Encoded but Not Express: Probing Subtle Speech-Conditioned Emotion in Large Audio Language Models

Jingxi Liang[1,*] **Pengze Guo**[1,*] **Zhiwen Xie**[1,2,†]
Qifeng Wang[1] **Yujia Huo**[3] **Derek F. Wong**[1,†]

[1]NLP[2]CT Lab, Department of Artificial Intelligence, University of Macau
[2]School of Computer Science, Central China Normal University
[3]School of Data Science and Information Engineering,
Guizhou Minzu University, Guizhou, China
nlp2ct.{jingxi1,pengze,qifeng}@gmail.com
zwxie@ccnu.edu.cn, huo.yujia@gzmu.edu.cn, derekfw@um.edu.mo

Abstract

Large Audio Language Models (LALMs) are increasingly used for speech understanding, yet it remains unclear whether they meaningfully utilize subtle speech-conditioned emotional cues beyond coarse acted emotions. In this work, we investigate whether the evaluated LALMs exhibit measurable sensitivity to fine-grained emotional variation in speech through controlled perturbation experiments, fine-grained emotion elicitation tasks, and representation geometry analysis. Our results show that vocal delivery variations systematically influence both model behavior and hidden representations, indicating that the evaluated LALMs encode non-trivial speech-conditioned affective information. However, the evaluated models struggle under emotionally ambiguous settings, particularly for semantically related fine-grained emotions. Representation analysis further suggests that subtle emotional structure is partially preserved in their hidden states despite weak downstream utilization. Overall, our findings suggest that the evaluated LALMs do not completely ignore subtle speech emotion cues, but instead utilize them inconsistently under ambiguous affective conditions.[1]

1 Introduction

In natural communication, emotion is often conveyed through subtle variations in speech delivery, including prosody, hesitation, timing, and vocal rhythm. These paralinguistic cues can substantially modify or even contradict literal semantic content, making them important for affective interpretation beyond coarse emotion categories. As Large Audio Language Models (LALMs) become increasingly deployed in conversational and affective applications, understanding whether they utilize such cues has become an important problem (Lian et al., 2025; Kim et al., 2024; wen Yang et al., 2026).

Prior work shows that speech input can improve affect-related tasks compared to text-only settings (Zhu et al., 2021; Min et al., 2025). However, recent studies also suggest that acoustic information weakens in later language-model layers (Yang et al., 2025a; Zhao et al., 2026), and that the LALMs examined in those studies do not consistently utilize speech delivery attributes during downstream inference (Yang et al., 2024a). Importantly, attenuation of general acoustic information does not necessarily imply the disappearance of emotionally relevant speech structure. This leaves an important unresolved question: when LALMs perform poorly on subtle emotional understanding tasks, does this reflect absent encoding, weak encoding, or failure to robustly utilize already-encoded affective information? Distinguishing between these possibilities is important because they imply different limitations and future directions for multimodal language models. To study this question, we combine controlled speech perturbation experiments, fine-grained emotion elicitation, and representation analysis across two open-source LALMs, Qwen2-Audio-Instruct(Chu et al., 2024) and Audio-Flamingo-3(Goel et al., 2025). We additionally include a closed-source model Gemini-2.5-flash as comparison(Comanici et al., 2025). We construct controlled perturbations involving emotional intensity variation and paralinguistic cues while preserving lexical content, and evaluate fine-grained emotional interpretation under varying levels of ambiguity.

Across the two evaluated open-source models, controlled perturbations produce measurable behav-

[*]Equal contribution. [†]Corresponding author.
[1]Link to the project: https://github.com/NLP2CT/Ene

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5809–5835
November 6-10, 2026 ©2026 Association for Computational Linguistics

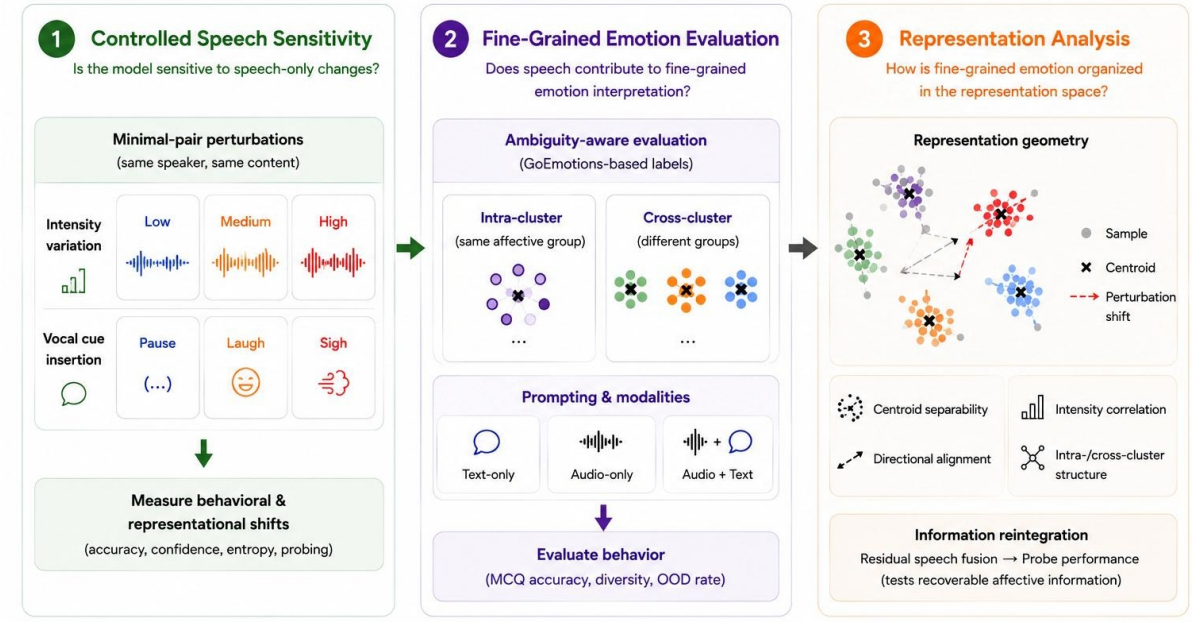


Figure 1: Overview of our evaluation framework. We examine (1) behavioral sensitivity to controlled speech perturbations, (2) fine-grained emotion recognition under ambiguity-aware prompting and modality settings, and (3) internal representation structure through geometry and speech reintegration analysis.

ioral and representational shifts. However, behavioral discrimination weakens substantially under semantically overlapping fine-grained emotional settings. Representation analyses nevertheless suggest that non-random affective organization remains partially preserved even when downstream utilization becomes inconsistent. Overall, our findings suggest that the evaluated LALMs do not completely ignore subtle speech-conditioned emotional cues, but instead exhibit a gap between internal affective organization and robust behavioral utilization under ambiguity.

Our contributions are summarized as follows:

1. We formulate the representation–utilization gap hypothesis for speech-conditioned affective processing in LALMs under ambiguity.
2. We introduce a controlled evaluation framework combining perturbation sensitivity, ambiguity-aware elicitation, and representation geometry analysis.
3. We show that, in the evaluated models, fine-grained speech-conditioned affective structure remains partially organized within hidden representations even when behavioral discrimination weakens.
4. We show that, in the evaluated models, weaker behavioral discrimination under ambiguity coexists with partially preserved affective structure.

2 Fine-Grained Emotion in Speech

Many LALMs combine pretrained audio encoders with large language models that have already acquired broad emotional and social concepts from large-scale textual corpora. Together with speech-oriented instruction tuning, this raises the question of whether LALMs internally organize speech-conditioned affective information beyond coarse emotional categories, and whether such information consistently influences downstream interpretation under ambiguity.

Why study LALMs? We investigate how speech-conditioned emotional information is preserved and utilized after multimodal fusion in the evaluated LALMs, where it can support reasoning and response generation. This interpretability question remains meaningful even if dedicated speech encoders achieve higher emotion recognition accuracy. Identifying the architectural or training factors underlying the observed behavior remains future work.

2.1 Controlled Speech Sensitivity

Emotion is conveyed not only through lexical content, but also through vocal delivery, including prosody, timing, intensity, hesitation, and other

paralinguistic cues. Before evaluating fine-grained emotional interpretation, we therefore first examine whether the evaluated LALMs exhibit measurable sensitivity to speech-conditioned variation itself. To isolate vocal delivery from lexical semantics, we construct controlled minimal-pair perturbations that preserve utterance content and speaker identity while modifying prosodic or paralinguistic properties. Fully spontaneous emotional interaction does not permit isolation of speech-conditioned affective variation because emotional state simultaneously alters lexical choice, discourse structure, pacing, and conversational context. Controlled perturbation settings are therefore necessary for causally attributable analysis of speech-conditioned sensitivity. Rather than focusing solely on recognition accuracy, we examine whether these perturbations systematically alter both model behavior and hidden representations relative to baseline conditions. This setting provides a controlled evaluation of speech sensitivity before moving to more ambiguous fine-grained emotional interpretation tasks.

2.2 Fine-Grained Emotion Settings

After establishing speech sensitivity, we evaluate whether this sensitivity extends to more naturalistic fine-grained emotional interpretation. In these samples, vocal delivery, lexical content, and context jointly contribute to affective meaning. We therefore examine speech-conditioned processing under this realistic ambiguity, complementing the lexical control provided by the perturbation experiments.

We additionally compare text-only and multimodal prompting settings to examine modality contribution. Following GoEmotions (Demszky et al., 2020), annotators may assign multiple emotional labels to the same sample. Rather than collapsing annotations into a single majority label, we retain the union of labels selected across annotators in order to preserve ambiguous and semantically overlapping affective structure. Enforcing a single definitive label would artificially remove the ambiguity central to the evaluation setting.

We partition samples according to emotional ambiguity: (1) basic emotion samples containing coarse emotional labels only, (2) mixed subtle emotion samples containing both coarse and fine-grained emotional labels, and (3) pure subtle emotion samples containing fine-grained emotional labels only. For representation analysis, we additionally separate samples into intra-cluster and cross-cluster emotional settings depending on whether emotional labels belong to semantically related or semantically distinct affective groups. Finally, we analyze hidden-state geometry to examine whether fine-grained speech-conditioned affective organization persists internally even when downstream utilization remains weak under ambiguity.

3 Methodology

We examine whether the evaluated LALMs exhibit speech-conditioned fine-grained affective sensitivity through three complementary stages: (1) evaluating model sensitivity to controlled speech perturbations, (2) evaluating fine-grained emotional interpretation under different prompting settings, and (3) analyzing whether fine-grained emotional structure remains internally organized even when behavioral discrimination weakens. Throughout the study, we combine behavioral evaluation with representational analysis using layer-wise probing and embedding-based geometry analysis.

3.1 Speech Sensitivity Evaluation

We first examine whether the evaluated LALMs exhibit measurable sensitivity to affective variation in speech. Rather than evaluating only final prediction accuracy, we analyze whether controlled speech perturbations systematically influence both behavioral outputs and intermediate representations.

Task 1: Emotion Intensity Variation. We construct matched speech pairs that preserve speaker identity, semantic content, and coarse emotion category while varying only emotional intensity.

Task 2: Paralinguistic Cue Perturbation. We additionally evaluate pauses, laughter, sighs, and hesitations to examine sensitivity to speech delivery cues beyond lexical content.

For both settings, we analyze prediction confidence, entropy, accuracy variation, and layer-wise hidden-state sensitivity using lightweight probe classifiers. After establishing controlled speech sensitivity, we evaluate whether these effects extend to fine-grained emotional interpretation under multimodal prompting settings.

3.2 Fine-Grained Emotion Elicitation

We evaluate whether the evaluated LALMs can behaviorally distinguish fine-grained emotional states under different prompting and modality settings. We consider open-ended prompting, fine-grained prompting, and constrained multiple-choice prompting. For multiple-choice evaluation,

we construct cross-cluster (CC-MCQ) and same-cluster (SC-MCQ) candidate sets following GoEmotions. Same-cluster settings contain emotionally related distractors and provide a more ambiguity-sensitive evaluation condition. Behavioral evaluation includes exact-label matching, prediction diversity, out-of-distribution response rate, and prompting variation. We additionally compare text-only and audio-text prompting settings separately.

Finally, we conduct representation geometry analysis to examine whether speech-conditioned emotional variation produces systematic organization within hidden-state space.

3.3 Representation Geometry Analysis

To analyze internal emotional organization, we perform representation-level analysis through centroid-based embedding comparison. For each emotional condition c, we compute a centroid representation:

$$\mu_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathbf{h}_i$$

where $\mathbf{h}_i$ denotes the hidden representation of sample i assigned to condition c, and N_c is the number of samples belonging to that condition.

We then examine whether multilabel emotional conditions produce systematic representational displacement relative to corresponding single-label emotional conditions. For each pair of emotional conditions, we compute centroid shift magnitude:

$$\Delta(c_a, c_b) = \|\mu_{c_a} - \mu_{c_b}\|_2$$

where c_a and c_b denote different conditions.

This analysis allows us to examine whether fine-grained emotional annotation structure corresponds to non-random geometric organization within hidden representation space. Larger centroid displacement indicates that additional emotional annotation structure is associated with systematic representational reorganization rather than remaining collapsed within coarse emotional categories.

To evaluate whether these geometric relationships exceed incidental embedding variation, we additionally construct randomized baselines by permuting emotional labels and recomputing centroid distances under shuffled conditions. Comparing observed centroid shift magnitudes against randomized label assignments allows us to estimate whether internal emotional organization exceeds chance-level structure.

Nearest-Centroid Representation Separability. We evaluate leave-one-out nearest-centroid matching at LM2. Multilabel annotations are reduced to one coarse target using the fixed-priority rule described in Appendix H. Given representation vectors x_i and coarse targets $y_i \in \mathcal{Y}$, we exclude the evaluated example from its class centroid. Let $S_k^{(-i)} = \{j : y_j = k,\ j \neq i\}$. The centroid used to predict example i is

$$c_k^{(-i)} = \frac{1}{|S_k^{(-i)}|} \sum_{j \in S_k^{(-i)}} x_j.$$

Classes with no remaining examples, or a zero centroid vector, are ineligible for that prediction. We compute cosine distance to each eligible centroid,

$$d(x_i, c_k^{(-i)}) = 1 - \frac{x_i \cdot c_k^{(-i)}}{\|x_i\| \|c_k^{(-i)}\|},$$

and assign $\hat{y}_i$ to the class with the smallest distance. Separability is

$$\mathrm{SepAcc} = \frac{1}{N} \sum_{i=1}^{N} I[\hat{y}_i = y_i].$$

This probe summarizes coarse single-target geometry, complementing the multilabel behavioral evaluation. Observed and label-permuted scores use the same leave-one-out procedure; bootstrap intervals resample the fixed correctness indicators, as detailed in Appendix H.

Fine-Grained Directional Shift Analysis. We compare intensity-induced CREMA-D shifts with EMOS-derived coarse-to-subtle emotion directions at LM2 within each model. Unlike the raw L2 analysis, we first unit-normalize pooled representations, $\widetilde{\mathbf{h}} = \mathbf{h}/\|\mathbf{h}\|$. For a medium-intensity reference s and its high- or low-intensity counterpart t, matched by speaker, utterance, and coarse emotion, we compute

$$\Delta\mathbf{h} = \frac{\widetilde{\mathbf{h}}_t - \widetilde{\mathbf{h}}_s}{\|\widetilde{\mathbf{h}}_t - \widetilde{\mathbf{h}}_s\|}.$$

Given normalized, annotation-weighted EMOS centroids $\mathbf{c}_b$ and $\mathbf{c}_e$ for a coarse emotion and a mapped subtle emotion, the reference direction is

$$\mathbf{d}_{b,e} = \frac{\mathbf{c}_e - \mathbf{c}_b}{\|\mathbf{c}_e - \mathbf{c}_b\|}.$$

Each transition pair receives one alignment score, averaged over its mapped subtle directions E_b:

$$a_i = \frac{1}{|E_b|} \sum_{e \in E_b} \Delta\mathbf{h}_i \cdot \mathbf{d}_{b,e}.$$

Positive and negative scores indicate similarly and oppositely oriented displacement, respectively; the score measures orientation rather than centroid distance. Appendix H provides centroid construction, emotion mappings, and speaker-level tests of alignment and matched intensity contrasts.

3.4 Residual Speech Reintegration

To further examine whether speech representations contain complementary affective information that may be underutilized during multimodal processing, we conduct a lightweight residual reintegration experiment on the subtle emotion recognition task.

We first extract the selected language-layer representations from the multimodal models and treat them as the baseline representation. Speech information is then reintroduced through residual integration:

$$H' = H + \alpha S$$

where H denotes the language representation, S denotes the projected speech representation after the projection layer, and α controls the contribution of speech information. We use LM2 for both Audio-Flamingo-3 and Qwen2-Audio, together with their projector representations. A lightweight multilabel probe head is trained on the resulting representations, and Macro-F1 changes are evaluated across five matched training seeds. The full protocol is given in Appendix I.

4 Experimental Setup

Dataset Curation We use CREMA-D (Cao et al., 2014) for the controlled speech sensitivity experiments in Task 1 and Task 2. CREMA-D contains professionally recorded emotional speech where speakers utter matched sentences under controlled emotion and intensity conditions. Because speaker identity and semantic content are largely preserved across perturbations, the dataset allows controlled analysis of speech delivery variation on model behavior and representations. For fine-grained emotion evaluation, we construct a multimodal subset from EMOS (Guo et al., 2026) containing approximately 5,200 samples. Unlike controlled perturbation datasets, EMOS reflects naturally entangled speech, semantic, and contextual affective information. Accordingly, we evaluate speech-conditioned processing of fine-grained emotion under realistic ambiguity, where vocal and lexical information jointly contribute. Additional dataset construction details are provided in Appendix B.

Model Selection We evaluate two open-source Large Audio Language Models (LALMs), Qwen2-Audio-Instruct and Audio-Flamingo-3, selected to cover different language backbones and audio integration designs. We additionally include Gemini-2.5-flash as a closed-source comparison to assess whether related behavioral trends also appear outside the two open-source systems.

Evaluation Protocol For behavioral evaluation, we report exact-match accuracy, prediction confidence, and prediction entropy across tasks. Under controlled perturbation settings, we analyze metric shifts between matched baseline and manipulated speech conditions in order to quantify the behavioral effect of speech variation. Except for the open-ended fine-grained emotion elicitation setting, all experiments are evaluated using constrained multiple-choice prompting following prior multimodal evaluation work (Wang et al., 2025a). For representation analysis, we extract intermediate hidden-state representations from selected model layers and perform centroid-based geometric analysis and lightweight probing experiments as described in Section 3. Prompt templates and implementation details are in Appendix C. Appendix H reports uncertainty estimates, inferential tests, and their scope.

5 Results and Analysis

Sensitivity to Speech Perturbations. Figure 3 shows that speech style perturbations, including delivery variation and inserted vocal cues, produce measurable behavioral and representational shifts across both models from audio encoder to projector and LLM layers. In contrast, aggregated emotional intensity manipulation produces substantially weaker overall effects (detailed perturbation breakdown shown in Figure 14 in Appendix G). However, separating intensity into Low (LO) and High (HI) conditions relative to Medium (MD) reveals partially opposing trends across several layers (Figure 4), particularly within audio encoder and projector components. These effects become weaker in later LLM layers, suggesting partial attenuation during multimodal processing. Although perturbation magnitudes remain relatively small, non-zero emotion shift rates indicate that controlled speech variation systematically alters internal emotion-

related representations while preserving lexical content. Figure 2 further shows that higher emotional intensity consistently improves prediction accuracy across both models.

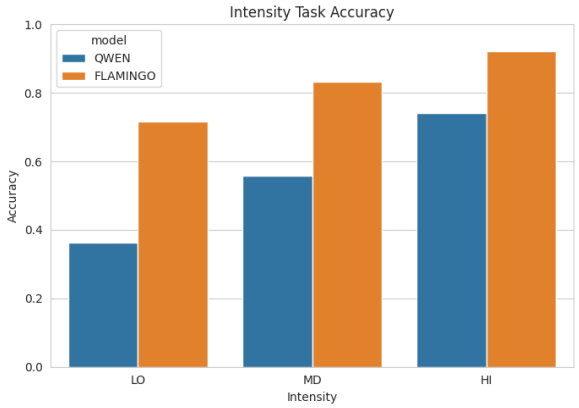


Figure 2: Emotion classification accuracy under different emotional intensity conditions on the CREMA-D dataset. Higher emotional intensity consistently improves classification accuracy across both models.

Fine-Grained Emotion and Elicitation. Under open-ended prompting, both models frequently produce responses that cannot be mapped to the 28-label GoEmotions taxonomy. Appendix E reports the corresponding valid-label rates and stratified results by prompting condition, label composition, and annotation salience. In the audio-only multiple-choice setting, observed SC-MCQ accuracy is lower than CC-MCQ accuracy for both Qwen2-Audio (0.507 versus 0.601) and Audio-Flamingo-3 (0.464 versus 0.621), consistent with greater difficulty when the candidate emotions are semantically related. The text-only and audio–text SC-MCQ conditions show modest, model-dependent differences and no common directional benefit from adding audio. Moreover, candidate sets and assigned targets are not identical across all prompting settings. These aggregate accuracies are therefore treated as descriptive comparisons rather than paired estimates of modality benefit. Because targets originate from multilabel annotations, a prediction mismatch may also reflect selection among multiple human-supported interpretations rather than a wholly unsupported emotion.

Representation Geometry. Table 2 shows the strongest nearest-centroid separation for basic emotions in both models. At the nominal, unadjusted 0.05 threshold, label-permutation tests support basic-emotion separation in both models and intra- and cross-cluster separation in Audio-Flamingo-3. Qwen2-Audio's subtle–intra and subtle–cross point estimates exceed their null means but do not reach this threshold ($p = 0.0594$ and $p = 0.1089$, respectively; Table 16). Together, these results support organized affective representations, with the strength of separability evidence varying across models and emotion conditions.

Model	Setting	Overall	Pure subtle
Qwen2	SC-MCQ (Audio)	0.507	0.464
Qwen2	CC-MCQ (Audio)	0.601	0.714
Qwen2	SC-MCQ (Text)	0.516	0.554
Qwen2	SC-MCQ (Audio+Text)	0.495	0.554
AF3	SC-MCQ (Audio)	0.464	0.518
AF3	CC-MCQ (Audio)	0.621	0.679
AF3	SC-MCQ (Text)	0.516	0.679
AF3	SC-MCQ (Audio+Text)	0.544	0.536
Chance	3-way MCQ	0.333	0.333

Table 1: Fine-grained emotion accuracy under different prompting and modality settings. SC-MCQ and CC-MCQ use same-cluster and cross-cluster candidate sets, respectively. The analysis excludes singleton-label samples and neutral targets.

Model	Basic	Intra	Cross
Qwen2	0.448 (0.065)	0.278 (0.102)	0.293 (0.170)
AF3	0.838 (0.059)	0.359 (0.096)	0.351 (0.173)

Table 2: Leave-one-out nearest-centroid separability at LM2. Each cell reports observed SepAcc followed by the mean randomized-label baseline in parentheses. Basic uses 649 samples; the intra- and cross-cluster subsets each use 800 samples. Confidence intervals and permutation results are reported in Table 16.

Basic-emotion separability exceeds the intra- and cross-cluster subtle-emotion estimates in both models, consistent with weaker geometric separation under affective ambiguity. Audio-Flamingo-3 yields higher point estimates in each condition, a descriptive contrast that complements the within-model tests.

Modality Representation Control. To examine whether the EMOS representation geometry is primarily determined by the transcript, we extract representations for the same 1,000 utterances under speech-only, text-only, and speech-and-text input. Table 3 shows that the joint representation at LM2 is nearly identical to its speech-only counterpart for both models, whereas its similarity to the text-only representation is substantially lower.

These results indicate that the early multimodal representation analyzed in our geometry experiments remains predominantly conditioned by the speech input rather than being replaced by the separately supplied transcript. This speech-conditioned

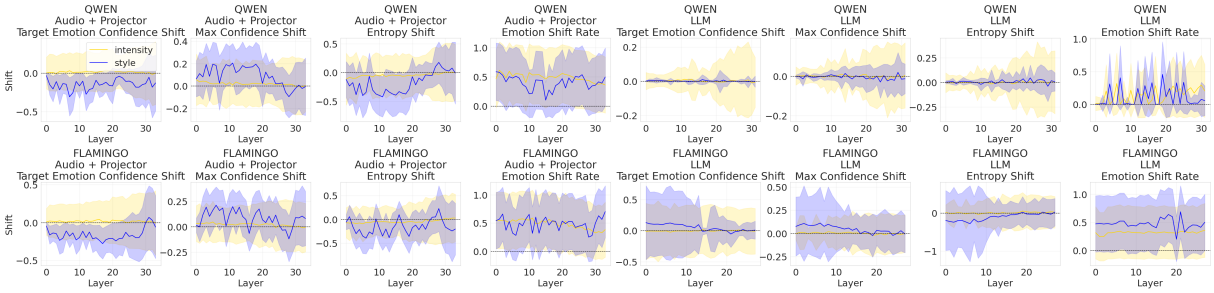


Figure 3: Layer-wise perturbation sensitivity across audio encoder, projector, and language model layers under speech style and emotional intensity manipulations. Style perturbations consistently produce larger behavioral and representational shifts than aggregated emotional intensity changes across both Qwen and Flamingo. Sensitivity patterns further evolve across multimodal processing stages, indicating that speech-conditioned affective information is not restricted to early acoustic representations alone.

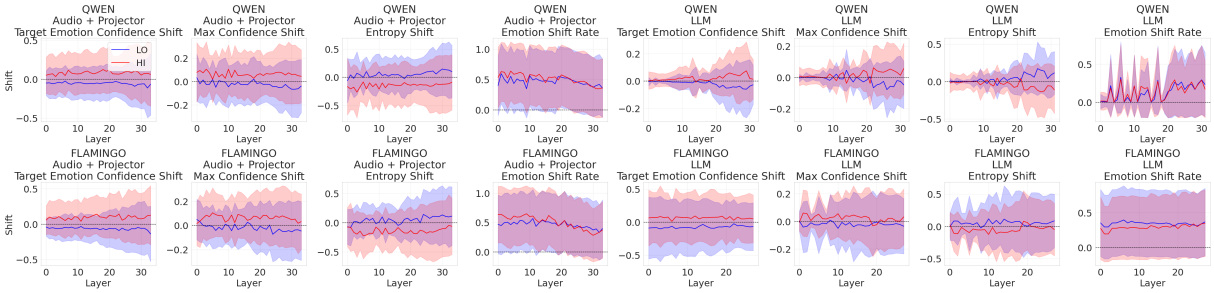


Figure 4: Layer-wise comparison between Low (LO) and High (HI) emotional intensity perturbations relative to Medium (MD) baseline speech across audio encoder, projector, and language model layers. LO and HI perturbations exhibit partially opposing representational and behavioral trends, particularly in early multimodal processing stages, suggesting that emotional intensity variations remain distinguishable throughout the representation hierarchy.

Model	S–T	S–(S+T)	T–(S+T)
Qwen2	0.3984	0.9993	0.4291
AF3	0.4947	0.9997	0.5074

Table 3: Mean cosine similarity between speech-only (S), text-only (T), and speech-plus-text (S+T) representations at LM2 over 1,000 matched EMOS utterances.

organization can reflect both vocal delivery and the lexical information carried by speech. Layer-wise similarity trajectories are reported in Appendix F.

Figure 5 shows L2 displacement from single-label base-emotion centroids to centroids of multi-label conditions containing the base and added emotion. Mean displacement across the 20 ordered conditions exceeds the shuffled-label baseline in both models ($p_{\text{perm}} = 0.0099$ for each; Appendix H). These tests support annotation-related organization in the global displacement pattern, while individual heatmap cells remain descriptive.

Relation between Speech Perturbation and Fine-Grained Emotion Encoding. Figure 6 reveals emotion-specific alignment between intensity shifts and coarse-to-subtle directions. After Holm correction, both models show positive MD→LO alignment for anger and negative MD→HI alignment for fear; Qwen2-Audio also shows positive MD→LO alignment for happiness. In matched HI/LO subsets, alignment is significantly higher for MD→LO for anger and fear in Audio-Flamingo-3, and for fear and happiness in Qwen2-Audio.

These findings link controlled intensity variation to fine-grained reference directions, supporting structured speech-conditioned affective organization. The pattern is emotion- and model-dependent, with smaller alignment magnitudes in Qwen2-Audio. Alignment quantifies orientation relative to these directions; its inferential scope and the full results, including nonsignificant conditions, are detailed in Appendix H, Tables 17 and 18.

Within-Dataset CREMA-D Control. As a within-dataset control, we measure matched intensity-induced changes in cosine similarity to emotion centroids estimated from a separate CREMA-D reference set. After Holm correction across 20 tests, higher intensity moves fear representations closer to their centroid in both models, while lower intensity moves anger representations farther from their centroid in Audio-Flamingo-3

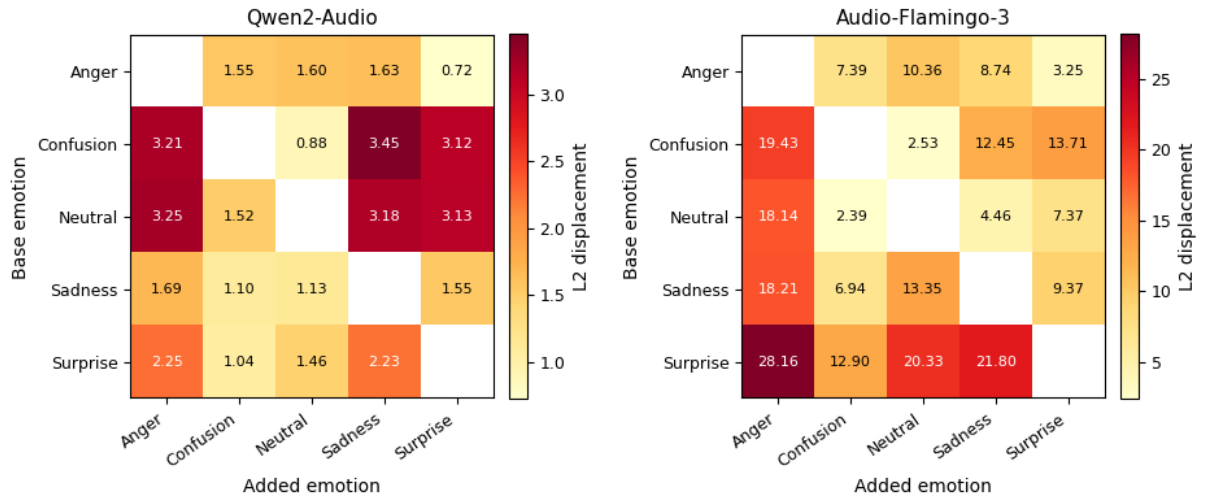


Figure 5: L2 centroid displacement for Qwen2-Audio (left) and Audio-Flamingo-3 (right), from single-label base-emotion conditions to multilabel conditions containing the base and added emotion. The mean randomized-label baselines are 0.972 and 3.979, respectively. The global mean across 20 ordered conditions exceeds its baseline in each model ($p_{\text{perm}} = 0.0099$); individual cells are descriptive. Color scales are model-specific.

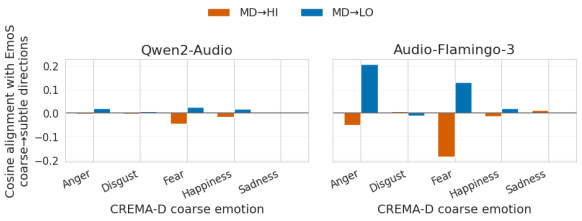


Figure 6: Mean cosine alignment of CREMA-D intensity shifts with EMOS coarse-to-subtle directions at LM2. Each bar averages 42 transition pairs from 42 speakers across mapped subtle directions. Signs indicate orientation, not centroid distance. Full uncertainty estimates and corrected tests appear in Appendix H; the within-dataset control is discussed in Appendix F.

Model	α	Δ Macro-F1	p
Qwen2	0.10	$+0.0057 \pm 0.0025$	0.0625
	0.20	$+0.0170 \pm 0.0028$	0.0625
AF3	0.10	$+0.0001 \pm 0.0052$	1.0000
	0.20	$+0.0129 \pm 0.0168$	0.1875

Table 4: Residual reintegration at LM2: mean $\pm$ sample SD of paired Macro-F1 changes across five seeds relative to $\alpha = 0$, with two-sided Wilcoxon p-values. Qwen2: Qwen2-Audio; AF3: Audio-Flamingo-3. Test sample size: 778.

(all $p_{\text{Holm}} \leq 0.0162$). These findings support emotion-specific sensitivity to controlled intensity variation within one dataset. Full effects, uncertainty estimates, and remaining conditions are reported in Appendix F.2.

Reintegration Results. Table 4 reports five-seed residual reintegration at LM2. At $\alpha = 0.20$, mean Macro-F1 increases by 0.0170 for Qwen2-Audio and 0.0129 for Audio-Flamingo-3, suggesting a modest contribution from complementary speech features. Both seed-level tests remain nonsignificant ($p = 0.0625$ and $p = 0.1875$, respectively), with greater seed variability in Audio-Flamingo-3. The full interpolation sweep, bootstrap intervals, and tests appear in Appendix H; additional settings and protocol details are in Appendix I.

Comparison to closed-source model. Table 5 reports same-cluster multiple-choice (SC-MCQ) performance across modality settings using Gemini, while Table 6 presents sensitivity analysis under manipulated speech conditions relative to neutral baselines. All fine-grained emotion recognition conditions remain above the 3-way chance level, indicating that the model can partially distinguish semantically related emotions under conversational ambiguity. Transcript-based prompting achieves the strongest performance, whereas audio-only prompting remains substantially lower, and multimodal prompting yields similar performance to transcript-only evaluation. In the sensitivity analysis, cue-based manipulations produce the largest increases in both prediction accuracy and confidence, while low-intensity manipulations induce comparatively small directional shifts. Together, these findings suggest that the model is responsive to salient vocal cues, yet relies more heavily on linguistic information for nuanced fine-grained emotional interpretation, consistent with our earlier ob-

Model	Setting	Overall	Pure subtle
Gemini	SC-MCQ(Audio)	0.476	0.457
Gemini	SC-MCQ(Text)	0.577	0.500
Gemini	SC-MCQ(Audio+Text)	0.574	0.500
Chance	3-way	0.333	0.333

Table 5: Fine-grained emotion recognition accuracy under different modality settings using same-cluster multiple-choice prompting (SC-MCQ). Transcript-based conditions outperform audio-only evaluation, suggesting stronger reliance on linguistic information for subtle emotion interpretation(temp=0).

Condition	Accuracy	Δ Acc	Confidence	Δ Conf
Cues	0.842	+0.214	0.781	+0.126
Low Int.	0.571	-0.057	0.601	-0.054
High Int.	0.694	+0.066	0.648	-0.007
Base.	0.628	–	0.655	–

Table 6: Sensitivity analysis under manipulated speech conditions relative to neutral baselines. Cue-based manipulations produce the largest shifts in prediction accuracy and confidence, while low-intensity manipulations induce comparatively smaller deviations(temp=0).

servations that speech information may be partially encoded but not fully utilized during downstream affective reasoning.

6 Conclusion

In this work, we investigate whether the evaluated Large Audio Language Models (LALMs) exhibit measurable sensitivity to subtle speech-conditioned emotional information without task-specific emotional fine-tuning. Across controlled speech perturbation experiments, fine-grained emotion elicitation settings, and representation-level analyses, we find consistent evidence that vocal emotional variation systematically influences the behavior and hidden representations of the evaluated models. However, emotional ambiguity substantially increases difficulty, particularly when these models must distinguish semantically related fine-grained emotions under ambiguity-sensitive settings.

Taken together, our findings suggest a representation–utilization gap in the evaluated LALMs: fine-grained affective structure remains partially organized within hidden representations even when downstream emotional interpretation becomes unstable. Residual reintegration provides complementary, exploratory evidence that speech features can improve prediction in selected settings, although gains are modest and vary across configurations and seeds. These findings motivate studying how preserved affective information can be used more consistently under ambiguity. Overall, our results highlight the importance of ambiguity-sensitive evaluation for understanding how the evaluated LALMs encode and utilize subtle emotional information in speech.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62266013, 62402194), the Science and Technology Development Fund of Macau SAR (Grant Nos. FDCT/0007/2024/AKP, EF2024-00185-FST), the UM and UMDF (Grant Nos. MYRG-GRG2024-00165-FST-UMDF, MYRG-GRG2025-00236-FST, MYRG-GRG2026-00018-FIC, MYRG-CRG2026-00007-FIC), the Dr. Stanley Ho Medical Development Foundation (Grant No. SHMDF-AI/2026/001), and the Guizhou Provincial Basic Research Program (Grant No. MS[2026]276), the Macao Young Scholars Program (Grant No. AM2024015), and the China Postdoctoral Science Foundation under Grant Number 2024M751062.

Limitations

Our controlled perturbation and representation analyses characterize speech-conditioned affective sensitivity across the evaluated models. A fuller causal account of how emotional cues propagate through model components remains an important next step. In particular, extending the matched text-only, audio-only, and joint-modality controls across more layers and experimental settings would help separate the contributions of vocal delivery, lexical content, and multimodal integration.

Second, our experiments are conducted on a limited set of LALMs and datasets. In particular, CREMA-D consists primarily of acted emotional speech recorded under controlled settings, which may not fully reflect spontaneous real-world emotional interaction. However, the controlled nature of the dataset is also important for reducing semantic variation and isolating the influence of speech-related emotional cues.

Third, fine-grained emotional states are inherently difficult to isolate cleanly at the semantic level. Many subtle emotions overlap in both linguistic meaning and vocal expression, and no dataset can provide perfectly disentangled emotional categories. While we employ weighted annotation strategies and controlled perturbations to reduce

ambiguity, latent confounding variables may still contribute to the observed representation shifts.

In addition, our representation analyses rely on geometric probes such as centroid separability and directional alignment. These probes characterize internal representational organization; relating that organization to human emotional understanding requires additional behavioral validation. Similarly, hidden-state pooling may discard temporal prosodic dynamics that could be important for subtle affective reasoning.

Finally, we do not systematically analyze the effect of prompt formulation on emotional representations or decision behavior. Since prompting can substantially influence multimodal reasoning, future work may benefit from jointly examining prompt sensitivity, representation geometry, and speech-conditioned affective processing under more diverse interaction settings.

Ethical Considerations

Our study analyzes emotional information in speech using existing research datasets and controlled synthesized-speech perturbations. Emotional labels should be interpreted as annotator-dependent affective judgments rather than objective measurements of a speaker's internal state, and our results should not be used for high-stakes profiling or decision-making about individuals.

We use CREMA-D and EMOS according to their respective access and licensing conditions and do not redistribute restricted source audio. Synthesized speech is employed solely to create controlled paralinguistic perturbations for research analysis. Where human verification is used, we follow the annotation and verification procedures described in the paper and report the relevant protocol details.

References

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. In *Advances in Neural Information Processing Systems*, volume 37, pages 56898–56918. Curran Associates, Inc.

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.

Carlos Busso, Reza Lotfian, Kusha Sridhar, Ali N. Salman, Wei-Cheng Lin, Lucas Goncalves, Srinivas Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, Huang-Cheng Chou, and Pravin Mote. 2026. The MSP-Podcast Corpus . *IEEE Transactions on Affective Computing*, 17(03):3065–3083.

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.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and EngSiong Chng. 2025. Audio large language models can be descriptive speech quality evaluators. In *The Thirteenth International Conference on Learning Representations*.

Xize Cheng, Ruofan Hu, Xiaoda Yang, Jingyu Lu, Dongjie Fu, Zehan Wang, Shengpeng Ji, Rongjie Huang, Boyang Zhang, Tao Jin, and Zhou Zhao. 2025. Voxdialogue: Can spoken dialogue systems understand information beyond words? In *The Thirteenth International Conference on Learning Representations*.

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. *Preprint*, arXiv:2407.10759.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, and Irwin King. 2025. VoxEval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16735–16753, Vienna, Austria. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, and 3 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *Preprint*, arXiv:2505.17589.

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. *Preprint*, arXiv:2412.10117.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. 2025. Benchmarking open-ended audio dialogue understanding for large audio-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4763–4784, Vienna, Austria. Association for Computational Linguistics.

Arushi Goel, Sreyan Ghosh, 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. *Preprint*, arXiv:2507.08128.

Pengze Guo, Jingxi Liang, Zhiwen Xie, Qifeng Wang, and Derek F. Wong. 2026. Emos: A high-fidelity multimodal benchmark for fine-grained streaming emotional understanding. *Preprint*, arXiv:2605.08847.

Guanyu Hou, Jiaming He, Yinhang Zhou, Ji Guo, Yitong Qiao, Rui Zhang, and Wenbo Jiang. 2025. Evaluating robustness of large audio language models to audio injection: An empirical study. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25671–25687, Suzhou, China. Association for Computational Linguistics.

Heeseung Kim, Soonshin Seo, Kyeongseok Jeong, Ohsung Kwon, Soyoon Kim, Jungwhan Kim, Jaehong Lee, Eunwoo Song, Myungwoo Oh, Jung-Woo Ha, Sungroh Yoon, and Kang Min Yoo. 2024. Paralinguistics-aware speech-empowered large language models for natural conversation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Minzhi Li, William Held, Michael J. Ryan, Kunat Pipatanakul, Potsawee Manakul, Hao Zhu, and Diyi Yang. 2025. Mind the gap: Static and interactive evaluations of large audio models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8749–8766, Vienna, Austria. Association for Computational Linguistics.

Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, Ziyang Ma, Xiaojiang Peng, Xie Chen, Ya Li, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2025. Mer 2025: When affective computing meets large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 13837–13842, New York, NY, USA. Association for Computing Machinery.

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):1–35.

Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2025. Speech-integrated modeling for behavioral coding in counseling. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 152–158, Avignon, France. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025a. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316, Albuquerque, New Mexico. Association for Computational Linguistics.

Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, and Chao Zhang. 2025b. QualiSpeech: A speech quality assessment dataset with natural language reasoning and descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23588–23609, Vienna, Austria. Association for Computational Linguistics.

Siyin Wang, Wenyi Yu, Yudong Yang, Changli Tang, Yixuan Li, Jimin Zhuang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, and Chao Zhang. 2025c. Enabling auditory large language models for automatic speech quality evaluation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zihan Wang, Xiaocui Yang, Yongkang Liu, Shi Feng, Daling Wang, and Yifei Zhang. 2025d. Muse: A multimodal conversational recommendation dataset with scenario-grounded user profiles. *Preprint*, arXiv:2412.18416.

Shu wen Yang, Ming Tu, Andy T. Liu, Xinghua Qu, Hung yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. 2026. Paras2s: Benchmarking and aligning spoken

language models for paralinguistic-aware speech-to-speech interaction. In *The Fourteenth International Conference on Learning Representations*.

Chih-Kai Yang, Neo Ho, Yi-Jyun Lee, and Hung-yi Lee. 2025a. Audiolens: A closer look at auditory attribute perception of large audio-language models. pages 1–8.

Chih-Kai Yang, Neo S. Ho, and Hung-yi Lee. 2025b. Towards holistic evaluation of large audio-language models: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10144–10170, Suzhou, China. Association for Computational Linguistics.

Hao Yang, Lizhen Qu, Ehsan Shareghi, and Reza Haf. 2024a. Towards probing speech-specific risks in large multimodal models: A taxonomy, benchmark, and insights. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10957–10973, Miami, Florida, USA. Association for Computational Linguistics.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024b. AIR-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998.

Xiutian Zhao, Björn Schuller, and Berrak Sisman. 2026. Discovering and causally validating emotion-sensitive neurons in large audio-language models. *Preprint*, arXiv:2601.03115.

Youxiang Zhu, Abdelrahman Obyat, Xiaohui Liang, John A. Batsis, and Robert M. Roth. 2021. WavBERT: Exploiting Semantic and Non-Semantic Speech Using Wav2vec and BERT for Dementia Detection. In *Interspeech 2021*, pages 3790–3794.

A Related Work

Recent multimodal emotion datasets have increasingly moved beyond controlled acted settings toward more human-centered and ambiguity-aware emotional understanding. One major direction involves the collection of in-the-wild conversational and spontaneous emotional data rather than laboratory-acted speech. Another direction focuses on fine-grained emotional interpretation through multilabel annotation, emotional captioning, and semantically overlapping affective categories. Related work on streaming and temporally localized emotion recognition further examines emotional dynamics through frame-level segmentation and emotional change point analysis. Together, these developments reflect a broader shift from coarse categorical recognition toward more realistic and ambiguous emotional understanding settings.

Despite these advances in emotional datasets, recent Large Audio Language Model (LALM) evaluation primarily focuses on objective audio understanding tasks, robustness analysis, and general speech perception benchmarks. Existing studies suggest that the LALMs they evaluate exhibit limited reliability in subjective emotional judgment and often underutilize speech prosodic information during downstream reasoning. At the same time, recent probing and representation studies indicate that affective information may remain partially encoded within internal multimodal representations even when behavioral performance is weak. Motivated by these observations, our work examines both behavioral prediction patterns and representation-level affective structure under ambiguity-sensitive fine-grained emotional settings. Table 7 summarizes representative emotional datasets relevant to our study.

Recent work has increasingly focused on establishing rigorous evaluation frameworks for Large Audio Language Models (LALMs) (Yang et al., 2025b). Existing benchmarks now cover a broad range of capabilities, including general audio understanding and spoken dialogue interaction (Wang et al., 2025a; Gao et al., 2025; Hou et al., 2025; Ao et al., 2024; Cheng et al., 2025), speech knowledge and reasoning (Cui et al., 2025; Li et al., 2025), perceptual audio assessment (Chen et al., 2025; Wang et al., 2025c,b), and instruction-following ability (Yang et al., 2024b). Together, these studies suggest that the LALMs evaluated in them can reliably assess many objective speech and audio properties while generating coherent multimodal reasoning. However, substantially less is understood about whether these capabilities extend to ambiguity-sensitive and fine-grained emotional interpretation.

B Dataset Details

B.1 EmoS Analysis Subset and Label Processing

Our EmoS experiments use the 5,295 short-utterance records available in the analysis manifest rather than the complete published corpus. The subset contains both English and Chinese samples. We retain the original transcripts, multilabel human annotations, and annotator counts without introducing new labels or collapsing each item to a single

Dataset	Mod.	Script Ctrl	Conv	Fine	Multi	Stream
CREMA-D (Cao et al., 2014)	S	High	N	N	N	N
RAVDESS (Livingstone and Russo, 2018)	S+V	High	N	N	N	N
IEMOCAP (Busso et al., 2008)	S+T	Moderate	Y	P	N	N
MELD (Poria et al., 2019)	S+T+V	Low	Y	P	N	P
MSP-Podcast (Busso et al., 2026)	S	Low	P	Y	Y	N
MuSE (Wang et al., 2025d)	S+T+V	Low	Y	Y	Y	Y
EmoS (Guo et al., 2026)	S+T	Low	Y	Y	Y	Y
GoEmotions (Demszky et al., 2020)	T	N/A	Y	Y	Y	N

Table 7: Representative datasets relevant to speech-conditioned emotional understanding. CREMA-D and RAVDESS provide strong scripted semantic control through repeated or fixed utterances rendered under different emotional delivery conditions, making them suitable for controlled speech sensitivity analysis. In contrast, recent datasets such as EmoS emphasize fine-grained, ambiguity-aware, and temporally continuous emotional understanding in more naturalistic conversational settings.

majority label.

We treat `unsure` as an abstention rather than an emotion category. When it co-occurs with another label, it is removed while the remaining labels are retained. The single record containing only `unsure` is excluded, leaving 5,294 usable items. Fine-grained behavioral analyses retain the complete set of remaining labels from the 28-label GoEmotions taxonomy. For representation analyses, annotator counts are normalized after removing `unsure` and used as soft label weights. The ambiguity-focused results reported in the main comparison additionally exclude singleton-label items and cases in which the sampled target is `neutral`; these restrictions isolate examples containing multiple non-neutral human-supported interpretations.

The cleaned 5,294-item subset contains 2,796 English utterances and 2,498 Chinese utterances. These language counts refer to the usable subset after the exclusion described above. For the original annotation protocol, dataset-level label-distribution statistics, and inter-annotator agreement, we refer readers to the EmoS publication (Guo et al., 2026). No new emotion annotation was conducted for this study.

B.2 Mapping to Seven Coarse Emotion Categories

The coarse taxonomy follows the Ekman mapping released with GoEmotions (Demszky et al., 2020).[2] The 27 non-neutral labels are grouped into six emotion categories, while `neutral` is retained as a seventh category. Table 8 lists all 28 label-to-category correspondences. HAP denotes the coarse joy/happiness category. This mapping specifies the coarse taxonomy; the overlapping candidate pools used to construct prompting distractors are documented separately in Appendix C.2.

[2]The mapping is provided in the official GoEmotions repository.

Coarse label	**GoEmotions labels**
Anger (ANG)	anger, annoyance, disapproval
Disgust (DIS)	disgust
Fear (FEA)	fear, nervousness
Joy (HAP)	joy, amusement, approval, excitement, gratitude, love, optimism, relief, pride, admiration, desire, caring
Sadness (SAD)	sadness, disappointment, embarrassment, grief, remorse
Surprise (SUR)	surprise, realization, confusion, curiosity
Neutral (NEU)	neutral

Table 8: GoEmotions mapping from 28 fine-grained labels to six Ekman emotion categories plus neutral. Each fine-grained label belongs to exactly one coarse category.

B.3 Controlled-Sensitivity Pair Construction

We distinguish among source recordings, available matched comparisons, and rows retained in the fixed evaluation manifest. Each manifest row represents one reference–target comparison rather than one distinct reference recording. The full fixed manifest contains 990 comparisons: 420 naturally recorded intensity pairs, 496 controlled TTS cue pairs, and 74 separate subtle-emotion TTS pilot pairs. The main controlled-sensitivity experiments in Tasks 1 and 2 exclude the pilot block and therefore contain 916 comparisons derived from 504 distinct reference recordings. Table 9 provides the complete accounting.

Naturally recorded intensity pairs. The intensity experiment is independent of the TTS pipeline and uses only original CREMA-D recordings. We

identified 455 medium-intensity (MD) recordings for which both a high-intensity (HI) and a low-intensity (LO) version exists with the same speaker, sentence, and coarse emotion. This yields 910 available MD→HI or MD→LO comparisons. The pool covers all 91 speakers and five non-neutral emotions, with 91 MD anchors per emotion. All intensity items use the same lexical item (CREMA-D content code IEO), maximizing lexical control across speakers and conditions.

The fixed evaluation manifest retains 420 comparisons. It is exactly balanced across the ten emotion×target-intensity cells: 42 MD→HI and 42 MD→LO comparisons for each of anger, disgust, fear, happiness, and sadness. These comparisons contain 319 distinct MD anchors from all 91 speakers. Among these anchors, 101 occur with both intensity directions, while 109 occur only with an HI target and 109 only with an LO target. The intensity-shift vectors in the reported CREMA-D directional analyses are computed from this same 420-pair subset. The additional 1,000-utterance CREMA-D reference set used for probe training and within-dataset centroid estimation is separate from these counts.

Controlled TTS cue pairs. For the cue experiment, we use CozyVoice3(Du et al., 2025) with its default generation settings. We first generate a neutral reference using the instruction "Speak this sentence in a neutral tone." This instruction was refined on a 10-utterance development set until the generated speech consistently followed a descriptive and emotionally neutral delivery. We then insert one target vocal cue or apply one delivery modification while preserving the transcript and speaker identity. The eight analyzed conditions are cough, emphasis, laughter, sigh, louder, quieter, faster, and slower speech.

The broader generation pool contained 500 neutral references that passed manual screening conduct by three researchers. For paired evaluation, the candidate inventory contains 192 neutral anchors spanning 16 speaker identities and all 12 CREMA-D sentences, each represented under the eight analyzed conditions. This yields 1,536 available comparisons. The fixed condition-balanced subset retains 62 comparisons per condition, or 496 comparisons in total, spanning 185 distinct neutral anchors.

Human verification proceeded in two stages. First, each neutral reference was required to remain calm, descriptive, and free of unintended emotional expression. Second, its manipulated counterpart was required to preserve the lexical content and otherwise-neutral speaking style while clearly realizing the intended cue without introducing additional unintended cues. Samples failing either screen were regenerated before inclusion. Cue locations were manually selected when insertion at a particular position was required.

Synthesis choice and artifact limitations. CozyVoice3 was selected following preliminary comparisons with CozyVoice2(Du et al., 2024) because it followed the requested cue instructions while preserving speaker identity in our trials. These comparisons informed the choice of synthesis system for controlled cue manipulation. Cue locations were selected for linguistically natural realizations; different placements may produce different representational effects, but our objective is to examine the presence versus absence of a controlled cue rather than optimize its position.

The TTS experiment compares a synthesized neutral reference with a counterpart generated under the same neutral speaking style and lexical content, with the intended cue or delivery modification as the controlled difference. Its purpose is to probe sensitivity to this intervention rather than benchmark emotion recognition on synthetic speech. Manual verification does not fully exclude condition-dependent synthesis artifacts. Additional synthesis systems and stronger artifact controls would therefore strengthen this analysis. Comparisons with naturally occurring human cues or non-emotional synthetic perturbations are possible extensions, not controls performed here. The separate intensity experiment uses original human recordings and does not depend on the TTS pipeline.

The manifest additionally contains 74 matched neutralized-versus-original subtle-emotion pilot comparisons. These items are retained for separate diagnostic analyses but are not included in the 916-pair main controlled-sensitivity experiments. Figure 7 reports model accuracy by demographic group for the selected CREMA-D recordings.

C Prompt

For probe head training, representations are extracted using the prompt: "What is the emotion in this speech?" The same prompt is used for geometry analysis and speech reintegration experiments to maintain consistent representational con-

Source	Condition	Available pairs	Retained pairs	Retained refs.	Cell balance
Naturally recorded CREMA-D *intensity comparisons*					
CREMA-D	Anger	182	84	65	42 HI + 42 LO
	Disgust	182	84	66	42 HI + 42 LO
	Fear	182	84	67	42 HI + 42 LO
	Happiness	182	84	59	42 HI + 42 LO
	Sadness	182	84	62	42 HI + 42 LO
Intensity total	Five emotions	**910**	**420**	**319**	210 HI + 210 LO
Controlled CozyVoice3 cue comparisons					
CozyVoice3	Cough	192	62	62	62 per condition
	Emphasis	192	62	62	62 per condition
	Laughter	192	62	62	62 per condition
	Sigh	192	62	62	62 per condition
	Louder	192	62	62	62 per condition
	Quieter	192	62	62	62 per condition
	Faster	192	62	62	62 per condition
	Slower	192	62	62	62 per condition
Cue total	Eight conditions	**1,536**	**496**	**185**	62 per condition
Separate TTS pilot	Subtle emotion	74	74	74	Excluded from main analysis

Table 9: Construction and retention counts for the controlled-sensitivity manifest. "Available pairs" denotes reference–target comparisons in the candidate pairing inventory, whereas "retained pairs" denotes rows in the fixed evaluation manifest. A reference is a unique MD recording for the intensity block and a unique neutral TTS anchor for the cue blocks. Reference totals are deduplicated within each block and therefore need not equal the sum of the condition-level rows.

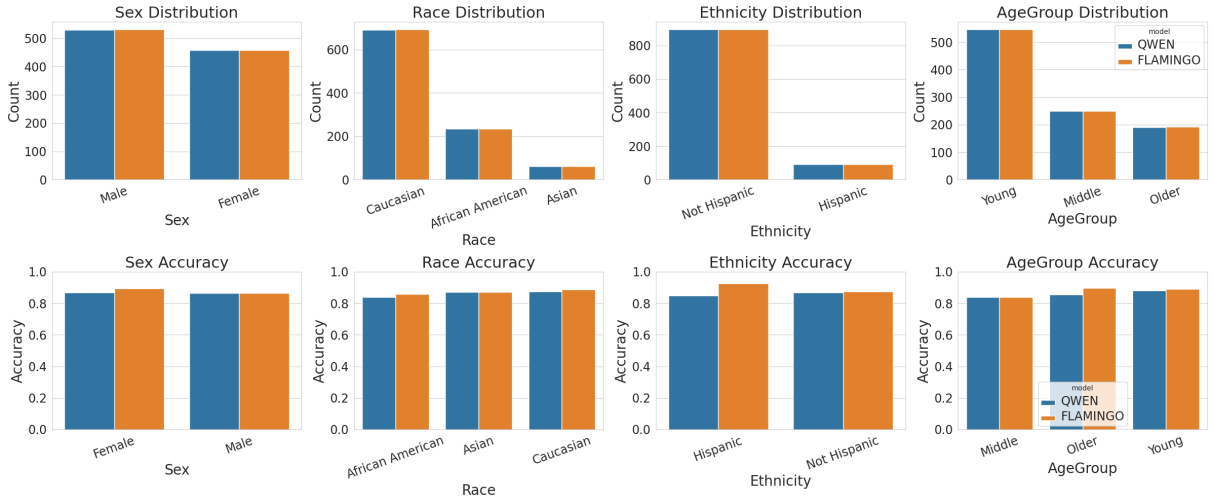


Figure 7: Demographic analysis of models output

ditions across fine-grained emotion evaluations. In contrast, sensitivity analysis employs task-specific prompts aligned with the corresponding behavioral objectives (e.g., neutral vs. emotional discrimination or intensity-based classification). Because prediction shifts are always computed relative to matched prompting conditions, the resulting directional changes remain internally comparable despite prompt differences.

C.1 Gemini API Evaluation

Gemini-2.5-Flash evaluations were conducted through the official API using the latest available version of that model at the time of experimentation, approximately May 24, 2026. Decoding temperature was set to 0.

C.2 Target Construction and Output Scoring

The prompting experiments use a fixed 1,000-utterance subset. During prompt generation, one target was sampled uniformly from the stored human-supported labels of each utterance, without weighting the draw by annotator frequency. Consequently, the selected target need not be the label with the largest annotator count.

Each constrained MCQ instance contains the sampled target, one distractor, and neutral. Same-cluster MCQ selects the distractor from an affectively related candidate pool, whereas cross-cluster MCQ selects it outside the target pool. Labels al-

Prompt Template (Speech Sensitivity)

System Prompt

You are a speech emotion classifier.
Determine whether the speaker sounds neutral or emotional.
Choices:
[neutral, emotional]
Return ONLY valid JSON:
{"label": "<neutral or emotional>", "confidence": <0.0-1.0>}

User Prompt

Classify the speech.

Figure 8: Prompt template used for speech sensitivity evaluation under neutral versus emotional delivery conditions.

Prompt Template (Emotion Intensity)

System Prompt

You are a speech emotion classifier.
Determine the emotion of the speaker.
Choices:
[neutral, {target_emotion}, {related_emotion}]
Return ONLY valid JSON:
{"label": "<...>", "confidence": <0.0-1.0>}

User Prompt

Classify the emotion of the speech.

Figure 9: Prompt template used for intensity-based emotion classification with target and semantically related emotion labels.

Fine-Grained Emotion Prompt Template

System Prompt

You are a speech emotion classifier.
Choose the most appropriate fine-grained emotion expressed in the speaker's voice.
You must choose ONLY ONE option.
Return ONLY valid JSON:
{"label": "<one of the choices>", "confidence": <0.0-1.0>}

User Prompt

Which of the following best describes the speaker's emotion?
Choices:
A. {emotion 1}
B. {emotion 2}
C. neutral

Figure 10: Shared prompting template used for fine-grained emotion evaluation. Candidate emotions were varied according to evaluation setting.

ready present in the utterance's human annotation set are excluded from the distractor pool. When no eligible related distractor is available, the generation implementation falls back to another taxonomy label not present in the ground-truth set.

The candidate pools used during generation may overlap. They consist of anger–annoyance–disapproval; disgust–disapproval; fear–nervousness; joy–amusement–excitement–pride–optimism–love–admiration–relief– gratitude–caring–desire; confusion–curiosity–realization–surprise; sadness–disappointment–remorse–grief; and surprise–realization. Targets without an explicit eligible pool use the fallback rule described above.

For camera-ready analysis, "unsure" is removed from the reference labels. An utterance retaining another emotion label remains in the evaluation, whereas an "unsure"-only utterance is excluded. If a saved target is "unsure", it is replaced by the retained label with the greatest annotator support; ties follow the original label order. This deterministic cleanup is recorded in the released target-resolution audit.

Predictions are parsed by extracting the first standalone token belonging to the 28-label inventory. An output containing no inventory label is counted

Setting	Candidate Emotion Construction
Subtle	Open-ended subtle emotion prompting without constrained candidate labels.
SC-MCQ	Distractor emotion sampled from the same semantic emotion cluster as the target emotion.
CC-MCQ	Distractor emotion sampled from a different semantic emotion cluster than the target emotion.

Table 10: Construction of candidate emotion labels for each prompting condition.

as out of distribution. In the open-ended condition, accuracy is exact match against the sampled target, while valid-label rate records whether any taxonomy label was produced. A different human-supported label is therefore valid but incorrect under the sampled-target metric.

Result files are aligned with the full metadata using normalized audio paths, rather than the subset-local numeric identifiers produced after dataframe resetting. Duplicate paths, missing matches, and disagreement between saved and source annotation sets cause the analysis to stop.

Targets and distractors were sampled independently for several prompting and modality runs. Aggregate accuracies across those runs are therefore descriptive condition summaries, not fully paired modality comparisons. A paired comparison is valid only for utterances that share both the sample and sampled target.

D Sensitivity Analysis

D.1 Baseline Representation Variability

All evaluated LALMs use a decoding temperature of 0. To estimate inference-time representational variability, we repeated inference three times on identical neutral speech inputs and computed cosine distances between corresponding hidden representations across selected early, middle, and late layers.

Across both Qwen2-Audio and Audio Flamingo, repeated neutral inference produced extremely small representational deviations, typically ranging from approximately 10^{-9} to 10^{-4} cosine distance depending on layer depth. Audio Flamingo exhibited near-deterministic behavior across both audio and language layers, while Qwen2-Audio showed slightly larger but still minimal variability within several language-model layers.

Importantly, these baseline deviations were substantially smaller than perturbation-induced representational shifts observed in the sensitivity experiments under the same cosine similarity calculation(Figure 11), which commonly ranged from approximately 10^{-3} to 10^{-2} and reached approximately 5×10^{-2} within multimodal projector representations. This large separation suggests that the observed hidden-state shifts are unlikely to arise solely from inference stochasticity.

Layer	Qwen2-Audio	Audio Flamingo
audio_0	1.61×10^{-8}	-4.98×10^{-9}
audio_31	7.03×10^{-7}	8.65×10^{-9}
lm_15	7.52×10^{-5}	8.89×10^{-9}
projector	1.76×10^{-6}	-1.04×10^{-8}

Table 11: Mean cosine distance between repeated inference runs on identical neutral inputs across representative hidden layers.

E Contextualizing fine-grained emotion tasks

Fine-grained emotional interpretation is inherently ambiguous, particularly when multiple semantically related affective states may plausibly describe the same utterance. Unlike coarse categorical emotion recognition, subtle emotional understanding often does not admit a single objectively correct label, and even human agreement remains limited in existing fine-grained emotion datasets. To better contextualize model behavior under this ambiguity, we evaluate multiple prompting formulations with different levels of semantic constraint.

Across both models, CC-MCQ achieves the highest accuracy (Qwen: 0.622; Flamingo: 0.651), whereas SC-MCQ remains more difficult (Qwen: 0.540; Flamingo: 0.537). Open-ended prompting produces substantially lower accuracy (Qwen: 0.063; Flamingo: 0.203) and lower valid-label rates (Qwen: 0.381; Flamingo: 0.680). Because accuracy is computed over all 1,000 samples and invalid outputs are counted as incorrect, the open-ended results reflect both adherence to the 28-label taxonomy and emotion-selection performance.

These results suggest that much of the observed difficulty arises not from complete failure to perceive emotional information, but from ambiguity among semantically neighboring emotional states. Both models remain substantially above chance under same-cluster discrimination, indicating partial sensitivity to subtle emotional distinctions despite reduced separability between closely related affec-

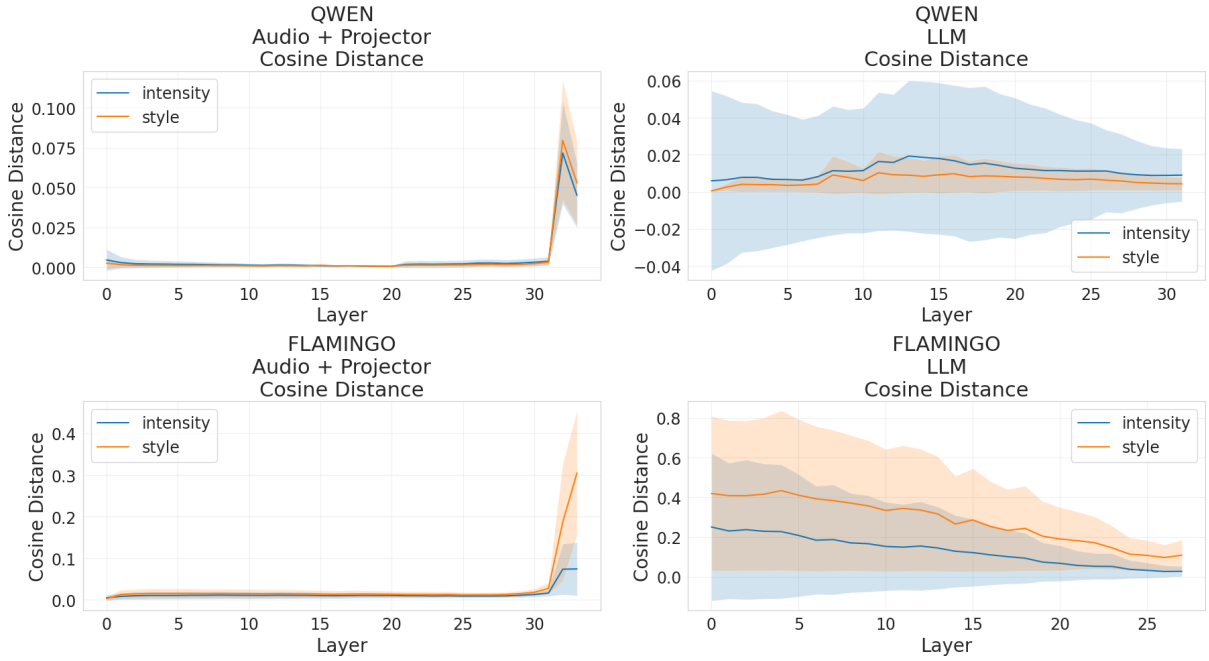


Figure 11: Mean cosine distance between baseline and manipulated inference runs on identical neutral inputs across representative hidden layers.

Model	Prompt Type	Accuracy	Valid Rate
Qwen	CC-MCQ	0.622	0.994
Qwen	SC-MCQ	0.540	0.994
Qwen	Open-ended	0.063	0.381
Flamingo	CC-MCQ	0.651	0.993
Flamingo	SC-MCQ	0.537	0.994
Flamingo	Open-ended	0.203	0.680

Table 12: Comparison of fine-grained emotion prompting settings over 1,000 aligned samples per model and setting. CC-MCQ and SC-MCQ denote cross-cluster and same-cluster multiple-choice prompting, respectively. Open-ended denotes unconstrained prompting without candidate labels. Outputs outside the 28-label inventory are treated as invalid and incorrect.

Model	Prompt Type	Subtle Type	Accuracy
Qwen	CC-MCQ	Basic-only	0.604
Qwen	CC-MCQ	Mixed	0.591
Qwen	CC-MCQ	Pure subtle	0.714
Qwen	SC-MCQ	Basic-only	0.333
Qwen	SC-MCQ	Mixed	0.523
Qwen	SC-MCQ	Pure subtle	0.464
Flamingo	CC-MCQ	Basic-only	0.612
Flamingo	CC-MCQ	Mixed	0.616
Flamingo	CC-MCQ	Pure subtle	0.679
Flamingo	SC-MCQ	Basic-only	0.275
Flamingo	SC-MCQ	Mixed	0.474
Flamingo	SC-MCQ	Pure subtle	0.518

Table 13: Fine-grained emotion recognition accuracy across different emotion-label subsets after excluding singleton-label samples and neutral targets. Basic-only contains basic emotion labels only, mixed contains both basic and fine-grained labels, and pure subtle contains exclusively fine-grained labels.

tive categories. The strong degradation observed under unconstrained prompting further suggests that open-ended fine-grained emotional interpretation may require coherent affective reasoning rather than strict single-label classification. Together, these findings motivate interpreting fine-grained emotion performance as a spectrum of semantic ambiguity and representational uncertainty rather than as a purely objective recognition problem.

Table 13 further separates performance according to the composition of the human label set. CC-MCQ outperforms SC-MCQ within every subset for both models, supporting the conclusion that discrimination among semantically neighboring candidates is more difficult than discrimination across broader emotion clusters. However, accuracy does not decrease monotonically with the presence of fine-grained labels. Under CC-MCQ, the pure-subtle subset produces the highest accuracy for both Qwen (0.714) and Flamingo (0.679), while the relative ordering of subsets under SC-MCQ differs between models. Thus, the clearest difficulty pattern concerns the semantic proximity of candidate labels, while the effect of fine-grained label composition varies with the prompting condition.

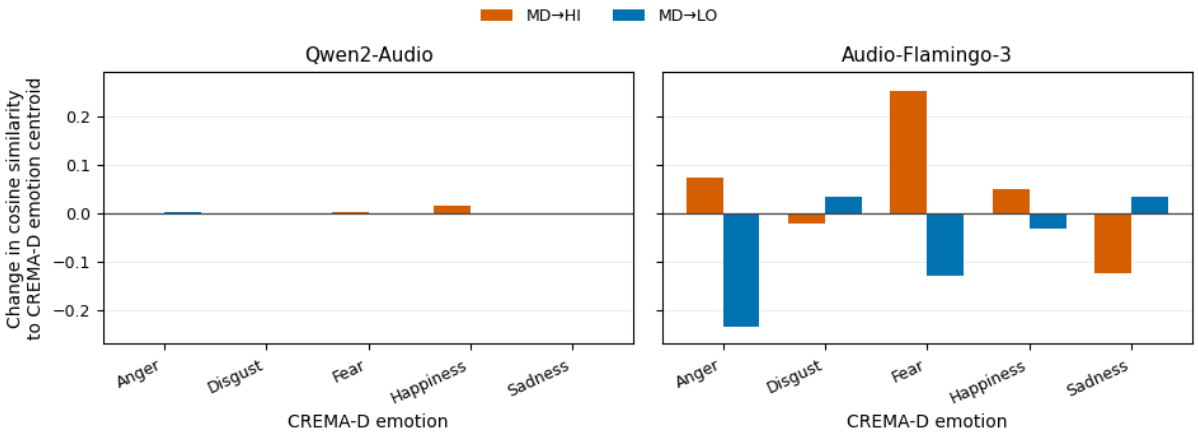


Figure 12: Within-CREMA-D control: changes in cosine similarity to the corresponding emotion centroid at LM2. Positive values indicate movement toward the centroid and negative values indicate movement away from it. Table 15 reports uncertainty and Holm-adjusted tests. Its MD−LO contrast reverses the sign of the medium-to-low bars shown here.

Model	Prompt	Strong Dominant	Dominant Salient	Secondary Salient	Diffuse Trace
Flamingo	CC-MCQ	0.740 (119)	0.616 (258)	0.586 (362)	0.607 (28)
Flamingo	SC-MCQ	0.475 (99)	0.450 (260)	0.458 (380)	0.633 (30)
Flamingo	Open-ended	0.139 (101)	0.096 (261)	0.033 (368)	0.029 (34)
Qwen	CC-MCQ	0.717 (113)	0.608 (273)	0.557 (341)	0.595 (37)
Qwen	SC-MCQ	0.571 (112)	0.565 (253)	0.437 (371)	0.636 (33)
Qwen	Open-ended	0.141 (99)	0.064 (282)	0.054 (369)	0.000 (31)

Table 14: Accuracy across target-label salience levels, with the number of evaluated samples shown in parentheses. A strong-dominant target is the unique highest-supported label accounting for more than half of the retained annotation support. Dominant-salient labels occupy the highest support rank otherwise, including ties; secondary-salient labels occupy the second distinct support rank; and diffuse-trace labels comprise the remaining lower-ranked interpretations.

E.1 Consensus Structure and Emotional Salience Distribution

To contextualize fine-grained emotion performance, we categorize each evaluation target according to its support within the human annotation distribution. A unique top-ranked label accounting for more than half of the retained support is categorized as strong dominant. Top-ranked labels without such a unique majority, including tied labels, are categorized as dominant salient. Labels at the second distinct support level are secondary salient, and the remaining lower-ranked labels are diffuse trace.

As shown in Table 14, secondary-salient targets constitute the largest category, whereas diffuse-trace targets are relatively rare. CC-MCQ exceeds SC-MCQ for strong-dominant, dominant-salient, and secondary-salient targets in both models. The diffuse-trace cells show the reverse numerical ordering, but contain only 28–37 samples per condition and are therefore treated descriptively. More generally, accuracy does not vary monotonically with annotator support rank, indicating that support frequency and candidate-set difficulty capture distinct aspects of ambiguity.

Open-ended prompting remains markedly less accurate than either multiple-choice condition across every salience category. This result suggests that unconstrained fine-grained prediction combines difficulty in selecting a human-supported interpretation with difficulty adhering to the requested label inventory.

F Additional Representation Geometry Controls

F.1 Layer-Wise Modality Similarity

For 1,000 matched EmoS utterances, we extract representations using speech-only (S), text-only (T), and speech-plus-text (S+T) inputs. Cosine similarity is computed between every matched modality pair. The LM2 values are reported in Table 3, while Figure 13 shows their progression across model depth as a descriptive comparison of the

input pathways conditioning the representations.

Scope of speech-conditioned representations. The principal geometry analyses and the matched modality control both use LM2. The near-identity of speech-only and speech-plus-text representations indicates that supplying a separate transcript introduces little additional change at this layer, supporting our characterization of the analyzed states as predominantly speech-conditioned. Because speech itself conveys words, these representations can carry both lexical and non-lexical affective information. The control identifies the dominant input pathway; separating the contributions of vocal delivery and lexical semantics to emotion clustering remains a complementary question, while the matched CREMA-D perturbations provide evidence under lexical control.

F.2 Within-CREMA-D Centroid Control

Rationale for the cross-dataset analysis. The directional analysis in Figure 6 intentionally combines intensity-induced shifts from CREMA-D with coarse-to-subtle emotion directions estimated from EMOS, using LM2 representations from the same evaluated model. EMOS supplies fine-grained labels such as annoyance, disappointment, and nervousness, whereas CREMA-D provides controlled intensity variation but only coarse emotion labels. The cross-dataset design examines whether emotion directions estimated from conversational speech are informative about perturbation trajectories in an independent acted-speech corpus. The within-CREMA-D analysis below complements this transfer-based comparison with a common-dataset control.

We construct reference emotion centroids from the separate 1,000-utterance CREMA-D set and evaluate the fixed 420 medium-to-high and medium-to-low intensity pairs at LM2 in each model. Every emotion-by-transition cell contains 42 pairs from 42 speakers. The evaluated emotions are anger, disgust, fear, happiness, and sadness. Surprise is absent from the CREMA-D reference inventory, and neutral has no eligible MD/HI/LO pairs in this evaluation; these categories are outside the tested conditions.

For an utterance representation $\mathbf{h}$ and its corresponding CREMA-D emotion centroid $\mathbf{c}_e$, let $s = \cos(\mathbf{h}, \mathbf{c}_e)$. The two primary paired contrasts are $s_{\mathrm{HI}} - s_{\mathrm{MD}}$ and $s_{\mathrm{MD}} - s_{\mathrm{LO}}$. Positive estimates support movement toward the centroid with increased intensity and away from it with decreased intensity, respectively. The second contrast reverses the sign of the MD→LO changes in Figure 12.

Raw p-values use two-sided speaker-level sign-flip tests with 19,999 draws and the absolute pair-weighted mean difference as the statistic. A common sign is applied to all observations from a speaker, and $p = (r + 1)/20{,}000$, where r counts simulated statistics at least as extreme as the observed statistic. Pointwise 95% percentile-bootstrap intervals use 9,999 draws of whole speakers. These procedures share the speaker-independence and sign-symmetry assumptions described for directional alignment in Appendix H. Inference is conditional on the fixed CREMA-D reference centroids. Holm correction covers the 20 tests (two models, five emotions, and two contrasts) at family-wise level 0.05, separately from the cross-dataset alignment family; HI-versus-LO is not tested in this control. Pointwise interval exclusion of zero and Holm-adjusted significance need not coincide.

Three comparisons survive correction, all in the expected direction: fear HI−MD in Audio-Flamingo-3 (mean 0.250948, $p_{\mathrm{Holm}} = 0.00285$) and Qwen2-Audio (0.001698, 0.00100), and anger MD−LO in Audio-Flamingo-3 (0.233920, 0.01620). These results support emotion-specific intensity sensitivity within CREMA-D, with a small absolute fear effect in Qwen2-Audio. The other 17 contrasts remain unresolved under the corrected tests. Table 15 reports all conditions.

The two analyses measure complementary quantities. The cross-dataset analysis measures alignment with coarse-to-subtle directions, whereas the within-CREMA-D control measures the change in cosine similarity to the corresponding coarse-emotion centroid. The latter examines intensity-dependent changes within one dataset at the coarse-emotion level, complementing the finer reference directions supplied by EMOS.

G Representation Sensitivity Details

Figure 14 shows layer-wise representation sensitivity across different speech cue manipulations. Overall, both models exhibit consistent sensitivity trends across layers, although the magnitude of representation shifts varies across perturbation types and model depth.

For probing analysis, we train a lightweight

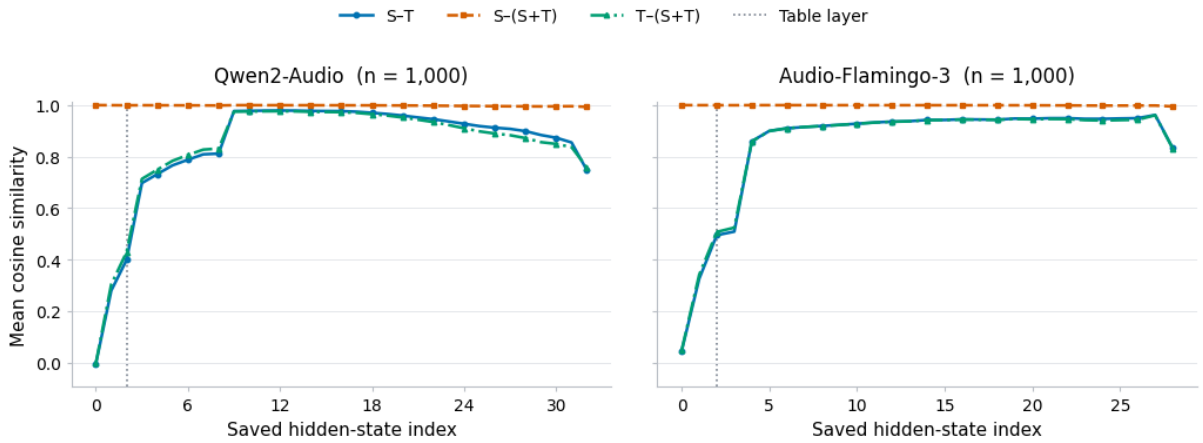


Figure 13: Layer-wise mean cosine similarity among speech-only (S), text-only (T), and speech-plus-text (S+T) representations over 1,000 matched EMOS utterances. The dotted line marks LM2, the layer used for the principal geometry analyses.

linear probe independently on representations extracted from each layer. Probe training uses an additional 1,000 CREMA-D samples that are excluded from the main probing evaluation set.

After training, the same probe is applied without further optimization to both the reference and perturbed representations. The resulting comparisons measure relative representational changes under matched probe parameters rather than differences caused by separately fitting classifiers to the two conditions. The matched design controls probe fitting, while potential synthesis artifacts remain a limitation of the TTS cue experiment.

To examine probe stability, we additionally repeat probe training using three different random seeds. Figure 15 shows the resulting probe accuracy across layers. We observe relatively small variation between runs, suggesting that the overall probing trends remain stable across random initialization.

H Statistical Inference and Robustness

Reporting convention. The main text emphasizes effect directions and magnitudes and distinguishes resampling-based inference from descriptive diagnostics. For the 30 directional-alignment comparisons, we report raw and Holm-adjusted p-values and assess significance using $p_{\text{Holm}} \leq 0.05$. This correction controls the family-wise error rate within that directional-alignment family. The 20 within-CREMA-D centroid-control tests form a separate Holm-corrected family; neither correction covers all analyses in the paper. Centroid separability, centroid displacement, and residual-reintegration tests retain their comparison-specific, unadjusted p-values and are interpreted at the nominal 0.05 threshold. Confidence intervals are pointwise 95% percentile-bootstrap intervals unless otherwise stated; they are not simultaneous intervals.

Scope of statistical analysis. Inferential analyses comprise randomized-label permutation tests for centroid separability and mean centroid displacement; speaker-level sign-flip tests and bootstrap intervals for cross-dataset directional alignment and matched intensity contrasts; the 20 paired within-CREMA-D centroid-control comparisons in Appendix F.2; and paired bootstrap intervals with five-seed Wilcoxon tests for residual reintegration. The controlled-perturbation behavioral curves, prompting comparisons, modality-similarity measurements, and Gemini results remain descriptive. Statistical conclusions concern the tested quantities and conditions.

Centroid separability. Separability is evaluated at LM2 on the EMOS representations. The label *unsure*, which lies outside the 28-emotion inventory, is removed before grouping; the single row left without a retained emotion is excluded. For this coarse geometric probe, labels are mapped to the seven categories listed in Appendix B.2, and a fixed-priority rule reduces each multilabel annotation to one coarse target. We evaluate leave-one-out nearest-centroid accuracy, ensuring that an evaluated example does not contribute to its own class centroid.

The confidence interval resamples the fixed leave-one-out correctness indicators 5,000 times.

Model	Emotion	Contrast	Mean	95% CI	p_{raw}	p_{Holm}
AF3	Anger	HI−MD	0.071960	[−0.003917, 0.163654]	0.21410	1.00000
AF3	Anger	MD−LO	0.233920	[0.108314, 0.361680]	0.00090	**0.01620**
AF3	Disgust	HI−MD	−0.020964	[−0.124077, 0.079203]	0.57225	1.00000
AF3	Disgust	MD−LO	−0.032823	[−0.137112, 0.070983]	0.62350	1.00000
AF3	Fear	HI−MD	0.250948	[0.139751, 0.354956]	0.00015	**0.00285**
AF3	Fear	MD−LO	0.128357	[0.028183, 0.226998]	0.01500	0.24905
AF3	Happiness	HI−MD	0.049712	[−0.018680, 0.128332]	0.23330	1.00000
AF3	Happiness	MD−LO	0.032250	[−0.079104, 0.142648]	0.56710	1.00000
AF3	Sadness	HI−MD	−0.124625	[−0.218502, −0.039933]	0.01465	0.24905
AF3	Sadness	MD−LO	−0.033655	[−0.112492, 0.042089]	0.32410	1.00000
Qwen2	Anger	HI−MD	0.000016	[−0.000587, 0.000636]	0.96245	1.00000
Qwen2	Anger	MD−LO	−0.000951	[−0.001982, −0.000101]	0.04985	0.59820
Qwen2	Disgust	HI−MD	−0.000115	[−0.002964, 0.002740]	0.89390	1.00000
Qwen2	Disgust	MD−LO	0.000463	[−0.002166, 0.003079]	0.74670	1.00000
Qwen2	Fear	HI−MD	0.001698	[0.001094, 0.002280]	0.00005	**0.00100**
Qwen2	Fear	MD−LO	0.000577	[0.000100, 0.001047]	0.02365	0.35475
Qwen2	Happiness	HI−MD	0.014779	[−0.002683, 0.047898]	0.97785	1.00000
Qwen2	Happiness	MD−LO	0.000009	[−0.001172, 0.001762]	0.99865	1.00000
Qwen2	Sadness	HI−MD	−0.001009	[−0.001854, −0.000124]	0.02430	0.35475
Qwen2	Sadness	MD−LO	0.000620	[0.000091, 0.001146]	0.02760	0.35880

Table 15: Within-CREMA-D centroid control at LM2. Each row contains 42 matched pairs from 42 speakers. AF3 denotes Audio-Flamingo-3 and Qwen2 denotes Qwen2-Audio. Estimates are mean paired cosine-similarity differences to the corresponding coarse-emotion centroid; positive values support the expected intensity ordering for both contrasts. Confidence intervals are pointwise 95% speaker-bootstrap intervals; raw p-values use two-sided speaker-level sign-flip tests. Holm adjustment covers all 20 rows as one family; bold adjusted p-values are at most 0.05. All rows were tested. Surprise and neutral are outside the eligible conditions, and no HI-versus-LO test is included.

For the null distribution, coarse targets are permuted 100 times and the complete leave-one-out statistic is recomputed after each permutation. If r of the B permuted statistics are at least as large as the observed statistic, the Monte Carlo p-value uses the finite-permutation correction

$$p = \frac{r+1}{B+1}.$$

At the nominal 0.05 threshold, Qwen2-Audio's basic-emotion separation and all three Audio-Flamingo-3 conditions exceed their respective randomized-label distributions. Qwen2-Audio's subtle–intra ($p = 0.0594$) and subtle–cross ($p = 0.1089$) conditions do not reach this threshold. The Qwen subtle-emotion estimates suggest above-null organization, although support under this test is weaker than for its basic-emotion condition.

Centroid displacement. For each model, we compute the mean raw L2 displacement over the 20 ordered base-to-added emotion conditions. Label lists are then shuffled intact across representations to form the null distribution. Qwen2-Audio produces a mean displacement of 1.985, compared with a null mean of 0.972 ± 0.121 ($p_{\text{perm}} = 0.0099$). Audio-Flamingo-3 produces a mean displacement of 12.064, compared with 3.979 ± 0.559 under permutation ($p_{\text{perm}} = 0.0099$). These values support displacement beyond that expected from the shuffled-label construction. Because raw hidden-state scales differ across architectures, their absolute L2 magnitudes should not be compared directly between models. Inference from this test concerns global displacement across the 20 conditions. Individual heatmap cells remain descriptive, and directional alignment is evaluated independently below.

Directional reference construction. For EMOS emotion e, we compute an annotation-weighted mean and normalize the resulting centroid:

$$\mathbf{m}_e = \frac{\sum_i w_{ie} \widetilde{\mathbf{h}}_i}{\sum_i w_{ie}}, \qquad \mathbf{c}_e = \frac{\mathbf{m}_e}{\|\mathbf{m}_e\|}.$$

The weight w_{ie} is the fraction of retained annotation support assigned to e, after excluding *unsure*. Canonical coarse-emotion anchors $\mathbf{c}_b$ use the same weighting but retain only samples in which the canonical label has maximal support; ties are retained. Fine-emotion centroids use all samples supporting the corresponding label.

The mapped subtle directions are annoyance and disapproval for anger; disapproval for disgust; nervousness for fear; amusement, excitement, pride, optimism, love, admiration, relief, gratitude, car-

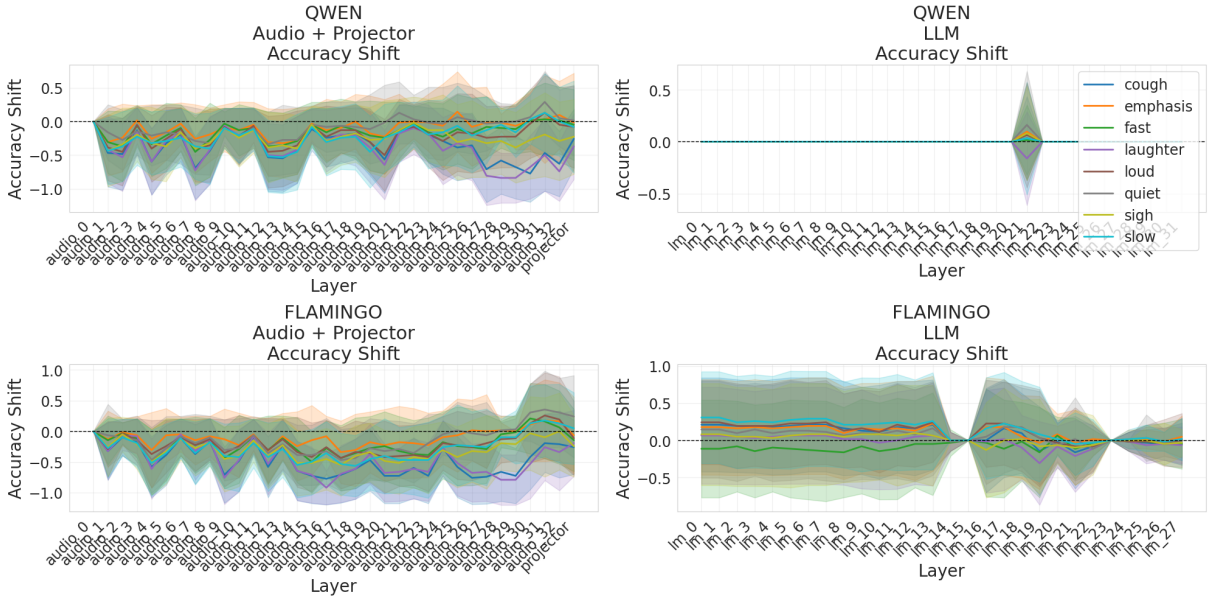


Figure 14: Layer-wise representation sensitivity under different speech cue perturbations. Different perturbation types produce distinct representation shift patterns across model layers, while overall sensitivity trends remain relatively consistent.

Model	Condition	N	SepAcc	95% CI	Permutation null	Permutation p
Qwen2-Audio	Basic	649	0.448	[0.411, 0.490]	0.065 ± 0.073	0.0099
Qwen2-Audio	Subtle–intra	800	0.278	[0.248, 0.309]	0.102 ± 0.074	0.0594
Qwen2-Audio	Subtle–cross	800	0.293	[0.261, 0.325]	0.170 ± 0.069	0.1089
Audio-Flamingo-3	Basic	649	0.838	[0.809, 0.866]	0.059 ± 0.111	0.0099
Audio-Flamingo-3	Subtle–intra	800	0.359	[0.325, 0.391]	0.096 ± 0.081	0.0099
Audio-Flamingo-3	Subtle–cross	800	0.351	[0.318, 0.385]	0.173 ± 0.072	0.0099

Table 16: Leave-one-out nearest-centroid separability at LM2. The permutation-null column reports the mean and standard deviation across 100 randomized-label permutations. Confidence intervals are obtained from 5,000 bootstrap resamples. Permutation p-values are comparison-specific and unadjusted.

ing, and desire for happiness; and disappointment, remorse, and grief for sadness. These direction sets are specific to this analysis and are distinct from the single-target coarse mapping used for separability.

Directional alignment. The cross-dataset analysis uses fixed EMOS-derived coarse-to-subtle directions and CREMA-D intensity transitions at LM2. Each transition pair contributes one cosine-alignment score, averaged over its mapped subtle directions; individual directions are not treated as independent observations. We test the mean score against zero for each of the two models, five coarse emotions, and two transitions, yielding 20 two-sided tests. Each condition contains 42 transition pairs from 42 distinct speakers.

We also test the mean within-anchor contrast $a_{i,\mathrm{HI}} - a_{i,\mathrm{LO}}$ for each model and emotion. Matching requires the same model, layer, coarse emotion, speaker, and utterance. The complete subsets contain 19, 18, 17, 25, and 22 anchors for anger, disgust, fear, happiness, and sadness, respectively, in each model. These total 101 matched emotion-specific anchors, consistent with Appendix B; they do not represent 101 distinct speakers across emotions. Unmatched observations remain in the alignment-versus-zero tests but do not enter the paired contrasts. The paired estimates therefore concern the matched subsets.

For each hypothesis, the statistic is the absolute pair-weighted mean alignment or matched difference. We generate 19,999 Monte Carlo sign patterns, applying a common sign to every observation from a speaker, and calculate $p = (r + 1)/20{,}000$, where r counts simulated statistics at least as extreme as the observed statistic, including numerical ties. The tests assume independent speaker clusters and joint sign symmetry of each speaker's score or difference vector under the null; a zero mean alone does not guarantee this assumption. For paired contrasts, sign reversal corresponds to swapping HI

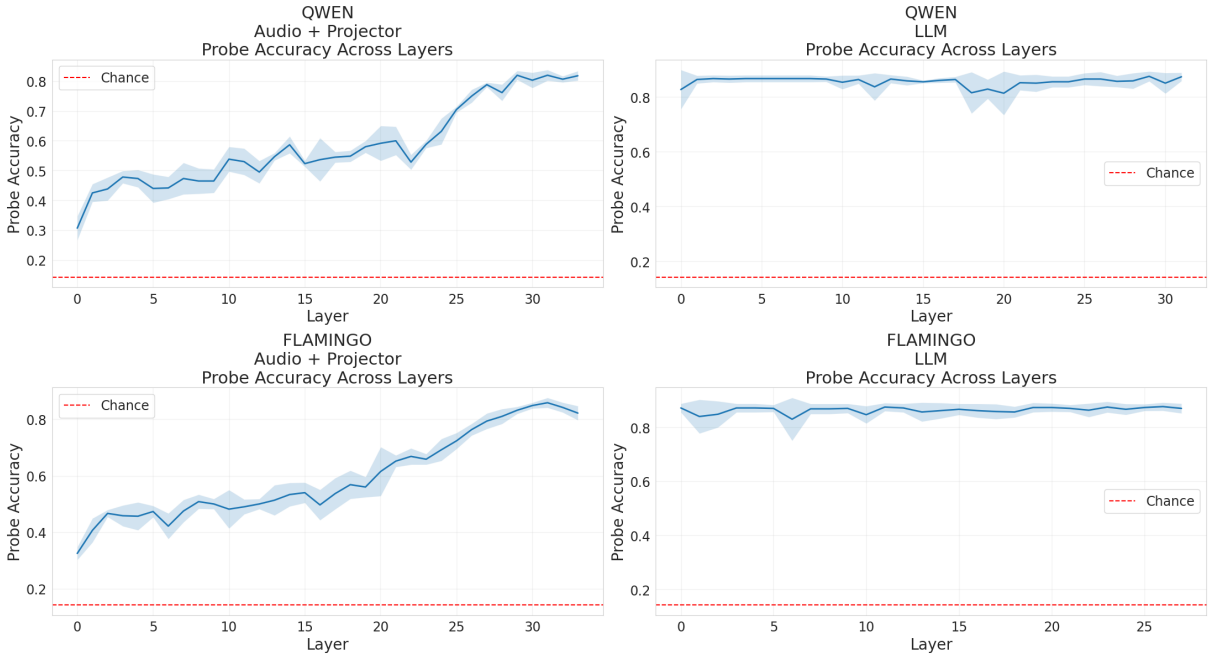


Figure 15: Probe accuracy across layers under three random seeds. Results show relatively small variance across runs, indicating stable probing behavior.

and LO jointly within each speaker. Confidence intervals use 9,999 bootstrap draws of whole speakers, retaining all their observations and recomputing the pair-weighted mean. The resampling seed is 20261002, with separate deterministic streams for each hypothesis and resampling method.

Holm correction is applied jointly to all 30 tests: 20 alignment-versus-zero hypotheses and 10 matched HI-minus-LO hypotheses. Significance is determined from the adjusted p-values. The percentile-bootstrap intervals are pointwise, are not adjusted for multiple comparisons, and are not inversions of the sign-flip tests; interval exclusion of zero and test significance need not coincide. Inference is conditional on the fixed EMOS reference directions and does not include uncertainty from estimating those centroids. These tests do not compare the semantic reference directions with randomly selected directions or directly compare the two models.

Directional-alignment results. Five alignment-versus-zero tests survive correction: positive anger MD→LO alignment in Audio-Flamingo-3 ($p_{\text{Holm}} = 0.01150$) and Qwen2-Audio (0.00150), negative fear MD→HI alignment in Audio-Flamingo-3 (0.00280) and Qwen2-Audio (0.00150), and positive happiness MD→LO alignment in Qwen2-Audio (0.01650). Four paired contrasts survive correction, all with higher alignment for MD→LO: Audio-Flamingo-3 anger (0.00750) and fear (0.00650), and Qwen2-Audio fear (0.00405) and happiness (0.00750). Qwen2-Audio's paired anger contrast does not survive correction (0.05145). Full results, including nonsignificant conditions, appear in Tables 17 and 18. The remaining estimates are reported with their uncertainty and are interpreted as unresolved under the corrected tests.

Residual reintegration. Reintegration robustness is evaluated at LM2 using the fixed 3,730/786/778 train/validation/test split and seeds 42–46. Within each seed and interpolation strength, a paired bootstrap resamples the same 778 held-out examples 5,000 times and recomputes the Macro-F1 difference relative to $\alpha = 0$. Seed-level robustness is evaluated using a two-sided Wilcoxon signed-rank test over the five matched differences.

Tables 4 and 19 report the same five-seed LM2 experiment. The earlier middle-layer single-run estimates are retained separately in Table 20. Baselines and gains are specific to each configuration; the bootstrap intervals and Wilcoxon p-values reported here apply to the LM2 experiment.

At $\alpha = 0.20$, all five Qwen2-Audio seed-specific paired-bootstrap intervals lie above zero: [0.0055, 0.0235], [0.0046, 0.0242], [0.0053, 0.0267], [0.0100, 0.0328], and [0.0082, 0.0275]. Nevertheless, with only five seeds the two-sided

Model	Emotion	Transition	Mean	95% CI	p_{raw}	p_{Holm}
AF3	Anger	MD→HI	−0.0524	[−0.1394, 0.0315]	0.27740	1.00000
AF3	Anger	MD→LO	0.2077	[0.1005, 0.3170]	0.00050	**0.01150**
AF3	Disgust	MD→HI	0.0056	[−0.0492, 0.0713]	0.86620	1.00000
AF3	Disgust	MD→LO	−0.0115	[−0.0643, 0.0335]	0.67275	1.00000
AF3	Fear	MD→HI	−0.1876	[−0.2714, −0.1059]	0.00010	**0.00280**
AF3	Fear	MD→LO	0.1303	[0.0497, 0.2154]	0.00430	0.08170
AF3	Happiness	MD→HI	−0.0152	[−0.0477, 0.0132]	0.36030	1.00000
AF3	Happiness	MD→LO	0.0193	[−0.0360, 0.0740]	0.49625	1.00000
AF3	Sadness	MD→HI	0.0106	[0.0004, 0.0212]	0.05750	0.86250
AF3	Sadness	MD→LO	−0.0023	[−0.0084, 0.0036]	0.48735	1.00000
Qwen2	Anger	MD→HI	−0.0032	[−0.0105, 0.0037]	0.39285	1.00000
Qwen2	Anger	MD→LO	0.0182	[0.0103, 0.0258]	0.00005	**0.00150**
Qwen2	Disgust	MD→HI	−0.0051	[−0.0132, 0.0031]	0.22795	1.00000
Qwen2	Disgust	MD→LO	0.0056	[−0.0016, 0.0128]	0.14005	1.00000
Qwen2	Fear	MD→HI	−0.0458	[−0.0554, −0.0358]	0.00005	**0.00150**
Qwen2	Fear	MD→LO	0.0256	[0.0101, 0.0405]	0.00335	0.06700
Qwen2	Happiness	MD→HI	−0.0188	[−0.0405, −0.0025]	0.02955	0.50235
Qwen2	Happiness	MD→LO	0.0168	[0.0073, 0.0255]	0.00075	**0.01650**
Qwen2	Sadness	MD→HI	−0.0002	[−0.0022, 0.0017]	0.84225	1.00000
Qwen2	Sadness	MD→LO	0.0020	[−0.0003, 0.0044]	0.10355	1.00000

Table 17: Alignment-versus-zero inference at LM2 for all 20 model/emotion/transition conditions. Each row uses 42 transition pairs from 42 speakers. AF3 denotes Audio-Flamingo-3 and Qwen2 denotes Qwen2-Audio. Effects are mean cosine alignments averaged over the mapped subtle directions. Confidence intervals are pointwise 95% speaker-bootstrap intervals. Raw p-values come from two-sided speaker-level sign-flip tests; Holm adjustment uses these 20 tests and the 10 contrasts in Table 18 as one family. Bold adjusted p-values satisfy $p_{\text{Holm}} \leq 0.05$.

Wilcoxon test has coarse resolution and returns $p = 0.0625$. For Audio-Flamingo-3, three of the five intervals at $\alpha = 0.20$ lie above zero, while two cross zero, producing greater seed variability and $p = 0.1875$. The bootstrap intervals quantify held-out-example variability conditional on a training seed, whereas the Wilcoxon test evaluates consistency across seeds. Thus, the LM2 results suggest a positive average contribution from reintegration at this interpolation strength, with more consistent gains across Qwen2-Audio seeds, although neither seed-level test reaches the 0.05 threshold. This qualification concerns the reliability of the reintegration gain in this setting.

Interpretation of the reintegration control. The modality-ablation results motivated examining reintegration at LM2, where the joint representation remains close to the speech-only representation. The positive mean changes in Table 19 are compatible with complementary affective information in projected speech features, although their seed-level uncertainty limits the strength of the performance claim. One possible explanation for the modest gains is that substantial speech-conditioned information already survives at this early layer, leaving limited additional information for simple residual reintegration to recover. Reintegration therefore complements the geometry and perturbation evidence as an exploratory analysis of information use; its effectiveness depends on the evaluated layer, interpolation strength, and training seed.

Descriptive analyses. The controlled-perturbation behavioral curves, prompting comparisons, modality-similarity measurements, and Gemini results characterize observed effect directions and magnitudes. Their scope is descriptive, complementing the dedicated inferential analyses of centroid geometry, cross-dataset alignment, within-CREMA-D centroid changes, and reintegration. Statistical conclusions are attached to the specific tested quantities and conditions; descriptive comparisons, including accuracy relative to the chance baseline, are interpreted at the level of the reported estimates.

I Representation Reinjection Details

Early- and middle-layer settings. Table 4 reports the five-seed LM2 follow-up for both models, matching the inferential analysis in Appendix H. Table 20 retains the original middle-layer single-run estimates alongside the LM2 results at $\alpha = 0.10$. The original middle-layer point gains differ from the corresponding five-seed LM2 mean gains, providing a descriptive comparison across the evaluated configurations. Layer and run protocol vary together: the LM2 follow-up uses validation-selected decision thresholds and seeds 42–46, while the middle-layer rows retain the ear-

Model	Emotion	N	HI−LO	95% CI	p_{raw}	p_{Holm}
AF3	Anger	19	−0.3893	[−0.5533, −0.2270]	0.00030	**0.00750**
AF3	Disgust	18	−0.0349	[−0.0799, −0.0058]	0.02240	0.40320
AF3	Fear	17	−0.3618	[−0.4954, −0.2326]	0.00025	**0.00650**
AF3	Happiness	25	−0.0670	[−0.1333, −0.0047]	0.05290	0.84640
AF3	Sadness	22	0.0133	[0.0010, 0.0266]	0.06920	0.96880
Qwen2	Anger	19	−0.0231	[−0.0361, −0.0108]	0.00245	0.05145
Qwen2	Disgust	18	−0.0115	[−0.0250, 0.0019]	0.12180	1.00000
Qwen2	Fear	17	−0.0589	[−0.0779, −0.0384]	0.00015	**0.00405**
Qwen2	Happiness	25	−0.0472	[−0.0836, −0.0221]	0.00030	**0.00750**
Qwen2	Sadness	22	−0.0018	[−0.0052, 0.0014]	0.32480	1.00000

Table 18: Matched MD→HI minus MD→LO alignment contrasts at LM2. N is the number of complete speaker/utterance/emotion matches and also the number of distinct speakers within each row. Negative estimates indicate higher alignment for MD→LO. Estimates concern the matched subsets, not differences between the full-condition means. AF3 denotes Audio-Flamingo-3 and Qwen2 denotes Qwen2-Audio. Confidence intervals are pointwise 95% speaker-bootstrap intervals. Bold adjusted p-values satisfy $p_{\text{Holm}} \leq 0.05$ in the joint family of 30 tests. These are the remaining 10 tests alongside Table 17.

Model	**Alpha**	**Mean Macro-F1 change**	**SD across seeds**	**Wilcoxon p**
Qwen2-Audio	0.01	−0.0001	0.0008	1.0000
	0.03	+0.0004	0.0009	0.6250
	0.05	+0.0029	0.0026	0.0625
	0.10	+0.0057	0.0025	0.0625
	0.20	+0.0170	0.0028	0.0625
Audio-Flamingo-3	0.01	−0.0003	0.0008	0.6250
	0.03	−0.0026	0.0066	1.0000
	0.05	+0.0002	0.0013	1.0000
	0.10	+0.0001	0.0052	1.0000
	0.20	+0.0129	0.0168	0.1875

Table 19: Five-seed residual-reintegration robustness at LM2. Changes are calculated relative to the matched no-reintegration condition within each seed. Wilcoxon p-values are obtained from two-sided signed-rank tests over the five seed-level changes.

lier single-run point estimates. Baselines and gains are therefore interpreted within each configuration, with statistical conclusions based on the five-seed LM2 experiment.

The five-seed LM2 follow-up uses the 5,294-utterance cleaned EmoS analysis subset and a fixed multilabel-stratified split of 3,730 training, 786 validation, and 778 test utterances. The split is constructed once with random seed 42 and is held fixed across interpolation strengths and probe seeds.

For each utterance, the saved projector and LM2 representations are mean pooled when necessary and standardized using the feature mean and standard deviation of that representation. Reintegration forms the sum of the LM2 representation and the standardized projector representation multiplied by the interpolation strength. We evaluate interpolation strengths of 0, 0.01, 0.03, 0.05, 0.10, and 0.20.

The probe is a single linear layer with 28 outputs. Normalized human annotation counts are used as soft training targets with binary cross-entropy loss on the logits. Training uses Adam with learning rate 0.001, batch size 64, and ten epochs. For each interpolation strength, training is repeated with seeds 42–46. The checkpoint with the highest validation Macro-F1 is retained. A shared decision threshold is selected on the validation set from values 0.05, 0.10, ..., 0.90 and is then applied once to the held-out test set.

Tables 21 and 22 retain the earlier exploratory direct-fusion and layer-probing results. These are separate configurations from the five-seed LM2 residual experiment; their point estimates provide descriptive comparisons within each table.

J Scope of Model Coverage and Future Validation

The evaluated systems cover different language backbones, audio integration designs, and open- versus closed-source access. This diversity allows us to characterize the representation–utilization gap across the studied configurations. Internal representation evidence comes from Qwen2-Audio and Audio-Flamingo-3, complemented by the behavioral comparison with Gemini-2.5-flash.

Our conclusions concern these evaluated mod-

Model	Layer	Run setting	Macro-F1		Micro-F1	
			$\alpha = 0$	$\alpha = 0.10$	$\alpha = 0$	$\alpha = 0.10$
Qwen2	LM2 (early)	Five seeds	0.2166 ± 0.0123	0.2223 ± 0.0107	—	—
Qwen2	LM15 (middle)	Original single run	0.217	0.235	0.403	0.448
AF3	LM2 (early)	Five seeds	0.2217 ± 0.0076	0.2217 ± 0.0081	—	—
AF3	LM13 (middle)	Original single run	0.212	0.222	0.438	0.442

Table 20: Residual reintegration in the early- and middle-layer settings. Qwen2 denotes Qwen2-Audio and AF3 denotes Audio-Flamingo-3. Five-seed Macro-F1 values are means $\pm$ sample SD across seeds 42–46, calculated from the saved per-seed outputs; dashes denote unreported Micro-F1 values. Original single-run rows retain the earlier middle-layer Macro-F1 and Micro-F1 point estimates. This descriptive comparison uses a common interpolation strength across separate run settings.

α	**Macro-F1**	**Micro-F1**
0.00	**0.216**	0.388
0.25	0.190	0.391
0.50	0.170	**0.404**
0.75	0.164	0.359
1.00	0.103	0.303

Table 21: Direct linear fusion results across different interpolation ratios α for Flamingo. Naive interpolation generally degrades performance relative to the language-only baseline.

Layer	**Macro-F1**	**Micro-F1**
Audio Encoder (32)	0.078	0.270
Projector	0.117	0.328
LM Layer 13	**0.212**	0.414
LM Layer 27	0.207	**0.439**

Table 22: Subtle emotion probing results across different Flamingo representation layers. Later language representations exhibit substantially stronger recoverable affective structure than early audio representations alone.

els and settings. Broader validation across architectures, training paradigms, and fusion strategies would establish how widely the observed pattern extends. Controlled comparisons of speech encoders, projector designs, and language backbones offer a next step toward explaining its architectural causes and improving the use of preserved affective information.