PREDICTING BRAIN RESPONSES TO NATURAL MOVIES WITH MULTIMODAL LLMS

Cesar Kadir Torrico Villanueva[m*] **Jiaxin Cindy Tu**[1,m*] **Mihir Tripathy**[2,m*]
Connor Lane[3,m*] **Rishab Iyer**[4,m] **Paul S. Scotti**[3,m]

[1]Psychological and Brain Sciences, Dartmouth College
[2]Core for Advanced Magnetic Resonance Imaging (CAMRI), Baylor College of Medicine
[3]Sophont [4]Princeton Neuroscience Institute [m]Medical AI Research Center (MedARC)

ABSTRACT

We present MedARC's team solution to the Algonauts 2025 challenge. Our pipeline leveraged rich multimodal representations from various state-of-the-art pretrained models across video (V-JEPA2), speech (Whisper), text (Llama 3.2), vision-text (InternVL3), and vision-text-audio (Qwen2.5-Omni). These features extracted from the models were linearly projected to a latent space, temporally aligned to the fMRI time series, and finally mapped to cortical parcels through a lightweight encoder comprising a shared group head plus subject-specific residual heads. We trained hundreds of model variants across hyperparameter settings, validated them on held-out movies and assembled ensembles targeted to each parcel in each subject. Our final submission achieved a mean Pearson's correlation of 0.2085 on the test split of withheld out-of-distribution movies, placing our team in fourth place for the competition. We further discuss a last-minute optimization that would have raised us to second place. Our results highlight how combining features from models trained in different modalities, using a simple architecture consisting of shared-subject and single-subject components, and conducting comprehensive model selection and ensembling improves generalization of encoding models to novel movie stimuli. Code is available at https://github.com/MedARC-AI/algonauts2025.

1 Introduction

The Algonauts Challenge is a biennial competition organized by the Algonauts Project in collaboration with the Conference on Cognitive Computational Neuroscience (CCN) aimed to advance our understanding of the human brain by encouraging researchers to develop the best encoding model for a given dataset. In computational neuroscience, encoding models refer to algorithms that take a stimulus as input and output predicted neural activations. Algonauts 2019–2023 focused on static images or short videos where models trained on deep convolutional networks were often adequate. However, humans experience the world through richly multimodal, temporally extended stimuli. The 2025 Algonauts challenge was designed to push the field beyond unimodal benchmarks by using long movies with synchronized video, audio and language. In collaboration with the CNeuroMod project, the organizers shared almost 80 hours of functional magnetic resonance imaging (fMRI) responses per subject.

Each of four subjects from the CNeuroMod project (sub-01, sub-02, sub-03, and sub-05) was scanned using fMRI while they watched 55 hours of Friends season 1 to 6 (*friends s1-s6*) and 10 hours across four movies, including The Bourne Supremacy (*bourne*), Hidden Figures (*figures*), BBC documentary Life (*life*), and The Wolf of Wall Street (*wolf*). Their brain responses as time series were spatially normalized to the Montreal Neurological Institute (MNI) template [1], and further downsampled to 1,000 functionally defined brain parcels spanning the left and right cerebrum [2]. The test set was collected and processed with the same paradigm but with the four subjects watching 1) the friends season 7 (*friends s7*) for the model-building phase and 2) six out-of-distribution (OOD) movie stimuli for the model-selection phase (for which fMRI responses were withheld). Each team was allowed up to ten submissions during the final model selection phase, which had a deadline of July 13, 2025.

* Equal contributions, order of first four authors was randomized

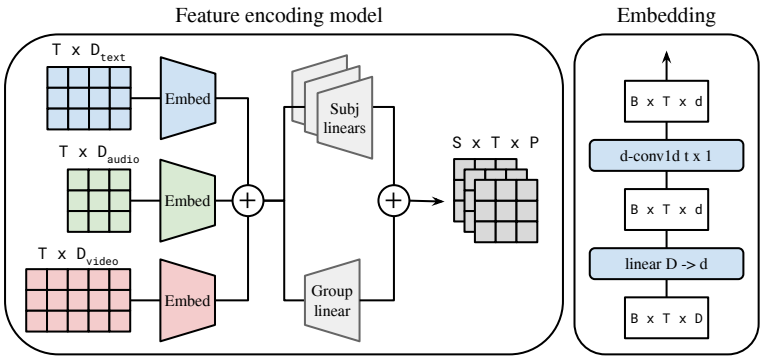

Figure 1: Encoding model architecture. The feature encoding model takes intermediate activations from a set of multimodal backbone models as inputs. These features are linearly projected and temporally aligned using independent "embedding" modules. The embeddings are then summed and passed through a linear fMRI prediction module consisting of a shared group head plus subject-specific heads. (`T` = number of time points, `D` = native feature dimension, `S` = number of subjects, `P` = number of parcels (1000), `d` = embedding dimension, `B` = batch size.)

2 Overview of our solution

Overall, our three main contributions for success in the challenge consisted of (1) multimodal feature extraction using pre-trained models (Section 3), (2) parameter-efficient encoding architecture with temporal alignment with shared-subject and subject-specific adaptation (Section 4), and (3) comprehensive model selection and ensembling to maximize generalization success to OOD movies (Section 5), including an ensemble model that was not submitted to competition in time which could have brought us to second place. Lastly, we discovered a strategy to use the returned Codabench scores for each movie to potentially increase the final scores in Section 6.

3 Feature Selection and Extraction

To build a comprehensive predictive model of brain activity, we extracted representations from five distinct, deep neural network models. Each model was chosen to capture different facets of the audiovisual and linguistic content of the stimuli. The feature sets were extracted independently as high-dimensional feature vectors for each fMRI time point (known as repetition time or TR = 1.49 seconds). The general pipeline for each model involved segmenting the stimulus data (video, audio, and text if available) into chunks aligned with fMRI TR and processing them through the model to capture intermediate activations from the model's prespecified layer. This was determined empirically through a systematic search A.

3.1 Multimodal Large Language Models (MLLMs)

We utilized two state-of-the-art multimodal LLMs, InternVL3 [3] and Qwen2.5-Omni [4] to capture integrated audiovisual and linguistic representation. The feature extraction for these models operates on a sliding window of the stimulus. For a given fMRI TR, a context window of the preceding 20 seconds of the video, audio (only for Qwen), and transcript data (if available) is processed in a single forward pass (the selection of this duration is justified in our ablation study in Appendix B.)

The intermediate activations from a pre-specified layer are captured for this entire sequence. The final feature vector for each TR is then derived by averaging the activation vectors over the token span corresponding to that specific TR's segment within the context window.

InternVL3 processes the 20-second context window as a sequence of image-text pairs. Specifically, the first frame of each TR, resized to 256 x 256 within the window, is selected and paired with its corresponding transcript chunk, if available. Qwen2.5-Omni samples the video segment at 2 frames per segment and resizes it to 256×384. The audio is sampled at 16 kHz audio waveform and fused with the available transcript to create a unified input representation.

3.2 Specialized Unimodal Models

To capture modality-specific information with high fidelity, we included features from specialized audio, language, and vision models. The general procedure for these models was to process their respective modality (video, audio, or text) to produce a single feature vector for every fMRI TR. This was achieved by passing the prepared input through the model, capturing activations from a target layer, and then applying a model-specific averaging or pooling strategy to produce a fixed-size vector per TR.

V-JEPA2 (Vision-based Joint Embedding Predictive Architecture) [5] is a non-generative, self-supervised vision model that learns an abstract understanding of world dynamics by predicting representations of masked-out video parts in a latent space. For each TR, its corresponding video segment was processed by selecting 15 frames, which were passed through the model's encoder. The final feature vector was produced by averaging the resulting image patch embeddings.

Whisper (Large-v3) [6] is an encoder-decoder transformer whose encoder learns robust representations of speech and general acoustics from large-scale pre-training for speech recognition. For each TR, its 16 kHz mono audio segment was passed through the Whisper encoder. The final feature vector was computed by averaging the temporal sequence of the resulting output activations.

Llama 3.2 (3B) [7] is a lightweight multilingual text-only language model. For this model, we processed the entire transcript for a stimulus in a single forward pass. The token spans corresponding to each TR were identified using character offsets, and the final feature vector for each TR was computed by averaging the activations of all tokens within its identified span.

4 Model Architecture

Our feature encoding model (Figure 1) is made up of a feature embedding stage and an fMRI prediction stage. In the feature embedding stage, the extracted features from each input backbone model are first linearly projected frame-by-frame from their native dimension to a small latent dimension. The projected feature time series are then temporally aligned using depth-wise 1D convolutions [8, 9], learned separately for each input feature. Finally, the embeddings from each backbone are summed together. The fMRI prediction stage consists of a group linear prediction head that is shared across the four subjects, as well as subject-specific "residual" heads. The group and subject heads are applied to the embedding time series frame-by-frame, and their results are added together.

Model training details. In this section, we focus on a representative "default" model to understand baseline model performance. In Section 5, we develop an ensemble of different model variants to achieve our final submission score. Our default model training setup uses all five backbone features described in Section 3. We used a default embedding dimension of 192 and convolution kernel width of 45 TRs (67 seconds). The total number of trainable parameters was 3.5M. Our default training data mixture included *friends s1-5*, *bourne*, and *wolf*. We held out *friends s6*, *figures*, and *life* for testing. We used *figures* as a validation set for early stopping. We trained our models using AdamW for a maximum of 1200 steps with a batch size of 16 and a temporal sequence length of 64 TRs. Wall clock training time was under 2 minutes using a single NVIDIA H100 GPU (<1 GB memory usage).

4.1 Default baseline model performance

Figure 3 shows the performance of our default baseline model, without ensembling, which achieves a final OOD test score of 0.203. We observe robust prediction performance in the visual cortex as well as lateral prefrontal and temporal brain areas. (See Figure 2 for reference.) Consistent with classic work mapping fMRI responses to naturalistic stimuli, e.g. Hasson et al. [12], Huth et al. [13], prediction performance is especially strong in temporal areas responsive to sound, speech, and language, as well as inferotemporal and lateral occipital areas responsive to people, places, and action. We observe reliable albeit weaker prediction in parts of the dorsal and frontoparietal attention networks. Interestingly, prediction performance is notably weaker in the movie *life* compared to *friends s6* and *figures*, especially in visual and frontoparietal areas. Unlike the other movies, *life* is a narrated nature documentary without human characters or storylines.

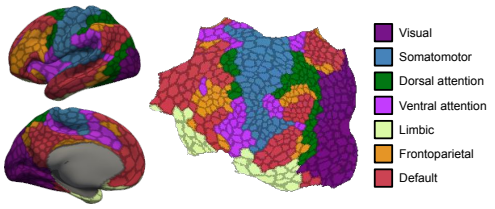

Figure 2: Schaefer 1000 parcellation with Yeo networks [2] shown on the HCP inflated surface [10] and pycortex flat map [11].

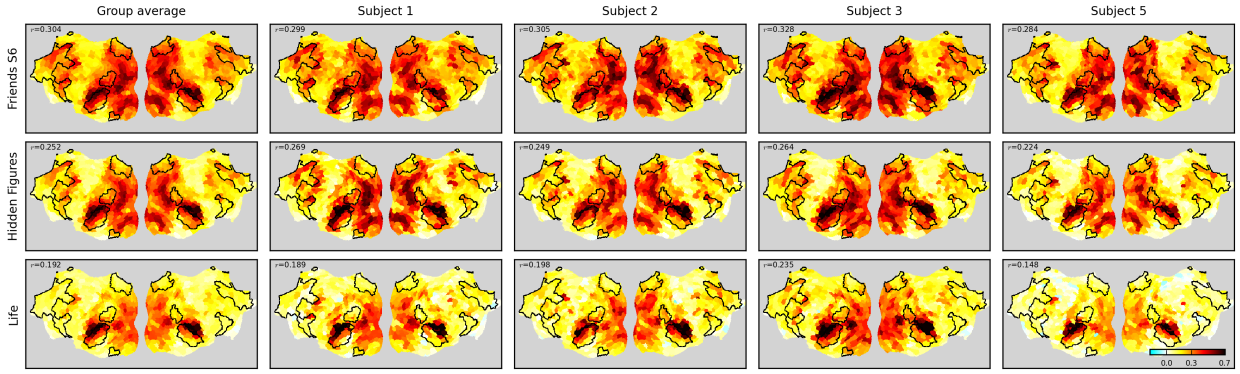

Figure 3: Feature encoding performance of our default model. Values are Pearson r encoding accuracy maps computed on held-out validation movies. Black outline indicates the boundary of the Yeo "default" resting-state network [2].

4.2 Ablation experiments

We evaluate the components of our default model in a set of ablation experiments summarized in Table 1.

Convolution kernel size. The fMRI signal has a temporally delayed and blurred response compared to the underlying neural signal and movie stimulus. Consistent with this, we find that temporally aligning the features with the fMRI is crucial for good performance (Table 1a). We find that a large kernel width of 45 TRs (67 seconds) is best. Interestingly, this is much longer than the typical hemodynamic response. This suggests that aggregating features over a long context window is beneficial for predicting fMRI.

Convolution kernel type. We explore a range of convolution kernel types in Table 1b. All kernel types use depthwise 1D convolution. The "causal" kernel uses a causal convolution mask so that predictions depend only on earlier (in time) features. The "positive" kernel constrains weights to be non-negative by applying absolute value to the weights. The "tied" kernel shares the same single-channel filter across all input embedding channels. We find that all kernel variants achieve accurate prediction, although the default standard depthwise convolution performs marginally better.

Embedding dimension. We find that an embedding dimension of 192 performs best (Table 1c). However, our model achieves strong performance even with much smaller embedding dimensions. This suggests that although the fMRI data are high-dimensional, most of the predictable stimulus-driven activity lies near a low-dimensional (linear) manifold.

Multi-subject training. A key aspect of our approach is to leverage the aligned fMRI responses of multiple subjects to the same stimuli in order to jointly train a multi-subject encoding model. In Table 1d, we find that this provides a $\Delta r = 0.01$ improvement over single subject training, with slightly more improvement for the OOD movie *life*.

fMRI prediction head architecture. Another aspect of our model architecture is the fMRI prediction head, which we factorized into a shared group head plus subject-specific residual heads. We find that the subject-specific residual heads provide a clear improvement over just the group head (Table 1e). Adding the group head does not provide any meaningful benefit on top of just the subject-specific heads. The group head alone still predicts well, suggesting that much of the predictable signal is shared across people.

Input feature set. In Table 1f, we test the impact of removing various features from the input feature set. In the first block of rows, we look at the performance of each backbone model feature individually. We observe that as expected, the multimodal modals InternVL3 (image, text) and Qwen2.5 (image, text, audio) achieve the best individual performance. Qwen2.5 performs especially well on the OOD movie *life*. V-JEPA2 and Llama 3.2 both achieve strong individual performance, while Whisper performs notably worse on its own. In the second block of rows, we look at the leave-one-model-out performance. We find that each backbone feature contributes to the overall encoding model performance (i.e. performance drops when each model is left out.) Llama 3.2 seems to have a strong, unique contribution, while Whisper has the weakest contribution.

5 Model Selection and Parcel-wise Ensemble Strategy

For competition purposes, we employed a straightforward model selection and parcel-wise ensemble strategy to enhance performance [14]. Initially, we defined a comprehensive hyperparameter search space, varying parameters such as

width	s6	figures	life
none	0.200	0.146	0.098
9	0.282	0.230	0.177
17	0.291	0.240	0.183
45	**0.304**	**0.252**	**0.192**
65	0.303	0.251	**0.192**

(a) **Kernel size**. Temporally aligning features with 1D convolution is key.

case	s6	figures	life
causal	0.301	0.244	0.192
positive	0.299	0.246	0.187
tied	0.302	0.251	**0.194**
default	**0.304**	**0.252**	0.192

(b) **Kernel type**. Similar performance for different conv1d kernel variants.

dim	s6	figures	life
32	0.298	0.249	0.190
64	0.298	0.252	0.191
128	0.303	**0.253**	0.194
192	**0.304**	0.252	0.192
256	0.300	0.252	**0.195**

(c) **Embed dimension**. Strong performance even for small embed dimension.

case	s6	figures	life
single sub	0.294	0.243	0.179
multi sub	**0.304**	**0.252**	**0.192**

(d) **Multi subject**. Jointly training on multiple subjects improves performance.

case	s6	figures	life
group only	0.273	0.232	0.183
sub only	0.302	**0.253**	**0.194**
sub + group	**0.304**	0.252	0.192

(e) **Prediction head**. Subject-specific heads are necessary (and sufficient).

internvl	qwen	vjepa	whisper	llama	s6	figures	life
✓	✗	✗	✗	✗	0.245	0.188	0.116
✗	✓	✗	✗	✗	**0.276**	**0.228**	**0.168**
✗	✗	✓	✗	✗	0.229	0.177	0.101
✗	✗	✗	✓	✗	0.173	0.106	0.091
✗	✗	✗	✗	✓	0.230	0.177	0.134
✗	✓	✓	✓	✓	0.300	0.251	0.188
✓	✗	✓	✓	✓	0.301	0.250	0.188
✓	✓	✗	✓	✓	0.298	0.247	**0.193**
✓	✓	✓	✗	✓	**0.303**	**0.253**	0.192
✓	✓	✓	✓	✗	0.290	0.243	0.181
✓	✓	✓	✓	✓	**0.304**	**0.252**	**0.192**

(f) **Feature set**. First block shows each feature's individual performance. Second block shows leave one feature out performance.

Table 1: **Ablation experiments** training on *friends s1-5*, *bourne*, *wolf*, and testing on *friends s6*, *figures*, *life*. *figures* was used as the validation set for early stopping. Default settings are marked in gray.

learning rate, weight decay, encoder kernel size, batch size, and embedding dimensions, among others. Additionally, multiple feature sets were prepared, including both individual model features and combinations thereof.

Subsequently, we randomly sampled multiple configurations from this hyperparameter space and trained each feature set using these configurations, reserving specific data (*life* and *bourne*) as the validation dataset (see Appendix C for an ablation on validation set choice). For each parcel in each subject, we selected the top-k models based on their performance on the validation dataset. We then averaged the fMRI activity predictions from these top-k models for the OOD movies to produce the final prediction for submission, a procedure shown theoretically and empirically to reduce variance and improve generalization [15]. More complex variants of this strategy (e.g., random ROI-ensembles, weighted averaging) have been successfully applied to fMRI decoding tasks [16, 17] and previous Algonauts editions [18]. The results are summarized in Figure 4.

The ensemble involved 49 models and provided a substantial increase in OOD performance of 0.011 for top-k=5, compared to the best single model, achieving a final leaderboard submission score of 0.208529.

5.1 Scaling Ensembles Further Improves Generalization

While our official submission employed a Top-$k = 5$ ensemble, increasing the ensemble size led to a consistent improvement in OOD generalization. In particular, a Top-$k = 20$ ensemble achieved a Pearson correlation of 0.211717 on the OOD set. This result, if submitted, would have placed our entry in second place overall.

Unfortunately, this optimization was not available in time for the official competition. Nevertheless, it highlights the value of large-scale ensembling when sufficient validation and model diversity are available. This finding aligns with classical ensemble learning theory, where increasing ensemble size often leads to reduced variance and enhanced generalization, provided that individual models contribute diverse errors [15].

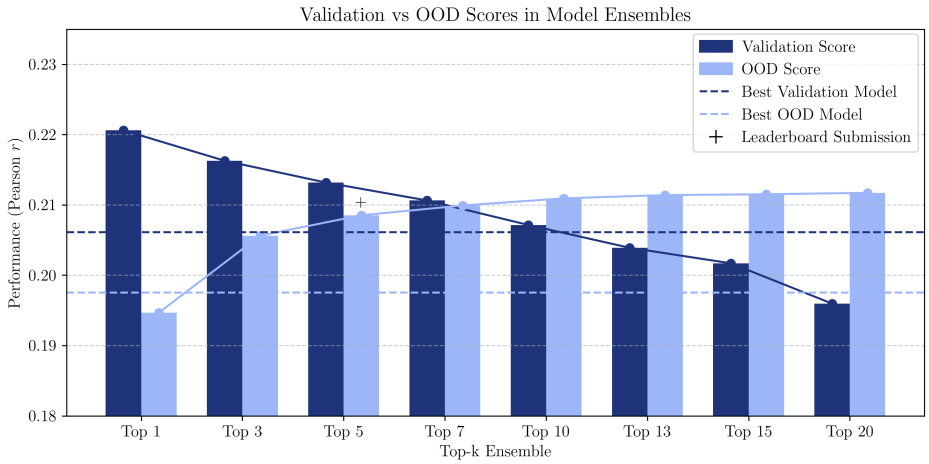

Figure 4: Comparison of ensemble performance across validation and out-of-distribution (OOD) data.[2] A total of 49 models were validated on *life* and *bourne*. Bars indicate Top-k ensemble scores. Dashed lines represent the performance of the best single model, while the '+' symbol denotes the best score submitted to the leaderboard.

6 Final thoughts: using Codabench scores to select the best model for each movie

The competition performance score on the leaderboard was calculated based on the average performance across six out-of-distribution movies that had some distinct characteristics from the training movies. For example, the movie "Chaplin" has no dialogue and is in black and white, in contrast to all training data. This posed a challenge to use the trained model with our previously identified optimal feature sets that include a language model. We first started by using dummy features for the Llama language model (using dummy transcripts where each line is just the time stamp). Using all "friends" and "movie10" movies as training data, our model's performance on "Chaplin" was 0.2205, 0.1467, 0.2057, and 0.1283 for the subjects 1, 2, 3, and 5, respectively. On a separate submission, we replaced the performance on Chaplin and trained with all "friends" and "movie10" movies without "Life" using a different model trained without Llama features, and saw the performance changed to 0.2394, 0.1412, 0.2004, and 0.1409, with a marked increase for subjects 1 and 5. In retrospect, the average performance score on "Chaplin" across subjects for the model without Llama features is higher than that of the top 3 submissions on the leaderboard.

Similarly, we noticed that the submissions from the ensemble solutions produced better scores for some movies, but not for others. Since the out-of-distribution data was so diverse, certain feature sets/training parameters may generalize better to specific out-of-distribution movies. Since Codabench provides the score for each movie, we were able to download the performance scores for each movie of each subject and select the best-performing model prediction among all our past submissions. This combined prediction was also not submitted in time for the competition, which could have resulted in a further increase in performance scores (average performance of r = 0.2105 across four subjects, 10% increase from our best submission performance r = 0.2085).

This optimization was only possible because Codabench provided movie-specific scores on the test set. While this approach was permitted under the competition rules, it partially circumvented the organizers' goal of a single generalizable encoding model. A future direction could explore whether systematic biases exist—e.g., if certain feature sets consistently generalize better to specific movie types—and whether feature optimization could be automated based on intrinsic properties of the test movie without access to the test scores.

7 Conclusion

This technical report summarizes the modeling approach and experimentation that went into MedARC's fourth-place submission to the Algonauts 2025 challenge. We use multimodal feature extraction, a shared-subject architecture, and ensemble strategy to optimize fMRI encoding model generalization to out-of-distribution stimuli. We describe our model and feature set ablations which may inform the development of future brain encoding models. This work contributes to our growing understanding of how the brain represents everyday, multimodal human experiences.

[2] Due to the 10-submission limit during the official competition, only two of the OOD results shown in this plot were actually submitted. The remaining results were retrospectively computed during the post-challenge phase.

References

[1] Matthew Brett, Ingrid S. Johnsrude, and Adrian M. Owen. The problem of functional localization in the human brain. *Nature Reviews Neuroscience*, 3(3):243–249, March 2002. ISSN 1471-0048. doi:10.1038/nrn756. URL https://www.nature.com/articles/nrn756. Publisher: Nature Publishing Group.

[2] 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.

[3] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[4] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

[5] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[6] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[7] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[8] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[9] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[10] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

[11] James S Gao, Alexander G Huth, Mark D Lescroart, and Jack L Gallant. Pycortex: an interactive surface visualizer for fmri. *Frontiers in neuroinformatics*, 9:23, 2015.

[12] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject synchronization of cortical activity during natural vision. *science*, 303(5664):1634–1640, 2004.

[13] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

[14] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, page 18. ACM, 2004.

[15] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[16] Takaaki Yoshimoto, Kai Tokunaga, and Junichi Chikazoe. Enhancing prediction of human traits and behaviors through ensemble learning of traditional and novel resting-state fmri connectivity analyses. *NeuroImage*, 303: 120911, 2024.

[17] Meenakshi Khosla, Keith Jamison, Amy Kuceyeski, and Mert R. Sabuncu. Ensemble learning with 3d convolutional neural networks for functional connectome-based prediction. *NeuroImage*, 199:651–662, 2019.

[18] Huzheng Yang, James Gee, and Jianbo Shi. Memory encoding model. *arXiv preprint arXiv:2308.01175*, 2023.

[19] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

[20] Oliver Schoppe, Nicol S Harper, Ben DB Willmore, Andrew J King, and Jan WH Schnupp. Measuring the performance of neural models. *Frontiers in computational neuroscience*, 10:10, 2016.

A Layer selection for feature encoding

This ablation study details the procedure for determining the optimal layer from each of the five models for our final encoding model. We first conducted a broad feature extraction, generating stimuli feature sets from multiple candidate layers spanning the architectural depth of each model. Each feature set was then used to independently train an encoding model on our primary training corpus (*Friends* S1-5, *wolf*, and *bourne*). The performance of each layer-specific model was subsequently evaluated on a held-out in-distribution dataset (*friends S6*) and two out-of-distribution (OOD) datasets (*life* and *figures*). The final layer reported for each model was the one whose features yielded the highest predictive performance in this ablation, ensuring our final feature space was composed of the most neurally-relevant representations from our extracted pool.

Table 2: Layer-wise Pearson Correlation Scores for all models

Model	**Layer**	**Pearson Correlation Score**		
		Friends S6	*Figures*	*Life*
Multimodal Large Language Models				
InternVL3-8B	10	0.2627	0.2061	0.1529
	15	0.2739	0.2197	0.1627
	20	**0.2753**	**0.2219**	**0.1685**
InternVL3-14B	20	0.2394	0.1784	0.0960
	30	**0.2431**	**0.1837**	**0.1102**
	40	0.2338	0.1822	0.1011
	47	0.2300	0.1764	0.0950
Qwen2.5-Omni 3B	10	0.2641	0.2117	0.1489
	15	0.2712	0.2185	0.1511
	20	**0.2746**	**0.2239**	**0.1671**
Qwen2.5-Omni 7B	10	0.2681	0.2095	0.1483
	15	0.2702	0.2181	0.1607
	20	**0.2703**	**0.2189**	**0.1609**
	25	0.2584	0.2025	0.1425
Specialized Unimodal Models				
VJEPA2 ViT-G	5	0.1721	0.1119	0.0449
	15	0.1910	0.1275	0.0562
	25	0.2256	0.1724	0.0964
	35	0.**2333**	0.1834	0.1089
	layernorm	0.2266	0.**1839**	0.**1177**
Llama 3.2 1B	**7**	**0.2221**	**0.1637**	**0.1246**
	11	0.2178	0.1595	0.1244
	15	0.2098	0.1589	0.1175
Llama 3.2 3B	7	0.2217	0.1603	0.1231
	11	**0.2274**	0.1722	**0.1303**
	15	0.2251	**0.1746**	0.1219
	19	0.2264	0.1714	0.1195
	23	0.2196	0.1664	0.1141
Whisper Large V3	**12**	**0.1759**	0.1073	0.0902
	25	0.1728	0.1111	0.0882
	31	0.1626	0.1056	0.0895
	layernorm	0.1739	**0.1117**	**0.0937**

B Context Length for Multimodal LLMs

This study was conducted to determine the optimal temporal context window for feature extraction from the M-LLMs. The primary context length used in our final models was 20 seconds, a decision initially guided by the computational

and time constraints of the challenge. To validate this choice and explore potential improvements, we extracted features using varied context lengths. For each model, these features were extracted exclusively from the single, most performant layer identified in our layer-wise ablation study.

For this specific ablation, we trained separate encoding models on a reduced corpus consisting of *Friends S1* and *Figures*. The performance of these models was then evaluated on a held-out in-distribution dataset (*Friends S6*) and an out-of-distribution dataset (*Life*). The results indicated that the 20-second window provided the most robust and highest overall performance across both M-LLM families. However, we noted a model-specific effect where the Qwen2.5-Omni model demonstrated a significant performance increase with a longer, 40-second context window.

Table 3: M-LLM Performance Across Varied Context Lengths

Model	**Context Length** (in seconds)	**Pearson Correlation Score**	
		Friends S6	*Life*
InternVL3-8B	10	0.2216	0.0948
	20	**0.2331**	**0.1392**
	30	0.2042	0.0899
	40	0.2052	0.0894
InternVL3-14B	10	**0.2161**	0.0928
	20	0.2082	**0.0955**
	30	0.2078	0.0939
	40	0.2081	0.0949
Qwen2.5-Omni 3B	10	0.2432	0.1401
	20	0.2430	0.1433
	30	0.2466	0.1477
	40	**0.2470**	**0.1517**
Qwen2.5-Omni 7B	10	0.2458	**0.1541**
	20	0.2315	0.1499
	30	0.2523	0.1521
	40	**0.2525**	0.1512

C Effect of Validation Set on Ensemble Performance

The choice of validation set plays a crucial role in determining which models are selected for the ensemble and can significantly affect generalization to out-of-distribution (OOD) data. To illustrate this, we compared two ensembling strategies: one using only *life* as the validation dataset, and another averaging performance across both *life* and *bourne*.

As shown in Figure 5, both strategies yield similar expected (validation) performance, but the impact on OOD generalization is notable. When using only *life* for validation, the Top-5 ensemble reached an OOD Pearson correlation of 0.2039. In contrast, incorporating both *life* and *bourne* increased the Top-5 OOD score to 0.2085. The gap is even more pronounced in the Top-3 and Top-1 ensembles, suggesting that more robust validation yields more transferable model selections.

These findings highlight the importance of choosing diverse and representative validation data when constructing ensembles—particularly in the presence of strong domain shifts between training and test distributions.

D Cross-subject encoding ceiling

To estimate a performance ceiling for our feature encoding model, we compare it with a *cross-subject* encoding model (Figure 6). The goal of the cross-subject encoding model is to jointly predict each subject's fMRI activity, using the activity from the remaining three other subjects as input. Similar to the feature encoding model, the cross-subject model is made up of an embedding stage and a prediction stage. In the embedding stage, the input fMRI activity time series of each subject is projected to a small latent dimension using a shared group plus subject-specific residual linear projection, similar to the fMRI prediction head above. To filter and temporally align each subject's embedding time series, we apply subject-specific depth-wise 1D convolutions with a small kernel width. We next compute a pooled latent embedding for each subject by leave-one-subject-out average pooling, so that the latent embedding for subject i

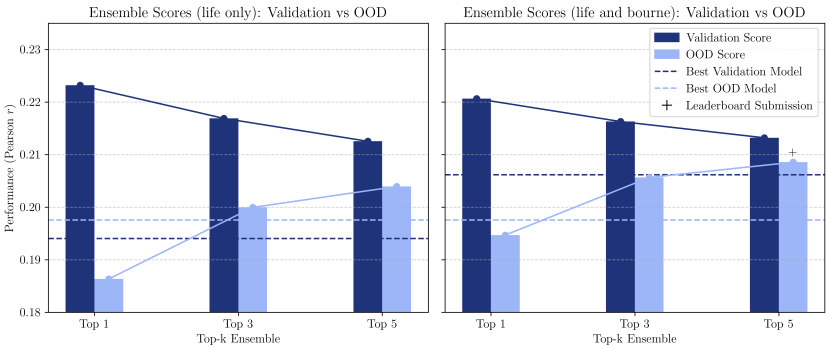

Figure 5: Impact of validation set selection on ensemble performance across validation and out-of-distribution (OOD) data. A total of 49 models were considered. The left panel shows ensembles selected using only *life*, while the right panel uses the average score across *life* and *bourne*. Bars indicate Top-k ensemble scores. Dashed lines represent the performance of the best single model, and the '+' symbol denotes the score submitted to the leaderboard.

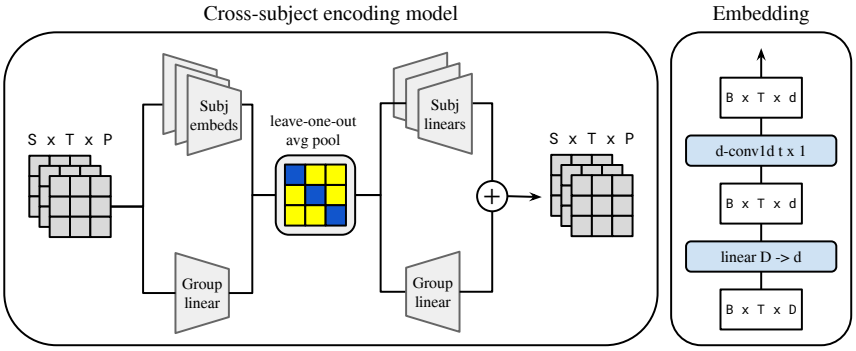

Figure 6: Cross-subject encoding model architecture. (`T` = number of time points, `S` = number of subjects, `P` = number of parcels (1000), `d` = embedding dimension, `B` = batch size.)

is computed by averaging the embedded activity for the other subjects $j \neq i$. Finally, we reconstruct the input fMRI activity for each subject using the same prediction head architecture as described in Section 4.

Ceiling comparison. In Figure 7 we compare our baseline model performance with two approaches for estimating a performance "ceiling' on the movie *life*. The first approach uses a simple split-half correlation estimate of the noise ceiling [19], taking advantage of the fact that a subset of the movies (*figures*, *life*) were shown twice. The second ceiling estimation uses our cross-subject encoding model described in Section 4. The intuition is that the aligned fMRI activity from other subjects should be an ideal "feature" for predicting a target subject's activity.

We observe that our baseline model outperforms the simple split-half correlation noise ceiling in most brain areas[3], except for parts of the peripheral early visual cortex and the dorsal attention network. By contrast, the cross-subject encoding model appears to be a more reliable ceiling estimate, outperforming the feature encoding model across the brain. In the difference map between the feature encoding and cross-subject encoding models, we observe significant unexplained variance in multiple brain areas, but especially in the dorsal attention network. A direction for future work is to investigate closing the gaps between feature encoding performance and the cross-subject encoding ceiling.

[3]This is not so surprising. The split-half correlation is a conservative lower bound of the true noise ceiling, since the repeat measurement is a noisy estimate of the true expected stimulus driven activity. Some corrections have been suggested, e.g. Schoppe et al. [20]. However these provide at best a loose upper bound.

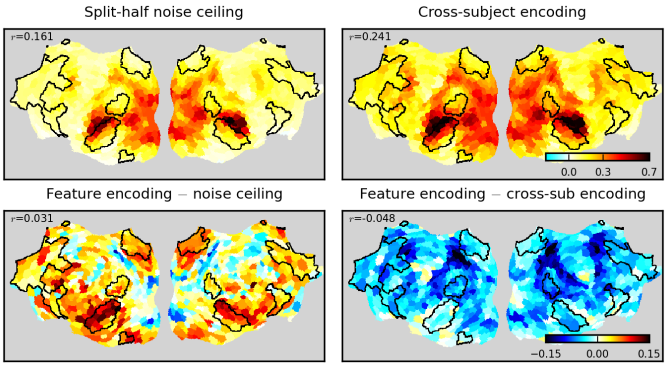

Figure 7: Comparison of performance ceiling estimates on the movie *life*. Top row shows the Pearson accuracy maps for each method. The bottom row shows the difference between feature encoding accuracy and the ceiling.