A MULTIMODAL SEQ2SEQ TRANSFORMER FOR PREDICTING BRAIN RESPONSES TO NATURALISTIC STIMULI

ABSTRACT

The Algonauts 2025 Challenge called on the community to develop encoding models that predict whole-brain fMRI responses to naturalistic multimodal movies. In this submission, we propose a sequence-to-sequence Transformer that autoregressively predicts fMRI activity from visual, auditory, and language inputs. Stimulus features were extracted using pretrained models including VideoMAE, HuBERT, Qwen, and BridgeTower. The decoder integrates information from prior brain states and current stimuli via dual cross-attention mechanisms that attend to both perceptual information extracted from the stimulus as well as narrative information provided by high-level summaries of the content. One core innovation of our approach is the use of sequences of multimodal context to predict sequences of brain activity, enabling the model to capture long-range temporal structure in both stimuli and neural responses. Another is the combination of a shared encoder with partial subject-specific decoder, which leverages common representational structure across subjects while accounting for individual variability. Our model achieves strong performance on both in-distribution and out-of-distribution data, demonstrating the effectiveness of temporally-aware, multimodal sequence modeling for brain activity prediction. The code is available at `https://github.com/Angelneer926/Algonauts_challenge`.

***K*eywords** neural encoding · sequence-to-sequence transformer · naturalistic movies · fMRI

1 Introduction

Understanding how the human brain responds to naturalistic, multimodal stimuli is a central goal in cognitive neuroscience. Recent advances in functional neuroimaging and the release of large-scale, richly annotated datasets have enabled the development of predictive models that map sensory input to brain activity across widespread cortical areas. The Algonauts 2025 Challenge offers a unique testbed for evaluating such models, inviting researchers across fields to participate in a 7-month long challenge to predict whole-brain fMRI responses of subjects viewing movie stimuli comprising synchronized visual, auditory, and linguistic streams[1].

Traditional approaches to modeling brain responses, such as finite impulse response (FIR) models or ridge regression, typically learn linear mappings that predict each timepoint in the fMRI signal independently from recent stimulus history. [2, 3]. While these methods have been successful in many settings, they often neglect the dynamic, autoregressive nature of neural responses and are limited in their ability to integrate multimodal inputs or adapt to individual differences across subjects. Inspired by recent studies demonstrating that deep neural networks can effectively model cortical responses to naturalistic language and vision [4, 5, 6], we propose a sequence-to-sequence Transformer-based architecture to better capture the temporal dependencies and multimodal interactions that shape fMRI signals over time. Similar to neural network-based machine translation models that map sequences of words from one language to another [7], our model translates sequences of audiovisual and linguistic stimuli into sequences of brain responses, conditioning each prediction on both the full input sequence and the history of prior neural activity.

To account for the rich, hierarchical structure of naturalistic stimuli, we extracted stimulus features from state-of-the-art pretrained models across multiple modalities: VideoMAE for temporal dynamics of visual motion [8], HuBERT for acoustic features [9], and Qwen for linguistic representations. Additionally, we incorporated sentence-level semantic features derived from BERT [10], using high-level summaries of the TV episodes and movies that constitute the training and testing data. This allows us to provide broader narrative context beyond the moment-to-moment stimulus input [11].

To capture joint visual-linguistic representations, we extracted cross-modal features from the stimuli using BridgeTower [12].

To address inter-subject variability in brain responses, we adopt a hybrid architecture with a shared encoder and subject-specific decoder heads, following the intuition that while perceptual representations may be broadly shared, how these representations are translated into neural activity can vary from person-to-person [13, 14]. This design improves robustness by training a shared encoder on pooled data, allowing it to learn generalizable stimulus representations. It also enables efficient adaptation to new participants by fine-tuning only the lightweight, subject-specific decoder head. Finally, our model leverages autoregressive decoding with teacher forcing and learnable BOS tokens to predict entire sequences of fMRI responses. We optimize our model using a combination of mean squared error and Pearson correlation loss to better align with the challenge's evaluation metric. Extensive experiments show that this architecture yields robust predictions and effectively integrates multimodal information over time.

2 Related Work

Most existing fMRI encoding models follow a standard pipeline: extracting features from continuous stimuli, concatenating them over a short temporal window, using linear models to predict voxel-wise brain responses, with regularization strength typically tuned via cross-validation. While this framework has proven effective, much of the recent progress has focused on optimizing the choice of features rather than altering the overall structure of the model. In the language domain, early approaches used static word embeddings to represent word meaning independent of context [15], whereas recent studies demonstrate that contextualized representations from large language models (LLMs) yield substantially better predictions of brain activity [16]. Encoding performance improves approximately logarithmically with model size, and larger models also reveal increasingly left-lateralized activation patterns, consistent with classic findings in language neuroscience[17].

In speech encoding, models have progressed from using handcrafted spectrogram and phoneme-level features [18] to self-supervised representations from models such as HuBERT and WavLM, which better capture the hierarchy of auditory information and show improved encoding with model size[19]. In vision, encoding models have evolved from CNNs trained on ImageNet to self-supervised video transformers such as VideoMAE, which capture temporal and motion features[20]. More recent approaches leverage representations from vision-language models such as CLIP, which align better with high-level visual cortex, though evidence suggests that these gains are primarily attributable to large-scale training data rather than to language supervision itself[21, 22]. To more explicitly integrate information across modalities, cross-modal transformers such as BridgeTower fuse language and visual features via co-attention, yielding representations that better match brain responses across multiple cortical systems[23].

However, despite these advances in feature selection and model scaling, the underlying encoding pipeline remain largely unchanged. While some recent work has explored architectural alternatives, such as fine-tuning pretrained auditory networks on subject-specific data to improve predictions in audiovisual regions[24], there remains limited exploration of end-to-end models that jointly capture temporal dynamics, multimodal integration, and individual variability. In this submission, we take a step in this direction by introducing a sequence-to-sequence transformer that frames fMRI encoding as an autoregressive generation task, incorporating temporally extended cross-modal attention and subject-specific output heads within a unified framework.

3 Dataset and Challenge

The Algonauts 2025 Challenge is based on a subset of the CNeuroMod dataset[25], and includes whole-brain fMRI recordings from four participants (sub-01, sub-02, sub-03, sub-05) while they viewed naturalistic multimodal movies. The training data includes approximately 65 hours of movie stimuli from all episodes of *Friends* Seasons 1–6 and four feature films (*The Bourne Supremacy, Hidden Figures*, *Life*, *The Wolf of Wall Street*), each aligned with time-resolved fMRI responses. fMRI responses have been preprocessed and parcellated into 1,000 cortical regions using the Schaefer atlas[26], sampled every 1.5 seconds and aligned to MNI space.

Model performance is evaluated using the Pearson correlation between predicted and ground-truth fMRI responses. The final score is obtained by averaging the correlation values across all subjects, test movies, and brain parcels. This score is calculated for the out-of-distribution (OOD) movies during the final model selection phase.

4 Approach

4.1 Feature Extraction

We extracted stimulus representations from multiple modalities to comprehensively capture the external input at each time point. See Figure 1. This section outlines the models and strategies used to obtain and enhance these features.

Visual Modality. For visual input, we employed VideoMAE, a masked autoencoder pretrained on large-scale video datasets. VideoMAE is well-suited for this task due to its ability to model fine-grained spatiotemporal dependencies, particularly motion dynamics and scene transitions. At each time point, we extracted a 768-dimensional embedding for the current frame by averaging token representations from the final (12th) encoder layer. We then concatenated this with the mean embedding of all preceding frames, resulting in a 1536-dimensional vector. This simple yet effective temporal smoothing provides a form of memory that helps the model maintain awareness of past visual context, which is especially beneficial for videos with continuous narrative flow.

Audio Modality. To encode the audio signal, we utilized HuBERT, a self-supervised model that learns hierarchical speech representations through masked prediction of latent units. We extracted a 1536-dimensional feature vector at each time point by concatenating the hidden states from the 3rd and 9th encoder layers, enabling the capture of both phonetic and prosodic information. While we experimented with incorporating past context (e.g., averaging over prior frames), we observed that such temporal aggregation degraded performance in this modality. This may be due to the higher temporal resolution of audio compared to visual signals, where frame-wise precision is critical for preserving short-term acoustic dynamics. Therefore, for each frame, we retained only the frame-level features without incorporating the previous time-points.

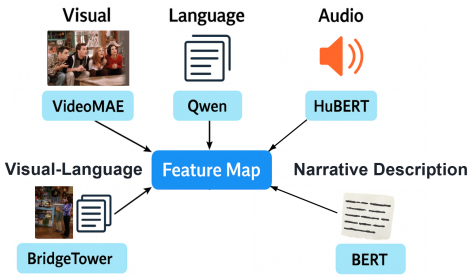

Figure 1: Feature Extraction

Language Modality. We used the Qwen model to embed the dialogue transcript. At each time point, the model receives the current utterance along with all previous utterances as textual context, truncated to the most recent 2048 tokens. From the hidden states of the 12th transformer layer, we computed two types of features: (1) the mean across all tokens in the full input sequence, and (2) the mean over the last 10 tokens only. Each is a 1024-dimensional vector, and their concatenation yields a 2048-dimensional representation that captures both global semantic context and localized information near the current utterance boundary.

Visual/Language Hybrid. To complement the unimodal features, we utilized BridgeTower, a pretrained cross-modal transformer architecture designed to integrate visual and textual signals. At each time point, the model receives a video frame and its corresponding utterance as input. We extract a 1536-dimensional fused representation from the pooler_output of the final layer, which corresponds to the [CLS] token after visual/text fusion. These features provide additional cross-modal alignment that is difficult to obtain from independent visual and language encoders, and serve as a valuable supplement to the modality-specific representations.

Narrative Summaries. While the original dataset provided dialogue-level annotations, the episodes and movies contained overarching narrative structures that are essential for interpretation. To compensate for this missing context, we manually sourced narrative summaries of each episode and movie.. Each summary is segmented into sentences and passed through a pretrained BERT model to obtain sentence-level embeddings. We explored two strategies to incorporate this additional context into the frame-wise representation learning process. The first is a Gaussian-weighted sum of sentence embeddings based on their temporal distance to each frame. The second is a cross-attention mechanism in the decoder, where the fMRI prediction at each time step attends to the sentence-level embeddings from the narrative summary. While the Gaussian-weighted approach imposes a smooth and interpretable temporal bias, cross-attention offers greater flexibility by allowing the model to learn dynamic alignments between frames and narrative content. Although the alignment between summaries and frames was approximate, this additional modality captures high-level thematic information that may guide interpretation, particularly in long-form narratives.

4.2 Model Architecture

We propose a Transformer-based encoder-decoder architecture to model the temporal mapping from multimodal stimulus features to fMRI responses. The design accounts for both the sequential dynamics of neural activity and

inter-subject variability. Departing from conventional static regression approaches, our model generates fMRI time series in an autoregressive manner, employing masked self-attention and cross-modal contextual integration from both low-level stimuli and high-level semantic representations. The overall framework is shown in Figure 2.

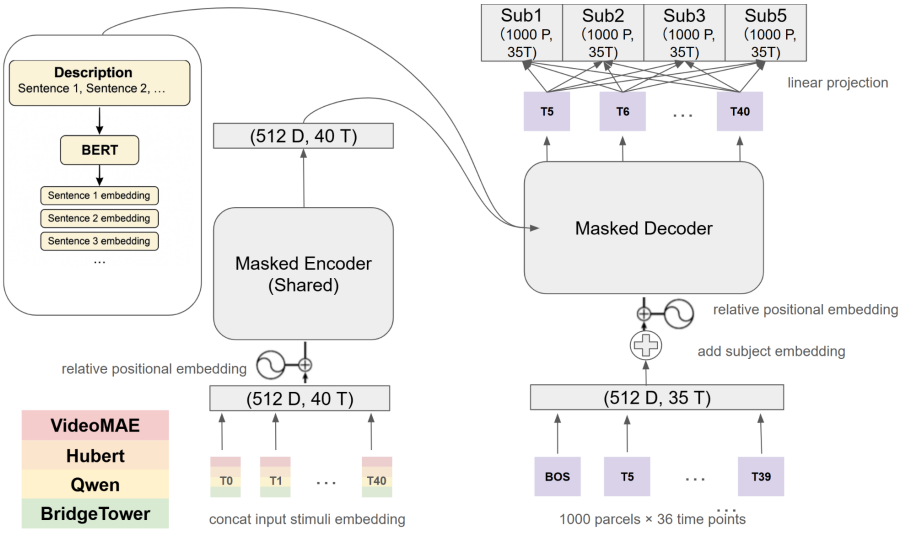

Figure 2: Model Architecture

The encoder is implemented as a Transformer with causal self-attention, such that each time step only attends to its current and preceding inputs. This departs from the standard bidirectional design of seq2seq Transformers used in machine translation and is motivated by the nature of neural processing: the brain encodes external stimuli sequentially as they unfold over time, without access to future information. By imposing this temporal constraint, the model better aligns with the dynamics of biological perception. Empirically, this modification led to improved predictive performance compared to an encoder with unrestricted attention. Before each encoder layer, we added a relative positional encoding to the input sequence, allowing the model to attend over time in a position-invariant fashion. Let $\mathbf{H}_{\text{enc}} = \text{Encoder}(\mathbf{X}) \in \mathbb{R}^{T \times d}$ denote the output of the encoder, where T is the number of time steps and d is the hidden dimension.

The decoder is a masked causal Transformer that generates fMRI signals autoregressively. Let $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_T]$ denote the predicted fMRI sequence. At each decoding step t, the decoder receives as query the embedding of the previously generated output $\hat{\mathbf{y}}_{t-1}$ and attends only to the past via masked self-attention, ensuring that predictions at time t depend only on information available at $t' \leq t$. This constraint simulates the temporal unidirectionality of how subjects viewed the stimuli.

To provide the decoder with high-level semantic context, we introduced an additional cross-attention module over episode-level descriptions $\mathbf{E}$, which are extracted using a frozen BERT model. This semantic memory is integrated after the standard cross-attention over stimulus features, allowing the decoder to attend to both perceptual and narrative information during fMRI prediction.

Formally, the decoder's output at time t is computed as:

$$
\begin{aligned}
\tilde{\mathbf{y}}_t &= \hat{\mathbf{y}}_{t-1}\mathbf{W}_{\text{dec}} + \mathbf{b}_{\text{dec}} \\
\mathbf{z}_t^{(0)} &= \tilde{\mathbf{y}}_t + \text{RelPosEnc}(t) \\
\mathbf{z}_t^{(l,1)} &= \text{LayerNorm}\left(\mathbf{z}_t^{(l-1)} + \text{SelfAttn}(\mathbf{z}_{\leq t}^{(l-1)})\right) \\
\mathbf{z}_t^{(l,2)} &= \text{LayerNorm}\left(\mathbf{z}_t^{(l,1)} + \text{CrossAttn}_{\text{stim}}(\mathbf{z}_t^{(l,1)}, \mathbf{H}_{\text{enc}})\right) \\
\mathbf{z}_t^{(l,3)} &= \text{LayerNorm}\left(\mathbf{z}_t^{(l,2)} + \text{CrossAttn}_{\text{desc}}(\mathbf{z}_t^{(l,2)}, \mathbf{E})\right) \\
\mathbf{z}_t^{(l)} &= \text{LayerNorm}\left(\mathbf{z}_t^{(l,3)} + \text{MLP}(\mathbf{z}_t^{(l,3)})\right) \\
\hat{\mathbf{y}}_t &= \mathbf{z}_t^{(L)}\mathbf{W}_{\text{out}} + \mathbf{b}_{\text{out}}
\end{aligned}
\tag{1}
$$

All attention modules are multi-head and include dropout and residual connections.

The training objective combines mean squared error (MSE) and negative Pearson correlation between the predicted and ground truth sequences:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{T}\sum_{t=1}^{T} \|\hat{\mathbf{y}}_t - \mathbf{y}_t\|_2^2, \quad \mathcal{L}_{\text{corr}} = -\frac{1}{T}\sum_{t=1}^{T} \rho(\hat{\mathbf{y}}_t, \mathbf{y}_t), \quad \mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_{\text{corr}} \tag{2}$$

where ρ denotes the Pearson correlation and λ is a weighting factor.

To support modeling multiple subjects with the same model, we introduced a hybrid architecture in which the encoder is shared across all subjects, while the decoder incorporates individual-specific components. This design reflects the fact that, even under identical stimuli, fMRI responses exhibit substantial inter-subject variability. At the same time, leveraging data from multiple subjects can improve model robustness and generalization. Each subject s is associated with a learnable embedding vector $\mathbf{e}_s$ that is concatenated to every time step of the decoder input. In addition, the final linear projection layer of the decoder head is specialized per subject, allowing personalized mapping from hidden representations to predicted fMRI outputs. This setup assumes that how stimuli are represented can be shared across individuals, while the corresponding brain responses remain subject-dependent. Formally, for subject s, two modifications were made to the original architecture:

$$\begin{aligned} \mathbf{z}_t^{(0)} &= \tilde{\mathbf{y}}_t + \mathbf{e}_s + \text{RelPosEnc}(t) \\ \hat{\mathbf{y}}_t^{(s)} &= \mathbf{z}_t^{(L)}\mathbf{W}_{\text{out}}^{(s)} + \mathbf{b}_{\text{out}}^{(s)} \quad \text{(for subject } s\text{)} \end{aligned} \tag{3}$$

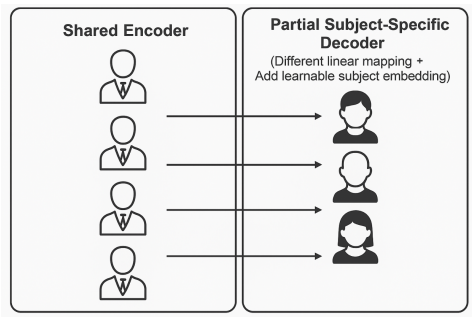

Figure 3: Hybrid architecture combining a shared encoder with partial subject-specific decoders

Notably, this architecture enables parameter-efficient adaptation to new subjects by fine-tuning only the subject-specific embeddings and output head, making it suitable for few-shot personalization in fMRI encoding tasks.

4.3 Training Strategy

To address the challenge of limited training data for sequence-level prediction of brain responses, we augmented the dataset using a sliding window approach. Specifically, we segmented the original stimuli into overlapping temporal chunks, each consisting of 40 consecutive time steps of multimodal input. The corresponding output is a 35-frame fMRI sequence, temporally aligned with the input window after applying a fixed hemodynamic delay. This strategy increases the effective number of training samples and ensures that each fMRI time point is predicted multiple times across overlapping windows, enabling more stable and robust learning.

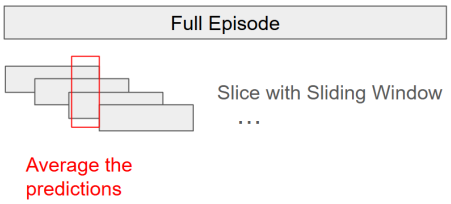

Figure 4: Sliding window strategy for sequence-Level fMRI prediction

To mitigate overfitting due to repeated context, we shuffled the order of training samples at the beginning of each epoch. This prevents the model from memorizing fixed stimulus-response mappings and encourages it to learn generalizable features across different temporal contexts. Both the window length and the hemodynamic delay were treated as tunable hyperparameters. While the current settings yield strong empirical performance, a comprehensive hyperparameter search remains an avenue for future work.

During training, we initialized the decoder with a learnable `[BOS]` token and employed the teacher forcing strategy to guide autoregressive generation. At each decoding step t, the model received either the ground-truth output $\mathbf{y}_{t-1}$ or its own previous prediction $\hat{\mathbf{y}}_{t-1}$ as input, selected according to a teacher forcing ratio $\gamma \in [0, 1]$. With probability γ, the ground truth is used to stabilize training and accelerate convergence; with probability $1 - \gamma$, the model samples from its own output distribution, exposing it to potential prediction errors and improving robustness during inference. The teacher forcing ratio is annealed over training epochs to gradually transition from supervised guidance to self-conditioned generation.

5 Results

We evaluate our model using both validation and test sets, which include both in-distribution (ID) and out-of-distribution (OOD) data.

To assess performance during model development, we constructed a validation set consisting of all b-part episodes from *Friends* Season 6, as well as the first two chunks from each Movie10 film (*The Bourne Identity*, *Hidden Figures*, *Life*, and *The Wolf of Wall Street*). This split includes a mix of Friends and movie content, enabling reliable evaluation across different content types. On this validation set, the model achieved an average Pearson correlation of 0.301 on the held-out Friends episodes and 0.225 on the held-out *Movie10* chunks. See Figure 5.

The performance gap between *Friends* and *Movie10* is likely due to the training set containing a larger proportion of Friends data, allowing the model to more effectively learn patterns in that stimulus set. In contrast, *Movie10* clips differ substantially in both visual style and narrative structure, posing a greater challenge for generalization. The lower performance on *Movie10* is thus expected and reflects the challenge of transferring representations learned from episodic television from a single series to a diverse set of feature films.

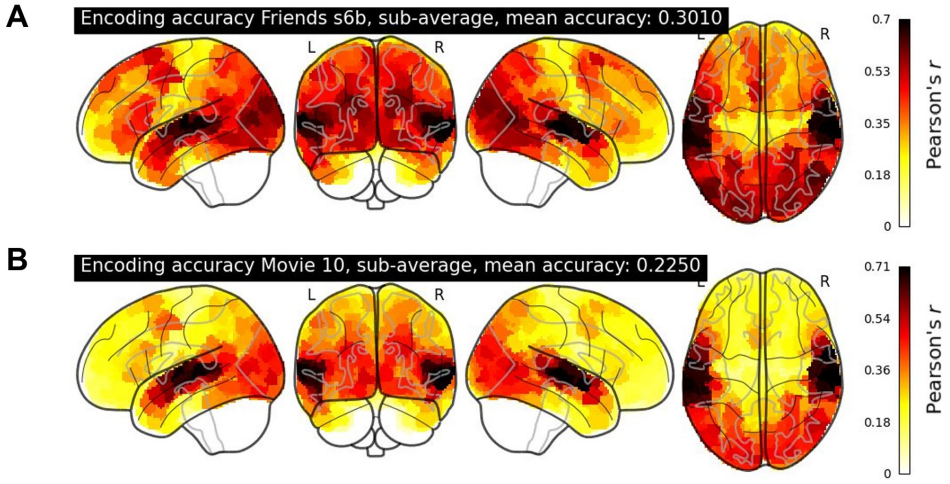

Figure 5: Modal validation accuracy for (A) *Friends* Season 6b (B) held-out movie segments

We then evaluated our final model on the official test data provided by the Algonauts 2025 Challenge, which is divided into two phases. In the **model building phase**, test performance is measured on *Friends* Season 7 episodes, representing an in-distribution (ID) evaluation setting. In the subsequent **model selection phase**, models are evaluated on a held-out set of unseen feature films, constituting a true out-of-distribution (OOD) generalization task..

Our model achieved an average Pearson correlation of 0.305 on Friends Season 7 and 0.199 on the OOD movie set. See Figure 6. The strong performance on Season 7 indicates that the model generalizes reliably within the domain it was primarily trained on. Similar to our validation results, performance on the movies was lower, indicating that the model did not transfer as effectively to out-of-distribution stimuli.

6 Conclusion

In this work, we present a multimodal sequence-to-sequence Transformer model for predicting full-brain fMRI responses to naturalistic audiovisual and linguistic stimuli. By leveraging pretrained encoders across multiple modalities, integrating semantic-level context from narrative summaries, and modeling brain activity autoregressively with subject-specific adaptations, our approach captures both the temporal and individual complexity of neural responses. The model demonstrates strong performance on both in-distribution and out-of-distribution data, highlighting the benefits of temporally-aware, cross-modal integration. Moreover, the sliding window training strategy, combined with teacher forcing, enables efficient learning despite limited data. Future directions include incorporating more precise alignment between high-level semantic context and neural dynamics, scaling the model to broader subject cohorts, and exploring zero-shot generalization across unseen individuals and stimuli. Our results support the promise of generative, multimodal models as a powerful framework for bridging perception and brain activity in real-world settings.

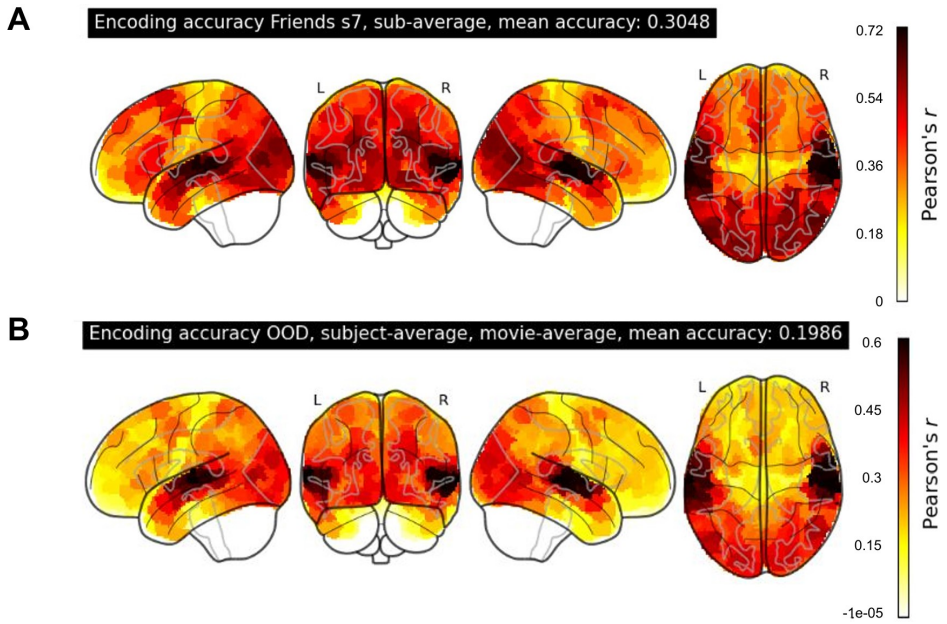

Figure 6: Modal testing accuracy for (A) Model Building phase (*Friends* Season 7) (B) Model Selection phase (out-of-distribution movies)

Acknowledgments

This work was completed with resources provided by the University of Chicago Data Science Institute's High Performance Computing Cluster.

References

[1] Alessandro T. Gifford, Domenic Bersch, Marie St-Laurent, Basile Pinsard, Julie Boyle, Lune Bellec, Aude Oliva, Gemma Roig, and Radoslaw M. Cichy. The algonauts project 2025 challenge: How the human brain makes sense of multimodal movies, 2025.

[2] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

[3] Tom Dupré la Tour, Matteo Visconti di Oleggio Castello, and Jack L Gallant. The voxelwise encoding model framework: a tutorial introduction to fitting encoding models to fmri data. *Imaging Neuroscience*, 3:imag_a_00575, 2025.

[4] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.

[5] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

[6] Haibao Wang, Lijie Huang, Changde Du, Dan Li, Bo Wang, and Huiguang He. Neural encoding for human visual cortex with deep neural networks learning "what" and "where". *IEEE Transactions on Cognitive and Developmental Systems*, 13(4):827–840, 2020.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[8] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[9] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[11] Shan Gao, Ryleigh Nash, Shannon Burns, and Yuan Chang Leong. Predicting whole-brain neural dynamics from prefrontal cortex functional near-infrared spectroscopy signal during movie-watching. *Social cognitive and affective neuroscience*, 20(1):nsaf043, 2025.

[12] Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10637–10647, 2023.

[13] Ma Feilong, Samuel A Nastase, Guo Jiahui, Yaroslav O Halchenko, M Ida Gobbini, and James V Haxby. The individualized neural tuning model: Precise and generalizable cartography of functional architecture in individual brains. *Imaging Neuroscience*, 1:1–34, 2023.

[14] Ido Tavor, O Parker Jones, Rogier B Mars, SM Smith, TE Behrens, and Saad Jbabdi. Task-free mri predicts individual differences in brain activity during task performance. *Science*, 352(6282):216–220, 2016.

[15] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014.

[16] Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *Advances in neural information processing systems*, 31, 2018.

[17] Laurent Bonnasse-Gahot and Christophe Pallier. fmri predictors based on language models of increasing complexity recover brain left lateralization. *Advances in Neural Information Processing Systems*, 37:125231–125263, 2024.

[18] Xue L Gong, Alexander G Huth, Fatma Deniz, Keith Johnson, Jack L Gallant, and Frédéric E Theunissen. Phonemic segmentation of narrative speech in human cerebral cortex. *Nature communications*, 14(1):4309, 2023.

[19] Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36:21895–21907, 2023.

[20] Jacob Yeung, Andrew F Luo, Gabriel Sarch, Margaret M Henderson, Deva Ramanan, and Michael J Tarr. Reanimating images using neural representations of dynamic stimuli. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5331–5343, 2025.

[21] Guoyuan Yang, Mufan Xue, Ziming Mao, Haofang Zheng, Jia Xu, Dabin Sheng, Ruotian Sun, Ruoqi Yang, and Xuesong Li. Clip-msm: A multi-semantic mapping brain representation for human high-level visual cortex. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9184–9192, 2025.

[22] Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, 15(1):9383, 2024.

[23] Jerry Tang, Meng Du, Vy Vo, Vasudev Lal, and Alexander Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in Neural Information Processing Systems*, 36:29654–29666, 2023.

[24] Maelle Freteault, Maximilien Le Clei, Loic Tetrel, Lune Bellec, and Nicolas Farrugia. Alignment of auditory artificial networks with massive individual fmri brain data leads to generalisable improvements in brain encoding and downstream tasks. *Imaging Neuroscience*, 3:imag_a_00525, 2025.

[25] Julie Boyle, Basile Pinsard, Valentina Borghesani, Francois Paugam, Elizabeth DuPre, and Pierre Bellec. The courtois neuromod project: quality assessment of the initial data release (2020). In *2023 Conference on Cognitive Computational Neuroscience*, pages 2023–1602, 2023.

[26] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.