# Multimodal Recurrent Ensembles for Predicting Brain Responses to Naturalistic Movies (Algonauts 2025)

**Semih Eren**[*1,2] [†] **Deniz Kucukahmetler**[1,3] **Nico Scherf**[1,4]

[1]Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[2]TU Dresden, Dresden, Germany
[3]School for Embedded and Composite AI (SECAI),Dresden/Leipzig, Germany
[4]Center for Scalable Data Analytics & AI (ScaDS.AI), Dresden/Leipzig, Germany

## Abstract

Accurately predicting distributed cortical responses to naturalistic stimuli requires models that integrate visual, auditory and semantic information over time. We present a hierarchical multimodal recurrent ensemble that maps pretrained video, audio, and language embeddings to fMRI time series recorded while four subjects watched almost 80 hours of movies provided by the Algonauts 2025 challenge. Modality-specific bidirectional RNNs encode temporal dynamics; their hidden states are fused and passed to a second recurrent layer, and lightweight subject-specific heads output responses for 1000 cortical parcels. Training relies on a composite MSE–correlation loss and a curriculum that gradually shifts emphasis from early sensory to late association regions. Averaging 100 model variants further boosts robustness. The resulting system ranked third on the competition leaderboard, achieving an overall Pearson r = 0.2094 and the highest single-parcel peak score (mean r = 0.63) among all participants, with particularly strong gains for the most challenging subject (Subject 5). The approach establishes a simple, extensible baseline for future multimodal brain-encoding benchmarks.

## 1 Introduction

Understanding how complex, naturalistic stimuli drive human brain activity is a core question in cognitive computational neuroscience [17, 23, 15, 9]. Recent advances in deep learning and systems neuroscience have produced models that map rich stimulus features onto distributed cortical responses. The Algonauts Project 2025 aims at supporting progress in this area by providing a large open dataset and a clear community benchmark for predicting fMRI BOLD signals [4, 12] from naturalistic movie stimuli. With a common dataset, standardized metrics and a collaborative challenge format, the project allows researchers to compare ideas head-to-head and drives methods development towards more accurate, biologically grounded accounts of how the brain interprets the real world.

**Our contribution.** In this report, we present the encoding model that earned third place in the Algonauts 2025 challenge. Our key contributions are (i) a multimodal RNN architecture that fuses visual, auditory, and textual embeddings from large pretrained models and maps each input sequence to a sequence of BOLD responses for every region of interest (ROI); (ii) a curriculum-weighted loss that gradually refocuses the training objective from early visual and somatomotor regions to higher-order areas, mirroring the brain's hierarchical processing dynamics; and (iii) a 100-model ensemble whose averaged predictions maximise robustness and accuracy. This approach achieved an overall score of $r = 0.2094$ (correlation between predicted and measured brain responses). It further attained an across-subject mean single-parcel peak score of $r = 0.63$, surpassing the nearest

[*]Corresponding author: semih.eren@mailbox.tu-dresden.de
[†]Code available at `https://github.com/erensemih/Algonauts2025_ModalityRNN`.

competitor's peak of $r = 0.60$ by an absolute margin of 0.03, and it performed best on the most challenging participant (Subject 5).

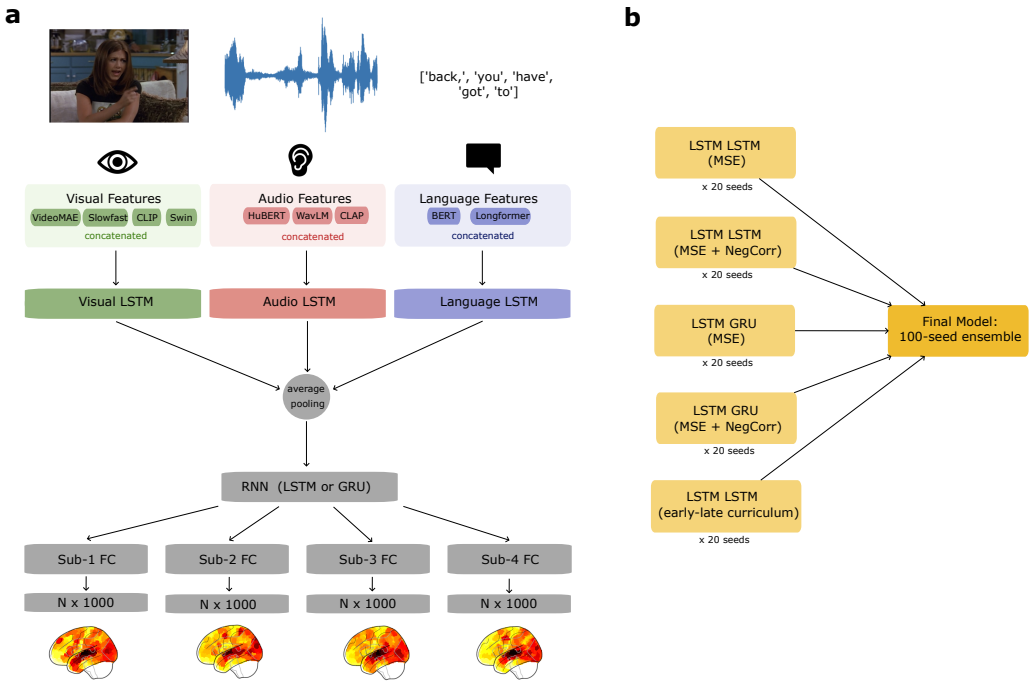


**Figure 1: a.** Method Illustration. **b.** Final Model Ensemble Components.

## 2 Related Work

Recent advances in computational neuroscience have shown that deep learning models provide powerful tools for predicting brain responses to multimodal, naturalistic stimuli [17, 23]. While initial studies primarily encoded fMRI responses to isolated sensory inputs such as static images or speech [15, 22], recent research highlights the benefits of integrating vision, audio, and language into unified brain-encoding models [12]. Models combining audiovisual features have notably improved prediction accuracy in higher-order cortical regions [7].

Capturing temporal dynamics is essential since natural stimuli induce brain activity evolving over extended timescales. Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) [13] and gated recurrent units (GRUs) [6], have successfully modeled temporal dependencies in brain data. Transformer-based architectures, capable of encoding longer temporal contexts, have also gained attention, though RNN-based methods remain prevalent due to their effectiveness and simplicity [1].

Benchmarking initiatives such as the Algonauts Project significantly accelerated progress through standardized datasets and evaluation methods [12]. Participants frequently adopt ensemble learning—averaging predictions from multiple diverse models—to enhance accuracy and robustness [18]. Curriculum learning, progressively shifting training focus from simpler to more complex brain regions, has similarly improved model stability and predictive power [3]. Our approach integrates these strategies to provide an effective multimodal baseline for future brain-encoding benchmarks.

## 3 Algonauts 2025 Challenge and Dataset

The Algonauts Project is an open challenge platform to bring together researchers in neuroscience and AI to build computational encoding models that accurately predict human brain responses to

rich naturalistic stimuli, fostering cross-disciplinary insights into biological and artificial intelligence. The Algonauts 2025 challenge dataset consists of whole-brain fMRI BOLD responses to naturalistic video stimuli with time-aligned audio and text. It consists of 65 hours of training data (55 hours of Friends seasons 1-6 and Movie10 set which consists of 4 movies). During the model-building phase, leaderboard scores were based on Friends Season 7, whereas in the model-selection phase, they were based on a two-hour-long out-of-distribution movie set. Details can be found at `https://algonautsproject.com/challenge.html`.

Neuroimaging data were acquired at a repetition time (TR) of 1.49 s from video clips from movies. Brain responses are summarised into $V = 1000$ cortical parcels covering early sensory and higher-order association areas. The evaluation metric is the Pearson correlation between predicted and actual parcel time series averaged across all parcels and subjects.

# 4 Method

## 4.1 Model

Our model has a three-stage approach that processes multiple input modalities, integrates them into a common representation and predicts subject-specific time-resolved BOLD signals (Figure 1.a). We extract features from each modality via frozen, pretrained models, feed them into separate LSTMs, average their hidden states and feed those into another RNN that forms a unified latent representation of the multimodal inputs. Finally, this joint embedding is routed through parallel, subject-specific prediction heads to map latent features to each individual's brain responses, accounting for personal scaling and idiosyncrasies. In essence, a global model is trained to predict every subject's brain responses, using an output gating mechanism to direct the shared representation to the appropriate subject-specific outputs.

### 4.1.1 Multimodal Feature Extraction

To obtain rich representations of each stimulus modality, we leverage state-of-the-art pretrained models. For the visual input, we extract features from each video using four complementary encoders—SlowFast [11], VideoMAE [21], Swin Transformer [16], and CLIP [19]—each trained on large-scale video or image datasets. Visual features are computed on 1.49-second clips and then time-aligned to the fMRI time points. For the auditory stream, we employ multiple pretrained audio models—HuBERT [14] and WavLM [5] for self-supervised speech/audio representations, and CLAP [10] for semantic audio embeddings—also extracted from the same 1.49-second windows and aligned to the fMRI. The text input is represented using two language models applied to dialogue transcripts: a base BERT [8] model for local semantic features (by feeding the last $n$ tokens based on the maximum token length of the language model) and a Longformer [2] for longer-range context. To ensure continuity of the transcripts of each episode and boost the performance of the first time stamps of the segments (e.g. episode split), we prepend the previous episode's transcripts (if available) when extracting language features.

### 4.1.2 Recurrent Modality Encoding and Fusion

Each modality's feature sequence $\mathbf{x}_m(t)$ (where $m$ indexes modality) is fed into a dedicated bi-directional RNN subnetwork (Figure 1.a):

$$\begin{aligned}
\mathbf{h}_m^{\rightarrow}(t) &= \mathrm{RNN}_m^{\rightarrow}\big(\mathbf{x}_m(t),\, \mathbf{h}_m^{\rightarrow}(t-1)\big),\\
\mathbf{h}_m^{\leftarrow}(t) &= \mathrm{RNN}_m^{\leftarrow}\big(\mathbf{x}_m(t),\, \mathbf{h}_m^{\leftarrow}(t+1)\big),\\
\mathbf{h}_m(t) &= \big[\mathbf{h}_m^{\rightarrow}(t)\,;\, \mathbf{h}_m^{\leftarrow}(t)\big] \in \mathbb{R}^{2H},
\end{aligned} \tag{1}$$

where each single-layer RNN has hidden size $H$ (set to 768). The concatenated hidden states $\mathbf{h}_m(t)$ form a sequence for each modality.

To combine information across the $M$ modalities, we average them elementwise:

$$\bar{\mathbf{h}}(t) = \frac{1}{M}\sum_{m=1}^{M} \mathbf{h}_m(t), \tag{2}$$

yielding $\bar{\mathbf{h}}(t) \in \mathbb{R}^{2H}$, an integrated representation at time $t$. (We found this simple average both effective and regularizing; learned weights or attention did not improve performance.)

The fused sequence $\bar{\mathbf{h}}(t)$ is then fed into a second RNN (denoted as "RNN (LSTM or GRU)" in Figure 1.a), which captures cross-modality temporal structure:

$$\mathbf{z}(t) = \mathrm{RNN}_{\mathrm{post}}\big(\bar{\mathbf{h}}(t),\, \mathbf{z}(t-1)\big), \tag{3}$$

producing $\mathbf{z}(t) \in \mathbb{R}^H$.

At each time $t$, the final hidden state $\mathbf{z}(t)$ goes into one of four subject-specific linear heads:

$$\mathbf{y}_s(t) = W_s\, \mathbf{z}(t) + b_s, \quad W_s \in \mathbb{R}^{1000\times H},\ \mathbf{y}_s(t) \in \mathbb{R}^{1000}. \tag{4}$$

During training, we pick the head matching the sample's subject $s$ and minimize

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}\Bigg[\underbrace{\|\mathbf{y}_s(t) - \hat{\mathbf{y}}_s(t)\|_2^2}_{\mathrm{MSE}} - \underbrace{\frac{\sum_{i=1}^{N}\big(y_{s,i}(t) - \bar{y}_s(t)\big)\big(\hat{y}_{s,i}(t) - \bar{\hat{y}}_s(t)\big)}{\sqrt{\sum_{i=1}^{N}\big(y_{s,i}(t) - \bar{y}_s(t)\big)^2}\,\sqrt{\sum_{i=1}^{N}\big(\hat{y}_{s,i}(t) - \bar{\hat{y}}_s(t)\big)^2}}}_{r}\Bigg], \tag{5}$$

where $\mathbf{y}_s(t)$ and $\hat{\mathbf{y}}_s(t)$ are the model's predicted and the true fMRI activations over $N = 1000$ parcels, respectively, and

$$\bar{y}_s(t) = \frac{1}{N}\sum_{i=1}^{N} y_{s,i}(t), \qquad \bar{\hat{y}}_s(t) = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_{s,i}(t). \tag{6}$$

At inference, routing through each $W_s$ yields subject-specific predictions. Sharing all recurrent layers but using separate output layers is both parameter-efficient and sufficiently flexible to capture individual response patterns.

### 4.2 Data Cleaning

To ensure that our subsequent analyses focus only on reliable, informative data, we applied a model-based filtering step as follows. First, we trained the model on the complete set of movie segments. Next, we used this trained model to generate predictions for each individual segment and computed the performance metric. We then excluded the segments with near-zero correlation scores from the dataset. The result is a pruned collection of segments on which the model can learn and predict with sufficient precision, thereby reducing noise and improving the robustness of downstream results. We removed the Friends season 6 18b, 19a, and 19b segments for subject 1 and Friends season 5 13a and Movie10 bourne01 segment for subject 2.

### 4.3 Training

Using Python 3.10 and PyTorch 2.7, we trained a single sequence-to-sequence model in PyTorch with separate subject-specific heads, using each movie episode paired with its corresponding fMRI time-series as one sample. To determine the optimal number of epochs, we employed early stopping based on cross-validation scores. Training was performed with the Adam optimizer at a fixed learning rate of $10^{-3}$ and a batch size of four to ensure balanced subject representation within each batch.

### 4.4 Validation

For validation, we used a group-wise cross-validation scheme, where each fold was defined by an individual movie from Movie10. All of our model improvements were assessed using this framework and throughout the competition our cross-validation score improvements remained closely correlated with our public leaderboard results.

### 4.5 "Early-vs-late" curriculum (Loss Weighting)

One training heuristic we explored was inspired by the hierarchical processing in the brain. Primary sensory areas encode basic stimulus features, whereas higher-order regions accumulate and integrate

information over much longer timescales. As a result, low-level sensory signals tend to be more predictable, while forecasting the responses of higher-order areas is more challenging. We hypothesised that emphasising early sensory ROIs at the start of training could guide the model to first learn low-level stimulus-response mappings, before focusing on more abstract regions. To implement this, we devised a dynamic loss weighting scheme that gradually shifts focus from early to late-processing ROIs over training epochs. Concretely, we predefined a set of "early-processing" parcels (covering Visual and Somatomotor Network from Schaefer 2018 parcellation [20]) based on the provided ROI labels. During training, we split the loss into two components: $\mathcal{L}_E$ for the early-ROI subset and $\mathcal{L}_L$ for the remaining late-ROI set. We then applied a time-varying weight: at the beginning, the loss weight $w_E$ for $\mathcal{L}_E$ is higher (0.55) and $w_L = 1 - w_E$ is lower (0.45), biasing training towards fitting early ROIs. As epochs progress, we linearly anneal these weights towards a balanced emphasis (e.g. $w_E$ down to 0.5 and $w_L$ up to 0.5). By the end of training, late ROIs receive equal or greater attention. In our experiments, this strategy improved the model's convergence on challenging ROIs and yielded a small boost in overall correlation performance (especially for mid-level and higher cortical areas) compared to a uniform loss weighting.

### 4.6 Ensemble Strategy

To further improve prediction accuracy, we employed an ensemble of models that achieved similar performances. Besides training models with different random seeds, we increased ensemble diversity by varying the model architecture and training objectives in five distinct ways: (1) using our base RNN architecture with MSE-only loss, (2) training a variant with a combined MSE and correlation loss, (3) swapping the recurrent units in the post-LSTM to GRU with MSE loss, (4) using the GRU version with combines MSE and correlation loss, (5) using the early-vs-late ROI weighted training curriculum described above (Figure 1.b).

For each of these five configurations, we trained 20 independent models with different random initialisations (seeds) for 5 epochs. In total, our ensemble comprised 100 models (5 configs × 20 seeds). At prediction time, we averaged the outputs of all ensemble members for each subject and timepoint. This simple averaging yielded a noticeable performance gain ($\sim 2\%$) over any single model.

## 5 Results and Experiments

In this section, we summarise the insights that guided our model's development and ultimately secured third place in the competition with an OOD score of 0.2094. All experiments employ leave-one-movie-out (LOMO) cross-validation on the Movie10 dataset: for each fold, we train on Friends Seasons 1–6 plus all movies except the held-out title. We believe LOMO provides a useful diagnostic for evaluating how well our method generalises to unseen data. All of the models are evaluated using the Pearson correlation coefficient as done in the challenge.

**Model Properties Impact on Results**

We investigated the impact of key architectural choices on our model's performance by conducting three ablation experiments in comparison to our model with multimodal RNN with subject heads (Figure 2.b). First, Unified modality & Single Head: we removed both subject-specific output heads and modality-specific encoders—replacing them with a single RNN that processes all modalities and a single prediction head for every subject. Second, Single Head: we retained separate modality encoders but collapsed all subject-specific heads into one shared output head. Third, Unified-Modality: we kept subject-specific heads intact, but replaced all modality-tailored encoders with a single, shared RNN.

The ablation experiments were successful: our RNN-based modality multi-head model significantly outperformed every ablated variant. (Figure 2.b).

**Final Models Results**

In Figure 2.c, we present the performance of our five final models—each trained with 20 different random seeds—and their ensemble. Although individual models achieve nearly identical scores, the ensemble consistently outperforms every individual model. Notably, however, in the visual areas

the ensemble lags slightly behind our early-late curriculum model, which shows the power of the curriculum approach in the visual areas.

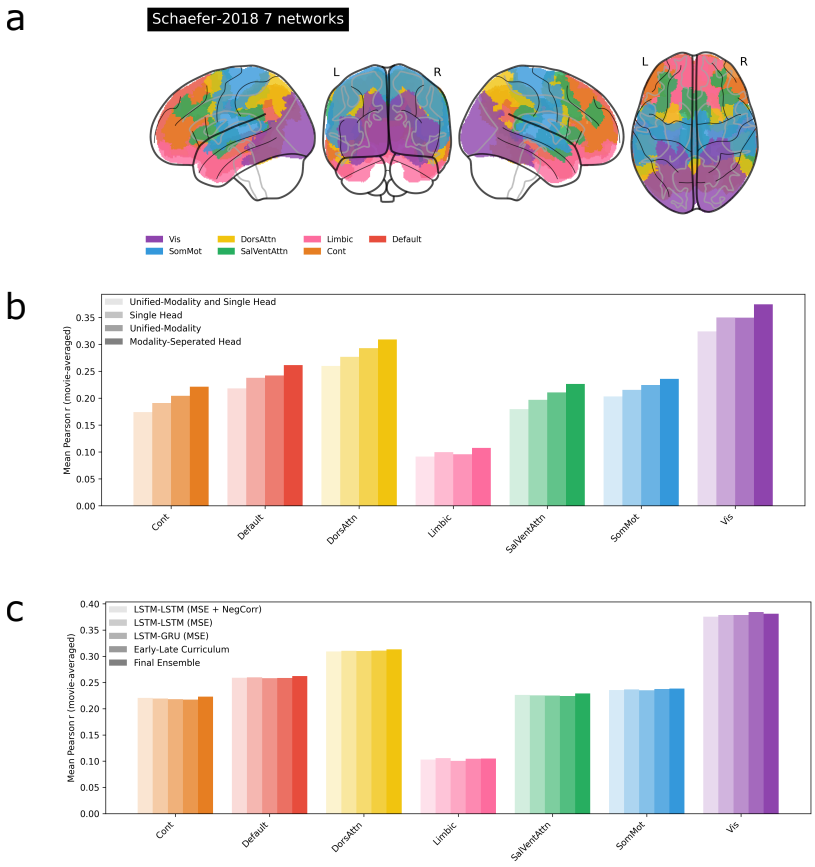


**Figure 2:** (a) Glass-brain rendering of the Schaefer-2018 1000-parcel atlas, highlighting seven functional networks: Visual (Vis), Somatomotor (SomMot), Dorsal Attention (DosAttn), Salience/Ventral Attention (SalVentAttn), Limbic, Frontoparietal (Cont), and Default. Smaller networks are rendered last to ensure full visibility, and the legend maps each color to its corresponding network. (b) Ablation study comparing three variants—excluding multi-head LSTM ("Single Head"), excluding multi-modal LSTM ("Unified-Modality"), and excluding both ("Unified-Modality and Single Head")—against our proposed Modality-Separated Head model. (c) Performance metrics for the final models and their ensemble.

| | Subject-1 | Subject-2 | Subject-3 | Subject-5 | Average |
|---|---|---|---|---|---|
| LSTM-GRU MSE+NegCorr | 0.283 | 0.255 | 0.276 | 0.242 | 0.264 |
| LSTM-LSTM MSE | 0.284 | 0.255 | 0.278 | 0.242 | 0.265 |
| LSTM-GRU MSE | 0.284 | 0.254 | 0.278 | 0.241 | 0.264 |
| LSTM-LSTM MSE+NegCorr | 0.283 | 0.254 | 0.277 | 0.242 | 0.264 |
| Early-late curriculum | 0.285 | 0.254 | 0.279 | 0.242 | 0.265 |
| Ensemble | 0.289 | 0.260 | 0.283 | 0.247 | 0.270 |

**Table 1:** Final Model Ensembling Performances Per Subject

## 6 Discussion and Limitations

Our model demonstrates particularly strong performance on the auditory task (as observed from the prediction visualisations on the challenge platform), which we attribute to its multimodal architecture's ability to leverage a greater number and variety of features heuristically demonstrated in Figure

2.b. Although it does not achieve the highest average score overall, it produces remarkably consistent results across subjects. In an effort to further improve performance and ensemble diversity, we experimented with incorporating a transformer backbone, but it failed to deliver meaningful gains and was therefore abandoned—an omission that may partly explain why we did not obtain the absolute top scores. Notably, while our model achieves the highest peak parcel accuracy, its weakest area lies in the prefrontal cortex (PFC). We hypothesized that this shortcoming stems from insufficient language features, and sought to address it by extending the window of our long-context language feature extractor to include previous episodes, with the goal of capturing more complex, temporally extended brain activity. Although this modification yielded a modest improvement, it did not produce the substantial performance jump we had hoped for. To further address this limitation, we introduced a dynamic ROI-weighting curriculum inspired by the brain's hierarchical processing of sensory signals. In this scheme, the model is first encouraged to concentrate its learning on primary sensory ROIs, and only in later stages to reallocate representational weight toward parcels exhibiting more complex, temporally extended dynamics. Although this curriculum produced modest gains in some regions and contributed to the final ensemble, it again fell short of delivering the substantial performance jump we had hoped for in the prefrontal cortex. We further investigated the use of a learning-rate scheduler. While scheduling afforded substantial gains in single-model performance, we observed that our seed ensemble without scheduling actually outperformed its scheduled counterpart. Consequently, we opted to use a relatively large fixed learning rate—likely promoting greater ensemble diversity—and achieved superior overall results.

## 7 Acknowledgements

We would like to thank Ariel Iporre Rivas, Pilou Bazin, Katja Seelinger and Kajal Singla for their support and insightful discussions. We also thank the Max Planck Computing and Data Facility (MPCDF) for providing GPU resources.

D.K. is supported by BMFTR in DAAD project 57616814 (SECAI).

N.S. is supported by BMBF (Federal Ministry of Education and Research) through ACONITE (01IS22065) and the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI.) Leipzig and by the European Union and the Free State of Saxony through BIOWIN.

## References

[1] H. Adeli, M. T. Rezazadeh Sereshkeh, and A. C. Connolly. Transformer-based brain encoding models, 2023. Online preprint.

[2] I. Beltagy, M. Peters, and A. Cohan. Longformer: the long-document transformer, 2020. Online preprint.

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the International Conference on Machine Learning*, 2009.

[4] J. Boyle, B. Pinsard, V. Borghesani, F. Paugam, E. DuPre, and P. Bellec. The courtois neuromod project: quality assessment of the initial data release (2020). In *Proceedings of the Conference on Cognitive Computational Neuroscience*, 2023.

[5] S. et al. Chen. Wavlm: large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 10:1254–1268, 2022.

[6] J. Chung, C. Gülçehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. Online preprint.

[7] R. M. Cichy, J. Kriegeskorte, and T. Dwivedi. The algonauts project: a platform for collaborative research in brain encoding and decoding, 2021. Online preprint.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding, 2019. Proceedings of NAACL-HLT.

[9] A. Doerig, R. P. Sommers, K. Seeliger, B. Richards, J. Ismael, G. W. Lindsay, K. P. Kording, T. Konkle, M. A. J. van Gerven, N. Kriegeskorte, and T. C. Kietzmann. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, pages 1–20, 2023.

[10] B. et al. Elizalde. Clap: learning audio–text joint embedding from noisy text supervision, 2023. Online preprint.

[11] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6202–6211, 2019.

[12] A. T. et al. Gifford. The algonauts project 2025 challenge: How the human brain makes sense of multimodal movies. arXiv preprint arXiv:2501.00504, 2025.

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[14] W. et al. Hsu. 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.

[15] N. Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446, 2015.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030, 2021.

[17] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant. Encoding and decoding in fmri. *NeuroImage*, 56(2):400–410, 2011.

[18] D. et al. Nguyen. Multiobjective ensemble for neural encoding, 2023. Online preprint.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020, 2021.

[20] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. T. Yeo. Local–global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral Cortex*, 28(9):3095–3114, 2018.

[21] Z. et al. Tong. Videomae: masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022. Online preprint.

[22] H. Wen, K. Shi, J. Chen, E. Liu, and Z. Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, 28:4136–4160, 2018.

[23] D. L. K. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19:356–365, 2016.