The One Where They Brain-Tune for Social Cognition: Multi-Modal Brain-Tuning on *Friends*

Nico Policzer[1,2,3] **Cameron Braunstein**[1,2] **Mariya Toneva**[2]
[1]Saarland University, Saarbrucken, Germany
[2] MPI for Software Systems, Saarbrucken, Germany
[3] University of British Columbia, Vancouver, Canada
npoliczr@student.ubc.ca
braunstein@cs.uni-saarland.de
mtoneva@mpi-sws.org

Abstract

Recent studies on audio models [27, 15] show *brain-tuning*–fine-tuning models to better predict corresponding fMRI activity–improves brain alignment and increases performance on downstream semantic and audio tasks. We extend this approach to a multimodal audio-video model to enhance social cognition, targeting the Superior Temporal Sulcus (STS), a key region for social processing, while subjects watch *Friends*. We find significant increases in brain alignment to the STS and an adjacent ROI, as well as improvements to a social cognition task related to the training data–sarcasm detection in sitcoms. In summary, our study extends brain-tuning to the multi-modal domain, demonstrating improvements to a downstream task after tuning to a relevant functional region.

1 Introduction

Recent works in fine-tuning audio models to human fMRI data, specifically language and auditory areas, show improvements to brain alignment, as well as increases to performance on semantic and audio evaluations [15, 27, 29]. However, frontier AI models are increasingly multi-modal [8, 41]. These models are uniquely posed to model human social cognition, *i.e.*, inferring a perceived person's internal state, which requires integrating information across modalities [9, 5, 3] and is critical as AI becomes more integrated in our daily lives [6]. However, a recent study [17] identified a major gap in AI models' abilities to match human social perception, as well as encode brain activity in the lateral stream, a processing stream proposed for social cognition [32]. The Superior Temporal Sulcus (STS), the end point of the lateral stream, is a brain-region that has been shown to encode features of social interaction relevant to social cognition [26, 21, 39, 32, 19, 1, 13]. We therefore investigate whether brain-tuning an audio-video model to the STS can 1) improve brain encoding of the STS and other lateral stream ROIs, and 2) increase downstream performance on social cognition tasks.

Concretely, we brain-tune the joint audio-video transformer model, TVLT [37], to the STS using data from n=6 subjects from the Courtois Neuromod Dataset [7], while subjects watch the sitcom *Friends*. This significantly increases alignment to both the STS (our tuning target), and a further (non-targetted) lateral-stream ROI.

To evaluate social cognition, we first test whether tuning improves performance in a context similar to the *Friends* training data, and report significantly increased performance on a sarcasm detection task containing data from sitcoms (MUStARD). We then test whether these improvements generalize to a social cognition task in a markedly different context, emotion and sentiment prediction on CMU-MOSEI, but find no significant increase in performance from brain-tuning, suggesting that

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Interpreting Cognition in Deep Learning Models.

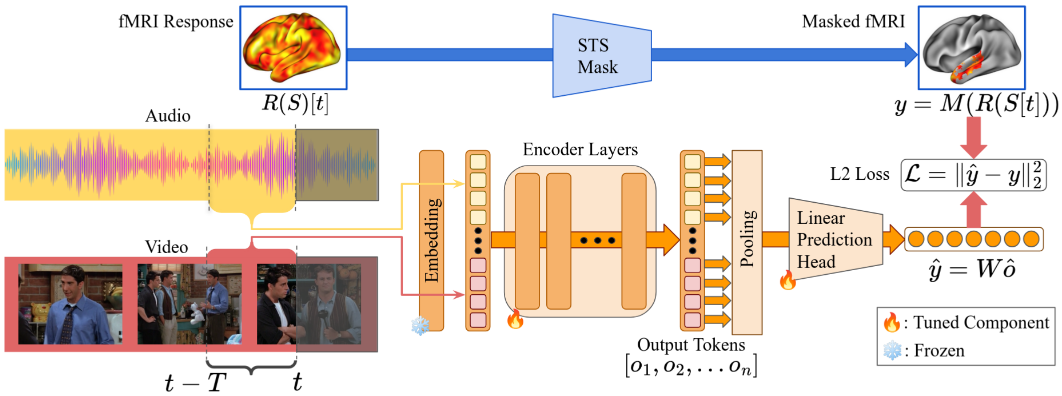


Figure 1: Our audio-video brain-tuning approach. Audio-video stimuli are perceived by the subject, and input to the model, and we fine-tune the model and projection head to better predict corresponding brain activation.

tuning improves social cognition performance in related contexts but does not generalize to contexts not represented in training.

Our main contributions are as follows: We extend the brain-tuning methodology to a multi-modal audio-video domain, and show, for the first time, that brain-tuning a model to an ROI involved in social cognition can improve its performance on a related social cognition task. This provides further evidence [27, 15] that targeted brain tuning to specific functional ROIs can increase alignment and improve performance to related downstream tasks.

2 Related Work

Lateral Stream & Superior Temporal Sulcus. The lateral stream has been recently proposed as a third visual processing stream specialized for dynamic social processing ([32]), in addition to the classical ventral and dorsal streams. Its endpoint, the Superior Temporal Sulcus (STS), robustly encodes features of social interaction allowing for the processing of the intentions and inner states of others. [26, 21, 39, 32, 19, 1, 13]. This motivates our use of STS activity as a tuning signal for social cognitive tasks. See the appendix for a visualization of the STS on the whole brain.

Prior work in Brain Alignment and Brain Tuning. There is a large body of work measuring brain alignment in neural models [30, 31, 14, 28, 24], however, few [27, 15] studies fine-tune a pretrained model to increase alignment. Unlike these prior works [27, 15] which fine-tune audio-only models to late language regions and evaluate on downstream auditory and semantic tasks, we instead tune our multi-modal model to the STS, and evaluate on downstream social cognition tasks. Our work differs from recent multi-modal brain encoder work [12], which trains a dedicated deep network for brain prediction across regions. In contrast, we tune an existing model to a specific functional region, and aim to improve alignment to that region and performance on a related task.

3 Method

3.1 Model and Stimulus

Model Selection. Recent works in Video-Language multimodal models are broadly split into LLM-based methods ([20, 23, 22, 35, 16, 36, 11, 40, 25]) and feature encoder-based methods ([42, 18, 37]). We chose to tune the *Textless Vision Language Transformer* (TVLT) [37], due to architectural similarities with the models brain-tuned in [27] including number of encoders layers (12), embedding size (768), and total number of parameters ($\sim$ 90M). It is pretrained on around 130K hours of audio-video with a joint masked auto-encoding and vision-audio matching objective. An initial embedding layer embeds each 16x16 patch of each video frame, and converts the audio to a log-mel spectrogram, which are then jointly encoded through the transformer layers.

fMRI Data. We use a subset of the preprocessed fMRI data from the 2022-alpha release of the Courtois Neuromod Dataset [7], containing n=6 subjects watching seasons 1-4 of the sitcom *Friends* (seasons 1-3 for training, 4 for evaluation). It is one of the largest available fMRI dataset of participants watching audio-video stimuli, and has previously been used for brain-tuning an audio model [15]. More information about this dataset can be found in the appendix.

Cross Subject Prediction Accuracy Estimation. Noise in the fMRI data–both natural fMRI noise as well as signal unrelated to the stimulus–can impair both our brain-tuning and evaluation procedures. To estimate the level of noise present in each voxel, we follow recent studies [31], [14] in adapting [34]'s method to estimate cross-subject prediction accuracy for each voxel. See A.2 for technical details. Following previous brain-tuning studies [27], we filter out voxels with a low cross-subject prediction accuracy to tune only on voxels reliably related to the stimulus. We attempt to reach the threshold of 0.4 used in prior brain-tuning [27], but find that beyond a threshold of 0.25, all STS voxels are removed for some subjects, preventing training (see appendix A.2). Therefore, we set our threshold to 0.25, leaving subjects with 100-700 STS brain-tuning target voxels. We also use cross-subject prediction accuracy to normalize voxel activations when computing normalized brain alignment (more in section 3.3).

3.2 Brain Tuning

Training Objective. Following [27], we fine-tune our pretrained model to predict the fMRI voxels in the STS with a high cross subject prediction accuracy. Formally, let S be the synchronized audio–video stimulus, and $R(S)[t]$ the recorded fMRI response at time t. We define a voxel masking function M such that:

$$y = M(R(S)[t]),$$

where $y \in \mathbb{R}^m$ is the STS-masked fMRI vector of m voxels. Let T be the length of the temporal receptive field, approximately 12s in our case. We take an audio–video clip from $t - T$ to t, denoted $S[t - T : t]$, and process it with TVLT to obtain output tokens $[o_1, o_2, \ldots, o_n] \in \mathbb{R}^{n \times 768}$. We mean-pool the tokens: $\hat{o} = \frac{1}{n} \sum_{i=1}^{n} o_i$. A linear projection layer $W \in \mathbb{R}^{m \times 768}$ maps $\hat{o}$ to the predicted fMRI vector:

$$\hat{y} = W\hat{o}.$$

We minimize the L2 loss, $\mathcal{L}$, between the predicted voxel activations $\hat{y}$ and true activations y:

$$\mathcal{L} = \|\hat{y} - y\|_2^2,$$

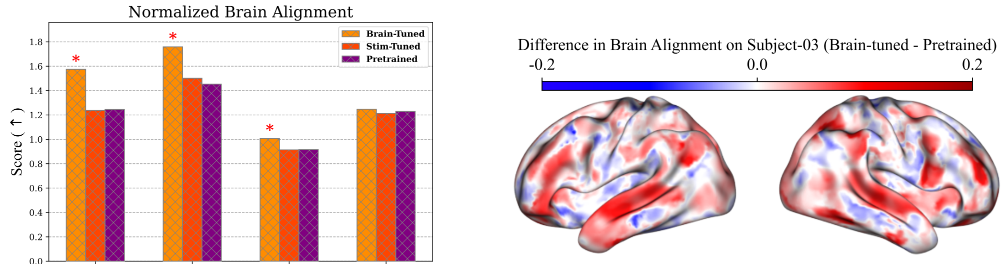
and backpropagate $\mathcal{L}$ through both the projection layer W and the TVLT transformer layers. The overall process is illustrated in fig. 1.

Training Details. To predict each fMRI snapshot, we give the model the previous 8 TR-lengths ($T = 11.92$s) of audio-video stimulus. This finite response window is similar to that used in prior work [27], and is in line with the average hemodynamic response cycle of 12s [38]. Following [15], we train our model on the first three seasons (68,063 TRs, TR=1.49s) of *Friends*, and evaluate on season four. Following the finding by [15] that individual models often outperformed models tuned to multiple subjects at once, we tune one model to each subject's (n=6) brain activity. Due to compute limits, we restrict our tuning to 10 epochs. For each 11.92s clip, we evenly sample 8 frames from the video following [37], and sample audio at the standard 44,100 Hz. We optimize with Adam with a constant learning rate of 1.0×10^{-6}. Brain-tuning each model uses 1 H100 GPU and 16 AMD EPYC 9654 CPUs on 244 GB of RAM, and takes approximately 70 hours on an H100 GPU. Each evaluation uses identical compute specs, and takes approximately 90 minutes.

3.3 Evaluation Procedure

Comparison Models. Following [27], we compare against a stimulus-tuned and a pretrained baseline on both brain-encoding and downstream evaluations. The pretrained baseline is the original pretrained TVLT model introduced in [37]. The stimulus-tuned baseline is trained using the original TVLT joint training objective, with the same video data and learning hyperparameters as the brain-tuned model. This baseline tests whether changes in performance are the result of simply training on the Friends dataset, or are due to the fMRI training objective used in brain-tuning.

Encoding Evaluation. Following [27], we use standard voxel-wise encoding models ([2] [31], [30]) to evaluate the change in brain alignment between our brain-tuned and baseline models. We follow



a) b)

Figure 2: **a:** Average change in alignment to lateral ROIs after brain-tuning over subjects. We find significant increases in the pSTS, aSTS, and LOC. **b:** Change in alignment before and after tuning on Subject-03. Differences for all subjects can be found in the appendix.

the same steps as during brain tuning to create TR-video pairs where each fMRI TR is paired with the 8 TR-lengths (11.92s) of video that precede it. This video is input into each model, and a voxel wise ridge regression model is learned to predict the fMRI activations for that TR, from the concatenation of the [CLS] and mean pooled output tokens. For training and testing, we use data from season 4 of Friends which was unseen during brain-tuning, using 8298 TRs to train and 2630 to test.

Normalized Brain Alignment. Following [27, 31] prediction performance of this encoding model on the test data is computed by voxel-wise Pearson correlation between the predicted fMRI activations, and the corresponding real brain responses. To account for different levels of noise between voxels, this voxel-wise correlation is then divided by the voxel-wise cross subject prediction accuracy, and averaged across all voxels in each ROI to provide a standardized measure for alignment between the model and different ROIs. We report normalized brain-alignment scores for two subdivions of the STS–the anterior STS (aSTS), and posterior STS (pSTS), as well as to two adjacent ROIs in the lateral stream (LOC, EBA). For each subject, we visualize the difference in normalized alignment between our brain-tuned models and pretrained (brain-tuned - pretrained) over the entire brain surface. Following [27], to test whether the brain-tuned models have significantly improved alignment to an ROI compared to our baselines, for each baseline we perform a wilcoxon signed rank test over the alignment of our brain-tuned models compared to the baseline models' alignment. We indicate significant differences ($p < 0.05$) with an asterisk *.

3.4 Downstream Evaluation

Sarcasm Detection. We first evaluate our brain-tuned and baseline models on MUStARD [10], an audio-video sarcasm detection database consisting of clips from various sitcoms. Because our models are brain-tuned to stimulus from a sitcom, this measures how our model's performance changes on a social cognition task with stimuli similar to the stimulus seen during brain-tuning. Each clip contains an utterance, accompanied by conversational context and is labeled for the sarcasm of the utterance. Because some MUStARD clips are from Friends, we train and test our classifier separately on both the full dataset, and a subset of the dataset with all Friends clips removed. Due to the small size of the dataset, models are evaluated on their mean performance across 10-fold cross validation.

Sentiment and Emotion Detection. To probe social cognition on our baseline and brain-tuned models' in a task markedly different from the *Friends* training data, we evaluate on CMU-MOSEI sentiment and emotion prediction [4], a dataset containing clips of people speaking into the camera from YouTube, and manually labeled for scalar sentiment, and the presence of each of six emotions (happy, sad, anger, surprise, disgust, fear). We use the original 15,288/4,830 train-test split provided by the original TVLT paper [37].

Evaluation Protocol. For both tasks, we a train a linear binary classifier on a concatenation of the [CLS] token and mean pooled tokens from the last layer. Since we brain-tune models through mean pooled tokens, but pretrained TVLT typically probes its [CLS] token for classification tasks, we concatenate both to fairly compare to baselines. We report A2 accuracy and F1 score for our binary classification tasks (sentiment, sarcasm), and weighted A2 accuracy and F1 score for emotion, averaged across n=6 for our brain-tuned models. We use a one-sided one sample t-test over the change in performance of our n=6 subject models compared to each baseline to test for significance,

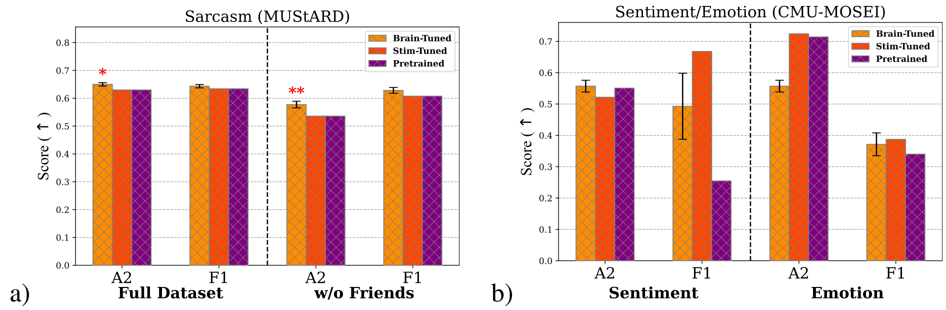


Figure 3: Brain-tuned and baseline performance on downstream social perception benchmarks. We find significant improvements on MUSTtARD A2 scores both including *Friends* clips ($p < 0.05$) and omitting them ($p < 0.01$).

indicating significant improvements (p < 0.05) with an asterisk *, and highly significant improvements (p < 0.01) with a double asterisk ** in our graphs. Error bars report SEM across n=6 brain-tuned models.

4 Results

Brain Alignment Results. We plot the change in alignment compared to the pretrained model (brain-tuned - pretrained) over the entire cortex for subject 03 in fig. 2, with other subjects plotted in fig. 5. Using cross-subject prediction accuracy underestimates the true noise ceiling, as some biological signal that varies between subjects is treated as noise. This leads to some normalized brain alignment scores above 1.0 for baseline and brain-tuned models, but their relative performance is unaffected by this scaling. Compared to both pretrained and stimulus tuned baselines, our n=6 brain-tuned models show significant improvements to brain alignment across various lateral stream ROIs (fig. 2). We report significantly increased alignment (p<0.05) to both subregions (aSTS, pSTS) of the STS (tuning target), and to one of two neighboring ROIs in the lateral stream (LOC). We observe no significant changes in alignment between our pretrained and stimulus-tuned baselines, confirming that increased brain-alignment in our brain-tuned models is not merely due to stimulus exposure.

Downstream Tasks Results. Our brain-tuned models significantly outperform baselines on both the full MUStARD sarcasm detection dataset (p<0.05), as well as the dataset after removing all *Friends* clips (p<0.01) (fig. 3a). In contrast, we observe no improvements or decreased performance on the sentiment and emotion prediction task (CMU-MOSEI). These results suggest our model improves performance on a social cognition task similar to the training stimulus (MUStARD), but that these increases do not generalize to a markedly different context (CMU-MOSEI). In the appendix, we break down our emotion classification results by individual emotion.

5 Conclusion

Our findings demonstrate that brain-tuning a multimodal audio-video model to a social cognition region (STS) not only increases alignment to the target area but also extends improved alignment to an adjacent lateral stream ROI. This increased alignment is accompanied by significant gains on a related social cognition task when the evaluation context resembles the training stimulus, sarcasm detection in sitcoms. However, these gains do not generalize to sentiment and emotion prediction in markedly different contexts, suggesting a limitation in the transferability of brain-tuning effects to contexts unseen during training. While our study was limited to a single model and a small number of evaluations, the results serve as a proof of concept for targeted brain-tuning as a means to enhance both brain alignment and task performance in relevant domains. We suggest future researchers experiment with larger LLM based multi-modal architectures, as well as more diverse evaluation and training datasets.

6 Funding and Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 “Neuroexplicit Models of Language, Vision, and Action” - project number 471607914.

The Courtois project on neural modelling was made possible by a generous donation from the Courtois foundation, administered by the Fondation Institut Gériatrie Montréal at CIUSSS du Centre-Sud-de-l’île-de-Montréal and University of Montreal. The Courtois NeuroMod team is based at “Centre de Recherche de l’Institut Universitaire de Gériatrie de Montréal”, with several other institutions involved. See the cneuromod documentation for an up-to-date list of contributors (`https://docs.cneuromod.ca`).

The authors gratefully acknowledge support from the DAAD RISE (Research Internships in Science and Engineering) program. The authors greatfully acknowledge support from MPI for Software Systems and MPI for Informatics. The authors thank Omer Moussa, Dota Dong, and Subba Oota Reddy for their helpful advice and guidance during the development of this study.

References

[1] Truett Allison, Aina Puce, and Gregory McCarthy. Social perception from visual cues: role of the sts region. *Trends in cognitive sciences*, 4(7):267–278, 2000.

[2] 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.

[3] Dahye Bae and Jun Soo Kwon. Pt735. multisensory integration of social interaction. *International Journal of Neuropsychopharmacology*, 19(Suppl 1):67, 2016.

[4] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, 2018. Association for Computational Linguistics.

[5] Stefania Benetti, Ambra Ferrari, and Francesco Pavani. Multimodal processing in face-to-face interactions: A bridging link between psycholinguistics and sensory neuroscience. *Frontiers in Human Neuroscience*, 17:1108354, 2023.

[6] Samuele Bolotta and Guillaume Dumas. Social neuro ai: Social interaction as the “dark matter” of ai. *Frontiers in computer science*, 4:846440, 2022.

[7] Julie A Boyle, Basile Pinsard, Amal Boukhdhir, Sylvie Belleville, Simona Brambatti, Jeni Chen, Julien Cohen-Adad, André Cyr, Pierre Rainville, Pierre Bellec, et al. The courtois project on neuronal modelling-first data release. In *26th Annual Meeting of the Organization for Human Brain Mapping (OHBM)*. Organization for Human Brain Mapping (OHBM), 2020.

[8] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey, 2024.

[9] Salvatore Campanella and Pascal Belin. Integrating face and voice in person perception. *Trends in cognitive sciences*, 11(12):535–543, 2007.

[10] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy, 2019. Association for Computational Linguistics.

[11] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms, 2024.

[12] Stéphane d’Ascoli, Jérémy Rapin, Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Tribe: Trimodal brain encoder for whole-brain fmri response prediction, 2025.

[13] Ben Deen, Kami Koldewyn, Nancy Kanwisher, and Rebecca Saxe. Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral cortex*, 25(11):4596–4609, 2015.

[14] Dota Tianai Dong and Mariya Toneva. Vision-language integration in multimodal video transformers (partially) aligns with the brain, 2023.

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

[16] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, Haoyu Cao, Di Yin, Long Ma, Xiawu Zheng, Rongrong Ji, Yunsheng Wu, Ran He, Caifeng Shan, and Xing Sun. Vita: Towards open-source interactive omni multimodal llm, 2025.

[17] Kathy Garcia, Emalie McMahon, Colin Conwell, Michael Bonner, and Leyla Isik. Modeling dynamic social vision highlights gaps between deep learning and humans. In *The Thirteenth International Conference on Learning Representations*, 2025.

[18] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

[19] Grit Hein and Robert T Knight. Superior temporal sulcus—it's my area: Or is it? *Journal of Cognitive Neuroscience*, 20(12):2125–2136, 2008.

[20] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024.

[21] Leyla Isik, Kami Koldewyn, David Beeler, and Nancy Kanwisher. Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 114(43):E9145–E9152, 2017.

[22] Yadong Li et al. Baichuan-omni-1.5 technical report, 2025.

[23] Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model, 2025.

[24] Haoyu Lu, Qiongyi Zhou, Nanyi Fei, Zhiwu Lu, Mingyu Ding, Jingyuan Wen, Changde Du, Xin Zhao, Hao Sun, Huiguang He, and Ji-Rong Wen. Multimodal foundation models are better simulators of the human brain, 2022.

[25] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration, 2023.

[26] Haemy Lee Masson and Leyla Isik. Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage*, 245:118741, 2021.

[27] Omer Moussa, Dietrich Klakow, and Mariya Toneva. Improving semantic understanding in speech language models via brain-tuning, 2025.

[28] Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Q. Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. Unveiling multi-level and multi-modal semantic representations in the human brain using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20313–20338, Miami, Florida, USA, 2024. Association for Computational Linguistics.

[29] Anuja Negi, Subba Reddy Oota, Manish Gupta, and Fatma Deniz. Brain-informed fine-tuning for improved multilingual understanding in language models. *bioRxiv*, pages 2025–07, 2025.

[30] Subba Reddy Oota, Akshett Jindal, Ishani Mondal, Khushbu Pahwa, Satya Sai Srinath Namburi, Manish Shrivastava, Maneesh Singh, Bapi S Raju, and Manish Gupta. Correlating instruction-tuning (in multimodal models) with vision-language processing (in the brain). *arXiv preprint arXiv:2505.20029*, 2025.

[31] Subba Reddy Oota, Khushbu Pahwa, Mounika Marreddy, Maneesh Singh, Manish Gupta, and Bapi S. Raju. Multi-modal brain encoding models for multi-modal stimuli, 2025.

[32] David Pitcher and Leslie G Ungerleider. Evidence for a third visual pathway specialized for social perception. *Trends in cognitive sciences*, 25(2):100–110, 2021.

[33] Ilana Porter, Bar Galam, and Roni Ramon-Gonen. Emotion detection and its influence on popularity in a social network-based on the american TV series friends. *Soc. Netw. Anal. Min.*, 13(1), 2023.

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

[35] Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding, 2023.

[36] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models, 2024.

[37] Zineng Tang, Jaemin Cho, Yixin Nie, and Mohit Bansal. Tvlt: Textless vision-language transformer. In *NeurIPS*, 2022.

[38] Michelle Voss. *The Chronic Exercise–Cognition Interaction*, pages 187–209. Elsevier Academic Press, 2016.

[39] Jon Walbrin, Paul Downing, and Kami Koldewyn. Neural responses to visually observed social interactions. *Neuropsychologia*, 112:31–39, 2018.

[40] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025.

[41] Yuan Yuan, Zhaojian Li, and Bin Zhao. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*, 57, 2025.

[42] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound, 2022.

A Appendix

This appendix contains the following sections:

In appendix A.1, we provide additional information on the subjects used in the fMRI data collection.

In appendix A.2, we visualize the effect of the accuracy threshold on the number of voxels we can brain-tune on.

In appendix A.3, we display the changes in brain alignment for all subjects.

In appendix A.4, we expand on the CMU MOSEI evaluation results across all emotions.

In appendix A.5, we show the STS region mask.

In appendix A.6, we discuss the potential broader impacts of our work.

In appendix A.7, we provide information on how to access our code and relevant data.

In appendix A.8, we include all licensing information.

A.1 Participants

Six healthy participants (aged 31 to 47 years at the time of recruitment in 2018), three women (sub-03, sub-04, and sub-06) and three men (sub-01, sub-02, and sub-05) were recruited to participate in the Courtois Neuromod Project for at least 5 years. All subjects provided informed consent to participate in this study, which was approved by the ethics review board of the "CIUSS du centre-sud- de- l'île- de- Montréal" (under number CER VN 18-19-22). Three of the participants reported being native franco- phone speakers (sub- 01, sub-02, and sub-04), one as being a native anglophone (sub-06), and two as bilingual native speakers (sub-03 and sub-05). All participants reported the right hand as being their dominant hand and reported being in good general health. Exclusion criteria included visual or auditory impairments that would prevent participants from seeing and/or hearing stimuli in the scanner and major psychiatric or neurological problems. Standard exclusion criteria for MRI and MEG were also applied. Lastly, given that all stimuli and instructions are presented in English, all participants had to report having an advanced comprehension of the English language for inclusion. The above boilerplate text is taken from the cNeuroMod documentation [7], with the express intention that users should copy and paste this text into their manuscripts unchanged. It was released by the Courtois NeuroMod team under the CC0 license. For more details regarding fMRI acquisition, stimuli presentation

A.2 Cross Subject Prediction Accuracy Calculation

We follow recent studies [31], [14] in adapting [34]'s method to estimate cross-subject prediction accuracy for each voxel. For each subject, we generate all possible subsets of the remaining 5 subjects, and for each subset we use a voxel-wise encoding model (see Sec. 5) to predict one participant's response from the others. As in previous studies [14, 31], the final value is calculated as an average at the group level. These cross-subject encoding models are trained using nine episodes (7700 TRs) from the first season of Friends, and tested on three other episodes from the same season (2872 TRs). In fig. 4, we display the number of viable voxels based on the cross subject prediction accuracy threshold. We observe that Subject-05 has no remaining voxels above 0.25, and thus deem this as our cut-off.

A.3 Differences in Normalized Brain Alignment For all Subjects

In fig. 5, we display the differences in normalized brain alignment for all subjects. Most subject models show improved alignment in and around the STS, but these improvements do not consistently extend to other regions.

A.4 CSU MOSEI Complete Emotion

In fig. 6, we breakdown the performance of the model across each emotion aggregated in the CMU MOSEI evaluation (See fig. 3, rightmost chart). Notably, sadness is the only emotion with significantly improved F1 score - however, we also observe decreased accuracy (A2). Although sadness occurs in

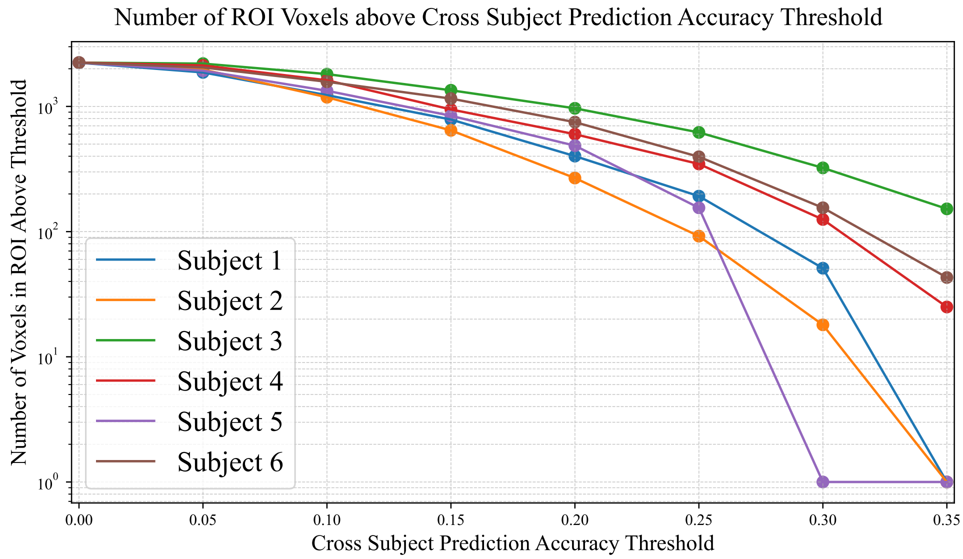


Figure 4: A subject (Subject 5) has no voxels in the STS above a cross subject prediction accuracy threshold of 0.25, and thus we cannot perform brain-tuning.

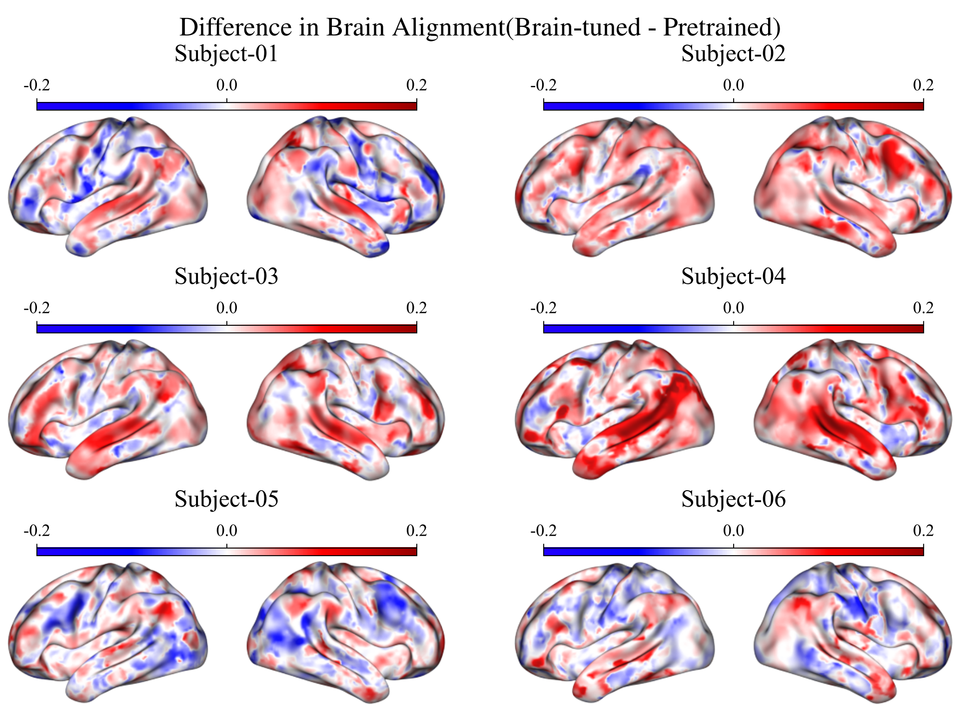


Figure 5: Differences in Normalized Brain Alignment before and after brain-tuning.

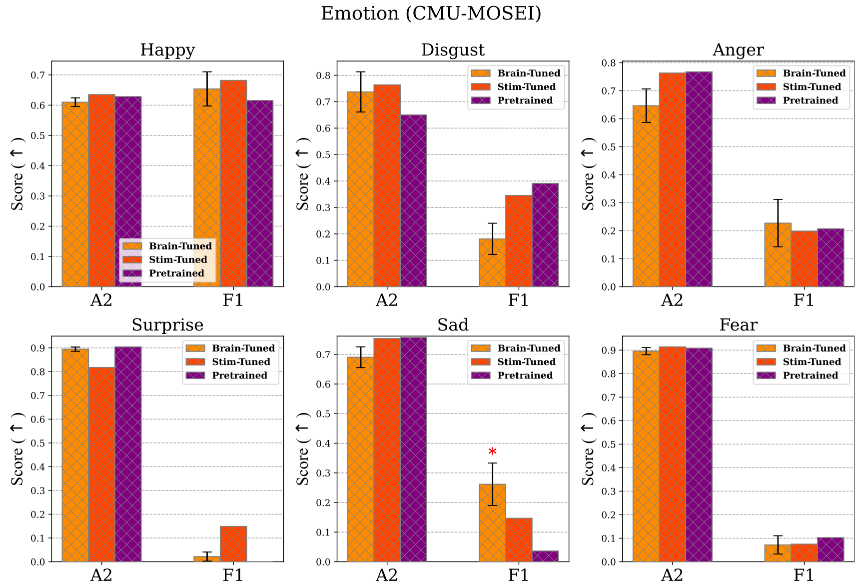


Figure 6: A breakdown of the performance of the model across each emotion aggregated in the CSU MOSEI evaluation (See fig. 3, rightmost chart).

Friends, it is it not the dominant emotion in the show [33]. In future work, we hope to investigate this finding further for an explanation.

A.5 Visualization STS Region

We display the voxel mask of the STS region which we tune our model to in fig. 7.

A.6 Broader Impacts

Our work can be an initial step towards creating AI models with better understanding of human social cognition using brain activity as a tuning target. This could have positive impacts, such as improving AI-human communication or potential uses in AI-assisted therapy. Further, an AI model which can replicate human social cognition may be a useful in-silico model helping us understand social cognition in humans. On the other hand, this could enhance the abilities of AI for human manipulation. We urge future researchers to consider these pros and cons as they continue investigating this topic.

A.7 Data and Code Availability

The fMRI data used to perform the brain tuning are openly available through registered access at link `https://www.cneuromod.ca/access/access/`.

To get our code, and for exact instructions on how to replicate or results, please visit `https://huggingface.co/AnonymousSubmission43/mmbt`, download and unzip all files, and follow the instructions on the README.md in mmbt-anon.

Due to privacy concerns, we do not release model weights or cross subject prediction accuracies, as these are derived from subjects' brain data.

A.8 Licenses

Method Diagram: Our method diagram in fig. 1 includes an audio sound wave, licensed under the public domain. You can find the url here: `https://www.pngfind.com/maxpin/bwRwR/`

Models:

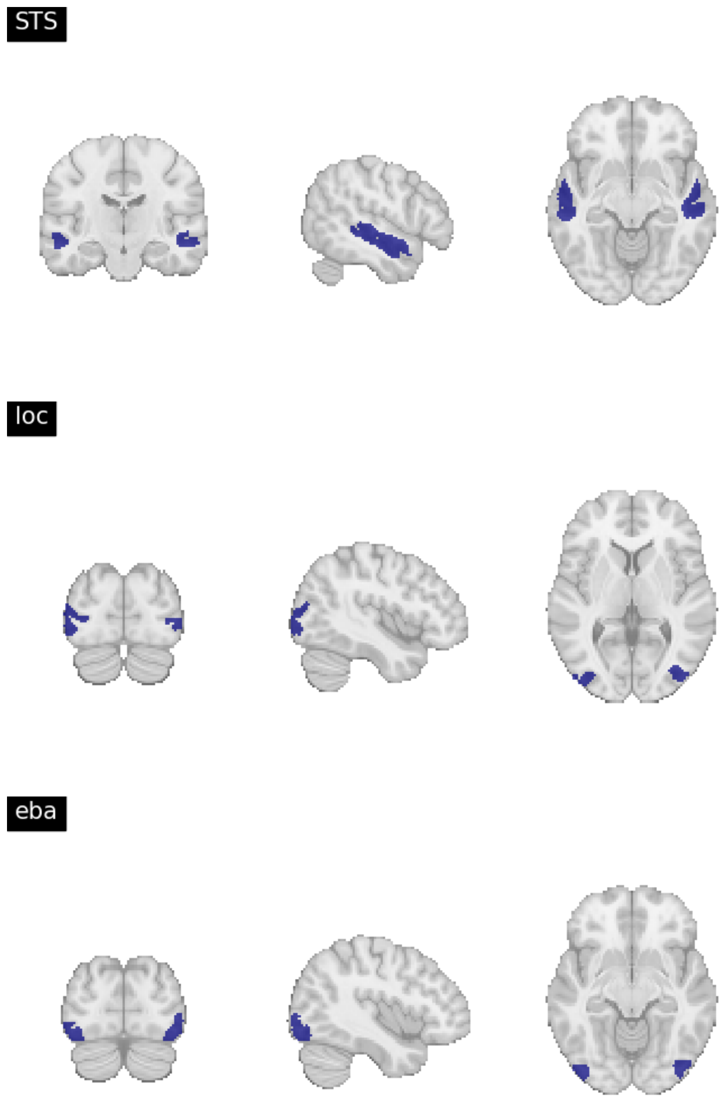


Figure 7: The STS, LOC, and EBA regions from coronal (left), lateral (middle), and horizontal (right) views.

- TVLT: MIT https://github.com/zinengtang/TVLT/blob/main/LICENSE

Packages:

- nilearn: BSD https://github.com/nilearn/nilearn/blob/main/LICENSE
- matplotlib: BSD https://github.com/nilearn/nilearn/blob/main/LICENSE

Datasets:

- CMU-MOSEI: MIT https://github.com/CMU-MultiComp-Lab/CMU-MultimodalSDK?tab=MIT-1-ov-file
- MUSTARD: MIThttps://github.com/soujanyaporia/MUStARD/blob/master/LICENSE
- Courtois NeuroMod: CC0 https://creativecommons.org/publicdomain/zero/1.0/legalcode

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and intro, we make claims which are backed up by our experimental results - we claim increases to alignment and improvements to one of two social perception tasks, which accurately reflect the results we find through statistical testing.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, see conclusion (section 5) for a discussion of the studies limitations (we use single model and a small number of evaluations).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We do not include any theoretical results in our paper that require justification.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In the methods section (section 3), we fully describe our tuning architecture, the hyper parameters used during training, the specific datasets we use, how we calculate the cross subject prediction accuracies, and make all code publicly available. We do not release any weights or cross subject prediction accuracies derived from fMRI due to privacy concerns, but you can reproduce our experiments following the instructions in our paper and our code. See appendix A.7.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide open access to our code, with instructions describing how to run the main experiments in the documentation, see appendix A.7.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In section 3.2 we provide details on training and testing, including the data splits, optimizer type, the choice of hyperparameters, and how they were chosen. Furthermore, one can view our provided code documentation for these settings, see appendix A.7.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We carry out significance testing as described in the methods (section 3) section of the paper, for both alignment and downstream tasks. For our downstream tasks (fig. 3), we also include error bars on the brain-tuned models. These are calculated using scipy.stats.sem on the metric scores across our 6 brain-tuned models. We do not include error bars in our brain-alignment plots (section 3.2) - as all values are paired within participants, conventional error bars for independent samples are not applicable. We perform a paired (wilcoxon) statistical test and indicate significance in the figure instead.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: At the end of section 3.2, we provide these details (Each brain-tuning uses 1 H100 GPU and 16 AMD EPYC 9654 CPUs on 244 GB of RAM, and takes approximately 70 hours on an H100 GPU. Each evaluation uses the same compute resources, and takes approximately 90 minutes.)

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We do not collect any data involving human participants, nor publish any novel dataset, nor expose any personal information. The CNeuromod dataset involving human subjects was collected with the approval the review board "CIUSS du centre-sud-de-l'île-de-Montréal" (under number CER VN 18-19-22).

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In appendix A.6, we discuss potential broader impacts of our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks (no novel data or models being released that have a possibility of misuse)

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In appendix A.8 cite all existing assets used with their respective version.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release our code, alongside details about training and evaluation of our brain-tuned model see appendix A.7.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We include details of stimulus presentation as is relevant to the project (subjects watched Friends half-episodes), and don't include more detail as we feel it is not necessary to understand our method or results. Subject details can be found at appendix A.1. Compensation information is not provided in the original cneuromod dataset. Further details can be found in the cneuromod documentation at `https://docs.cneuromod.ca/en/2020-alpha2/MRI.html#stimuli`

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The study did not itself collect any data from human subjects, but uses the cneuromod 2022-alpha release of the Friends dataset. All subjects provided informed consent to participate in the cneuromod data release, which was approved by the ethics review board of the "CIUSS du centre-sud-de-l'île-de-Montréal" (under number CER VN 18-19-22). For more details, see the cneuromod documentation: `https://docs.cneuromod.ca/en/latest/DATASETS.html#friends`

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Our research does not involve any LLMs as any important, original or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.