When Vocal Tone and Literal Meaning Diverge: An Acoustic–Semantic Incongruity Study for Large Audio–Language Models

Yu-Wen Chen[1,*] **William Ho**[1,*] **Maxim Topaz**[2] **Zoran Kostic**[1] **Julia Hirschberg**[1]
[1]The Fu Foundation School of Engineering and Applied Science, Columbia University, USA
[2]School of Nursing, Columbia University, USA
{yu-wen.chen, william.ho}@columbia.edu

Abstract

Affective cues across modalities may be incongruous (e.g., sarcasm or mocking praise), potentially leading to misinterpretation when relying on a single modality. Large Audio–Language Models (LALMs) have recently gained popularity and been applied to multimodal emotion recognition, but their ability to disentangle acoustic and semantic cues, especially in incongruent cases, remains underexplored. To address this gap, we introduce CREMA-ASIS, a dataset specifically created to investigate incongruence between acoustic emotion and semantic sentiment cues. It pairs acoustic emotion labels with semantic sentiment polarities. Using this dataset, we evaluate LALM biases within a multitask framework and conduct a layer-wise analysis to identify modality dominance across layers. Our findings reveal that LALMs struggle with semantic-acoustic incongruent cases, rarely predicting incongruity, and that LALMs are predominantly influenced by semantic information. However, supervised fine-tuning significantly improves LALM performance on our CREMA-ASIS test set while preserving transcription accuracy and joint emotion recognition. Results demonstrate potential for enhancing both acoustic and semantic understanding on out-of-domain data.

1 Introduction

Human affect is naturally conveyed through multiple modalities, including speech and text. Capturing emotional cues from both modalities provides deeper insight into the intent of the speaker (Du et al., 2025). Recent advances in Large Audio–Language Models (LALMs) have attracted increasing attention, enabling a unified framework supporting multiple audio-text tasks within a single model, including emotion recognition (Chu et al., 2024; Ding et al., 2025; Ghosh et al., 2025; Tang et al., 2024; Deshmukh et al., 2023). However, most studies assign a single unified emotion label across modalities, thus overlooking both the heterogeneity and the complementary information provided by each. Moreover, researchers have observed that current LALMs may exhibit bias toward textual descriptions. For example, Wang et al. (2025) showed that LALMs tend to prioritize textual information in *prompts* over acoustic evidence when the sources conflict. Beyond prompt-audio conflicts, mismatches can occur within the audio signal itself when acoustic and semantic content diverge. Such mismatches frequently occur in real-world communication, as in sarcasm, mocking praise, or concealed sadness. Dominance of one modality can bias interpretation, leading listeners to misread or overlook the speaker's true intent.

Recent studies have begun exploring LALM bias in emotion recognition. Chen et al. (2026) investigated how different instructions (listening, reading, or both) affect LALM performance on emotion classification. Corrêa et al. (2026) evaluated LALMs using audio samples generated by a text-to-speech (TTS) system from GPT-generated sentences and speech emotion samples. Both studies observe a consistent lexical dominance. However, these studies mainly highlight LALMs' bias toward semantic cues in audio, without further exploring potential strategies for mitigating this behavior. In addition, the datasets used in both works are limited, making them insufficient to guide models to disentangle acoustic and semantic cues. While previous studies focus on single-task settings, a more challenging yet practical approach is a multitask framework that generates separate predictions for each modality, enabling finer-grained analysis of speaker intent through comparison of modality-specific affective cues.

While prior work uses emotion categories or sentiment scores for both modalities, we consider pairing acoustic emotion categories (e.g., happy, angry) with semantic sentiment labels (positive, neutral,

[*]These authors contributed equally.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 702–717
October 24-29, 2026 ©2026 Association for Computational Linguistics

negative), which provides a more appropriate and informative setting. This choice was made to differentiate between a speaker's emotion expressed through acoustic cues and the general consensus sentiment conveyed in the semantics. For example, "He had a car accident." can express neutral (factual), sad (sympathetic), or even happy (malicious) emotions; however, for the lexical meaning, the general consensus sentiment of the event is negative. Acoustic signals are represented using emotion categories rather than sentiment to better capture the richness of human expressive states.

Our contributions are summarized as follows (1) We introduce CREMA-ASIS, an Acoustic–Semantic Incongruity Study dataset built upon CREMA-D (Cao et al., 2014). The dataset is created using a state-of-the-art (SOTA) text-to-speech (TTS) system, allowing controlled manipulation of acoustic and semantic labels and enabling systematic evaluation of how LALMs interpret multimodal affective cues. (2) We analyze the performance of several SOTA LALMs on CREMA-ASIS, revealing their strengths and limitations across various acoustic–semantic pairings. (3) Leveraging the scale of CREMA-ASIS, we conduct Supervised Fine-Tuning (SFT) experiments where models are trained to predict both acoustic emotion and semantic sentiment simultaneously. (4) We empirically demonstrate a pronounced acoustic–semantic gap in LALMs through layer-wise linear probing (Alain and Bengio, 2017) and show that SFT on CREMA-ASIS improves the linear decodability of acoustic information in deeper layers while largely preserving semantic processing, narrowing the gap without sacrificing generalization. These findings provide diagnostic insight toward understanding how LALMs represent multimodal affective cues and highlight the potential of leveraging synthetic data to mitigate the acoustic–semantic gap[1].

2 Method

We have designed CREMA-ASIS that explicitly targets affects which are inconsistent between acoustic and semantic information, specifically when the acoustic emotional valence contradicts the semantic sentiment expressed in the utterance. We used the dataset both to evaluate LALMs' biases across acoustic and semantic modalities and to improve their ability to differentiate between the two via SFT. The models are instructed to perform acoustic emotion recognition and semantic sentiment prediction simultaneously (detailed in Appendix A.1). By explicitly pairing inconsistent acoustic and semantic signals, we create conditions where predicting acoustic emotion from lexical content alone yields systematically incorrect predictions. Successful learning therefore requires attending to and preserving acoustic features predictive of spoken emotion. Figure 1 illustrates the study overview.

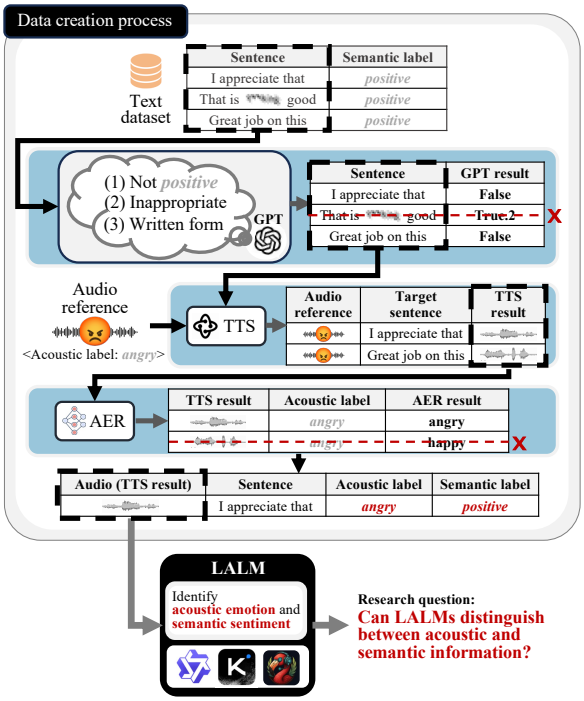


Figure 1: Study overview. The data creation process illustrates generating audio with an angry voice and positive semantic content.

2.1 CREMA-ASIS data creation

Recent advances in TTS systems have enabled high-quality, natural-sounding speech generation. Among several open-source models supporting emotion expression and voice-cloning (Du et al., 2024), IndexTTS2 (Zhou et al., 2026) was selected based on manual listening tests. Its voice cloning capabilities allow it to replicate tone from the reference audio, enabling control of both acoustic emotion and semantic content in the generated speech by providing an emotion-specified reference sample.

2.1.1 Audio Reference and Sentence Selection

We focused on emotion categories that commonly exhibit acoustic emotion which is incongruous to

[1]Code and data: https://github.com/yuwchen/CREMA-ASIS

semantic sentiment in real-world scenarios (Table 1). Specifically, a happy vocal tone paired with negative semantics often arises when speakers use a cheerful tone to complain, mask frustration, or make jokes. A disgusted or angry tone with positive semantic content may be delivered when praise or gratitude is expressed mockingly or as an ironic commendation. Lastly, speakers may convey positive messages using a sad tone to conceal personal distress and avoid causing concern for the listener. We also generated data under acoustic-neutral and semantic-neutral conditions for comparison. CREMA-ASIS covers five acoustic emotion categories: happy, neutral, angry, disgust, and sad, and three sentiment categories: negative, neutral, and positive.

Acoustic emotion	Semantic sentiment	Example
Happy	Negative	The client canceled the entire project after we worked all weekend 😀💥
Disgust	Positive	Wow, you're a genius 🙄
Angry	Positive	I appreciate that 😡
Sad	Positive	I am fine 🙂

Table 1: Acoustic–semantic incongruence scenarios.

To generate controlled acoustic–semantic pairs, we required reference audio for each emotion category and sentences for each sentiment group.

Audio Reference Selection: Speech emotion datasets include acted (Burkhardt et al., 2005) and spontaneous emotions (Lotfian and Busso, 2017; Bujnowski et al., 2024). While spontaneous emotions are more natural, we consider acted emotion datasets more appropriate for generating synthetic speech signals in this study. First, many spontaneous emotion datasets are derived from TV shows, where licensing restrictions often limit the distribution of derivative works. Second, most spontaneous datasets provide only overall labels, whereas generating audio references requires labels that specifically reflect the acoustic signal, independent of the semantic content. In contrast, acted datasets convey emotion solely through acoustic expression, keeping the semantic content neutral. In addition, acted emotions are often more exaggerated, making them easier for models to recognize. Failure to correctly classify speech synthesized from acted emotions indicates an even greater challenge in recognizing speech from spontaneous emotions. Therefore, we opted for an acted emotion dataset. Among these, CREMA-D [2] was selected due to its relatively large number of unique speakers. Noticing some non-stationary noise in the dataset, we applied a speech enhancement model (Chao et al., 2024) to improve audio quality before using the samples as references in the TTS system.

Sentence Selection: We did not use a Large-Language Model (LLM) to generate sentences because it often produces similar structures and repetitive words especially when more than a hundred of sentences are needed (Holtzman et al., 2020; Yao et al., 2025). Instead, we used sentences from the GoEmotions[3] (Demszky et al., 2020) dataset, which are collected from Reddit and include human-annotated emotion categories. We focused on categories that express opinions, as acoustic–semantic incongruence often arises when people do not want to express their opinions directly to mitigate social tension, avoid offending others, or employ humor and irony (Brown and Levinson, 1987). For positive semantic sentiment, we selected the *approval*, *admiration*, and *gratitude* categories. For negative semantic sentiment, we selected *annoyance*, *disapproval*, and *disappointment*. Sentences in the *neutral* category are also included to serve as control samples. We further used GPT-4o (Hurst et al., 2024) to filter the sentences based on the following criteria: (1) The sentence does not express the target sentiment. (2) The sentence contains content that is strongly racist, very offensive, or inappropriate for public use. (3) The sentence includes words or expressions that people typically would not say aloud (i.e., phrases that exist only in written form).

2.1.2 Acoustic Filtering of Generated Speech

We applied an acoustic-based emotion recognition (AER) model (Yang et al., 2021) to identify and filter out samples whose conveyed emotion is unlikely to match the acoustic label. Although AER has limitations, it is effective in removing clear mismatches, e.g., when an angry target is rendered as sounding clearly happy, thereby improving the dataset's overall emotional consistency.

3 Experimental Settings

3.1 Synthetic Data Generation

We use the GoEmotions dataset as the source of spoken content, randomly selecting a subset to limit

[2] Open Data Commons Attribution License (ODC-By) v1.0
[3] Apache License 2.0

experimental complexity and keep the size of synthetic data manageable. The resulting dataset consists of 2,255 negative, 1,944 neutral, and 2,267 positive sentences. These sentences are randomly split into training, validation, and test sets, and are randomly sampled during synthetic data generation. CREMA-D is used as the acoustic reference and consists of 7,442 speech clips produced by 91 actors. Each actor performs speech expressing basic emotional states, while the spoken content is restricted to 12 emotionally neutral sentences. We follow the original training, validation, and test splits provided with the dataset. The CREMA-D dataset is processed using SEMamba (Chao et al., 2024). The AER is from Hugging Face [4]. Because AER is imperfect, we retained not only samples whose recognized emotion matches the target but also those with perceptually similar emotion categories (Appendix A.4). After acoustic filtering, we apply *whisper-base.en* (Radford et al., 2023) to transcribe the generated speech and compare it with the specified spoken content. The resulting word error rate (WER) of 0.1 indicates that IndexTTS2 can reliably generate speech that closely matches the intended text in most cases. To further ensure evaluation reliability, we exclude samples with a WER greater than 0.5 in the test set.

3.2 Supervised Fine-Tuning

Three well-known LALMs were evaluated in this study: Qwen2-Audio (Chu et al., 2024) (*Qwen2-Audio-7B-Instruct)*, Kimi-Audio (Ding et al., 2025) (*Kimi-Audio-7B-Instruct*), and Audio-Flamingo3 (Ghosh et al., 2025)(*audio-flamingo-3-hf*). In addition to CREMA-ASIS, we include the original acoustic-neutral pairs from the CREMA-D and MELD (Poria et al., 2019) datasets during SFT. MELD, collected from TV series, exposes the model to more naturalistic emotional expressions, enhancing its generalizability. We use emotion and sentiment labels in MELD as acoustic emotion and semantic sentiment targets. Including these helps prevent the fine-tuned model from forgetting emotion categories it can initially recognize and from focusing solely on categories present in the synthetic data. SFT is performed using LoRA (Hu et al., 2022; Mangrulkar et al., 2022), with parameter settings provided in Appendix A.5.

[4] firdhokk/speech-emotion-recognition-with-openai-whisper-large-v3

3.3 Evaluation Metric

We report acoustic accuracy (Acc_{acou}, emotion classification), semantic accuracy (Acc_{sem}, sentiment classification), and dual-condition accuracy (Acc_{dual}), which requires both predictions to be correct for a single instance. In addition to CREMA-ASIS and MELD test set, we evaluate generalizability on LISTEN_full (Chen et al., 2026), where samples are annotated with either acoustic or semantic emotion categories. We use its test split[5]. Since LISTEN_full is collected from several open-source emotion datasets, we remove samples originating from CREMA-D and MELD to avoid domain overlap with our training data. Labels are manually mapped to the categories used in our dataset, and samples with ambiguous mappings are excluded from evaluation (details in Appendix A.7). After filtering, we retain 1,123 samples for acoustic emotion analysis and 651 samples for semantic sentiment analysis.

3.4 Layer-wise Acoustic-semantic Probing

To better understand how SFT affects acoustic vs. semantic information processing, we conduct layer-wise linear probing experiments (Alain and Bengio, 2017) on the three LALMs from Section 3.2 using CREMA-ASIS. We evaluate the discriminative power of hidden states across model depth for acoustic emotion (Acc_{acou}) and semantic sentiment (Acc_{sem}), and analyze the acoustic-semantic gap $G = |Acc_{\text{acou}} - Acc_{\text{sem}}|$. We compute changes between base and SFT models as $\Delta G = G^{\text{SFT}} - G^{\text{base}}$ (negative values indicate gap reduction), $\Delta Acc_{\text{acou}} = Acc_{\text{acou}}^{\text{SFT}} - Acc_{\text{acou}}^{\text{base}}$, and $\Delta Acc_{\text{sem}} = Acc_{\text{sem}}^{\text{SFT}} - Acc_{\text{sem}}^{\text{base}}$.

For each model, we extract representations from the audio encoder (Whisper), multi-modal projector, six LLM layers spanning early, middle and late depth ([1, 2, 9, 17, 25, 32] for Qwen2-Audio's 32 layers, [1, 2, 7, 14, 21, 28] for Kimi-Audio and Audio-Flamingo3's 28 layers), and Kimi-Audio's MIMO module. After mean pooling over the sequence dimension, we train linear classifiers to independently predict one of 5 acoustic or 3 semantic categories. Linear probes ensure results reflect information encoded in hidden states rather than classifier capacity. Training parameter settings are given in Appendix A.6.

[5] https://huggingface.co/datasets/VibeCheck1/LISTEN_full/blob/main/data/test-00000-of-00001.parquet

4 Results

4.1 CREMA-ASIS Dataset

4.1.1 Dataset Statistics and Composition

Table 2 presents CREMA-ASIS statistics: acoustic–semantic incongruous, congruous, and neutral-associated cases (where either or both the acoustic or semantic label is neutral). The incongruous pairs in which semantic sentiment opposes acoustic valence are challenging and common in real-world scenarios but underexplored in prior work. Congruous samples align acoustic and semantic cues, making them easier to recognize. Samples with exactly one neutral modality fall between incongruous and congruous, producing partially misaligned signals. Including congruous and neutral-associated samples helps prevent LALM from assuming acoustic and semantic cues are always inconsistent. Although the dataset distribution is slightly uneven due to post-processing, each category contains enough samples for effective SFT, so strict class balance was not enforced.

		Train	Val	Test	Total
Sample size		65534	6146	5879	77559
# of Actors		63	14	14	91
Acoustic	Semantic				
Acoustic-semantic Incongruous					
disgust	positive	4524	437	373	5334
angry	positive	6434	575	513	7522
sad	positive	3821	334	330	4485
happy	negative	3919	376	358	4653
Acoustic-semantic Congruous					
disgust	negative	4993	485	382	5860
angry	negative	7100	618	549	8267
sad	negative	3764	302	334	4400
happy	positive	3449	376	354	4179
Neutral-associated					
disgust	neutral	2814	284	282	3380
angry	neutral	3790	379	388	4557
happy	neutral	2057	217	258	2532
sad	neutral	4031	384	385	4800
neutral	negative	6108	534	530	7172
neutral	neutral	3230	333	357	3920
neutral	positive	5500	512	486	6498

Table 2: CREMA-ASIS Dataset statistics.

4.1.2 Subjective Human Validation of Acoustic Emotion Control

We randomly sampled 150 instances (10 per category) from the test set for a subjective evaluation of the TTS system's ability to reproduce the reference audio's acoustics. Three evaluators performed two tasks: rating acoustic sentiment on a scale from –1 (negative) to 1 (positive) and assigning an acoustic emotion category. Figure 2 shows violin plots of the intended emotions of the reference audio and their corresponding subjective sentiment scores, indicating that median perceived sentiment aligns with the intended acoustic valence (e.g., lowest for angry, highest for happy). Accuracy of acoustic categorical emotion recognition, computed using majority voting, is 64.6% with inter-annotator agreement Krippendorff's α 0.54. Figure 3 shows the intended-versus-perceived emotion confusion matrix of CREMA-ASIS and that reported from CREMA-D paper. The confusion patterns in our validation closely follow those observed in the original data. In particular, the neutral and disgust categories remain challenging: emotions with weak acoustic cues are often perceived as neutral and disgust is inherently difficult to recognize. In addition, when we listened to samples in which the perceived emotion diverged from the intended emotion of the reference audio, the synthetic and reference audio remained acoustically similar. Overall, these results suggest that the TTS system can reliably produce samples with controlled acoustic emotion and semantic sentiment.

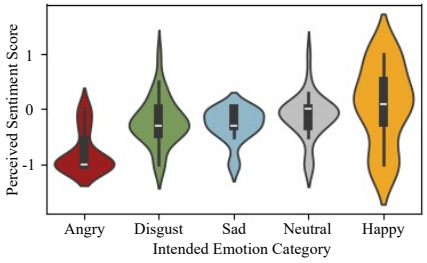


Figure 2: Results of subjective acoustic sentiment task.

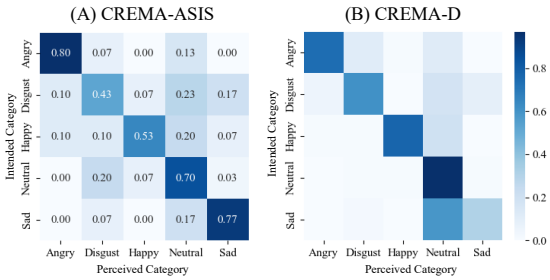


Figure 3: Confusion matrices of intended versus perceived emotions, showing similar patterns despite matrix B being based on audio-visual signals and matrix A on acoustic signals only. B is replotted from Figure 9 in (Cao et al., 2014); *fear* is excluded to match CREMA-ASIS categories, and numerical values are omitted as they were computed with fear included.

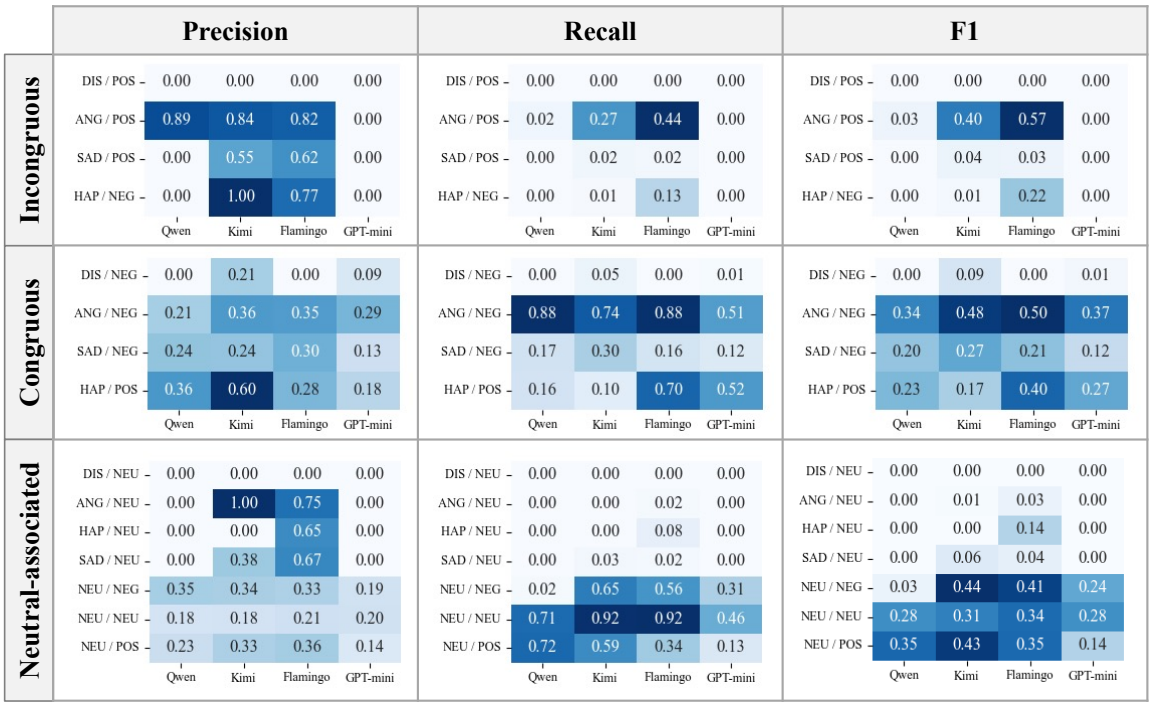


Figure 4: LALM performance across acoustic emotions and semantic sentiment pairs. Each figure label is an abbreviation representing the corresponding acoustic emotion category and semantic sentiment. For example, DIS/POS indicates acoustic disgust and semantic positive.

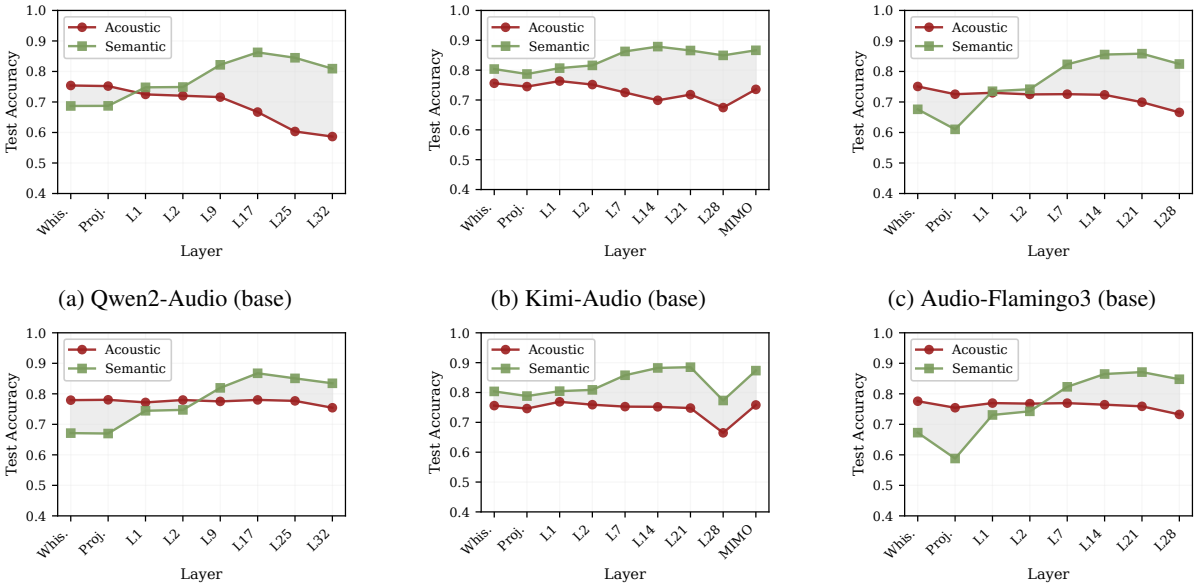
4.2 LALM Performance Across Acoustic Emotions and Semantic Sentiments

We examined dual-condition performance of LALMs, including Qwen2-Audio, Kimi-Audio, Audio-Flamingo3, and GPT-audio-mini across acoustic emotion category and semantic sentiment settings (Figure 4). These results demonstrate that incongruous cases pose a significant challenge for existing LALMs, which rarely produce predictions in which acoustic and semantic are incongruous, resulting in very low recall and, consequently, very low F1 scores. Incongruous predictions are made only when both acoustic and semantic cues are strong, resulting in relatively high precision. Among the four incongruous scenarios, angry–positive performs best, consistent with previous listening results showing the highest accuracy for angry acoustic cues. Compared to incongruous cases, higher recall observed in congruous cases reflects that LALMs are more likely to produce predictions with consistent semantic and acoustic cues, generally resulting in higher F1 scores across most conditions. As judged by human listeners, disgust is the most difficult emotion to identify, probably because its emotion cues are subtle. For neutral-associated conditions, LALMs struggle particularly with the semantic-neutral setting, where the acoustic signal carries emotion but the semantic content is neutral. In most such cases, the models ignore the acoustic cues and predict neutral–neutral, suggesting a strong bias toward semantic information. Further evidence comes from the acoustic-neutral setting, where models generally achieve higher performance. Overall, results indicate that both acoustic–semantic incongruent cases and semantic-neutral cases are particularly challenging for LALMs.

4.3 Impact of SFT on LALM Behavior

4.3.1 Layer-wise Acoustic-semantic Probing

We performed layer-wise linear probing of LALMs before and after SFT. Across all base models (Figure 5, top row), acoustic and semantic probing accuracy diverge with depth: semantic accuracy (green) rises above 85% in deeper layers, while acoustic accuracy (red) drops from approximately 75% at the Whisper encoder to approximately 67% (Kimi-Audio, Audio-Flamingo3) and 59% (Qwen2-Audio) at the final LLM layer, suggesting a progressive representational shift from acoustic cues to more semantic-oriented features. After SFT (Table 3 and Figure 5, bottom row), acoustic performance improves most where base models were weakest, with gains concentrated in later layers. Im-



(a) Qwen2-Audio (base) (b) Kimi-Audio (base) (c) Audio-Flamingo3 (base)

(d) Qwen2-Audio-SFT (e) Kimi-Audio-SFT (f) Audio-Flamingo3-SFT

Figure 5: Layer-wise probing accuracy for acoustic emotion (red) and semantic sentiment (green) across model depth using mean pooling over hidden states. Top row: base models showing acoustic-semantic divergence in deeper layers. Bottom row: SFT models showing reduced divergence with improved acoustic accuracy.

provements range from 5.3 percentage points (pp) (Kimi-Audio L14) to 17.4 pp (Qwen2-Audio L25), while early-layer gains remain comparatively modest, resulting in more uniform acoustic accuracy throughout layers.

Semantic accuracy after SFT remains mostly similar with changes within ± 2.5 pp across nearly all layers and models, with a notable exception in Kimi-Audio L28 (-7.6 pp). As this layer precedes Kimi-Audio's MIMO module, we hypothesize that the additional module introduces a distribution shift (details in Appendix C.2). This appears unique to Kimi-Audio. Since acoustic gains consistently exceed semantic changes (excluding Kimi-Audio L28), the reduced gap ΔG in Table 3 appears driven primarily by improved acoustic processing rather than semantic degradation.

These findings suggest that, after SFT on CREMA-ASIS, acoustic emotion information becomes more linearly decodable from deeper LLM representations while semantic sentiment remains comparably accessible, leading to a reduced acoustic-semantic gap across layers. Because probing reflects the information available in intermediate representations rather than the model's decision process, our analysis provides diagnostic evidence that SFT alters the representational balance between acoustic and semantic cues, rather than direct evidence about how the model uses these cues.

Layer	$\Delta Acc_{\text{acou}} \uparrow$	$\Delta Acc_{\text{sem}} \uparrow$	$\Delta G \downarrow$
Qwen2-Audio			
L1	0.047	-0.004	0.004
L17	0.113	0.005	-0.108
L25	0.174	0.006	-0.168
L32	0.168	0.025	-0.143
Kimi-Audio			
L1	0.005	-0.002	-0.008
L14	0.053	0.003	-0.050
L28	-0.010	-0.076	-0.067
MIMO	0.023	0.007	-0.016
Audio-Flamingo3			
L1	0.040	-0.005	0.034
L14	0.041	0.010	-0.032
L28	0.067	0.023	-0.043

Table 3: Impact of SFT on layer-wise accuracy and the acoustic-semantic gap G across models. Δ represents changes (SFT $-$ base). Negative ΔG values indicate reduced gap.

Complete probing results for mean pooling and last hidden state pooling, as well as analysis of the Kimi-Audio outlier are presented in Appendix C.

4.3.2 Performance on CREMA-ASIS Test Set

Table 4 shows the performance of LALMs on CREMA-ASIS test set. Before SFT, Audio-Flamingo3 consistently outperforms others across all evaluated tasks, demonstrating particularly strong semantic understanding. Despite this, all models show difficulty in correctly predicting both semantic and acoustic labels simultaneously, result-

ing in low Acc_{dual}. The relatively low Acc_{acou} of GPT-audio-mini may be primarily attributable to frequent audio processing failures. Specifically, the model fails to process 928 out of 5,879 samples, and these cases are treated as incorrect during evaluation, thereby substantially degrading its overall performance. After SFT, all open-source LALMs exhibit significant performance improvements. To verify whether SFT on CREMA-ASIS affects the models' original capabilities, we also evaluate transcription performance. Results show that WER of the transcription task for LALMs remains stable at approximately 0.07, indicating no noticeable degradation in transcription ability.

	Acc_{acou}	Acc_{sem}	Acc_{dual}
GPT-audio-mini	0.174	0.716	0.153
Qwen2-Audio	0.356	0.669	0.207
Kimi-Audio	0.421	0.722	0.285
Audio-Flamingo3	0.426	0.780	0.320
Qwen2-Audio-SFT	0.783	0.876	0.683
Kimi-Audio-SFT	0.752	0.896	0.673
Audio-Flamingo3-SFT	0.751	0.902	0.678

Table 4: Performance on CREMA-ASIS test set.

4.3.3 Evaluation on a Joint Setting

Unlike the prior section, evaluating LALMs using the same task-specific prompt as in SFT, this assesses the model under a joint recognition setting. LALMs are prompted to perform a single task without explicit specification of the modality or cue to attend to, aiming to assess whether SFT affects the models' general emotion and sentiment recognition and their robustness to prompts that differ from those used during SFT. In each run, the model is prompted with either "Identify the emotion of the given audio." or "Identify the sentiment of the given audio.", without imposing explicit acoustic or semantic constraints. Table 5 reports results on the MELD dataset, showing that emotion recognition performance improves across all models. However, for Audio-Flamingo3, sentiment performance decreases despite improved emotion recognition, perhaps due to the relatively limited linguistic diversity in CREMA-ASIS constraining a model originally possessing strong semantic understanding.

[6]The difference between our Kimi scores and those reported in the original paper is due to prompt variation: the original study outputs category codes (A, B, C), while we output the actual emotion labels (happy, sad, etc.).

	Emotion	Sentiment
Qwen2-Audio	0.518	0.545
Qwen2-Audio-SFT	**0.529**	**0.585**
Kimi-Audio	0.342[6]	0.476
Kimi-Audio-SFT	**0.536**	**0.542**
Audio-Flamingo3	0.517	**0.629**
Audio-Flamingo3-SFT	**0.532**	0.603

Table 5: Performance on the MELD test set (Acc).

4.3.4 Evaluation on Out-of-domain Dataset

We evaluate the generalizability of the fine-tuned models on the out-of-domain LISTEN dataset (Table 6). Results indicate that, after SFT on CREMA-ASIS, both Qwen2-Audio and Audio-Flamingo3 achieve improved acoustic task performance. Performance degradation of Kimi-Audio's Acc_{acou} may be attributed to the original Kimi model having previously been SFT on RAVDESS (Livingstone and Russo, 2018), SAVEE (Jackson and Haq, 2014), and IEMOCAP (Busso et al., 2008) datasets, which also constitute part of the LISTEN evaluation data. These were not included during our subsequent SFT stage, leading to a shift away from the fitted dataset and thus reduced performance on LISTEN. For the semantic task, SFT Qwen2-Audio also improved semantic performance, while Kimi-Audio and Audio-Flamingo3 decreased. However, we note that Acc_{sem} in Table 6 is reported using the labels provided by the LISTEN dataset, which may not reflect transcript-level sentiment. We provide further analysis of Acc_{sem} for this dataset in Appendix B.

	Acc_{acou}	Acc_{sem}
Qwen2-Audio	0.515	0.409
Qwen2-Audio-SFT	**0.570**	**0.470**
Kimi-Audio	**0.516**	**0.565**
Kimi-Audio-SFT	0.481	0.330
Audio-Flamingo3	0.607	**0.508**
Audio-Flamingo3-SFT	**0.620**	0.479

Table 6: Evaluation on out-of-domain data.

5 Conclusion

We introduce CREMA-ASIS, a synthetic dataset for evaluating LALMs' bias toward acoustic and semantic affective cues in speech. Subjective tests show that TTS models can reproduce the acoustic emotion of reference audio, enabling controlled speech generation. Using this dataset, we reveal that LALMs struggle with acoustic–semantic incongruence and exhibit a tendency to underweight acoustic cues when semantic content is neutral.

Layer-wise probing analysis shows that deeper layers exhibit reduced acoustic decodability while semantic decodability improves, widening the gap between modalities. Fine-tuning on our synthetic data mitigates this by improving the availability of acoustic emotion information in deeper layers while largely preserving semantic sentiment representations, reducing modality disparity. We hope this work draws attention to the fact that speech may inherently encode distinct acoustic and semantic information, and that existing LALMs have limited ability to disentangle the two. We further show that TTS-synthesized data may offer a promising direction to mitigate this limitation.

Limitations

Methodology limitations: The primary goal of this study is to provide a baseline evaluation of LALMs on the acoustic–semantic inconsistent task, and certain limitations follow from this scope. While we fine-tune all LALMs under the same SFT hyperparameter settings for fair comparison, we acknowledge that these settings may not be optimal for each model. Alternative adaptation strategies beyond standard SFT could further improve performance but are not the focus of this work. Additionally, linear probing reveals the accessibility of information in intermediate representations but not whether models use it during inference.

Dataset quality and evaluation limitations: We acknowledge several limitations of the proposed dataset. First, except for the 150 samples used in the subjective evaluation to verify the TTS model's ability, the labels correspond to intended rather than perceived emotion, as they are derived from actors who aimed to express specific emotions. Intended and perceived emotion may diverge, as perception can be influenced by both listener subjectivity and semantic content. Second, although our listening task, conducted with evaluators blinded to labels and audio references, revealed patterns consistent with the original references, our synthetic data may contain imperceptible artifacts that could affect learning. Lastly, two of the three evaluators are co-authors, with the third not being involved in dataset construction. Agreement between the independent annotator and co-authors was consistent, but a fully independent panel would further strengthen confidence.

Generalizability and scope limitations: Our dataset is generated from acted emotion references, which are often exaggerated and easier to identify than spontaneous speech. To assess generalizability, we evaluate SFT models on MELD (TV dialogue) and LISTEN (acted and spontaneous subsets), finding some transferability, but potential limitations remain. Notably, our results show that LALMs struggle with acoustic-semantic inconsistencies even for synthesized acted speech, so we view this as a necessary first step towards handling real-world spontaneous emotion. Although synthetic data can be generated without limit, we observe overfitting before a full epoch, suggesting diminishing returns from redundant training samples. Lastly, CREMA-ASIS covers only a subset of GoEmotions sentences and basic emotion categories. Future work could increase sentence diversity and explore additional emotion categories.

Acknowledgments

This work was supported by the National Institutes of Health under Grant No. 1R01AG081928-01. We thank other members of the Columbia School of Nursing for their fruitful discussion, including Pallavi Gupta and Zhihong Zhang, as well as members of VNS Health, including Sasha Vergez and Margaret McDonald. We also thank Amazon Web Services (AWS) for providing computing credits that supported the development of this work.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *Proceedings of the International Conference on Learning Representations (Workshop Track)*.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Paweł Bujnowski, Bartłomiej Kuźma, Bartłomiej Paziewski, Jacek Rutkowski, Joanna Marhula, Zuzanna Bordzicka, and Piotr Andruszkiewicz. 2024. SAMSEMO: New dataset for multilingual and multimodal emotion recognition. In *Proceedings of Interspeech*, pages 2925–2929.

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech. In *Proceedings of Interspeech*, pages 1517–1520.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S

Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Rong Chao, Wen-Huang Cheng, Moreno La Quatra, Sabato Marco Siniscalchi, Chao-Han Huck Yang, Szu-Wei Fu, and Yu Tsao. 2024. An investigation of incorporating Mamba for speech enhancement. In *Proceedings of the IEEE Spoken Language Technology Workshop*, pages 302–308. IEEE.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. 2026. Do audio LLMs really listen, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pages 5848–5877.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Pedro Corrêa, João Lima, Victor Moreno, Lucas Ueda, and Paula Dornhofer Paro Costa. 2026. Evaluating emotion recognition in spoken language models on emotionally incongruent speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 17952–17956.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 4040–4054.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, and 20 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. 2025. UniC: a dataset for emotion analysis of videos with multimodal and unimodal labels. *Language Resources and Evaluation*, 59(3):2857–2892.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the International Conference on Learning Representations*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 399 others. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Philip Jackson and Sanaul Haq. 2014. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations*.

Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *Proceedings of the International Conference on Learning Representations*.

Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. 2025. When audio and text disagree: Revealing text bias in large audio-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4878–4888.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech processing universal performance benchmark. In *Proceedings of Interspeech*, pages 1194–1198.

Junchi Yao, Shu Yang, Jianhua Xu, Lijie Hu, Mengdi Li, and Di Wang. 2025. Understanding the repeat curse in large language models from a feature perspective. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7787–7815.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2026. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35139–35148.

A Experiment Parameter Settings

A.1 LALM Prompt

Figure 6 shows the prompt used for SFT and evaluation of the LALMs. Notice that the prompt includes the acoustic emotion category fear while it does not appear in our CREMA-ASIS dataset. We found that the fear recordings in CREMA-D often lacked naturalness and failed to convey the intended acoustic cues of fear. Therefore, we decided to exclude it from the synthetic dataset. However, we still included the fear category in the prompt because datasets, such as MELD, contain it. This ensures that our evaluation of MELD remains consistent with previous studies, and that the raw LALM performance matches that reported in the original LALM papers.

```
You are an expert in annotation. Analyze the given audio and identify both its
**acoustic** emotion and **semantic** sentiment.

Definitions:
- *Acoustic emotion*: the emotion conveyed by voice, regardless of word meaning
or spoken content.
- *Semantic sentiment*: the sentiment expressed by the literal meaning of the words,
regardless of how they are spoken.

Respond strictly in the following JSON format (no explanations, no extra text):
{ "acoustic_emotion": "<emotion inferred solely from acoustic cues: happy | sad |
angry | fear | disgust | surprise | neutral | calm >",
  "semantic_sentiment": "<sentiment inferred solely from semantic cues: positive |
negative | neutral >"}
```

Figure 6: LALM prompt, instructing LALM to simultaneously predict acoustic emotion categories and semantic sentiment labels, with acoustic predictions including categories beyond those in CREMA-ASIS.

A.2 Sentence Selection Prompt

Figure 7 shows the prompt used for text sentence filtering.

```
Input: a single sentence.
Task: Return only two things, separated by a period:

a. True or False — depending on whether any of the following conditions are met
(1) The sentence does not express a neutral sentiment.
(2) The sentence contains content that is strongly racist, very offensive, or
inappropriate for public use.
(3) The sentence includes words or expressions that people typically would not say
aloud (i.e., phrases that exist only in written form).
b. If True, also return the reason number (1, 2, or 3) that applies.
If True, list all applicable reasons as comma-separated numbers after the period.
If False, just return False.

Input:
```

Figure 7: Prompt for sentence selection, where *neutral* is replaced with *positive* or *negative* when filtering sentences with corresponding sentiment.

A.3 IndexTTS2 Generation Settings

Synthetic speech is generated using IndexTTS2 (Zhou et al., 2026) in a zero-shot, speaker-conditioned setting. For each sample, a reference audio clip from the original CREMA-D corpus is used as the speaker prompt for voice cloning, and the target sentence from the filtered GoEmotions is passed as the text input. The audio clips are preprocessed using SEMamba (Chao et al., 2024) [7] with the *vd.pth* checkpoint for speech enhancement. We control acoustic emotions solely based on the reference voice and set the emotion vector to all zeros and disable the textual emotion prompt. Standard FP32 precision is used, and all outputs are written as WAV files.

A.4 Acoustic Filtering settings

The goal of acoustic filtering with AER is to remove samples whose conveyed emotion is clearly different from the target emotion. However, the AER model is not perfect and may fail to identify the correct emotion. If we retained only samples whose AER predictions exactly matched the target emotion, many valid samples would be removed even though the synthesized audio conveys an emotion similar to the reference audio. For example, sad audio is often identified as neutral by the AER, and it rarely predicts the sad label; even when it does, the audio often does not clearly convey sad emotion. Table 7 shows the target–prediction pairs retained in this study.

Audio Reference	AER Prediction
happy	happy, surprised, neutral
neutral	neutral, angry
sad	sad, neutral
disgust	surprised, neutral
angry	angry, surprised, neutral

Table 7: Target–prediction pairs retained.

A.5 SFT Parameter Settings

The models are fine-tuned using a rank-stabilized LoRA (rsLoRA) adapter with rank $r = 4$ and $\text{LoRA}_\alpha = 32$, applied to all linear layers in the `CAUSAL_LM` setting. We use a learning rate of 1×10^{-5} with a gradient accumulation step of 4 with LoRA dropout of 0.1. Checkpoints are saved every 2,500 steps, and the checkpoint achieving the best overall performance is selected. SFT is performed on NVIDIA L40, RTX A5500, and A100-40GB GPUs.

A.6 Layer-wise Linear Probing Training Parameters Settings

We use the data split from Table 2. We train all linear probes for 20 epochs using Adam (Kingma and Ba, 2015) with cosine annealing (Loshchilov and Hutter, 2017), grid-searching over learning rates $\{5 \times 10^{-2}, 1 \times 10^{-2}, 5 \times 10^{-3}, 1 \times 10^{-3}\}$ evaluated on the validation set. Each probe consists of a single linear layer and uses input dimensions matching the representation size (1280 for Whisper, 4096 for Qwen2-Audio, 3584 for Kimi-Audio and Audio-Flamingo3) with output dimensions of 5 and 3 for the acoustic emotion and semantic sentiment tasks respectively, which corresponds to the categories. We use no hidden layers in any of our probing experiments.

A.7 LISTEN Dataset Setup

We split the LISTEN dataset into an acoustic emotion analysis subset and a semantic sentiment analysis subset based on the question associated with each sample from the original dataset. A sample is assigned to the acoustic subset if the question asks about emotion conveyed through vocal prosody or tone (e.g., vocal expression, emotional tone of the speaker's voice), and to the semantic subset if the question asks about emotion conveyed through the literal textual content (e.g., the semantics of the spoken words). The full set of questions used for each subset is listed in Table 8.

When constructing the semantic subset, we need to be careful about a property of the LISTEN dataset. LISTEN assigns to each utterance a single emotion label taken from its source corpus, and in conditions where acoustic and lexical cues are treated as aligned, this same label is attached to the text-centric question (i.e., the LISTEN question within our semantic subset) as well. Because these source labels (e.g., in IEMOCAP) were annotated from audio or audiovisual delivery, they encode the emotion of the vocal performance rather than the semantics from the transcript alone. We further discuss the consequences of this in Appendix B.

For the semantic analysis, we map the resulting labels to positive, negative, and neutral sentiment following Table 9, and exclude samples with ambiguous mappings.

[7] https://github.com/RoyChao19477/SEMamba/blob/main/ckpts/vd.pth

Acoustic Emotion Analysis	Semantic Sentiment Analysis
What emotion is communicated through the speaker's vocal prosody?	Based solely on the text content, what emotion would you identify?
Based on the vocal expression, what emotion is the speaker feeling?	From the semantic content alone, what emotion is being expressed?
How would you describe the emotional tone of the speaker?	What emotional tone is conveyed by the literal meaning of this statement?
Listening to the voice, what emotion is being expressed?	What feeling is suggested by the meaning of these words?
What emotion is expressed in the speaker's voice?	Based on the content of this text, what emotion would the person likely be feeling?
What emotional state is reflected in the speaker's voice?	Reading this text, what emotional state does the speaker appear to be in?
What emotion does the speaker convey through their tone?	What emotion is conveyed by the words in this statement?

Table 8: Questions used to split LISTEN samples into acoustic emotion analysis and semantic sentiment analysis subsets.

Sentiment	Original Labels
positive	amusement, excitement, happiness, happy, pleasantly surprised
negative	anger, angry, annoyance, anxious, concern, concerned, confusion, contempt, disappointment, disgust, fear, frustration, ridicule, sad, sadness
neutral	calm, neutral

Table 9: Semantic sentiment label mapping from the LISTEN dataset to our setting.

B Supplementary Analysis of LISTEN Dataset

In Section 4.3.4, we observed low Acc_{sem} on LISTEN, and reduced Acc_{sem} after SFT for Kimi-Audio and Audio-Flamingo3, when scored against the labels provided by LISTEN. Upon manual inspection, this behavior appears to stem primarily from how LISTEN's labels are constructed rather than from the models themselves. LISTEN assigns each utterance a single emotion label taken from its source corpus, and attaches this label to the text-centric ("semantic") question as well. For corpora such as IEMOCAP, these labels were annotated from audio or audiovisual delivery, so for utterances whose transcript's sentiment is neutral or divergent from the vocal delivery, the text-centric label reflects the acoustic emotion rather than the sentiment expressed by the transcript. This makes these labels unsuitable targets for a purely text-based sentiment task.

Several examples within the "Matched-Emotion" for text setting (LISTEN's experiment type 2A) illustrate this mismatch. The transcript "Somebody please call 911" paired with the question "What emotion is conveyed by the words in this statement?" is labeled *happiness*, whereas the words alone convey distress or concern rather than the positive sentiment the label implies. Furthermore, "I would've loved to" and "As a matter of fact, that's perfectly true" are labeled *sadness* and *anger* respectively, while in both cases the text alone reads more neutral-to-positive. Rather than annotation errors, these examples are labeled based on the delivered acoustic emotions while being associated with a text-centric question under the assumption that lexical and acoustic emotional cues match. However, for these examples, the assumption does not hold, and the text-centric label and the transcript's semantic sentiment diverge.

To quantify this divergence, we feed the transcripts of the audios alone to a strong text-only classifier, Qwen3-32B (Yang et al., 2025), and compare its semantic sentiment predictions against LISTEN's text-centric labels using the mapping in Table 9. We treat Qwen3-32B's predictions here as a proxy-label for transcript sentiment, not as ground truth. On the three transcript examples above, Qwen3-32B predicts negative, positive, and positive, consistent with what a reader would infer from the transcript alone. However, across the full semantic subset (651 samples), Qwen3-32B's predictions agree with LISTEN's labels on only 36.9% of samples. This low agreement suggests that the text-centric labels encode acoustic affect rather than lexical. When we instead score the three LALMs' semantic predictions from Table 6 (which were inferred *from the audio*) against Qwen3-32B's *text-only* transcript-level predictions (see Table 10), their agreement with transcript-level sentiment is substantially higher than against LISTEN's labels. This effect is most pronounced for Kimi-Audio, whose agreement increases after SFT on CREMA-ASIS (0.594 to 0.731), with smaller

increases for Qwen2-Audio (0.622 to 0.662) and Audio-Flamingo3 (0.711 to 0.720). Notably, Kimi-Audio's semantic score moves in opposite directions depending only on the scoring reference. It drops under LISTEN's labels but rises under transcript-level references, even though the model's predictions are identical.

While our results suggest that SFT improves the models' alignment with transcript-level sentiment rather than degrading it, establishing a true semantic sentiment gold-label would require manual human annotation of the dataset, which we leave to future work. Therefore, we treat the weaker LISTEN Acc_{sem} in Table 6 with caution.

	Agreement with Qwen3-32B
Qwen2-Audio	0.622
Qwen2-Audio-SFT	**0.662**
Kimi-Audio	0.594
Kimi-Audio-SFT	**0.731**
Audio-Flamingo3	0.711
Audio-Flamingo3-SFT	**0.720**

Table 10: Agreement between each LALM's semantic predictions from audio inputs and Qwen3-32B's transcript-level sentiment predictions on LISTEN, before and after SFT on our CREMA-ASIS. Note that this is a reference-agreement measure, not accuracy against verified ground truth.

C Supplementary Layer-Wise Probing Results

C.1 Complete Probing Results

Table 11 shows the complete results of our layer-wise probing experiments across three models. In addition to the changes between base and SFT models (columns denoted with Δ), we also report the acoustic emotion ($Acc^{\text{base}}_{\text{acou}}$, $Acc^{\text{SFT}}_{\text{acou}}$), semantic sentiment accuracies ($Acc^{\text{base}}_{\text{sem}}$, $Acc^{\text{SFT}}_{\text{sem}}$), and acoustic-semantic gaps (G^{base}, G^{SFT}) for each layer probed across base and SFT models.

Table 11 also presents the same experiment using the last hidden state as features in place of mean pooled hidden states while Figure 8 shows the layer-wise trajectories for both acoustic emotion and semantic sentiment accuracies. We observe similar behavior for the last hidden state pooling as mean pooling: for base models, the acoustic-semantic gap widens the deeper we propagate through the layers as acoustic accuracies decrease while semantic accuracies increase. Similar to before, we observe that SFT improves the acoustic emotion accuracy the most where the base models performed the worst while the changes are modest in earlier layers, leading to more uniform acoustic performance throughout the layers. Again, semantic accuracy changes comparatively less with SFT, leading to gap reductions being largely driven by acoustic gains.

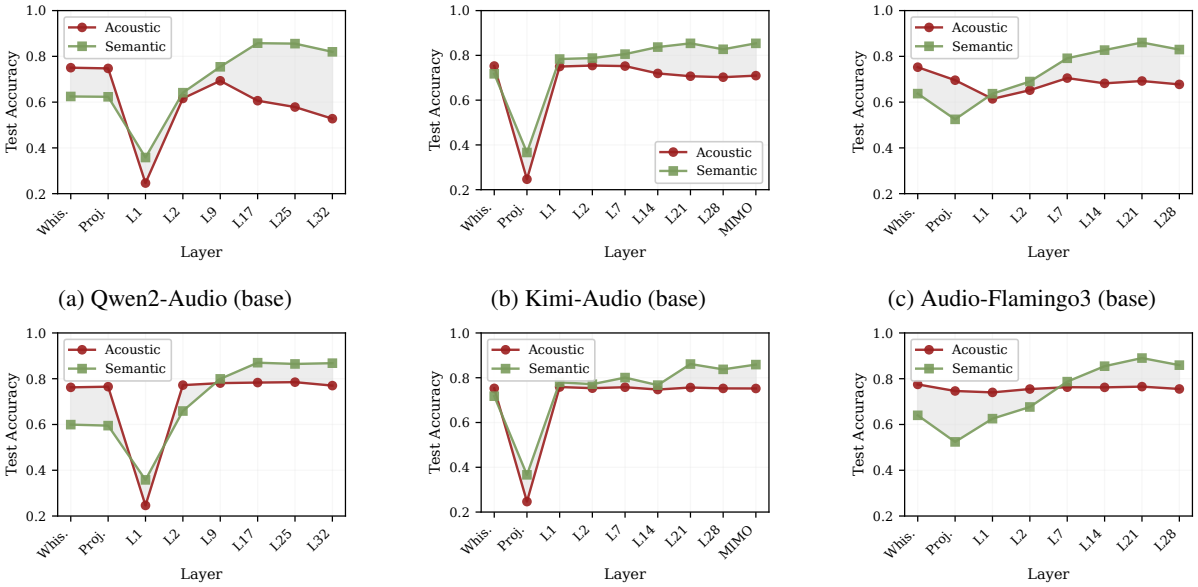
C.2 Kimi-Audio L28 Analysis

To further investigate the accuracy drop of Kimi-Audio L28 from Section 4.3.1, we extend our probing experiment to layers 22–28 of Kimi-Audio's LLM backbone and layers 1–5 of the MIMO module. We run all experiments with 5 different random seeds and present the mean accuracies in Table 12 and Figure 9. We omit reporting standard deviations as they are all < 0.005. For the base model, we see a steady decrease in acoustic accuracy through each layer of the LLM backbone followed by a sharp increase at the MIMO module, reaching levels comparable to early LLM layers. The same pattern can be observed for the SFT model, but this time, the drop at Kimi-Audio L28 is significantly sharper. Since the LLM backbone was initially pre-trained on text-only tasks, we hypothesize that there is a distribution mismatch between it and the MIMO module (consisting of a 1B audio transformers), which leads SFT to exacerbate the mismatch. We leave further investigation for future work.

Similar to Kimi-Audio L28, we also observe dips in accuracy with the last hidden state pooling for Qwen2-Audio L1, Kimi-Audio Projector (Table 11 and Figure 8) and Kimi-Audio MIMO 2 (Table 12 and Figure 9) at a much larger magnitude. For Qwen2-Audio L1 and Kimi-Audio Projector, SFT has no effect on the accuracy, while for Kimi-Audio L28 and MIMO 2 SFT either reduce the accuracy or barely increase it by 0.2 percentage points. While not conclusive, we note that all drops happen at the very start or end of the LLM layers, which may indicate domain mismatches between the modules. We leave further investigation for future work.

Layer	Pooling	$Acc_{\text{acou}}^{\text{base}}$	$Acc_{\text{sem}}^{\text{base}}$	G^{base}	$Acc_{\text{acou}}^{\text{SFT}}$	$Acc_{\text{sem}}^{\text{SFT}}$	G^{SFT}	$\Delta Acc_{\text{acou}} \uparrow$	$\Delta Acc_{\text{sem}} \uparrow$	$\Delta G \downarrow$
Qwen2-Audio										
Whisper	Mean	0.754	0.687	0.067	0.779	0.671	0.108	0.025	-0.016	0.041
Projector	Mean	0.752	0.687	0.065	0.780	0.670	0.111	0.028	-0.017	0.046
L1	Mean	0.725	0.748	0.023	0.772	0.744	0.027	0.047	-0.004	0.004
L2	Mean	0.721	0.749	0.029	0.780	0.747	0.032	0.059	-0.002	0.004
L9	Mean	0.716	0.822	0.106	0.775	0.820	0.044	0.059	-0.002	-0.062
L17	Mean	0.667	0.863	0.196	0.780	0.867	0.087	0.113	0.005	-0.108
L25	Mean	0.603	0.845	0.242	0.777	0.851	0.074	0.174	0.006	-0.168
L32	Mean	0.587	0.809	0.223	0.754	0.835	0.080	0.168	0.025	-0.143
Whisper	Last	0.750	0.625	0.126	0.762	0.599	0.163	0.012	-0.025	0.037
Projector	Last	0.747	0.623	0.124	0.765	0.595	0.170	0.018	-0.028	0.046
L1	Last	0.246	0.358	0.112	0.246	0.358	0.112	0.000	0.000	0.000
L2	Last	0.616	0.641	0.025	0.772	0.659	0.113	0.156	0.018	0.088
L9	Last	0.693	0.754	0.061	0.781	0.799	0.019	0.087	0.045	-0.043
L17	Last	0.606	0.858	0.251	0.783	0.870	0.087	0.177	0.012	-0.164
L25	Last	0.579	0.856	0.277	0.785	0.865	0.080	0.207	0.009	-0.198
L32	Last	0.528	0.820	0.292	0.770	0.868	0.098	0.242	0.048	-0.194
Kimi-Audio										
Whisper	Mean	0.756	0.804	0.048	0.756	0.804	0.048	0.000	0.000	0.000
Projector	Mean	0.745	0.787	0.042	0.746	0.788	0.042	0.001	0.001	0.000
L1	Mean	0.763	0.807	0.043	0.769	0.805	0.036	0.005	-0.002	-0.008
L2	Mean	0.752	0.816	0.064	0.759	0.809	0.050	0.007	-0.007	-0.015
L7	Mean	0.725	0.863	0.138	0.753	0.858	0.105	0.028	-0.005	-0.033
L14	Mean	0.699	0.879	0.180	0.752	0.882	0.130	0.053	0.003	-0.050
L21	Mean	0.718	0.866	0.148	0.748	0.885	0.137	0.030	0.019	-0.011
L28	Mean	0.675	0.850	0.175	0.665	0.773	0.108	-0.010	-0.076	-0.067
MIMO	Mean	0.736	0.867	0.131	0.758	0.873	0.115	0.023	0.007	-0.016
Whisper	Last	0.753	0.718	0.035	0.753	0.718	0.035	0.000	0.000	0.000
Projector	Last	0.247	0.366	0.120	0.247	0.366	0.120	0.000	0.000	0.000
L1	Last	0.750	0.784	0.034	0.759	0.780	0.021	0.008	-0.004	-0.012
L2	Last	0.755	0.788	0.033	0.753	0.771	0.018	-0.001	-0.017	-0.015
L7	Last	0.752	0.806	0.053	0.758	0.801	0.043	0.005	-0.005	-0.010
L14	Last	0.719	0.837	0.118	0.747	0.767	0.020	0.028	-0.070	-0.098
L21	Last	0.707	0.854	0.147	0.756	0.861	0.105	0.049	0.007	-0.042
L28	Last	0.703	0.827	0.125	0.752	0.837	0.085	0.050	0.010	-0.040
MIMO	Last	0.710	0.854	0.144	0.752	0.859	0.107	0.043	0.005	-0.038
Audio-Flamingo3										
Whisper	Mean	0.751	0.676	0.075	0.776	0.673	0.103	0.025	-0.003	0.028
Projector	Mean	0.726	0.610	0.115	0.754	0.588	0.166	0.029	-0.022	0.051
L1	Mean	0.730	0.735	0.005	0.770	0.731	0.039	0.040	-0.005	0.034
L2	Mean	0.725	0.742	0.017	0.768	0.743	0.025	0.043	0.001	0.008
L7	Mean	0.726	0.823	0.098	0.770	0.823	0.053	0.044	0.000	-0.044
L14	Mean	0.723	0.855	0.132	0.764	0.865	0.101	0.041	0.010	-0.032
L21	Mean	0.699	0.858	0.159	0.759	0.871	0.112	0.060	0.013	-0.047
L28	Mean	0.666	0.825	0.159	0.732	0.848	0.115	0.067	0.023	-0.043
Whisper	Last	0.753	0.637	0.116	0.775	0.640	0.135	0.022	0.003	0.020
Projector	Last	0.696	0.525	0.171	0.746	0.524	0.223	0.051	-0.001	0.051
L1	Last	0.614	0.637	0.023	0.740	0.625	0.115	0.126	-0.011	0.092
L2	Last	0.652	0.690	0.038	0.755	0.677	0.078	0.103	-0.013	0.041
L7	Last	0.705	0.792	0.087	0.763	0.787	0.024	0.058	-0.004	-0.062
L14	Last	0.682	0.827	0.145	0.762	0.855	0.093	0.080	0.027	-0.053
L21	Last	0.692	0.861	0.168	0.765	0.890	0.125	0.073	0.030	-0.043
L28	Last	0.678	0.830	0.152	0.755	0.859	0.104	0.078	0.030	-0.048

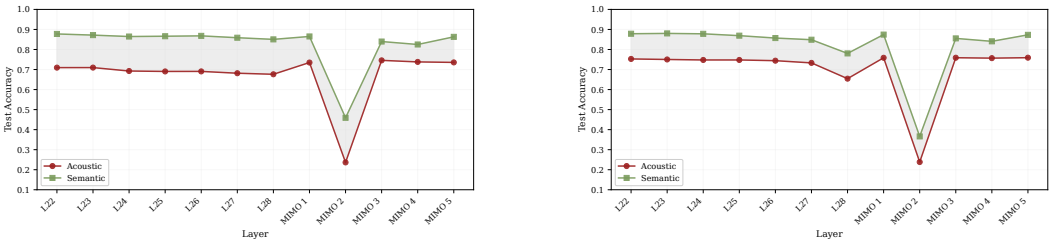
Table 11: Layer-wise acoustic emotion accuracy Acc_{acou}, semantic sentiment accuracy Acc_{sem}, and acoustic-semantic gap G between base and SFT LALMs from linear probing experiments. Δ represents changes (SFT$-$base). Negative ΔG values indicate reduced gap.



(a) Qwen2-Audio (base) (b) Kimi-Audio (base) (c) Audio-Flamingo3 (base)

(d) Qwen2-Audio-SFT (e) Kimi-Audio-SFT (f) Audio-Flamingo3-SFT

Figure 8: Layer-wise probing accuracy for acoustic emotion (red) and semantic sentiment (green) across model depth using the last hidden states. Top row: base models showing acoustic-semantic divergence in deeper layers. Bottom row: SFT models showing reduced divergence with improved acoustic accuracy.



(a) Kimi-Audio (base) (b) Kimi-Audio (Fine-tuned)

Figure 9: Layer-wise probing accuracy for acoustic emotion (red) and semantic sentiment (green) across model depth using the mean pooling for Kimi-Audio's late LLM layers and MIMO module. Left: base models. Right: SFT models. Accuracies are computed based on means from five runs with different random seeds.

Kimi-Audio									
Layer	$Acc_{\text{acou}}^{\text{base}}$	$Acc_{\text{sem}}^{\text{base}}$	G^{base}	$Acc_{\text{acou}}^{\text{SFT}}$	$Acc_{\text{sem}}^{\text{SFT}}$	G^{SFT}	$\Delta Acc_{\text{acou}} \uparrow$	$\Delta Acc_{\text{sem}} \uparrow$	$\Delta G \downarrow$
L22	0.710	0.878	0.168	0.753	0.879	0.126	0.043	0.001	-0.042
L23	0.710	0.872	0.162	0.750	0.880	0.130	0.041	0.009	-0.032
L24	0.693	0.865	0.172	0.748	0.878	0.131	0.055	0.014	-0.042
L25	0.691	0.866	0.176	0.748	0.869	0.121	0.057	0.003	-0.054
L26	0.691	0.868	0.177	0.744	0.857	0.113	0.054	-0.011	-0.065
L27	0.682	0.859	0.177	0.733	0.849	0.116	0.052	-0.010	-0.062
L28	0.676	0.851	0.174	0.654	0.780	0.126	-0.022	-0.070	-0.048
MIMO 1	0.735	0.865	0.130	0.759	0.874	0.116	0.024	0.009	-0.014
MIMO 2	0.236	0.459	0.223	0.238	0.366	0.128	0.002	-0.093	-0.095
MIMO 3	0.746	0.840	0.094	0.759	0.855	0.097	0.013	0.016	0.003
MIMO 4	0.738	0.825	0.087	0.757	0.841	0.084	0.019	0.016	-0.003
MIMO 5	0.736	0.863	0.128	0.759	0.873	0.114	0.024	0.010	-0.014

Table 12: Layer-wise acoustic emotion accuracy Acc_{acou}, semantic sentiment accuracy Acc_{sem}, and acoustic-semantic gap G between base and SFT LALMs from linear probing experiments. Δ represents changes (SFT$-$base). Negative ΔG values indicate reduced gap. Numbers are computed based on means from five runs with different random seeds.