional networks

Introduction

While neuroscience has long studied discrete sensory responses in isolated regions, this modular approach often overlooks the integrative nature of real-world cognition. Recent advances in computational neuroscience and artificial intelligence (AI) have made it increasingly feasible to model brain function in ecologically valid settings using multimodal stimuli. Naturalistic viewing, where visual and linguistic information unfold in rich temporal contexts, offers a promising bridge between neuroscience and AI.

One key challenge lies in unifying encoding approaches that have traditionally been modality-specific — mapping visual, auditory, or language inputs separately — into a single framework that reflects the inductive biases and integrative dynamics of the brain (Isik et al., 2017; Small et al., 2024). Prior work has shown that distinct yet overlapping regions in the bilateral superior temporal sulcus (STS) respond to visual (e.g., biological motion, faces) and linguistic (e.g., speech, theory of mind) social signals (Deen et al., 2015; Lee Masson & Isik, 2021). While these studies provided valuable maps of sensory specialization, they often used simplified, controlled stimuli that do not capture the complexity of natural experiences. In this work, we model neural responses to naturalistic movie-viewing using a unified encoding framework that incorporates visual, auditory, and language features extracted from the same stimulus masked to the subset of cortical parcels that contribute to functional connectivity (FC) networks, ensuring that our modeling is grounded in the brain's intrinsic network architecture. By applying feature-based encoding within this functionally coherent mask, we uncover the unique contributions of each modality to neural activation patterns and explore how high-level semantic processes are embedded within distributed functional networks.

Methods

Dataset and stimuli. We used the Algonauts 2025 dataset which includes functional magnetic resonance imaging (fMRI) responses recorded during naturalistic movie-viewing tasks in four individuals. The training data comprises multimodal stimuli and corresponding fMRI responses for each of the four subjects for all episodes of seasons one to five of the sitcom *Friends*. Evaluation data from season six of Friends enables assessment of model generalization. The stimuli presented to subjects consist of three distinct components: (i) movie visual frames, (ii) audio samples, and (iii) time-stamped language transcripts. Neural data was registered to the Montreal Neurological Institute (MNI) spatial template (Brett, Johnsrude, & Owen, 2002) and processed into time series. These signals are assigned to 1,000 functionally defined brain parcels based on the Schaefer parcellation scheme (Schaefer et al., 2018).

Encoding and predictive modeling. We used a banded ridge regression model (Dupre la Tour et al., 2022; Small et al., 2024) using 250 principal components of features extracted from pretrained visual, audio, and language models on the raw input stimuli. We extracted the visual features using the slow_r50 model, a 3D ResNet pretrained on Kinetics-400. We used activations from the blocks.5.pool layer, which captures high-level spatiotemporal patterns in the video. Language features were obtained from the movie's transcripts. We used the pretrained BERT model (bert-base-uncased) to embed each text snippet. Audio features were derived from the movie's audio track using Mel-frequency cepstral coefficients (MFCCs). All feature frames were

segmented into non-overlapping chunks of 1.49 seconds to match the fMRI repetition time (TR).

Results

Mapping Functional Networks. Mapping Functional Networks. We used the Schaefer 2018 brain atlas which divides the brain into 17 distributed networks. We computed the Pearson correlation coefficients across all predefined parcels (Figure 1). We then computed a mask that included only parcels with FC above their global median as a conservative yet functionally meaningful mask. Encoding within this FC-based mask achieved a mean R^2 of 0.2810 ± 0.0895.

We extracted example feature encoding predictions for the strongest feature among visual, auditory, and language and found the strongest predictive power (calculated by the explained variance R^2) in the expected visual, bilateral STS, and Broca's areas (Figure 2A). The full encoding model explained significant group-level variance ($p<0.01$, FDR corrected) in all parcels in the FC mask (Figure 2B). Applying the FC mask improved model performance by 45.39% across cortical parcels ($t = 65.69$) (Figure 2C).

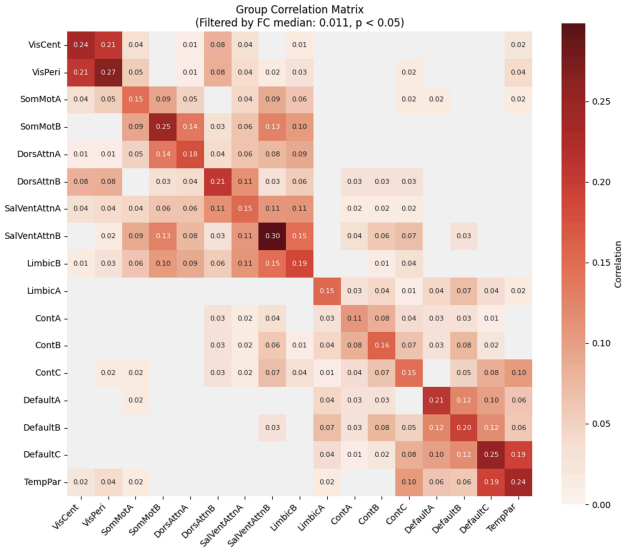

Figure 1: Group-level functional connectivity with correlation value above the median. See Schaefer 2018 for network abbreviations.

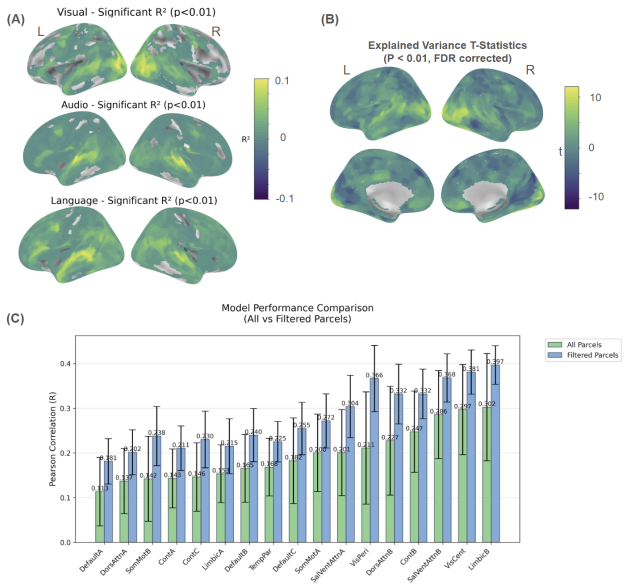

Figure 2: (A) Explained variance for example functional networks. (B) Group-level explained variance. (C) Model performance improvement by brain network after FC masking.

Discussion

We present a multimodal encoding model that integrates visual, auditory, and linguistic features within the brain's intrinsic functional architecture to predict neural responses during naturalistic movie viewing. By aligning feature-based modeling with a functionally defined cortical mask, we achieve a 45.39% improvement in prediction performance, supporting the value of biologically grounded constraints. Performance peaks in expected regions — including early visual cortex, bilateral STS, and language-sensitive frontal areas — highlight the model's sensitivity to both low- and high-level representations. While we used established models for feature extraction (slow_r50, BERT, MFCCs), future work will incorporate fused multimodal representations from state-of-the-art architectures to better capture integration. We also aim to analyze modality-specific contributions across networks, including ablation studies, to clarify where and how integration is most impactful for out-of-distribution generalizability. Our approach underscores the role of network organization and supports development of interpretable, brain-aligned AI.

References

Brett, M., Johnsrude, I. & Owen, A. The problem of functional localization in the human brain. Nat Rev Neurosci 3, 243–249 (2002). Doi: https://doi.org/10.1038/nrn756

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. Cerebral Cortex (New York, N.Y.: 1991), 25(11), 4596– 4609. doi: 10.1093/cercor/bhv111

Dupre la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. NeuroImage, 264, 119728. doi: 10.1016/j.neuroimage.2022.119728

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. Proceedings of the National Academy of Sciences, 114(43), E9145–E9152. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1714471114

Lee Masson, H., & Isik, L. (2021). Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. NeuroImage, 118741. doi: 10.1016/j.neuroimage.2021.118741

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cerebral Cortex, 28(9), 3095–3114. doi: https://doi.org/10.1093/cercor/bhx179

Small, H., Lee Masson, H., Wodka, E., Mostofsky, S. H., & Isik, L. (2024). *Ubiquitous visual representations during neural processing of a naturalistic movie*. Paper presented at the Cognitive Computational Neuroscience (CCN) 2024 Conference, Pasadena, CA, United States.