MODALITY-AGNOSTIC BRAIN-TUNING OF TEXT AND SPEECH REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The human brain has largely shared higher-level language representations during reading and listening, despite differences in early sensory processing. Recent "brain-tuning" studies have shown that fine-tuning language models with brain data can improve their semantic representations, but these studies have focused on a single modality. Here we ask whether brain-tuning can help multimodal models learn brain-relevant representations that are shared across modalities. We brain-tune two multimodal text-speech models using fMRI data collected during naturalistic reading and listening. We then evaluate how well the brain-tuned model encodes brain activity, both within and across modalities. We further evaluate the brain-tuned model's representation on several semantic and audio probing tasks. All analyses are conducted across model layers, on held-out participants. Our results show that brain-tuning improves both within- and cross-modality prediction accuracy across language-selective brain regions. These gains emerge primarily in middle and late layers, where brain-tuning also strengthens semantic and audio information. Notably, tuning with both modalities jointly yields greater improvements, for both encoding performance and probing tasks, than tuning with either modality alone. Together, these findings suggest that brain-tuning can strengthen brain-relevant language information shared across modalities, providing a way to ground multimodal language models in human brain representations.

1 INTRODUCTION

Recently, "brain-tuning" has emerged as a way to ground model representations in human brain representations by fine-tuning them directly with brain data(Schwartz et al., 2019). Brain-tuned models show improved brain alignmentVattikonda et al., 2025; Moussa et al., 2025; Negi et al., 2025, generalization to unseen participants (Negi et al., 2025; Moussa & Toneva, 2025b), improved semantic representations (Moussa et al., 2025; Vattikonda et al., 2025; Negi et al., 2025), and improved downstream task performance (Negi et al., 2025). However, this approach has so far been applied within a single modality: speech models have been brain-tuned with brain data during a listening task, and text models with brain data during a reading task. Neuroimaging evidence has shown that semantic representations during language comprehension are modality-invariant in the human brain. (Deniz et al., 2019) showed that, despite differences in early sensory processing, semantic representations are largely shared in language-selective regions during naturalistic reading and listening. Given that semantic representations in the human brain are largely modality-invariant, this raises a natural question: can brain-tuning be extended to ground language model representations in brain-relevant modality-invariant representations?

In this study, we use brain data collected while the same participants read and listened to the same naturalistic stories (Deniz et al., 2019) and investigate whether brain-tuning helps multimodal models improve brain-relevant semantic representations shared across reading and listening. We first develop an end-to-end, modality-agnostic brain-tuning pipeline that uses reading and listening brain responses to update a multimodal model's text and speech representations. We brain-tune two multimodal models, SONAR (Duquenne et al., 2023) and CLAP (Elizalde et al., 2023), which encode text and speech separately and map them into a shared representation space. We evaluate the brain-tuned models along two main axes. First, to test whether the learned brain-relevant information is modality-invariant, we evaluate encoding performance both within- and cross- modalities. Specifically, whether brain-tuning with reading brain responses improves listening-brain alignment, and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

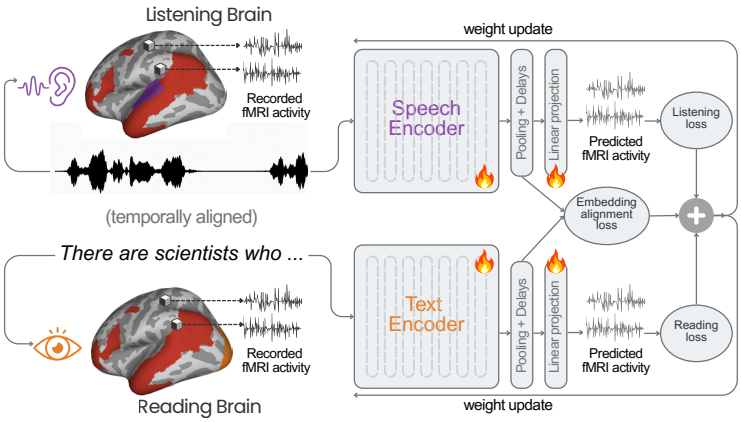

Figure 1: **Schematic of the modality-agnostic brain-tuning pipeline.** Participants read or listened to naturalistic stories in English while fMRI responses were recorded. The temporally aligned text transcript and speech waveform were passed through the text and speech encoders, respectively, of a multimodal language model. Representations from the final token of the last hidden layer were pooled and temporally delayed to match the fMRI acquisition. A linear layer then projected these representations to voxel space to predict brain activity for the corresponding modality (reading or listening). The prediction loss for each modality was summed with an embedding-space alignment loss between the text and speech encoders, and backpropagated through all layers of both encoders for full brain-tuning.

whether brain-tuning with listening brain responses improves reading-brain alignment. Second, we evaluate the brain-tuned representations on several semantic and audio probing tasks. We conduct these analyses across model layers and test generalization on held-out participants.

Our results show that brain-tuning improves brain encoding performance both within and across modalities: reading-based brain-tuning improves representations that predict brain activity during listening, and vice versa. These gains emerge primarily in middle and later layers, where brain-tuning also strengthens semantic and audio-related information. Brain-tuning with both modalities jointly yields greater improvements in both encoding performance and probing tasks than brain-tuning with either modality alone. Together, these findings suggest that brain-tuning can induce brain-relevant, modality-invariant language representations within a single multimodal model, offering a way to ground multimodal language models in human brain representations. We will release the code upon publication of this paper.

2 Related Work

Brain-tuning of speech- and text-based language models. Our work builds on the recent brain-tuning approach, in which language models are brain-tuned directly with brain data. So far, this has been done using modality-specific models and brain data: text models have been brain-tuned with reading-evoked brain responses (Schwartz et al., 2019; Negi et al., 2025), and speech models with listening-evoked brain responses (Moussa et al., 2025; Vattikonda et al., 2025; Vaidya et al., 2026). We extend this work by brain-tuning a multimodal text–speech model in a modality-agnostic fashion, using temporally aligned reading and listening brain data.

Brain alignment of text and speech representations. A complementary line of work has examined how well pretrained text- and speech-based models predict brain responses during naturalistic comprehension, using linearized encoding models. Text-based language models predict cortical responses during both reading and listening, particularly in higher-order semantic regions (Wehbe et al., 2014; Jain & Huth, 2018; Toneva & Wehbe, 2019; Caucheteux & King, 2022; Aw & Toneva, 2023; Oota et al., 2024b; Lamarre et al., 2022; Antonello et al., 2024), while self-supervised speech models capture acoustic, phonetic, and linguistic information during listening (Millet et al., 2022; Vaidya et al., 2022; Oota et al., 2023; 2024a). However, these studies have been conducted with unimodal models. We extend this line of work by evaluating a multimodal model's brain alignment within and across modalities, before and after brain-tuning.

3 METHODOLOGY

3.1 NATURALISTIC BRAIN IMAGING DATASET

We used a publicly available dataset in which the 8 participants listened to and read the same 11 naturalistic stories from the Moth Radio Hour (Deniz et al., 2019), while their blood oxygen level-dependent (BOLD) responses were collected using fMRI. The timing of the words is temporally matched in both conditions. We used 6 participants for brain-tuning, and held out 2 participants (P2&P9) for evaluation. Each story was 10-15mins long. 7 stories were used for brain-tuning. We held out the remaining 4 stories: 3 for training voxelwise encoding models and 1 reserved for testing.

3.2 MULTIMODAL MODELS

We brain-tuned two pretrained multimodal text-speech models, SONAR (Duquenne et al., 2023) and CLAP (Elizalde et al., 2023). Both models use separate parameterized text and speech encoders, but differ substantially in their backbone architectures and in how the two modalities are aligned during pretraining. In SONAR, the text and speech encoders map into a shared 1024-dimensional representation space via a teacher–student objective. Both encoders are 24-layer Transformers with a hidden dimension of 1024, initialized from NLLB (text) and Wav2Vec2-BERT 2.0 (speech). We also brain-tuned CLAP, an architecturally distinct dual-encoder model, to test whether our results generalize across dual-encoder text-speech backbones rather than being specific to a single model or pretraining objective. More details on the models are in Appendix A.

3.3 MODALITY-AGNOSTIC BRAIN-TUNING

We perform supervised full brain-tuning of two pretrained multimodal models using fMRI BOLD responses recorded during naturalistic reading and listening as training targets. A schematic of the brain-tuning pipeline is shown in 1. We do a multi-participant brain-tuning using fMRI responses from language-selective regions (Fedorenko et al., 2010). fMRI responses for each training participant were projected to the common *fsaverage* surface.

Model architecture. Text transcripts and corresponding speech waveforms from the narrative stimuli are passed as inputs to the text and speech encoders, respectively, of the pretrained multimodal model. For text, we extract representations for the last token from the last hidden layer, for each word, using a sequence length of 20 tokens. For speech, the waveform is split into 500ms target chunks with 16s of preceding acoustic context, and frame-level representations are pooled per chunk. Representations pass through dropout layers ($p = 0.2$) and are downsampled to the fMRI sampling rate via 3-lobe Lanczos interpolation. To model the delayed hemodynamic response, we learn a finite impulse response (FIR) filter by concatenating representations delayed by 2, 4, 6, and 8 seconds. Separate linear heads then predict reading- and listening-evoked BOLD responses in *fsaverage* space. This brain-tuning architecture accommodates the different temporal resolutions of text and speech and allows the brain-tuning loss to backpropagate end-to-end through the prediction heads and encoders.

Training protocol. We fully brain-tune all encoder layers together with their corresponding prediction heads, using 8-bit AdamW (Loshchilov & Hutter, 2017) (learning rate 10^{-4}, weight decay 10^{-3}) with mixed precision and gradient clipping. Text and speech batch sizes are eight and four, respectively. We train on six stories, validate on a seventh, and train for a maximum of 20 epochs. A ReduceLROnPlateau scheduler reduces the learning rate by a factor of 0.5 after two consecutive validation epochs without improvement. If either the reading or listening validation loss fails to improve for five consecutive epochs, training is stopped. CLAP-specific training details are in Appendix D.

Training objective. The joint training objective combines the brain-prediction loss with the embedding-space alignment loss:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{brain}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}, \qquad \lambda_{\text{align}} = 0.1. \tag{1}$$

The *brain-prediction loss* compares the predicted and actual BOLD response, for each modality (read/listen), for each of the $S = 6$ training participants separately:

$$\mathcal{L}_{\text{brain}} = \frac{1}{S}\sum_{s=1}^{S}\Big[\lambda_{\text{text}}\,\ell\big(\widehat{Y}^{\text{read}}, Y_s^{\text{read}}\big) + \lambda_{\text{speech}}\,\ell\big(\widehat{Y}^{\text{listen}}, Y_s^{\text{listen}}\big)\Big], \qquad \lambda_{\text{text}} = \lambda_{\text{speech}} = 1, \tag{2}$$

where ℓ is the negative Pearson correlation between predicted($\widehat{Y}$) and actual(Y) BOLD response, computed across time for each vertex in and averaged over the V vertices:

$$\ell(\widehat{Y}, Y) = -\frac{1}{V}\sum_{v=1}^{V}\text{corr}_t\big(\widehat{Y}_v,\, Y_v\big). \tag{3}$$

The *embedding-space alignment loss* helps align the representations of the modality-specific encoders. We instantiate it according to each backbone's pretraining objective. For SONAR, we align the TR-matched (after 3-lobe Lanczos downsampling and before FIR delays) shared 1024-dimensional embedding space (D) using mean-squared error:

$$\mathcal{L}_{\text{align}}^{\text{SONAR}} = \frac{1}{|\mathcal{T}_{\text{train}}|D}\sum_{t\in\mathcal{T}_{\text{train}}}\sum_{d=1}^{D}\Big(H_{td}^{\text{text}} - H_{td}^{\text{speech}}\Big)^2, \qquad D = 1024. \tag{4}$$

Here, $\mathcal{T}_{\text{train}}$ is the set of training TRs. $H^{\text{text}}, H^{\text{speech}} \in \mathbb{R}^{|\mathcal{T}_{\text{train}}|\times D}$ are the temporally aligned text and speech representations after Lanczos downsampling, with $H_{t_d}^{\text{text}}$ and $H_{t_d}^{\text{speech}}$ denoting their d-th feature dimension at training TR t. For CLAP, this alignment loss was based on symmetric NT-Xent.

Brain-tuning conditions. We brain-tune under four conditions: (1) joint brain-tuning, in which both encoders are optimized simultaneously using the full objective above with reading and listening data; (2) text-only brain-tuning, in which only the text encoder and reading-prediction head are optimized (speech pathway frozen, alignment loss disabled); (3) speech-only brain-tuning, analogous to (2); and (4) a shuffled baseline control, in which joint/modality-specific brain-tuning is performed using shuffled fMRI responses (to remove the temporal correspondence between stimuli and brain responses). All conditions otherwise share identical training settings.

4 Evaluation

4.1 Voxelwise Encoding Modeling

We evaluate whether brain-tuning improves the alignment between model representations and brain activity using voxelwise encoding models (VEMs). These estimate a linear mapping from an embedding to the recorded fMRI responses for each voxel (Huth et al., 2016; Deniz et al., 2019). We evaluate this alignment (i.e., encoding performance) both within the modality used for brain-tuning and across modalities: text representations predict reading and listening responses, and speech representations predict listening and reading responses. We use hidden representations from each model layer as embeddings. Embeddings and fMRI responses are z-scored for each story, downsampled via 3-lobe Lanczos interpolation, and delayed via an FIR filter (2, 4, 6, 8 seconds). Ridge regression then maps the embedding to each voxel's BOLD response (Wu et al., 2006; Naselaris et al., 2011), with the regularization parameter selected via 5-fold cross-validation. 3 stories are used for training VEM. Encoding performance is the Pearson correlation (r) between predicted and recorded responses on the test story, computed separately for each voxel, layer, held-out participant, modality, encoder, and brain-tuning condition.

4.2 Downstream probing tasks

To evaluate how brain-tuning changes the models' representations, we probe the encoders on several semantic and audio tasks before and after brain-tuning. For text, we use the 6 SentEval semantic tasks (Conneau et al., 2018) and 14 FlashHolmes tasks (Waldis et al., 2024): 7 world-knowledge tasks from EWoK (Ivanova et al., 2025) (physical, material, social, agent, and quantitative relations) and 7 linguistic tasks (negation, speculation, antonymy) from BioScope (Vincze et al., 2008) and

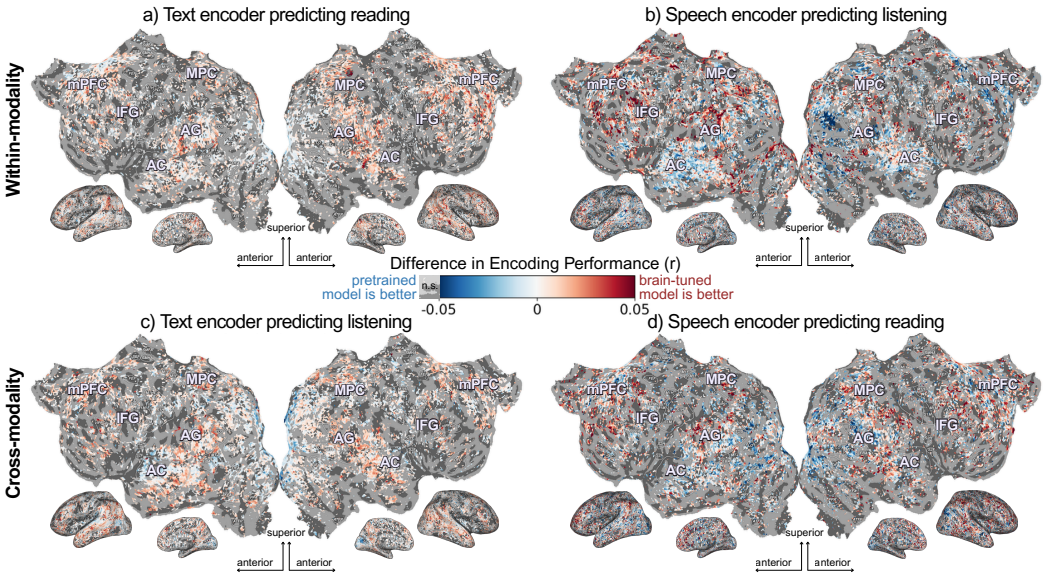

Figure 2: **Brain encoding performance before and after modality-agnostic brain-tuning.** Difference in voxelwise encoding performance between the joint brain-tuned and pretrained SONAR model, using the best layer of the text or speech encoder. Positive values (red) indicate voxels where brain-tuning improves encoding performance, while negative values (blue) indicate better performance for the pretrained model. Voxels shown in gray were not significantly predicted. Top: within-modality evaluation, with (a) text representations predicting brain responses during reading and (b) speech representations predicting brain responses during listening. Bottom: cross-modal evaluation, with (c) text representations predicting brain responses during listening and (d) speech representations predicting brain responses during reading. Results are shown on the cortical surface of a representative held-out participant (P2); with representative higher-level language and association regions labeled: mPFC (medial prefrontal cortex), MPC (medial parietal cortex), IFG (inferior frontal gyrus), AG (angular gyrus), and STS (superior temporal sulcus). Results for the held-out participant 4 are in Appendix C.

SemAntoNeg. For speech, we use 4 probes from Speech Commands (Warden, 2018) (keyword, speaker and their hard variants), phoneme, and emotion identification tasks. Appendix B contains further benchmark and metric details. All probes are linear classifiers trained on frozen encoder representations, so accuracy changes reflect changes in representation rather than added capacity. Sentences and speech segments are mean-pooled to a single vector per item. We fit an ℓ_2-regularized logistic regression, with regularization strength selected on a held-out split, and evaluate on the untouched test split.

5 Results

5.1 Brain-tuning Improves Brain Alignment Within and Across Modalities

We first tested whether brain-tuning improves the alignment between multimodal model representations and brain activity. All analyses in this section are conducted on held-out participants. We train and test the encoding model on stories held out during brain-tuning. This ensures that neither the brain-tuning procedure nor the encoding model evaluation has been exposed to that participant's data or these specific stories.

We evaluated four settings: text representations predicting reading brain responses and speech representations predicting listening brain responses (within-modality), and text representations predicting listening brain responses and speech representations predicting reading brain responses (cross-modality). Figure 2 shows the difference in encoding performance between the joint brain-tuned and the pretrained model on one representative held-out participant's cortical surface; positive values indicate improved alignment after brain-tuning. Corresponding results for the remaining held-out participant are shown in Appendix C.

Within-modality. For the text encoder, both joint brain-tuned and text-only brain-tuned models improve encoding performance during reading (Fig.2a and Appendix Fig. 6a) across distributed cortical regions, with improvements concentrated in higher-level language and association regions, including AG, IFG, mPFC, and MPC.

For the speech encoder, both joint brain-tuned and speech-only brain-tuned models improve encoding performance during listening (Fig. 2b and Appendix Fig.6b) in higher-level regions. Compared to the text encoder, these changes are larger in magnitude, concentrated in frontal regions (mPFC, IFG) and near AG, and more pronounced in the left hemisphere than the right. Because speech models lack brain-relevant semantic information (Oota et al., 2024a), they likely have more room for improvement via brain-tuning.

Cross-modality. More importantly, brain-tuning also improves encoding performance when the tuning and evaluation modalities differ. For the text encoder, both joint and text-only brain-tuned representations better predict listening responses(Fig. 2c and Appendix Fig.6c) in several cortical regions, including secondary AC, AG, MPC, and mPFC.

For the speech encoder, joint and speech-only brain-tuned representations also show cross-modal changes when used to predict reading responses (Fig. 2d and Appendix Fig.6d), with the largest improvements in near AG (anterior to the IPS) and near IFG. However, these effects are more spatially variable across participants than the text-to-listening transfer.

Effect across model layers. We next examined the effect of brain-tuning on brain encoding performance across model layers. Figure 3 reports the percentage improvement in encoding performance of joint and single-modality brain-tuned models relative to the pretrained model at each layer, averaged across held-out participants. We also report the corresponding shuffled baseline for each layer.

Within-modality, improvements for the text encoder predicting reading responses are small in early layers and increase around the middle layers (Fig. 3a), reaching approximately 4% around layers 20-23 for the joint brain-tuned model. For the speech encoder predicting listening responses, the trend is variable (Fig. 3b), with the largest gains around layers 7-13. In both cases, joint brain-tuning outperforms single-modality tuning: text-only brain-tuning produces smaller gains than joint tuning in later layers, and speech-only brain-tuning shows comparatively lower improvements across most layers, with a sharper decline than joint tuning in the final layers.

Cross-modally, the text encoder predicting listening responses shows little change in early layers, with clear improvements emerging from approximately layers 12–21 and reaching about 2–4% around the middle-to-late layers (Fig. 3c). This closely mirrors the within-modality layer profile of the text encoder predicting reading responses. It suggests that changes induced by reading-based brain-tuning robustly transfer to listening. The speech encoder predicting reading responses shows a more jagged pattern across layers (Fig. 3d), with improvements at several intermediate layers that vary substantially across participants. Here too, joint brain-tuning outperforms single-modality tuning: speech-only tuning declines sharply in the final layers, while text-only tuning follows trends similar to joint tuning but with reduced performance gains. The layerwise results on the in-distribution participants, follow similar trend across all conditions with higher gains; reported in Appendix Fig. 9.

Together, these results show that our modality-agnostic brain-tuning approach improves brain alignment both within and across modalities. Joint brain-tuning produces greater alignment than single-modality tuning, in both the within- and cross-modality setting. This suggests that joint brain-tuning strengthens shared information between text and speech that supports cross-modal transfer, and that this shared information also helps improve the within-modality alignment. The corresponding results for CLAP are reported in Appendix E.1).

5.2 Brain-tuning Improves Semantic and Audio Representations

Finally, we tested whether the changes induced by brain-tuning extend beyond brain encoding performance. We evaluated the pretrained and brain-tuned representations on semantic text probes from SentEval and FlashHolmes, and on several audio probes(Fig. 4). We report the change in task performance relative to the pretrained model; values above zero indicate improvement after brain-tuning.

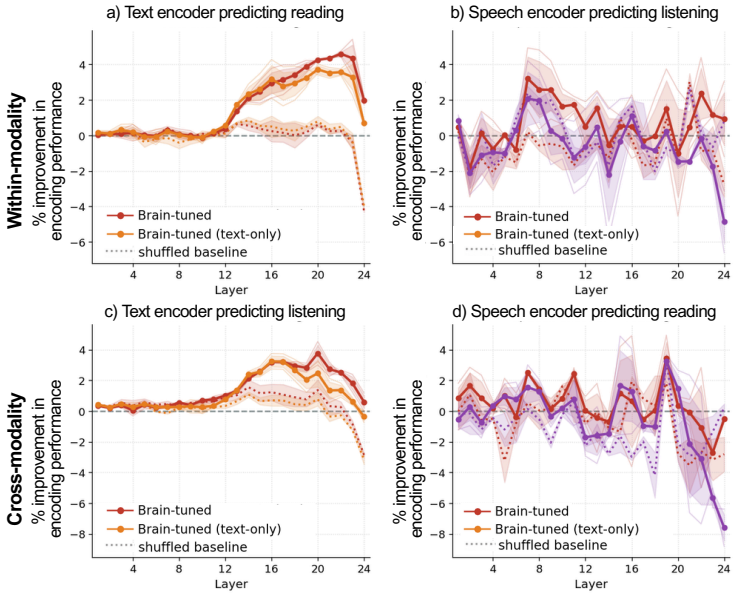

Figure 3: **Layerwise changes in brain encoding performance after modality-agnostic brain-tuning.** Percentage change in encoding performance after brain-tuning relative to the pretrained SONAR model, shown at each layer and averaged across held-out participants. Top: within-modality evaluation, with (a) text representations predicting reading responses and (b) speech representations predicting listening responses. Bottom: cross-modal evaluation, with (c) text representations predicting listening responses and (d) speech representations predicting reading responses. Red denotes joint brain-tuning using reading and listening brain data, orange denotes text-only brain-tuning, and violet denotes speech-only brain-tuning. Dashed colored lines show the corresponding shuffled baseline brain-tuned models. Thick solid lines show the mean across participants; shaded regions show the standard error of the mean.

Semantic probes. On SentEval (Fig. 4a), joint brain-tuning improves performance on Subject Number(+0.30), Object Number(+ 0.26), and SOMO (+0.38), with slight decreases for Tense (-0.08) and Coordination Inversion (-0.12). Text-only tuning yields a larger gain on SOMO (+0.48), smaller gains on Object Number (+0.16) and Tense (+0.04), but decreases performance on Subject Number (-0.08) and Coordination Inversion (-0.10). Averaged across probes, joint brain-tuning yields a gain of +0.15 over the pretrained model, larger than the average gain from text-only tuning (+0.10).

On FlashHolmes (Fig. 4b): joint brain-tuning improves most evaluated properties. Among the word-model probes, it improves physical dynamics information(+3.33), material property(+1.00), social interaction(+0.71), quantitative property(+0.66), social relation(+0.64), and agent property(+0.45). Among the linguistic probes, it improves BioOrgNeg(+0.50), negation(+0.26), specificity (+0.26), and BioNeg(+0.14). Only NegOrg(-0.20) and SpecOrg(-0.48) show a decrease in accuracy. Text-only tuning shows similar improvements on some tasks (e.g., PhysDyn, MatProp, NBioOrgNeg, negation), but shows no improvement or even a decrease on others (e.g., quantitative property, agent property).

This suggests that modality-agnostic brain-tuning improves higher-level semantic information in model representations, rather than only features specific to written language. Joint brain-tuning also produces higher and more consistent gains than single-modality tuning, particularly on the world-model probes.

Audio probes. We observe a complementary pattern for the audio probes (Fig. 4c). Joint brain-tuning improves keyword(+1.20), speaker(+6.72), keyword hard(+0.41) and phoneme(+3.37) task performance, with a small decrease for SpeakerHard(-0.40) task. Speech-only brain-tuning produces smaller or negative changes on most of these probes. Notably, joint brain-tuning strengthens audio-relevant information even more than speech-only tuning, suggesting that reading-based supervision does not come at the expense of speech-relevant representations.

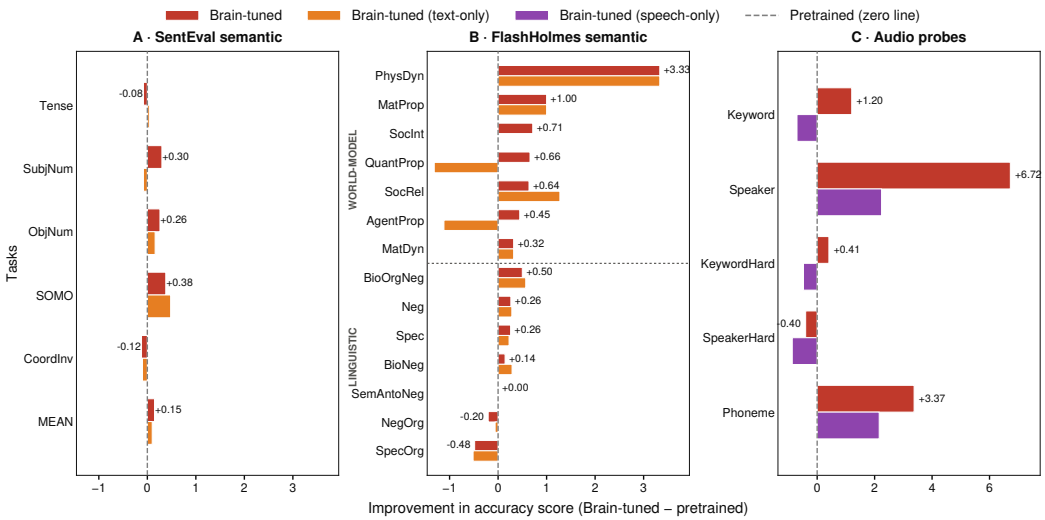

Figure 4: **Effect of brain-tuning on semantic and audio probing tasks.** Change in probing accuracy relative to the pretrained SONAR model, using representations from the final layer. (a) SentEval semantic probes, (b) FlashHolmes semantic probes, and (c) audio probing tasks. Red denotes joint brain-tuning, orange denotes text-only brain-tuning, and violet denotes speech-only brain-tuning. Positive values indicate improved probing performance after brain-tuning, while negative values indicate decreased performance. Brain-tuning improves several semantic- and audio-related properties, with joint brain-tuning showing broader gains across tasks than single-modality tuning.

Effects across model layers; We next examined downstream probing performance across the model layers. We group the 24 layers into three bands: early (L1-8), middle (L9-16), and late (L17-24), and report the change in probing accuracy relative to the pretrained model (Fig. 5).

For both SentEval and FlashHolmes, the semantic gains of brain-tuning progressively strengthen with model depth. On SentEval, joint brain-tuning shows almost no change in early layers(-0.01), a small gain in middle layers(+0.05), and the largest improvement in late layers(+0.24). A similar trend holds for FlashHolmes, where joint brain-tuning improves from +0.02 in early layers to +0.17 in middle layers and +0.53 in late layers. For both SentEval and FlashHolmes, text-only brain-tuning follows the same pattern on both benchmarks, with smaller gains than joint brain-tuning.

The audio keyword and phoneme probes show a similar depth-dependent pattern (Fig. 5). On the keyword probe, joint brain-tuning decreases performance in the early layers (-0.33) but improves it in the middle ($+0.38$) and late layers ($+1.11$). Speech-only tuning follows the same trend with smaller gains. In contrast, speaker and emotion information decreases across all bands. This is likely because the stories participants heard/read in the experiment were not optimized to elicit rich emotion or speaker-identity information, leaving little such structure for brain-tuning to learn from.

Overall, the layerwise analysis explains the final-layer probing results: the strongest semantic and keyword gains emerge primarily in the later parts of the encoder. This pattern is consistent with the idea that brain-tuning mainly changes higher-level language-related representations rather than uniformly affecting all layers (Moussa & Toneva, 2025a).

Taken together, the probing results further support the findings from the brain alignment evaluation. Modality-agnostic brain-tuning changes model representations in ways that extend beyond predicting fMRI responses, improving multiple semantic and audio properties relative to the pretrained model. The broader, more consistent gains from joint brain-tuning further suggest that combining reading and listening brain responses benefits information useful across text and speech. The corresponding results for CLAP are reported in Appendix E.2).

6 Discussion and Conclusion

In this work, we performed modality-agnostic brain-tuning of multimodal text–speech models using fMRI data recorded while participants read and listened to the same naturalistic stories (Fig. 1). We

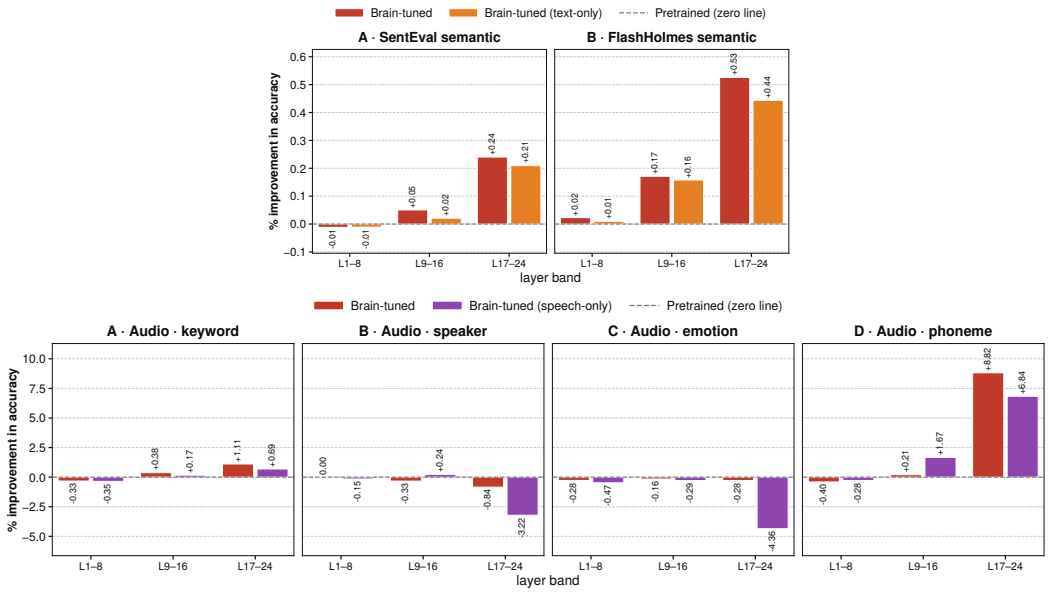

Figure 5: **Layerwise analysis of semantic and audio information after brain-tuning.** Change in probing accuracy relative to the pretrained SONAR model across early (L1-8), middle (L9-16), and late (L17-24) layer groups. Top: (a) SentEval semantic probes, (b) FlashHolmes semantic probes; Bottom: (a) audio, (b) speaker, (c) emotion, and (d) phoneme audio probes. Red, orange, and violet denote joint, text-only, and speech-only brain-tuning, respectively.

show that brain-tuning improves brain encoding performance both within and across modalities, and changes model representations in ways that improve its performance on several semantic and audio probing tasks. Our analyses reveal three key findings:

First, brain-tuning strengthens brain-relevant information shared across reading and listening. Text representations improve at predicting not only reading brain responses but also listening brain responses, and vice versa for speech (Fig. 2). This shows that brain-tuning is not confined to the modality used during training, but transfers across modalities. Notably, this shared information also improves the within-modality gains. Second, brain-tuning primarily reshapes higher-level representations in the model network, rather than modality-specific, low-level features. Improvements in both brain alignment and downstream probing tasks concentrate in middle and later model layers (Fig. 3). Lastly, joint brain-tuning generally produces larger and more consistent improvements than tuning with either modality alone, indicating that the two modalities offer complementary rather than redundant supervision.

Our results extend prior brain-tuning work, previously limited to unimodal text or speech models, to a multimodal setting. Inspired by neuroimaging evidence of modality-invariant semantic representations in the human brain, we propose a modality-agnostic brain-tuning approach to induce brain-relevant shared representations in multimodal models. This work further supports the growing body of work highlighting the potential of brain data as a rich supervisory signal for improving language model representations with biologically meaningful representations.

7 Limitations and Future Work

A key limitation is the limited availability of public human fMRI data for naturalistic language tasks in different modalities. In addition, we used only a subset of participants and stories for brain-tuning, holding out the remainder for evaluation. Hence observed improvements are constrained by this limited amount of training data. We focus on brain-tuning text–speech models with reading and listening brain data; future work could extend this approach to other modalities and models, such as brain-tuning vision-language models with movie-watching brain data.

8 AI USAGE DISCLOSURE

We did not use generative AI tools for any of the tasks requiring mandatory disclosure under the ICLR AI policy. We acknowledge the use of AI tools (Grammarly, OpenAI GPT-5.5, Claude Opus 4.8) for grammar correction, refining phrasing for better readability, and refining the presentation of figures. All AI-assisted edits were verified and finalized by the authors.

ETHICS STATEMENT

Our work uses publicly available fMRI data collected in prior studies while participants read and listened to naturalistic stories. We do not collect new human-participant data or acquire additional personal information.

REPRODUCIBILITY STATEMENT

We provide clear details on the fMRI dataset, the experimental setup, and the training procedure in the Methodology. We plan to release the brain-tuning pipeline code upon publication.

REFERENCES

Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36, 2024.

Khai Loong Aw and Mariya Toneva. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single! vector: Probing sentence embeddings for linguistic properties. In *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pp. 2126–2136. Association for Computational Linguistics, 2018.

Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 39(39):7722–7736, 2019.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*, 2023.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194, 2010.

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.

Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi U Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, et al. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *Transactions of the Association for Computational Linguistics*, 13:1245–1270, 2025.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *Advances in Neural Information Processing Systems*, 31, 2018.

Mathis Lamarre, Catherine Chen, and Fatma Deniz. Attention weights accurately predict language representations in the brain. *bioRxiv*, pp. 2022–12, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

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.

Omer Moussa and Mariya Toneva. Brain-tuned speech models better reflect speech processing stages in the brain. *arXiv preprint arXiv:2506.03832*, 2025a.

Omer Moussa and Mariya Toneva. Brain-tuning improves generalizability and efficiency of brain alignment in speech models. *Advances in Neural Information Processing Systems*, 39, 2025b.

Omer Moussa, Dietrich Klakow, and Mariya Toneva. Improving semantic understanding in speech language models via brain-tuning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 59823–59849, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/95b7a93e60fdfd10cc202f44fd6adf5f-Paper-Conference.pdf.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Anuja Negi, Subba Reddy Oota, Anwar O Nunez-Elizalde, Manish Gupta, and Fatma Deniz. Brain-informed fine-tuning for improved multilingual understanding in language models. *Advances in Neural Information Processing Systems*, 39, 2025.

Subba Reddy Oota, Khushbu Pahwa, Mounika Marreddy, Manish Gupta, and Bapi S Raju. Neural architecture of speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8503–8528. Association for Computational Linguistics, 2024a.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36, 2024b.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. Inducing brain-relevant bias in natural language processing models. *Advances in neural information processing systems*, 32, 2019.

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.

Aditya R Vaidya, Shailee Jain, and Alexander Huth. Self-supervised models of audio effectively explain human cortical responses to speech. In *International Conference on Machine Learning*, pp. 21927–21944. PMLR, 2022.

Aditya R Vaidya, Richard J Antonello, and Alexander G Huth. Fine-tuning language encoding models on slow fmri improves prediction for fast ecog. *arXiv preprint arXiv:2605.19224*, 2026.

Nishitha Vattikonda, Aditya R Vaidya, Richard J Antonello, and Alexander G Huth. Brain-wavlm: Fine-tuning speech representations with brain responses to language. *arXiv preprint arXiv:2502.08866*, 2025.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(Suppl 11):S9, 2008.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. Holmes a benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647, 2024.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS One*, (11), 2014.

Michael C-K Wu, Stephen V David, and Jack L Gallant. Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, 29(1):477–505, 2006.

Overview of Appendices

- Section A: Multimodal Model Details
- Section B: Probing Dataset Details
- Section C: Additional results for SONAR
- Section D: Brain-tuning with CLAP
- Section E: Brain-tuning results for CLAP

A Multimodal Model Details

This section provides additional architectural and input-processing details for the two multimodal models brain-tuned in this work, SONAR and CLAP, complementing the summary in Section 3.2; with key details summarised given in Table 1.

SONAR. We use SONAR's Duquenne et al. (2023) pretrained text and speech encoders. Both encoders consist of 24 Transformer layers with a hidden dimension of 1024. The text encoder is initialized from an NLLB-based Transformer encoder, while the speech encoder is initialized from Wav2Vec2-BERT 2.0. The text and speech encoders are separately parameterized, but produce representations in a common 1024-dimensional space. We use the English ($\texttt{eng}_{Latn}$) text and speech checkpoints in all experiments.

CLAP. We use CLAP's Elizalde et al. (2023) pretrained text and speech encoders. Its text tower is a RoBERTa encoder containing 12 Transformer layers with a hidden dimension of 768. The audio tower uses the Hierarchical Token-Semantic Audio Transformer (HTSAT), a hierarchical Swin Transformer operating on log-mel spectrograms. CLAP expects audio sampled at 48,kHz in fixed 10-s windows; each 500-ms target chunk and its preceding acoustic context are placed at the end of this window, with left zero-padding applied when fewer than 10,s of audio are available. The native text and audio representations have different dimensions (768 and 1024, respectively), so CLAP uses pretrained projection modules to map both modalities into a shared 512-dimensional contrastive space.

Table 1: Architectural and input-processing details of the multimodal models.

Property	**SONAR**	**LAION-CLAP**
Checkpoint	English text and speech encoders (`eng_Latn`)	`laion/larger_clap_music_and_speech`
Text backbone	NLLB-based Transformer encoder	RoBERTa encoder
Text architecture	24 Transformer layers; hidden dimension 1024	12 Transformer layers; hidden dimension 768
Speech/audio backbone	Wav2Vec2-BERT 2.0	Hierarchical Token-Semantic Audio Transformer (HTSAT)
Speech/audio architecture	24 Conformer blocks; hidden dimension 1024	Four hierarchical stages with depths $[2, 2, 12, 2]$; final hidden dimension 1024

B Probing dataset details

This section provides full details for the semantic and audio probing benchmarks summarized in Section 4.2. Table 2 gives an overview of all probing tasks; Table 3 lists per-task train/test sizes for the SentEval and FlashHolmes semantic probes; and Table 4 details the speech probing tasks.

Table 2: Probing benchmark summary. All probes are linear (L2-regularized logistic regression) on frozen, pooled encoder embeddings; [a]SentEval tasks subsample the official 100k/10k train/test split. [b]FlashHolmes task sizes vary by task; see Table 3.

Probe	**Source**	**Classes**	**#Tasks**	**Train N**	**Test N**	**Metric**
SentEval semantic	Conneau et al. (2018)	2	5	15,000[a]	5,000[a]	Acc. (%)
FlashHolmes – EWoK	Ivanova et al. (2025)	2	7	38–546[b]	60–896[b]	Acc. (%)
FlashHolmes – linguistic	Vincze et al. (2008)	2	7	834–3,261[b]	837–4,940[b]	Acc. (%)
Keyword ID	Warden (2018)	35	1	5,985	1,995	Acc. (%)
Speaker ID	Warden (2018)	40	1	5,595	580	Acc. (%)
Speaker ID (hard)	Warden (2018)	250	1	2,500	1,500	Acc. (%)

Table 3: Per-task train/test sizes. All 19 tasks are binary classification, scored by top-1 accuracy. EWoK tasks classify a sentence *pair*, featurized as $[\,|a-b|\,;\,a\odot b\,]$ from the two pooled embeddings (the standard SentEval pair-probe recipe). FlashHolmes ships train/dev/test splits with very small official train sets (as few as 20 examples); we pool train+dev and evaluate on the untouched test split.

Group	Task	Train+dev N	Test N	Input
SentEval	Tense (past/present)	15,000	5,000	single sentence
	SubjNum (subject number)	15,000	5,000	single sentence
	ObjNum (object number)	15,000	5,000	single sentence
	SOMO (odd-man-out)	15,000	5,000	single sentence
	CoordInv (clause inversion)	15,000	5,000	single sentence
EWoK (world-model)	physical-dynamics	38	60	sentence pair
	material-properties	62	100	sentence pair
	social-interactions	87	140	sentence pair
	quantitative-properties	92	152	sentence pair
	social-relations	384	628	sentence pair
	agent-properties	546	896	sentence pair
	material-dynamics	380	624	sentence pair
Linguistic	bioscope-org-negation	834	1,403	single sentence
	negation	3,260	4,940	single sentence
	speculation	1,920	3,099	single sentence
	bioscope-negation	838	1,390	single sentence
	SemAntoNeg	2,091	837	single sentence
	negation-org	3,261	4,897	single sentence
	speculation-org	1,882	3,105	single sentence

Table 4: Speech probing benchmarks. All tasks are linear (L2-regularized logistic regression) probes of a frozen, pooled encoder embedding, scored by top-1 accuracy. Keyword ID and Speaker ID share one Speech Commands download but split it differently: Keyword ID balances classes over a random clip split (train and test may share speakers); Speaker ID uses a closed set of the N most frequent speakers, with disjoint clips of the *same* speakers in train and test (standard closed-set speaker identification, not speaker generalization). The "hard" variants keep the same protocol with more classes (speaker) or less training data per class, to avoid the ceiling effect the easy variants show. [a]Of the 39 ARPAbet phone classes (stress digits stripped, e.g.IH0/IH1/IH2 $\rightarrow$ IH), 4 have fewer than 20 training examples and are dropped.

Task	Source	Classes	Train / Test N	Split	Clip
Keyword ID	Speech Commands v0.02	35	5,985 / 1,995	class-balanced, any speaker	$\leq$1 s word
Keyword ID (hard)	Speech Commands v0.02	35	2,485 / 1,470	class-balanced, any speaker	$\leq$1 s word
Speaker ID	Speech Commands v0.02	40	5,595 / 580	closed-set, disjoint clips	$\leq$1 s word
Speaker ID (hard)	Speech Commands v0.02	250	2,500 / 1,500	closed-set, disjoint clips	$\leq$1 s word
Emotion	CREMA-D	6	4,000 / 1,500	actor-disjoint (22 held out of 91)	1.3–5 s utterance
Phoneme	LibriSpeech dev-clean	35[a]	3,960 / 1,482	speaker-disjoint (10 held out of 40)	phone interval, $\geq$40 ms

C Additional results for SONAR

This section reports additional voxelwise encoding results complementing the main text. Figures 6 and 8 show text-only and speech-only brain-tuning effects for the two helod-out participants (Participant2 and Participant 9), complementing the joint brain-tuning results shown in Fig. 7 for Participant 9. Fig. 9 shows the layerwise encoding results from Fig. 3 recomputed on in-distribution (training) participants.

D Brain-tuning with CLAP

This section details the CLAP-specific training protocol and alignment loss, following the same overall brain-tuning framework described in Section 3.3.

CLAP training protocol. We use the same training data, subject split, brain-prediction heads, optimizer, learning rate, weight decay, batch sizes, validation split, learning-rate schedule, and early-stopping criterion for CLAP, allowing differences between SONAR and CLAP to be attributed primarily to their pretrained representations and encoder architectures. All layers of the active RoBERTa text tower and HTSAT audio tower are fully fine-tuned, including CLAP's original 512-dimensional contrastive projections. Because the CLAP audio pathway processes fixed 10-s, 48-kHz inputs and the training loop accumulates computational graphs across many speech chunks, we apply gradient checkpointing to the complete text and audio encoder forward passes to reduce memory consumption. In addition, HTSAT applies `BatchNorm2d` to its log-Mel inputs. Because the speech

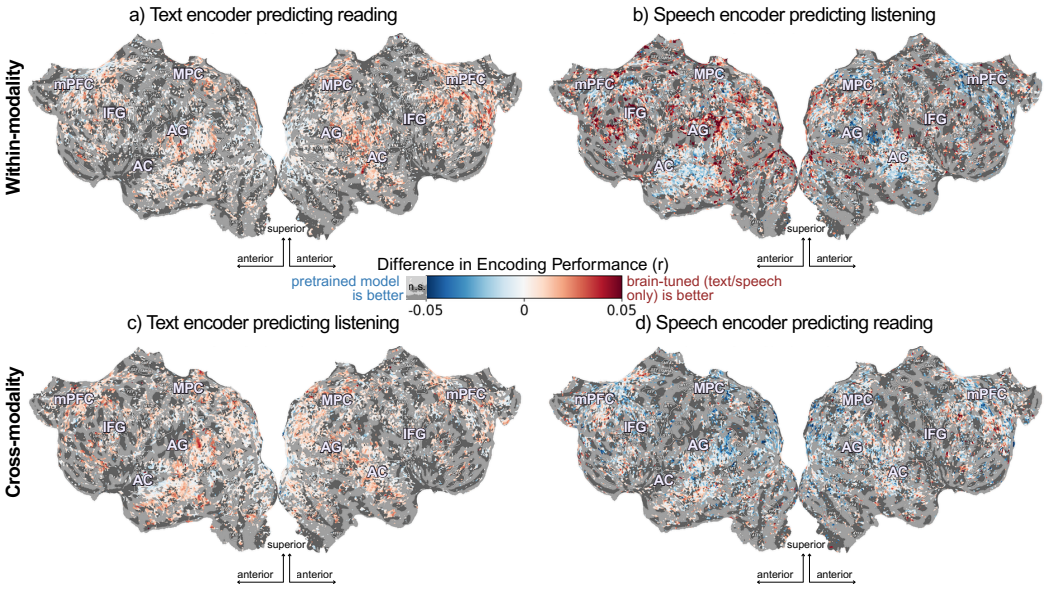

Figure 6: **Brain encoding performance before and after modality-agnostic brain-tuning (text/speech-only for Participant 2).** Difference in voxelwise encoding performance between the joint brain-tuned and pretrained SONAR model, using the best layer of the text or speech encoder. Positive values (red) indicate voxels where text/speech-only brain-tuning improves encoding performance, while negative values (blue) indicate better performance for the pretrained model. Voxels shown in gray were not significantly predicted. Top: within-modality evaluation, with (a) text representations predicting brain responses during reading and (b) speech representations predicting brain responses during listening. Bottom: cross-modal evaluation, with (c) text representations predicting brain responses during listening and (d) speech representations predicting brain responses during reading. Results are shown on the cortical surface of participant 2, with representative higher-level language and association regions labeled: mPFC (medial prefrontal cortex), MPC (medial parietal cortex), IFG (inferior frontal gyrus), AG (angular gyrus), and STS (superior temporal sulcus).

batch size is small(four), we keep all BatchNorm layers in evaluation mode and use their pretrained running statistics throughout fine-tuning, while all remaining HTSAT parameters continue to receive gradients. The text-only, speech-only, and joint conditions follow the same parameter-update scheme used for SONAR, with inactive encoder branches frozen and excluded from the optimizer.

CLAP alignment loss. CLAP produces native text and audio representations with different dimensionalities. For alignment only, we pass them through CLAP's pretrained text and audio projection functions, p_{text} and p_{audio}, which map both representations to the model's 512-dimensional contrastive space. The projected representations are then L_2-normalized:

$$z_i^{\text{text}} = \frac{p_{\text{text}}(H_i^{\text{text}})}{\|p_{\text{text}}(H_i^{\text{text}})\|_2 + \epsilon}, \qquad z_i^{\text{audio}} = \frac{p_{\text{audio}}(H_i^{\text{speech}})}{\left\|p_{\text{audio}}(H_i^{\text{speech}})\right\|_2 + \epsilon}. \tag{5}$$

For a batch of B temporally aligned text–speech pairs, the similarity matrix is

$$S_{ij} = \frac{\langle z_i^{\text{text}}, z_j^{\text{audio}} \rangle}{\tau}, \qquad \tau = 0.07. \tag{6}$$

We then minimize the symmetric NT-Xent loss

$$\mathcal{L}_{\text{align}}^{\text{CLAP}} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(S_{ii})}{\sum_{j=1}^{B} \exp(S_{ij})} + \log \frac{\exp(S_{ii})}{\sum_{j=1}^{B} \exp(S_{ji})} \right]. \tag{7}$$

A positive pair (i, i) consists of text and speech representations aligned to the same TR of the same story, whereas the other TRs in the batch serve as negatives. The pretrained CLAP projections are used only for the alignment objective; the cortical prediction heads operate directly on the native hidden states.

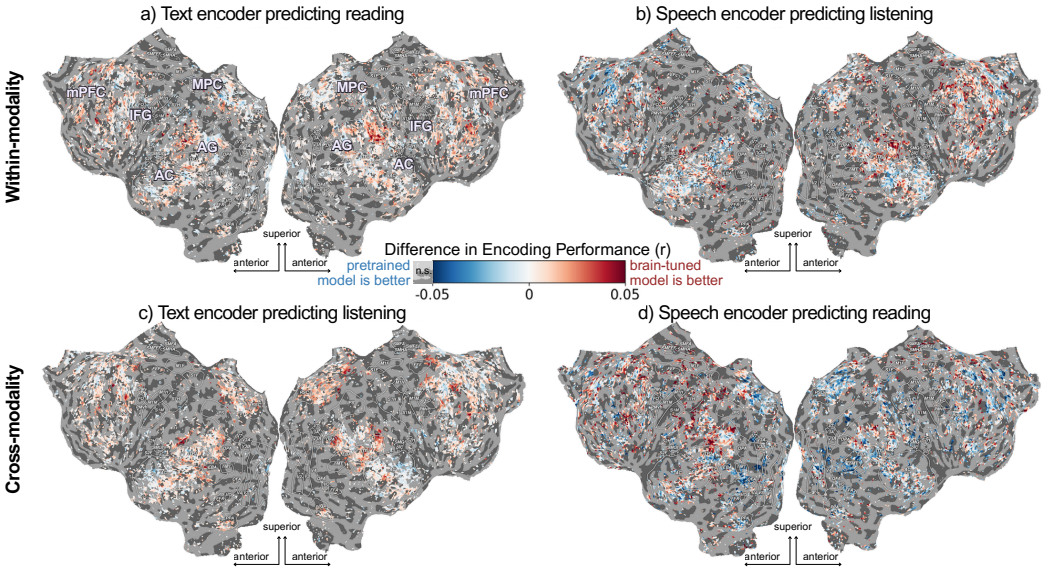

Figure 7: **Brain encoding performance before and after modality-agnostic brain-tuning (joint for Participant 9).** Difference in voxelwise encoding performance between the joint brain-tuned and pretrained SONAR model, using the best layer of the text or speech encoder. Positive values (red) indicate voxels where joint brain-tuning improves encoding performance, while negative values (blue) indicate better performance for the pretrained model. Voxels shown in gray were not significantly predicted. Top: within-modality evaluation, with (a) text representations predicting brain responses during reading and (b) speech representations predicting brain responses during listening. Bottom: cross-modal evaluation, with (c) text representations predicting brain responses during listening and (d) speech representations predicting brain responses during reading. Results are shown on the cortical surface of participant 9, with representative higher-level language and association regions labeled: mPFC (medial prefrontal cortex), MPC (medial parietal cortex), IFG (inferior frontal gyrus), AG (angular gyrus), and STS (superior temporal sulcus).

E BRAIN-TUNING RESULTS FOR CLAP

This section reports brain encoding and probing results for CLAP, complementing the SONAR results in the main text and testing whether our findings generalize across a second, architecturally distinct multimodal backbone (Section D).

E.1 ENCODING PERFORMANCE

We evaluate voxelwise encoding performance for CLAP following the same procedure used for SONAR (Section 4.1). Figure 10 shows improvement in brain alignment after brain-tuning on the cortical surface of held-out participant 2, and Figure 11 shows this across CLAP's encoder layers.

E.2 PROBING RESULTS FOR CLAP

We next examined whether brain-tuning changes the semantic and audio information represented by CLAP. We evaluated the pretrained and brain-tuned representations on the same SentEval and FlashHolmes semantic probes and audio probing tasks used for SONAR (Fig. 12). We report the change in probing accuracy relative to the pretrained CLAP model, where positive values indicate improved information after brain-tuning.

Semantic probes. On SentEval (Fig. 12a), joint brain-tuning improves all five probes: Object Number ($+2.40$), SOMO ($+1.28$), Subject Number ($+1.22$), Tense ($+1.06$), and Coordination Inversion ($+0.48$), for an average gain of $+1.29$. Text-only tuning also improves every probe (Object Number $+2.48$, Subject Number $+1.18$, SOMO $+0.72$, Tense $+0.60$, Coordination Inversion $+0.32$; mean $+1.06$), with joint tuning ahead on four of the five. These gains are considerably larger and more uniform than for SONAR, where the average SentEval gain was $+0.15$ and two probes decreased.

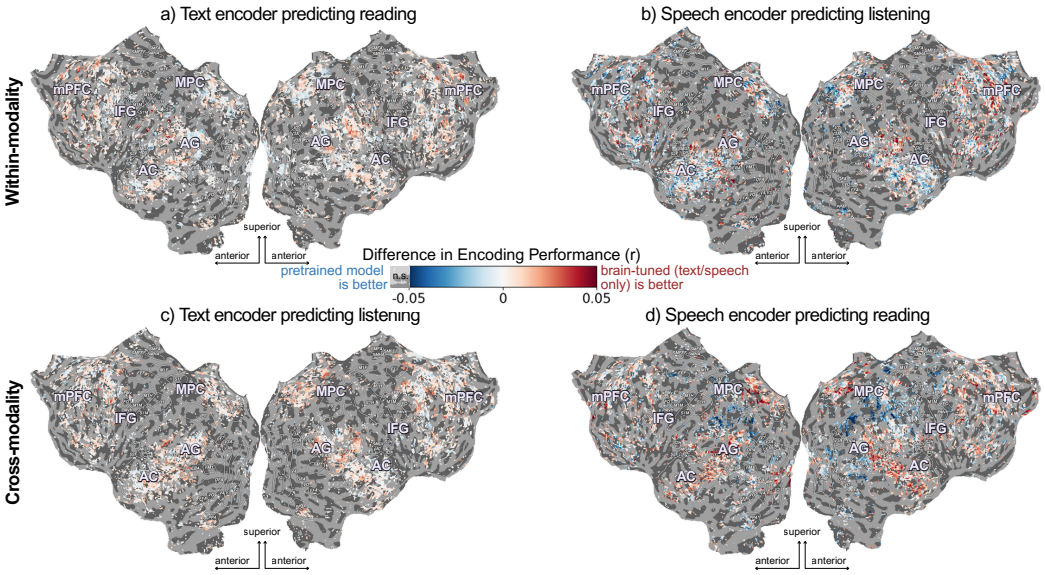

Figure 8: **Brain encoding performance before and after modality-agnostic brain-tuning (text/speech-only for Participant 9).** Difference in voxelwise encoding performance between the joint brain-tuned and pretrained SONAR model, using the best layer of the text or speech encoder. Positive values (red) indicate voxels where text/speech-only brain-tuning improves encoding performance, while negative values (blue) indicate better performance for the pretrained model. Voxels shown in gray were not significantly predicted. Top: within-modality evaluation, with (a) text representations predicting brain responses during reading and (b) speech representations predicting brain responses during listening. Bottom: cross-modal evaluation, with (c) text representations predicting brain responses during listening and (d) speech representations predicting brain responses during reading. Results are shown on the cortical surface of participant 9, with representative higher-level language and association regions labeled: mPFC (medial prefrontal cortex), MPC (medial parietal cortex), IFG (inferior frontal gyrus), AG (angular gyrus), and STS (superior temporal sulcus).

On FlashHolmes (Fig. 12b), the pattern is split by probe family. Joint brain-tuning improves every linguistic probe: BioOrgNeg ($+3.71$), Negation ($+3.04$), NegOrg ($+2.57$), Specificity ($+2.52$), BioNeg ($+1.80$), and SpecOrg ($+1.13$), with SemAntoNeg unchanged (0.00). The world-model probes are mixed: material property ($+1.00$), material dynamics ($+0.96$), social interaction ($+0.71$), and agent property ($+0.45$) improve, while physical dynamics (-1.67), social relation (-2.23), and quantitative property (-3.95) decrease. Text-only tuning matches joint tuning on the linguistic probes (BioOrgNeg $+2.92$, Specificity $+2.78$, Negation $+2.21$, NegOrg $+2.17$, BioNeg $+2.16$, SpecOrg $+0.58$) but is markedly worse on the world-model probes, with larger decreases on quantitative property (-7.24), physical dynamics (-3.34), and material property (-2.00), and no gain on social interaction (0.00). Thus, for CLAP, brain-tuning robustly strengthens linguistic properties, while world-model properties benefit from joint supervision specifically—the reverse of SONAR, where physical dynamics showed the largest gain.

Audio probes. Joint brain-tuning improves keyword ($+0.60$) and keyword hard ($+1.09$), leaves speaker unchanged (0.00), and decreases speaker hard (-1.00) and phoneme (-1.55) (Fig. 12c). Speech-only tuning is worse on keyword (-0.35) and keyword hard ($+0.54$), similar on speaker hard (-1.07), and less harmful on phoneme (-0.47). As with SONAR, joint brain-tuning yields the stronger keyword gains, but unlike SONAR—where speaker ($+6.72$) and phoneme ($+3.37$) improved substantially—CLAP shows no speaker gain and a phoneme loss, indicating that brain-tuning shifts the CLAP audio encoder toward lexical rather than acoustic–phonetic information.

Effects across model layers. We grouped the 12 text-encoder layers into early (L1–4), middle (L5–8), and late (L9–12) bands, and the 18 audio-encoder layers into early (L1–6), middle (L7–12), and late (L13–18) bands (Fig. 13). As with SONAR, semantic gains are concentrated in the late layers. On SentEval, joint brain-tuning shows a small decrease in early (-0.02) and middle layers (-0.16) and a large late-layer gain ($+0.93$); text-only tuning follows the same profile ($+0.05$, -0.11, $+0.76$). FlashHolmes shows a monotonic increase for joint tuning (-0.04, $+0.31$, $+0.81$),

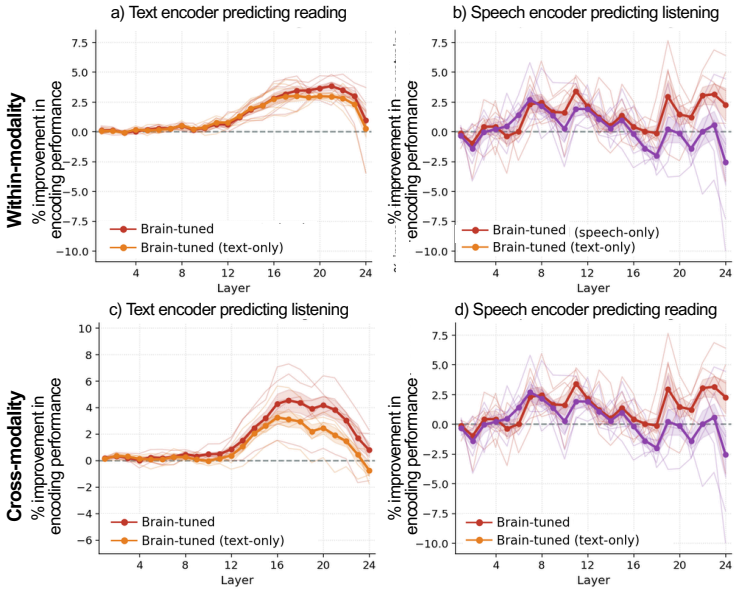

Figure 9: **Layerwise changes in brain encoding performance after modality-agnostic brain-tuning (in-distribution participants).** Percentage change in encoding performance after brain-tuning relative to the pretrained SONAR model, shown at each layer and averaged across in-distribution participants. Top: within-modality evaluation, with (a) text representations predicting reading responses and (b) speech representations predicting listening responses. Bottom: cross-modal evaluation, with (c) text representations predicting listening responses and (d) speech representations predicting reading responses. Red denotes joint brain-tuning using reading and listening brain data, orange denotes text-only brain-tuning, and violet denotes speech-only brain-tuning. Thick solid lines show the mean across participants; shaded regions show the standard error of the mean.

again with smaller text-only gains (-0.09, $+0.07$, $+0.44$). The keyword probe follows the same depth-dependent pattern for joint tuning (-1.23, -0.50, $+0.98$), whereas speech-only tuning barely recovers in late layers (-0.48, -0.71, $+0.09$). The phoneme probe, however, shows the opposite profile: both joint ($+1.42$, -0.21, -0.36) and speech-only tuning ($+1.26$, $+0.16$, $+0.43$) improve phoneme information in early layers, and joint tuning loses it in later layers, which explains the final-layer phoneme decrease. Emotion information improves slightly in early layers ($+0.61$ for both conditions) and decreases thereafter (joint: -0.34, -0.28; speech-only: -0.58, -0.97), and speaker information is essentially unchanged across depth.

Overall, the CLAP probing analysis shows that brain-tuning strengthens sentence-level and linguistic semantic information, with the gains again emerging in the later layers of the encoder, and that joint supervision is needed to improve world-model properties. Together with the SONAR results, this indicates that brain supervision reshapes linguistically relevant information across distinct multimodal architectures, while the acoustic properties that change (speaker and phoneme for SONAR, keyword for CLAP) depend on how information is organized within each pretrained model.

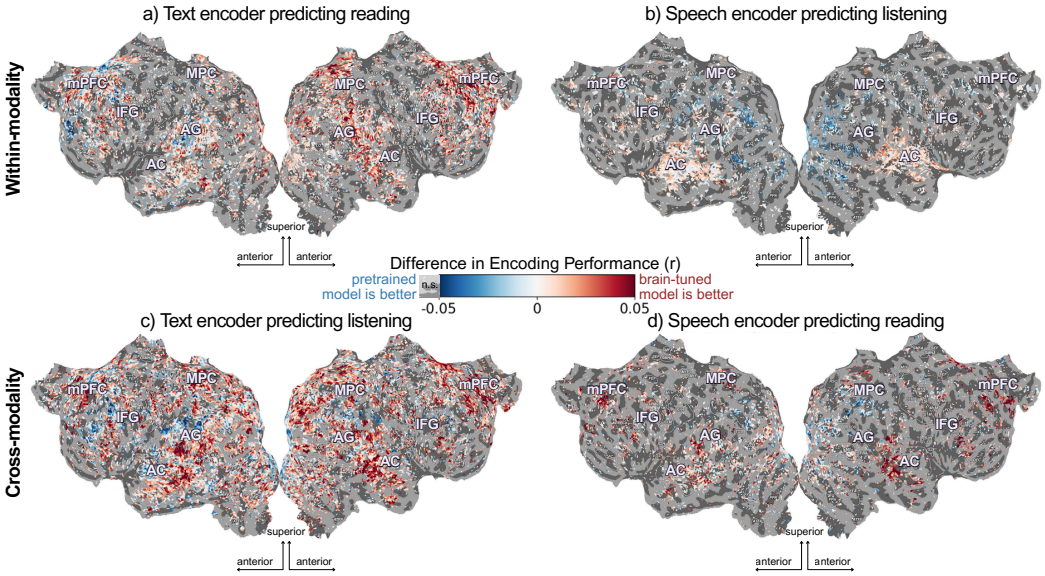

Figure 10: **Brain encoding performance before and after modality-agnostic brain-tuning of CLAP.** Difference in voxelwise encoding performance between the joint brain-tuned and pretrained SONAR model, using the best layer of the text or speech encoder. Positive values (red) indicate voxels where joint brain-tuning improves encoding performance, while negative values (blue) indicate better performance for the pretrained model. Voxels shown in gray were not significantly predicted. Top: within-modality evaluation, with (a) text representations predicting brain responses during reading and (b) speech representations predicting brain responses during listening. Bottom: cross-modal evaluation, with (c) text representations predicting brain responses during listening and (d) speech representations predicting brain responses during reading. Results are shown on the cortical surface of participant 2, with representative higher-level language and association regions labeled: mPFC (medial prefrontal cortex), MPC (medial parietal cortex), IFG (inferior frontal gyrus), AG (angular gyrus), and STS (superior temporal sulcus).

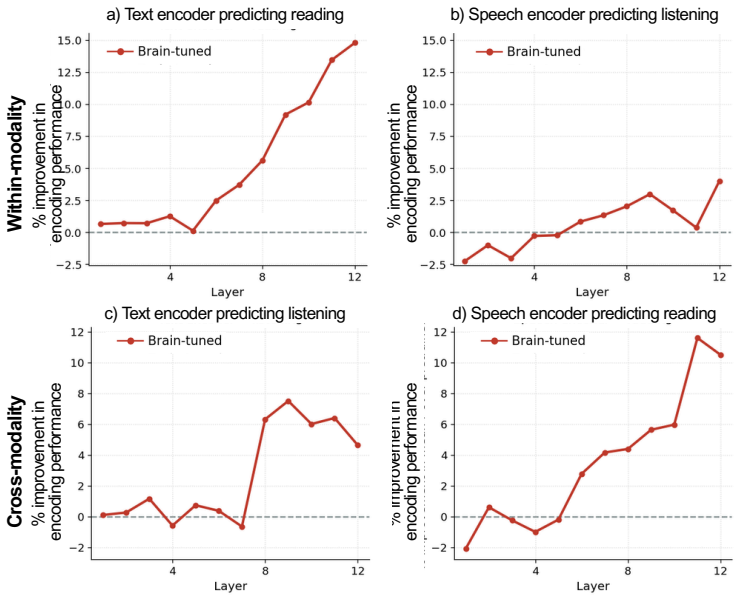

Figure 11: **Layerwise changes in brain encoding performance after modality-agnostic brain-tuning of CLAP.** Percentage change in encoding performance after brain-tuning relative to the pretrained SONAR model, shown at each layer for held-out participant 2. Top: within-modality evaluation, with (a) text representations predicting reading responses and (b) speech representations predicting listening responses. Bottom: cross-modal evaluation, with (c) text representations predicting listening responses and (d) speech representations predicting reading responses. Red denotes joint brain-tuning using reading and listening brain data, orange denotes text-only brain-tuning, and violet denotes speech-only brain-tuning.

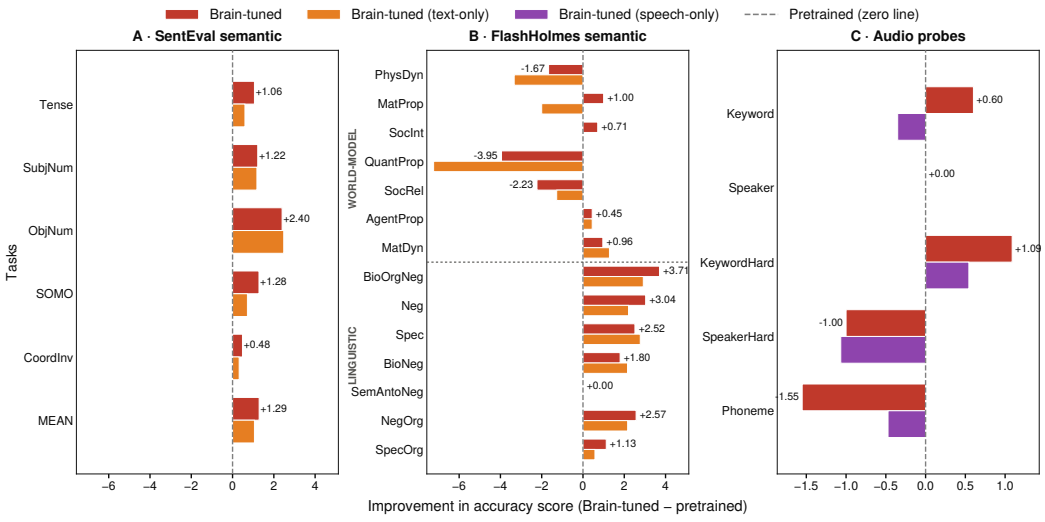

Figure 12: **Effect of brain-tuning on semantic and audio probing tasks.** Change in probing accuracy relative to the pretrained CLAP model, using representations from the final layer. (a) SentEval semantic probes, (b) FlashHolmes semantic probes, and (c) audio probing tasks. Red denotes joint brain-tuning, orange denotes text-only brain-tuning, and violet denotes speech-only brain-tuning. Positive values indicate improved probing performance after brain-tuning, while negative values indicate decreased performance. Brain-tuning improves several semantic- and audio-related properties, with joint brain-tuning showing broader gains across tasks than single-modality tuning.

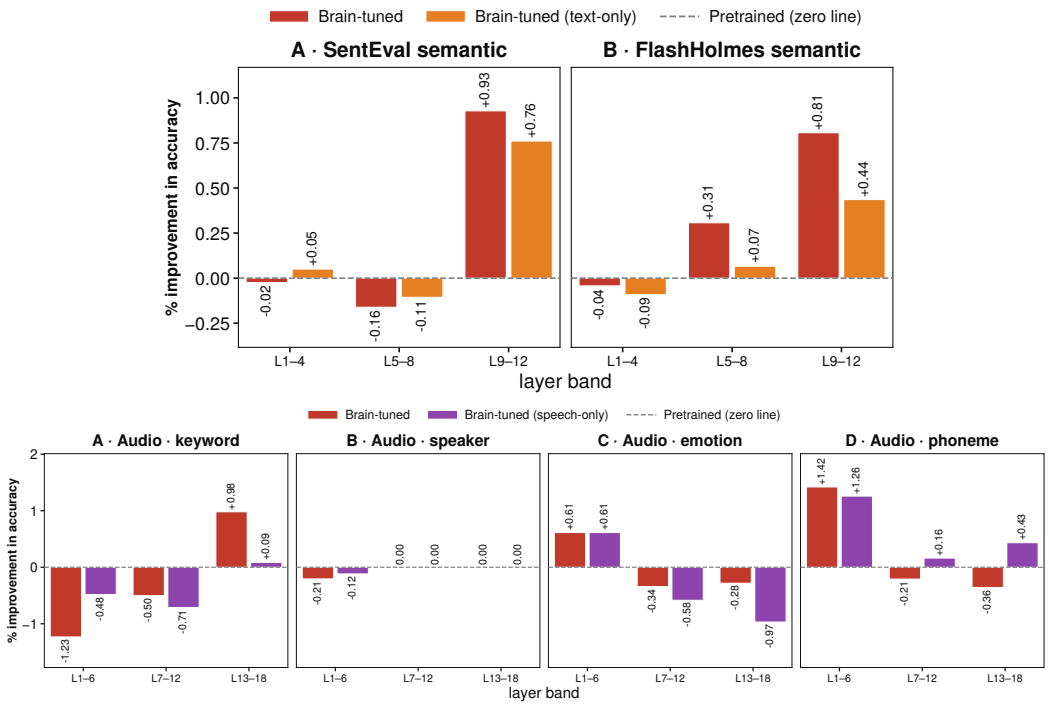

Figure 13: **Layerwise analysis of semantic information after brain-tuning.** Change in probing accuracy relative to the pretrained CLAP model across early (L1–8), middle (L9–16), and late (L17–24) layer groups for: (a) SentEval semantic probes, (b) FlashHolmes semantic probes, (c) audio keyword probes, and (d) audio speaker probes. Red denotes joint brain-tuning, orange denotes text-only brain-tuning, and violet denotes speech-only brain-tuning.