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: Foundation Models for the Brain and Body.

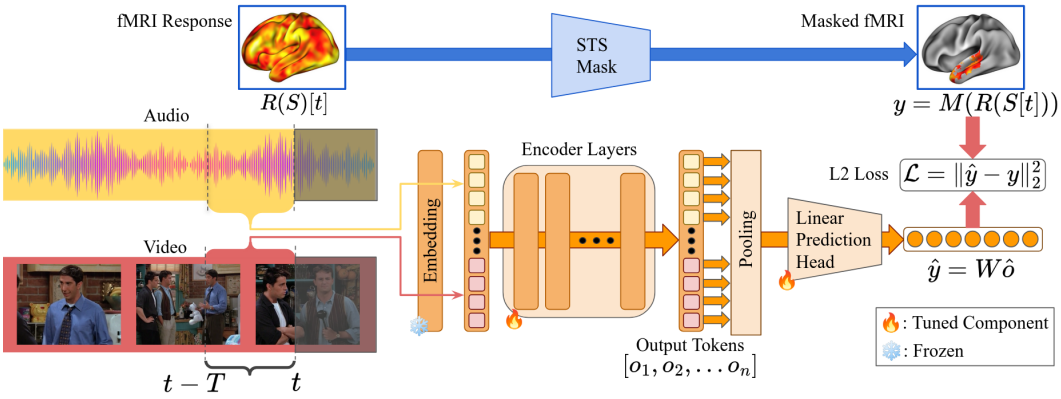

Figure 1: Our audio-video brain-tuning approach. Audio-video stimuli are perceived by the subject, and input to the model, and we fine-tune the model and projection head to better predict corresponding brain activation.

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

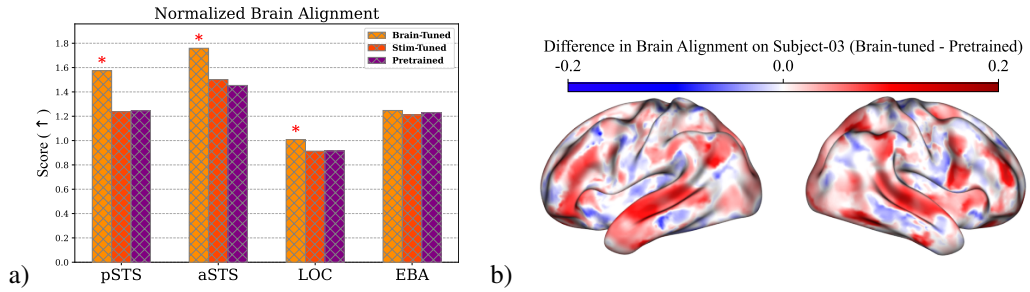

Figure 2: **a:** Average change in alignment to lateral ROIs after brain-tuning over subjects. We find significant increases in the pSTS, aSTS, and LOC. **b:** Change in alignment before and after tuning on Subject-03. Differences for all subjects can be found in the appendix.

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,

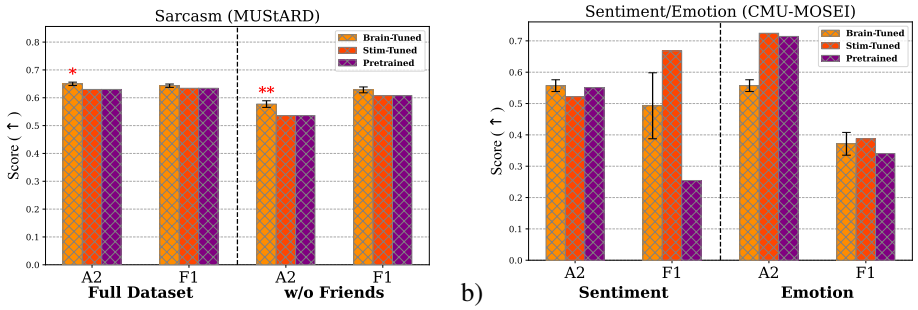

Figure 3: Brain-tuned and baseline performance on downstream social perception benchmarks. We find significant improvements on MUSTtARD A2 scores both including *Friends* clips ($p < 0.05$) and omitting them ($p < 0.01$).

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.