TRIBE: TRImodal Brain Encoder for whole-brain fMRI response prediction

Stéphane d'Ascoli
Meta AI
sdascoli@meta.com

Jérémy Rapin
Meta AI
jrapin@meta.com

Yohann Benchetrit
Meta AI
ybenchetrit@meta.com

Hubert Banville
Meta AI
hubertjb@meta.com

Jean-Rémi King
Meta AI
jeanremi@meta.com

Abstract

Historically, neuroscience has progressed by fragmenting into specialized domains, each focusing on isolated modalities, tasks, or brain regions. While fruitful, this approach hinders the development of a unified model of cognition. Here, we introduce TRIBE, the first deep neural network trained to predict brain responses to stimuli across multiple modalities, cortical areas and individuals. By combining the pretrained representations of text, audio and video foundational models and handling their time-evolving nature with a transformer, our model can precisely model the spatial and temporal fMRI responses to videos, achieving the first place in the Algonauts 2025 brain encoding competition with a significant margin over competitors. Ablations show that while unimodal models can reliably predict their corresponding cortical networks (e.g. visual or auditory networks), they are systematically outperformed by our multimodal model in high-level associative cortices. Currently applied to perception and comprehension, our approach paves the way towards building an integrative model of representations in the human brain. Our code is available at `https://github.com/facebookresearch/algonauts-2025`.

1 Introduction

Motivation. Progress in neuroscience has historically derived from an increasing specialization into cognitive tasks and brain areas. In the domain of vision, for instance, research focused on specialized cortical areas and their associated tasks, such as motion perception in V5 [1], face recognition in the fusiform gyrus [2], or the visual processing of written language in the visual word form area [3]. While this divide-and-conquer approach has undeniably yielded deep insights into the brain's mechanisms of cognition, it has led to a fragmented scientific landscape: How neuronal assemblies together construct and globally broadcast a unified representation of the perceived world remains limited to coarse conceptual models [4].

The fast progress in AI in the domains of language [5, 6], image [7], audio [8, 9] and video [10, 11] may help resolve this fragmentation challenge. Indeed, the representations learnt by these AI models have been shown to – at least partially – align with those of the brain [12, 13, 14, 15]. Motivated by this unexpected alignment, several teams have built encoding models to predict brain responses to natural stimuli from the activations of neural networks in response to images [12], speech [16] and text [13, 14, 17]. However, these encoding models are currently limited in three critical ways.

Preprint.

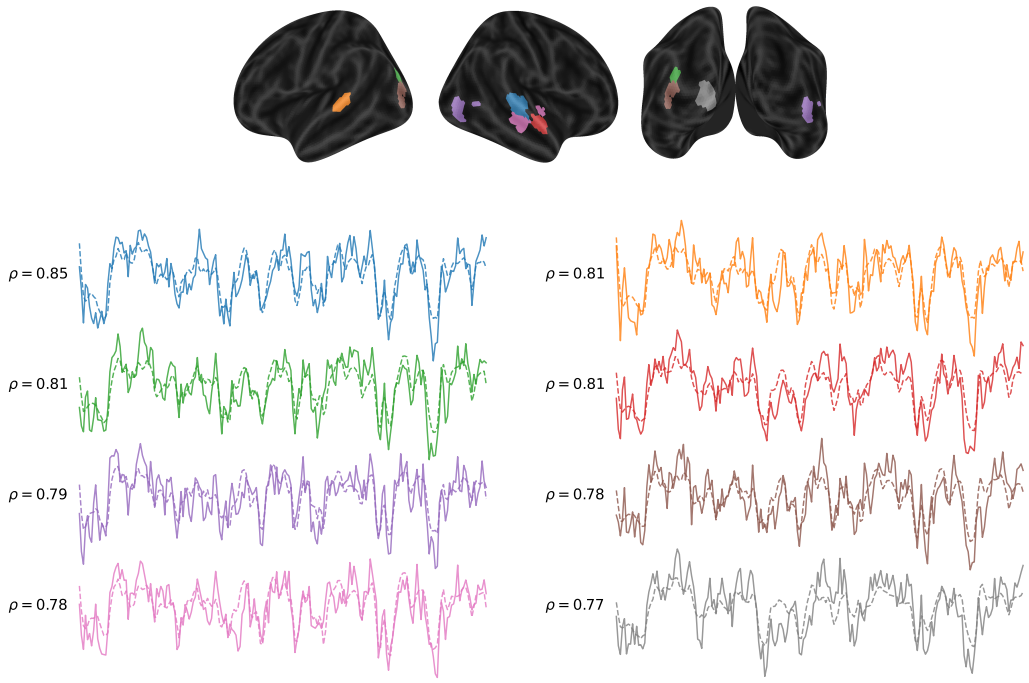


Figure 1: **TRIBE predicts brain responses to videos across diverse regions.** For eight brain parcels color-coded in the brain images, we report the BOLD response of the first participant to the first 5 minutes of a held-out movie in solid lines and our model's predictions in dashed lines, with the Pearson correlation of the two curves reported on the left.

First, *linearity*: existing encoding approaches typically rely on a ridge regression to map the AI model representations onto those of the brain. This assumes that these two sets of representations are linearly equivalent – a likely incorrect assumption [12, 14, 18].

Second, *subject-specificity*: due to large variability in brain responses from one subject to another, existing encoding approaches typically train a separate model for each subject, which prevents them from leveraging the similarities between brains (although [19]).

Third, *unimodality*: most existing encoding approaches predict brain responses from unimodal stimuli, which makes them incapable of capturing how the brain integrates information from multiple modalities [20]. This is particularly limiting as it has been shown that cross-modal interactions occur not only in specific multisensory areas [21, 22], but also in primary sensory areas [23, 24].

Contribution In this work, we introduce TRIBE, a novel deep learning pipeline to predict the fMRI brain responses of participants watching videos from the corresponding images, audio and transcript. This approach addresses the three limitations outlined above: our model learns how to capture the dynamical integration of modalities in an end-to-end manner across the whole brain, and from multiple subjects.

Our model achieves state-of-the-art results, reaching the first place out of 263 teams in the Algonauts 2025 competition on multimodal brain encoding. We demonstrate with ablation analyses the importance of the multimodal, multisubject and nonlinear nature of TRIBE. Finally, we observe that the benefit of multimodality is highest in associative cortices.

Related work While there has been recent research on deep learning for multimodal brain decoding [25, 26], there currently exists no equivalent for brain encoding. Some recent works suggest to train recurrent models to predict brain responses from frozen visual or linguistic features [27, 19], or fine-tune existing pretrained models using the brain encoding objective. While these relax the linearity assumption, they are restricted to a single sensory modality.

Conversely, a few recent studies have built encoding models on top of vision-language transformers, demonstrating gains compared to unimodal transformers [28, 29, 30, 31, 32]. However, these works rely solely on linear mappings to model brain responses from the activations of the multimodal transformers. We believe this can be suboptimal for two reasons. First, multimodal transformers are still relatively new: at rare exceptions [33, 34, 35], they often only integrate static images and text (audio and video being significantly more compute-intensive), and tend to lag behind the performance of unimodal transformers. Second, and more fundamentally, the way these models integrate information across modalities may be very different from how the human brain does such multimodal integration. An ideal encoding pipeline should thus *learn* how to best combine different modalities.

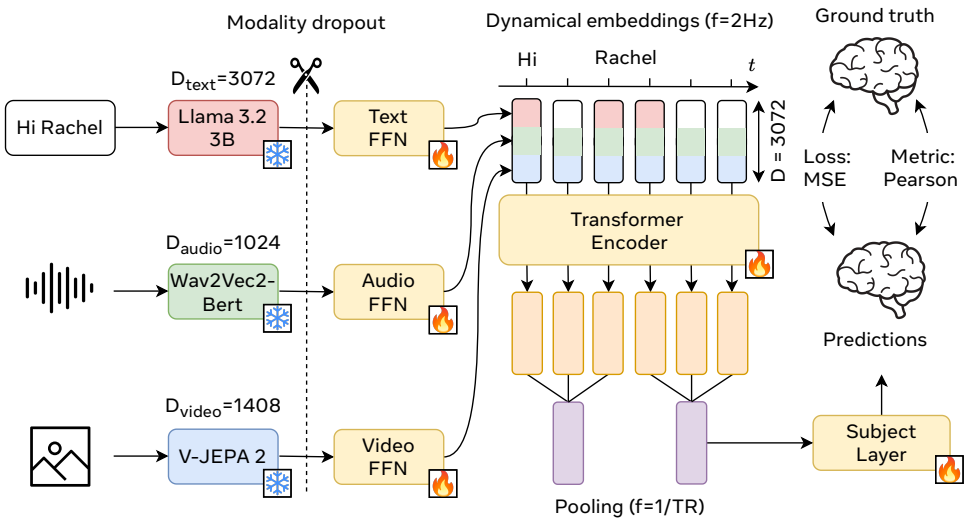


Figure 2: **Visual summary of our method.**

2 Methods

2.1 Overview

Task Our objective is to predict the brain activity of participants watching videos. This is framed as a regression task where the targets are the blood-oxygen-level-dependent (BOLD) signals detected at every repetition time (TR=1.49s) of a 3T fMRI recording device, separated into 1,000 non-overlapping parcels fig. 1.

For this, we take as input the video clip being viewed by the participant, as well as the corresponding audio file and transcript. From these, we extract high-dimensional embeddings from the intermediate layers of start-of-the art generative AI models along three modalities of interest: text, audio and video, which we feed to our deep encoding model, as illustrated in fig. 2.

Evaluation To assess performance, we evaluate for each parcel the Pearson Correlation ρ between predicted and ground truth fMRI responses across all TRs of the evaluation set. We then average these values over all parcels. We also refer to this metric as the "encoding score" of the model.

We hold out 10% of the recording sessions for the validation set, ensuring that the same videos are held out for each participant to prevent any form of data leakage.

2.2 Dataset

Data collection We train our encoding model on the Courtois NeuroMod dataset [36]. This dataset consists of six human participants who watched the same naturalistic videos, namely the

first six seasons of the popular TV series *Friends* as well as four movies: *The Bourne Supremacy, Hidden Figures, The Wolf of Wall Street* and *Life* (a BBC Nature documentary). This amounts to an unprecedently large recording volume of over 80 hours of fMRI per subject. In the present work, we focus on a subset of four subjects curated for the Algonauts 2025 competition [37].

Preprocessing We use the preprocessing pipeline provided by the Algonauts2025 competition. The whole-brain BOLD fMRI responses are preprocessed using fMRIprep [38] and projected to the Montreal Neurological Institute (MNI152NLin2009cAsym) standard space [39]. Functional images are then parcellated by averaging voxel-wise BOLD signals within each of the 1,000 parcels of the Schaefer atlas [40], yielding a single fMRI time series for each of the parcels. Finally, activations were z-scored per parcel across each session (corresponding to approximately 15 minutes of recording).

2.3 Model

Timed text embeddings We extract "timed" text embeddings from the timestamped transcripts of the videos. For each word w to embed, we prepend the preceding $k = 1,024$ words in the transcript, which we feed through Llama-3.2-3B [6]. For each intermediate layer l, we extract the token(s) overlapping with the word w and average them to obtain a contextualized word embedding of dimension $D_{\text{text}} = 2048$.

We then construct an evenly spaced grid at a frequency $f = 2\,\text{Hz}$, and for each time-bin τ, we sum the embeddings of words which overlap with the bin. This allows to temporally align the text features with the audio and video features.

Audio embeddings To obtain audio embeddings, we extract audio files from the videos, split them into 60-second chunks, then feed these through Wav2Vec-Bert-2.0 [9]. We then resample the hidden representations of the latter from 50 Hz to $f = 2\,\text{Hz}$. For each intermediate layer l, this yields time series of embeddings of dimension $D_{\text{audio}} = 1,024$.

Note that the resulting embeddings carry bidirectional information about both the past and future of the stimulus window, whereas text and video embeddings only contain information about the past.

Video embeddings For video embeddings, we again construct an evenly spaced grid at a frequency $f = 2\,\text{Hz}$, and for each bin of time, we feed 64 frames spanning the preceding 4 seconds to Video-JEPA 2 gigantic [11]. For each intermediate layer l, we compress the tensor of activations by averaging over all patch tokens, yielding a time series of embeddings of size $D_{\text{video}} = 1,280$.

Note that this spatial averaging step was necessary to keep the size of the tensor manageable. However, it comes at the cost of discarding positional information, which we expect to deteriorate encoding performance in low-level visual areas which exhibit a retinotopic mapping [41].

Combining the modalities For each of the three modalities m, the feature extraction described above leads to a time series of embeddings at $f = 2\,\text{Hz}$, with embeddings of shape $[L_m, D_m]$, where L_m and D_m are the number of layers and dimensionality of the transformer of modality m.

To compress these embeddings while retaining both low-level and high-level information, for each modality, we split the layers into L groups, then average the tensor per group along the layer dimension, compressing to a shape $[L, D_m]$.

We then concatenate the layers and feed the resulting vector through a linear layer with a shared output dimension $D = 1024$ followed by layer normalization. Finally, we concatenate the three modalities, leading to a time series of *multimodal* embeddings of shape 3×1024. This will be the input to our transformer encoder.

Transformer encoder We extract windows of duration $T = N \times TR$ from these embedding time series, add learnable positional embeddings and a learnable subject embedding, then feed the result through a Transformer encoder with 8 layers and 8 attention heads. This enables information to be exchanged between timesteps.

At the output of the transformer, we use an adaptive average pooling layer to compress the sequence from length fT to N, yielding one embedding per TR. Following [42], we then use a subject-conditional linear layer to project the latter to the 1,000-dimensional target space.

2.4 Training details

Modality dropout One desirable property of a multimodal encoding model is its ability to provide meaningful predictions in the absence of one or several modalities, for example for a silent movie or a podcast. To encourage this behaviour, while at the same time avoiding excessive reliance on one modality, we introduce modality dropout: during training, we randomly mask off each modality by zeroing out the corresponding input tensor with a probability p, ensuring that at least one modality is left unmasked.

Optimization We train our model for up to 15 epochs with the AdamW optimizer [43] using a batch size of 16. The learning rate is warmed up linearly to 10^{-4} over the first 10% of steps, then decayed following a cosine learning rate schedule. We use early stopping based on the validation Pearson score computed on a held-out set. To improve generalization, we use stochastic weight averaging [44], which involves averaging model weights obtained at the end of each epoch, once the validation metrics are near their plateau.

Ensembling To further improve generalization, we ensemble the predictions of $M = 1000$ models, whose initialization and shuffling seeds are all different. To strengthen ensemble diversity, for each model, we sample a set of hyperparameters uniformly in the grid specified in appendix A. For each parcel separately, we compute the encoding score of all models on the validation set, then compute a softmax distribution over models with temperature 0.3, which will determine the weight assigned to each model for this given parcel.

Implementation details We extract stimuli features from pretrained language, audio and video models available on the `HuggingFace` platform [45] and cache them as `Numpy` memmap arrays [46] for fast loading during the training of our encoding model. Feature extraction is completed in 24 hours on 128 V100 GPUs with 32GB of VRAM, and model training lasts 24 hours on a single such GPU. We use the transformer implementation from the `x-transformers` package[1]. We list the licenses of the assets used in this work in appendix B.

Rank	Team	Mean score	Subject 1	Subject 2	Subject 3	Subject 5
1	**Ours**	**0.2146**	0.2381	0.2105	0.2377	0.1720
2	NCG	**0.2096**	0.2353	0.2046	0.2268	0.1718
3	SDA	**0.2094**	0.2233	0.2072	0.2271	0.1798
4	MedARC	**0.2085**	0.2295	0.2003	0.2300	0.1743
5	CVIU-UARK	**0.2055**	0.2306	0.2010	0.2240	0.1662

Table 1: **Our model achieves first place in the Algonauts 2025 public leaderboard.** We report the results of the top five out of 263 teams.

OOD	Movie	Mean score	Subject 1	Subject 2	Subject 3	Subject 5
✗	Friends Season 7	**0.3195**	0.3419	0.3239	0.3346	0.2775
✓	Pulp Fiction	**0.2604**	0.2765	0.2611	0.2431	0.2610
✓	Princess Mononoke	**0.2449**	0.2816	0.2507	0.2851	0.1623
✓	Passe-partout	**0.2323**	0.2763	0.2587	0.2370	0.1573
✓	World of Tomorrow	**0.1924**	0.2210	0.1606	0.2196	0.1686
✓	Planet Earth	**0.1886**	0.1483	0.2029	0.2331	0.1699
✓	Charlie Chaplin	**0.1686**	0.2249	0.1289	0.2080	0.1128

Table 2: **Our model generalizes to highly out-of-distribution movies.** We provide the detailed results on the held-out datasets of the Algonauts 2025 competition.

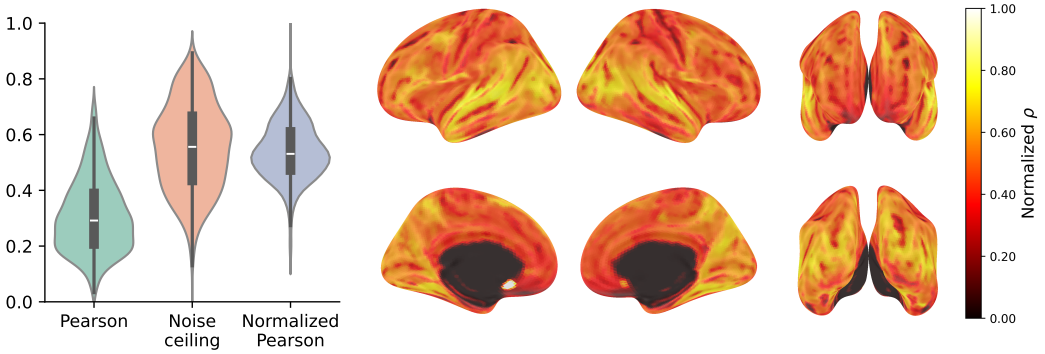
[1] `https://github.com/lucidrains/x-transformers`

3 Results

3.1 Algonauts 2025 competition results

Our model achieves the first place out of 262 teams in the Algonauts 2025 competition on multimodal brain encoding. As shown in table 1, we outperform competitors by a substantial margin: the gap between our model and the runner-up is larger than between the runner-up and the fifth.

We display the results across the various held-out datasets in table 2. In the in-distribution setting of the first phase of the competition (*Friends* season 7), our model achieves a mean score of 0.3195. Unsurprisingly, when tested in the out-of-distribution setting of the second phase of the competition, performance is lower, with an average of 0.2146. Remarkably, our model achieves robust scores even in the extreme out-of-distribution setting of cartoons (0.1924 for *World of Tomorrow*) wildlife documentaries (0.1886 for *Planet Earth*), and silent black-and-white movies (0.1686 for *Charlie Chaplin*).



(a) Distribution across parcels

(b) Cortical projection of the normalized scores

Figure 3: **Our model performs consistently across the whole brain.**
(a) We report a histogram of unnormalized and normalized encoding scores (see definition in eq. (1)) of the 1000 parcels, averaged across subjects.
(b) We project the normalized encoding scores of the first subject onto the `fsaverage5` cortical surface.

3.2 Noise ceiling analysis

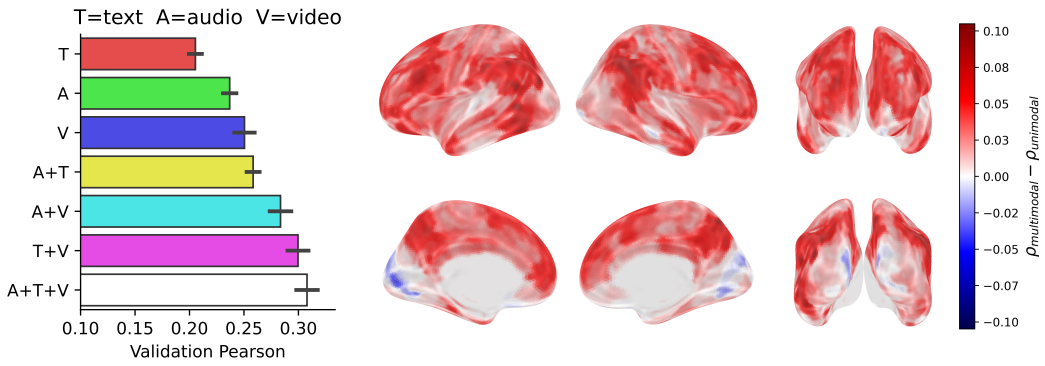
Following common practice in the brain encoding literature [47], we estimate the noise ceiling to investigate to what extent encoding errors are due to unexplainable randomness rather than model suboptimality. We achieve this by computing the inter-trial correlation ρ_{self} of the BOLD responses to the movies *Hidden Figures* and *Life*, which were viewed twice by each participant. We then define the normalized Pearson correlation of our model by dividing it by that of an ideal encoding model, following [47]:

$$\rho_{norm} = \frac{\rho}{\rho_{max}}, \quad \rho_{max} = \sqrt{\frac{2}{1 + \frac{1}{\rho_{self}}}} \tag{1}$$

Our model achieves a normalized Pearson correlation of 0.54 ± 0.1 across all parcels (fig. 3a): in other words, it captures 54% of explainable variance on average. The fairly small inter-quartile range reflects the fact that our model is rather consistent across brain areas. The highest values are achieved in the auditory and language processing cortices, where our model is near the noise ceiling (fig. 3b).

3.3 The benefit of multimodality

To what extent do the three modalities combined by TRIBE contribute to encoding performance? We address this question in fig. 4a, by assessing the encoding performance of TRIBE retrained with



(a) Modality ablations

(b) Multimodal versus best unimodal

Figure 4: **Multimodal encoders outperform unimodal encoders.**
(a) We compare the encoding scores of encoders trained on a subset of modalities in the same conditions. Error bars denote s.e.m across subjects.
(b) For each parcel of the first subject, we compute the difference in encoding score between the multimodal encoder and the best out of the three unimodal encoders, then project the results onto the `fsaverage5` cortical surface.

various modalities ablated. When training on a single modality, TRIBE achieves substantially lower encoding scores. We observe that the text modality achieves the lowest average encoding score with 0.22, followed by audio at 0.24 and video at 0.25. When combining any two modalities together, the encoding scores rise significantly compared to the unimodal models: the best bimodal model, obtained by combining text and video, achieves an encoding score of 0.30. Finally, combining the three modalities together yields an additional boost, bringing the value to 0.31. This hints to the fact that each modality plays a complementary role.

In which areas does multimodality yield the strongest gains? In fig. 4b, we compare for each parcel the encoding score of the multimodal encoder with that of the best of the three unimodal encoders. We observe that the multimodal encoder consistently outperforms the unimodal models, especially in associative areas such as the prefrontal or parieto-occipito-temporal cortices (up to 30% increase in encoding score). Interestingly, the multimodal model performs less well than the vision-only model in the primary visual cortex, which is highly specific to visual features.

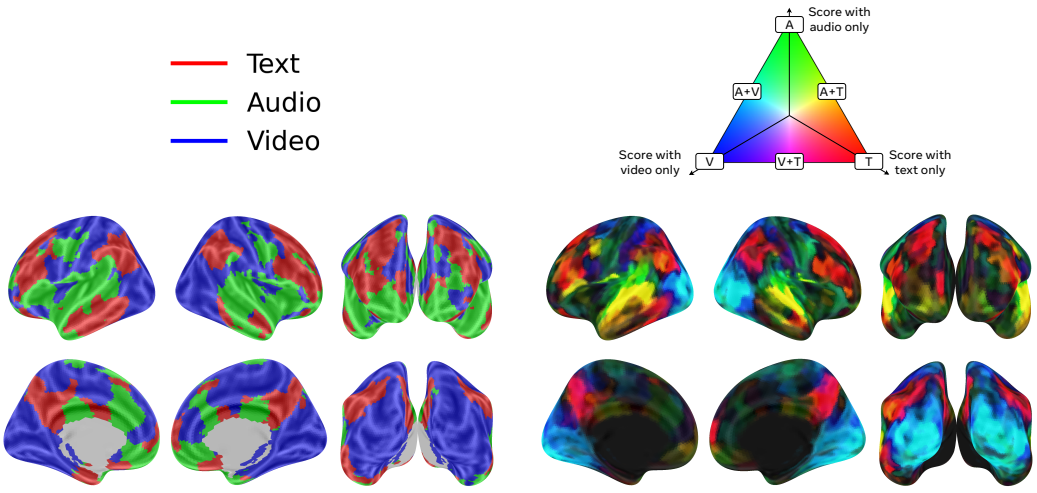
Together, these results demonstrate that our multimodal encoder effectively captures interactions between modalities, which improves whole-brain decoding.

3.4 The interplay between modalities

Which brain areas is dominated by unimodal or multimodal representations? To address this question, we single out modalities by probing our multimodal model with all modalities masked off except one (using the same procedure as for the dropout described in section 2).

In fig. 5a, we color each parcel of the cortex according to the modality that achieves the highest encoding score via this procedure. The three modalities each cover broad regions: audio predominates near the temporal gyrus, video predominates in the occipital cortex and parts of the parietal cortex, while the text feature, which presumably contains the most semantic information, predominates in large parts of the parietal and prefrontal lobes.

To study the interplay between modalities, we then overlay the contribution of the three modalities using an RGB encoding where red, green and blue respectively represent the encoding scores achieved solely with text, audio and video (fig. 5b). To make hues more visible, we restrict our analyses to unimodal and bimodal interactions by substracting the smallest contribution among the three modalities. We observe interesting bimodal associations in some key areas: in particular, text+audio (yellow) in can be observed in the superior temporal lobe and video+audio (cyan) can be observed in the ventral and dorsal visual cortices.



(a) Highest scoring modalities (b) Interplay between modalities

Figure 5: **The modalities and their interplay map onto the expected brain areas.**
(a) For each parcel of the first subject, we assess the encoding contribution of each modality by masking the two other modalities, and color-code according to the highest contribution.
(b) We color-code each parcel using an RGB encoding where red, green and blue intensities are determined by the encoding score of the model with text, audio and video unmasked, respectively. For readability, we limit ourselves to bimodal interactions by substracting the smallest of the three contributions. Red, blue and green areas correspond to unimodal areas well encoded by text, audio and video respectively, while magenta, yellow and cyan correspond to bimodal areas well encoded by text-video, text-audio and video-audio interactions respectively.

These observations provide new insights on how multisensory integration may occurs in the human cortex.

3.5 Ablations and scaling laws

In fig. 6a, we demonstrate the importance of using a transformer and a multi-subject training scheme: the encoding score drops from 0.31 to 0.29 when training separately for each subject, and down to 0.23 when removing the transformer.

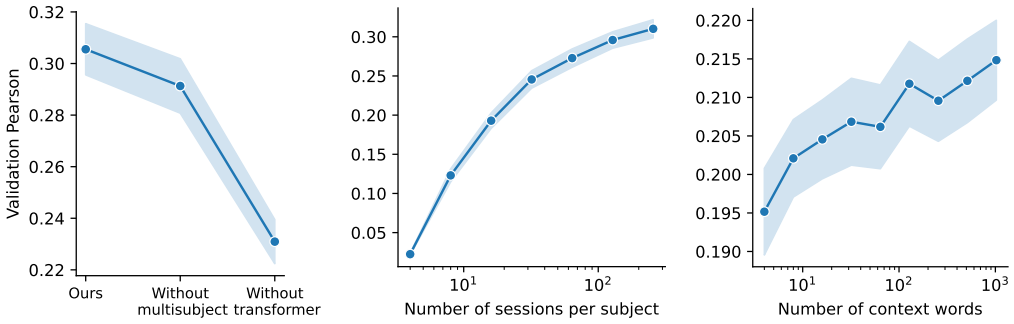
In fig. 6b, we show how the encoding score scales with the amount of recordings in the training set. We observe a strong increasing trend which has not reached a plateau, in line with recent work [48].

In fig. 6c, we show that increasing the context length used for the language model words strongly enhances encoding performance, without any plateau even at very long context lengths of 1024 words. This confirms that TRIBE effectively captures high-level semantic features, far beyond the word and sentence level.

4 Discussion

In this work, we make a step towards an integrative model of the brain during naturalistic perception by training an encoding model on an unprecedently-large fMRI dataset of participants watching videos. Crucially, our model is the first encoding pipeline which is simultaneously nonlinear, multisubject and multimodal: our ablations demonstrate the importance of these three aspects for encoding, especially in associative cortices. Our model achieves the first place in the Algonauts 2025 brain encoding competition, and scaling laws suggest that encoding performance increases systematically with the number of recordings, holding promise for further improvements with larger datasets.

Limitations There are several remaining limitations to our work. First, our approach currently operates on a coarse parcellation of brain areas – reducing hundreds of thousands of voxels to 1,000



(a) Model ablations (b) Number of sessions (c) Context length

Figure 6: **Ablations and scaling laws of our model.**
(a) We compare the results of our model with those achieved without multi-subject training or without the transformer model.
(b) We report the results obtained by the multimodal encoder as we increase the number of sessions used in the training set.
(c) We report the results obtained by the text-only encoder as we increase the context length of the language model used to extract features (Llama-3.2-3B).
In all panels, shaded regions indicate the s.e.m over the four subjects.

cortical parcels. This design choice, introduced by the Algonauts2025 challenge, likely increases the signal-to-noise ratio by smoothing out the responses spatially, and certainly reduces computational cost, which is important for the whole-brain prediction setting considered here. However, this approach limits the spatial resolution of our model, which inherently prevents it from capturing highly localized phenomena. Adapting our model for voxel-level prediction is an important avenue for future work. Second, our current approach is limited to fMRI data. Consequently, the precise temporal dynamics of neuronal activity, and the exact neural assemblies underlying these macroscopic signals, remain, here, unresolved. Third, while the volume of recording per participant in the study considered here is particularly large, only four participants were included: extending and improving our results on a larger pool of participants is an important next step. Finally, the present approach remains limited to perception and comprehension. Behavior, memory and decisions are other important cognitive components to integrate into the present model.

Broader impact Building a model able to accurately predict human brain responses to complex and naturalistic conditions is an important endeavour for neuroscience. Not only does this approach open the possibility of exploring cognitive abilities (e.g. theory of mind, humour) that are challenging to isolate with minimalist designs, but they will eventually be necessary to evaluate increasingly complex models of cognition and intelligence. In addition, our approach forges a path to (1) integrate the different sub-fields of neuroscience into a single framework and to (2) develop in silico experimentation [49], where in vivo experiments could be complemented and guided by the predictions of a brain encoder. While the exploration of this paradigm falls beyond the scope of this technical report, we believe that epistemology is ordered: interpretation and scientific insights are most legitimate if they derive from the model that makes the best prediction. In that regard, the first place achieved by TRIBE in the Algonauts 2025 competition gives scientific credit to our approach.

References

[1] Michael N Shadlen and William T Newsome. Neural basis of a perceptual decision in the parietal cortex (area lip) of the rhesus monkey. *Journal of neurophysiology*, 86(4):1916–1936, 2001.

[2] Nancy Kanwisher and Galit Yovel. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1476):2109–2128, 2006.

[3] Stanislas Dehaene and Laurent Cohen. The unique role of the visual word form area in reading. *Trends in cognitive sciences*, 15(6):254–262, 2011.

[4] George A Mashour, Pieter Roelfsema, Jean-Pierre Changeux, and Stanislas Dehaene. Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5):776–798, 2020.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] 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.

[7] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[8] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[9] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE, 2021.

[10] 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.

[11] 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.

[12] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[13] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

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

[15] 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.

[16] Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, Jean-Remi King, et al. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 35:33428–33443, 2022.

[17] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32, 2019.

[18] Drew Linsley, Pinyuan Feng, and Thomas Serre. Can deep neural networks learn biological vision? *arXiv preprint arXiv:2504.16940*, 2025.

[19] Omar Chehab, Alexandre Defossez, Jean-Christophe Loiseau, Alexandre Gramfort, and Jean-Remi King. Deep recurrent encoder: A scalable end-to-end network to model brain signals. *arXiv preprint arXiv:2103.02339*, 2021.

[20] Yu Hu and Yalda Mohsenzadeh. Neural processing of naturalistic audiovisual events in space and time. *Communications Biology*, 8(1):110, 2025.

[21] Chuanji Gao, Jessica J Green, Xuan Yang, Sewon Oh, Jongwan Kim, and Svetlana V Shinkareva. Audiovisual integration in the human brain: a coordinate-based meta-analysis. *Cerebral Cortex*, 33(9):5574–5584, 2023.

[22] Michael S Beauchamp. See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Current opinion in neurobiology*, 15(2):145–153, 2005.

[23] Jon Driver and Toemme Noesselt. Multisensory interplay reveals crossmodal influences on 'sensory-specific'brain regions, neural responses, and judgments. *Neuron*, 57(1):11–23, 2008.

[24] Barry E Stein and Terrence R Stanford. Multisensory integration: current issues from the perspective of the single neuron. *Nature reviews neuroscience*, 9(4):255–266, 2008.

[25] Simon Dahan, Gabriel Bénédict, Logan Zane John Williams, Yourong Guo, Daniel Rueckert, Robert Leech, and Emma Claire Robinson. Sim: Surface-based fmri analysis for inter-subject multimodal decoding from movie-watching experiments. In *The Thirteenth International Conference on Learning Representations*, 2025.

[26] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision*, pages 242–259. Springer, 2024.

[27] Umut Güçlü and Marcel AJ Van Gerven. Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in computational neuroscience*, 11:7, 2017.

[28] Dota Tianai Dong and Mariya Toneva. Vision-language integration in multimodal video transformers (partially) aligns with the brain. *arXiv preprint arXiv:2311.07766*, 2023.

[29] Subba Reddy Oota, Jashn Arora, Vijay Rowtula, Manish Gupta, and Raju S Bapi. Visio-linguistic brain encoding. *arXiv preprint arXiv:2204.08261*, 2022.

[30] Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*, 10, 2022.

[31] Aria Y Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *BioRxiv*, pages 2022–09, 2022.

[32] 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.

[33] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A General Architecture for Structured Inputs & Outputs. *arXiv*, July 2021.

[34] Siddharth Srivastava and Gaurav Sharma. Omnivec2 - a novel transformer based network for large scale multimodal and multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27412–27424, June 2024.

[35] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 Technical Report. *arXiv*, December 2024.

[36] Marie St-Laurent, Basile Pinsard, Oliver Contier, Katja Seeliger, Valentina Borghesani, Julie Boyle, Pierre Bellec, and Martin Hebart. cneuromod-things: a large-scale fmri dataset for task-and data-driven assessment of object representation and visual memory recognition in the human brain. *Journal of Vision*, 23(9):5424–5424, 2023.

[37] 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. *arXiv preprint arXiv:2501.00504*, 2024.

[38] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

[39] Matthew Brett. The mni brain and the talairach atlas. *www. mrc-Mrc-cbu. cam. ac. uk/Imaging/mnispace. Html*, 2002.

[40] 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.

[41] Brian A Wandell and Jonathan Winawer. Imaging retinotopic maps in the human brain. *Vision research*, 51(7):718–737, 2011.

[42] Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023.

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[44] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[45] Shashank Mohan Jain. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pages 51–67. Springer, 2022.

[46] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[47] Oliver Schoppe, Nicol S. Harper, Ben D. B. Willmore, Andrew J. King, and Jan W. H. Schnupp. Measuring the performance of neural models. *Frontiers in Computational Neuroscience*, Volume 10 - 2016, 2016.

[48] 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.

[49] Shailee Jain, Vy A Vo, Leila Wehbe, and Alexander G Huth. Computational language modeling and the promise of in silico experimentation. *Neurobiology of Language*, 5(1):80–106, 2024.

Appendices

A Hyperparameters

Hyperparameter	Base value	Other values used for ensembling
Number of epochs	15	
Number of TRs per window	100	
Window jitter	10s	
Learning Rate	10^{-4}	
Batch Size	16	
Optimizer	AdamW	
Scheduler type	OneCycleLR (cosine)	
Scheduler warmup phase	10%	
Stochastic weight average epochs	8	
Dropout	0	
Weight Decay	0	
Hidden Size	3072	
Text model	Llama-3.2-3B	
Audio model	Wav2Vec-Bert-2.0	
Video model	V-JEPA-2-Gigantic-256	
Loss	MSE	Pearson, SmoothL1, HuberLoss
Modality Dropout	0.2	0.0, 0.4
Layer groups	[0.5, 0.75, 1]	[0,0.5,1], [0.5, 1], [0, 0.2, 0.4, 0.6, 0.8, 1.]
Layer extraction	group by intervals	extract single layers
Layer aggregation	concatenate	average
Modality aggregation	concatenate	average
Use subject embedding	✓	✗

Table 3: Hyperparameters used in our model

B Licenses

HuggingFace models:

- Video-JEPA 2: Apache (`https://github.com/facebookresearch/vjepa2/blob/main/LICENSE`)
- Wav2Vec-Bert-2.0: MIT (`https://huggingface.co/datasets/choosealicense/licenses/blob/main/markdown/mit.md`)
- LLama-3.2-3B: llama3.2 (`https://huggingface.co/meta-llama/Llama-3.2-1B/blob/main/LICENSE.txt`)

Packages:

- `x-transformers`: MIT (`https://github.com/lucidrains/x-transformers/blob/main/LICENSE`)
- `nilearn`: BSD (`https://github.com/nilearn/nilearn/blob/main/LICENSE`)
- `pytorch`: `https://github.com/pytorch/pytorch/blob/main/LICENSE`

Datasets:

- Courtois NeuroMod: CC0 (`https://creativecommons.org/publicdomain/zero/1.0/legalcode`)