DO OMNI MODELS USE HOW WE SPEAK? DIAGNOSING AND IMPROVING PARALINGUISTIC UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Strong performance on speech and multimodal benchmarks does not establish whether omni models use how an utterance is spoken to guide their decisions. We investigate this question through controlled translation selection, asking models to choose the translation better synchronized with the audio while varying vocal expression, holding the transcript and candidate translations fixed, and testing both candidate orders. Evaluated models, including commercial closed-source models, often fail to adjust their choices correctly as vocal expression changes, and most choices persist even when speech is replaced with silence, indicating limited use of vocal information beyond the transcript. We introduce **P**aralinguistic **C**ounterfactual **T**raining (**PCT**), which combines same-transcript audio pairs with opposite target choices and human translation-quality preferences, using candidate-order augmentation and supervised low-rank adaptation. On Qwen2.5-Omni-3B and 7B, PCT increases paired correctness from 2.65% to 49.29% and from 8.17% to 50.71%, respectively, with gains also observed in four further post-trained omni models. Additional evaluations show transfer to general audio understanding and paralinguistic tasks, and largely preserved performance on other evaluated input modalities for Qwen2.5-Omni. These findings show that explicitly linking vocal expression to correct decisions can improve audio use beyond the supervision task.

1 INTRODUCTION

Omni models integrate text, images, audio, and video within a unified system, supporting tasks ranging from speech recognition and generation to multimodal understanding (Xu et al., 2025). Although advanced omni models perform well on benchmarks for speech recognition, speech generation, and multimodal understanding (Ghosh et al., 2025), these results do not establish whether their decisions use not only what we say but also how we speak (Schuller et al., 2013; Fujisaki, 2004). Figure 1 illustrates this distinction: identifying a speaker's anger does not necessarily lead to selecting a translation that preserves it.

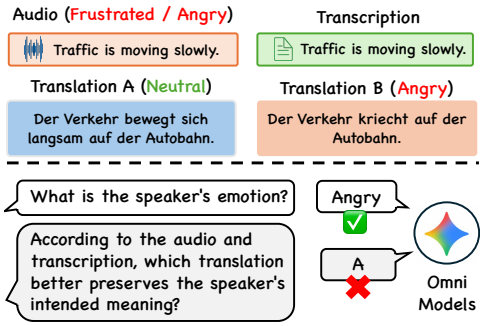


Figure 1: Illustration of the recognition–utilization gap.

We study this question through controlled translation selection. Using ContraProST examples (Tsiamas et al., 2024), we present recordings with the same transcript but contrasting vocal expressions that support different candidate translations. The transcript and candidate content remain fixed across audio conditions, allowing us to test whether the model selects the appropriate translation for each recording. This setting separates the ability to make plausible judgments from the requirement to respond correctly to task-relevant audio evidence.

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

Our diagnosis reveals that the evaluated models, including commercial closed-source models, frequently fail to make these audio-dependent distinctions. Their paired correctness is low, even when they achieve relatively high agreement with human preferences in general translation-quality evaluation. Moreover, replacing speech with silence while retaining all textual inputs leaves 60.1%–87.1% of their choices unchanged. Together, these observations indicate that the models do not reliably use vocal information beyond the transcript to guide translation selection. This behavioral limitation motivates targeted training to improve the use of vocal information while preserving the general capabilities of omni models.

We then investigate whether targeted supervision can improve this behavior through a simple intervention. We introduce Paralinguistic Counterfactual Training (PCT), which uses contrasting, same-transcript recordings to associate identical inputs with different correct judgments. With both audio conditions equally represented, a predictor that ignores audio cannot correctly resolve the complete pair. We combine these examples with candidate-order augmentation and human translation-quality preferences, and train using supervised objective with low-rank adaptation (Hu et al., 2021). This construction allows us to test whether supervision that makes vocal information necessary for the answer can improve the observed behavior without redesigning the model architecture.

Experiments show that PCT substantially improves the diagnosed behavior. On Qwen2.5-Omni-3B and 7B, paired correctness increases from 2.65% to 49.29% and from 8.17% to 50.71%, respectively, while agreement with human translation preferences also improves over the base models. Additional evaluations demonstrate transfer to general audio understanding and paralinguistic tasks, with Qwen2.5-Omni-3B gaining 5.49 percentage points on MMSU and 31.05 points on SAVEE emotion recognition. Layer-wise analysis further reveals a shift in vocal sensitivity toward deeper layers, suggesting that PCT changes the model's internal processing of vocal information. Performance on other evaluated input modalities is largely preserved for Qwen2.5-Omni, suggesting that targeted audio supervision can complement existing multimodal capabilities.

Our contributions are threefold:

- We diagnose unreliable use of vocal expression through text-controlled translation selection and silence replacement, revealing limitations across the evaluated open and commercial closed-source models.
- We introduce PCT, a simple counterfactual supervision scheme implemented with standard supervised training and LoRA, and evaluate its ability to improve the diagnosed behavior.
- We evaluate transfer beyond translation selection, demonstrating benefits on broader audio and paralinguistic tasks while identifying model-dependent limitations and largely preserved performance on other evaluated input modalities.

2 Diagnosing the Recognition–Utilization Gap

Recognizing vocal cues does not necessarily imply using them correctly in decisions. We assess whether models select the appropriate translation when vocal expression changes the intended meaning while the transcript and candidate translations remain fixed.

2.1 Controlled Task and Evaluation

We adapt ContraProST (Tsiamas et al., 2024) into a translation-selection task. Each example $g = (a_1, a_2, t, c_1, c_2)$ contains two recordings sharing transcript t, with a_r supporting translation c_r. The model receives one recording, the transcript, and both candidates, and returns their position label A/B or TIE for insufficient evidence. All examples have explicit A/B gold labels; TIE counts as incorrect.

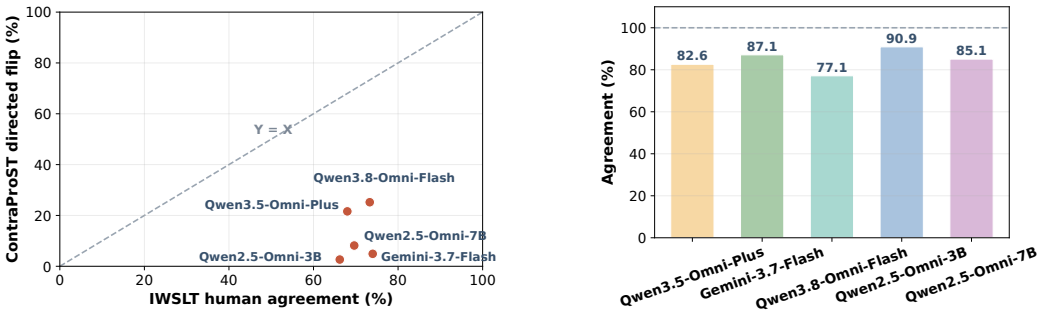
Crossing the two recordings with both candidate orders gives:

$$\mathcal{V}(g) = \begin{Bmatrix} ((a_1, t, c_1, c_2), A), & ((a_1, t, c_2, c_1), B), \\ ((a_2, t, c_1, c_2), B), & ((a_2, t, c_2, c_1), A) \end{Bmatrix}. \tag{1}$$

Switching audio changes the preferred translation; swapping candidate order changes only its position label.

```
Prompt: You are a strict reference-free evaluator of speech-to-text translation.
Judge whether the candidate translation preserves the meaning intended by the source speech.
...
Return exactly one JSON object and no markdown.
Compare two {target} translations of the same source utterance.
Listen to the attached source audio before deciding.
** Treat audio as the primary evidence for speech-only meaning. **
...
Candidate A: {candidate_a}
Candidate B: {candidate_b}
{output_instruction}
Choose TIE only when the available source evidence genuinely does notfavor either candidate.
```

Figure 2: Prompt template for conterfactual speech pair reformulated as a translation-selection task.



(a) Human agreement versus directed flip.

(b) Real–silence choice agreement.

Figure 3: Diagnosis of audio evidence use. (a) IWSLT human agreement versus ContraProST directed flip; the dashed diagonal marks numerical equality. (b) Choice agreement after replacing real audio with silence while keeping the instruction, transcript, and candidates unchanged.

Directed Flip (DF) measures whether both audio conditions are answered correctly at a fixed candidate order:

$$\mathrm{DF} = \frac{100}{2N} \sum_{i=1}^{N} \sum_{o \in \{\mathrm{AB},\mathrm{BA}\}} \mathbf{1}[\hat{y}_{i,1,o} = y_{i,1,o}]\mathbf{1}[\hat{y}_{i,2,o} = y_{i,2,o}], \tag{2}$$

where $\hat{y}_{i,r,o}$ and $y_{i,r,o}$ are the predicted and gold labels for example i, recording r, and order o. A deterministic text-only predictor cannot answer both conditions correctly because their textual inputs are identical.

The test split contains $N = 283$ language-specific examples from 154 source-sentence groups across English-to-German, English-to-Spanish, and English-to-Japanese translation, yielding 1,132 judgments and 566 paired events per model. Source-sentence groups and their cross-language variants are disjoint across splits. Prompt and split details appear in Figure 2 and Appendix C.1.

2.2 Quality Judgment Does Not Guarantee Audio-Conditioned Judgment

As shown in Figure 3a, the five evaluated models achieve IWSLT human-preference agreement of 66.21%–74.0%, while their DF scores range from 2.65% to 25.22%. Gemini-3.7-Flash and Qwen3.8-Omni-Flash achieve human agreement of 74.0% and 73.29%, respectively, but their DF scores are 4.9% and 25.22%. Thus, general translation-quality evaluation does not adequately characterize whether models select the appropriate translation based on vocal expression.

2.3 Choice Stability under Silence Replacement

We further replace each recording with a fixed 0.5-second silent waveform while keeping all textual inputs unchanged. As shown in Figure 3b, real–silence choice agreement ranges from 77.1% to 90.9% across the five evaluated models, reaching 90.9% for Qwen2.5-Omni-3B. Most choices therefore persist after speech is removed. Unchanged choices alone do not establish that a model entirely ignores audio. Together with the low DF scores, however, these results indicate that the

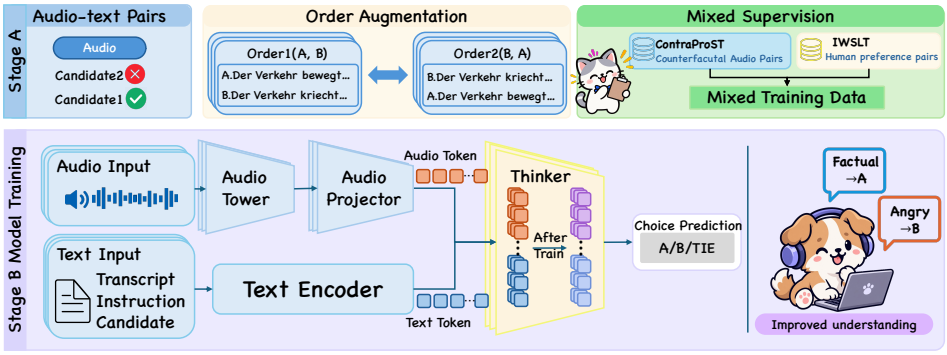


Figure 4: PCT combines counterfactual audio judgments and human quality preferences with candidate-order augmentation, then trains LoRA adapters while freezing pretrained weights.

evaluated models do not reliably adjust their judgments to changes in vocal meaning. This observation motivates directly supervising the relation between audio and choice: keeping text fixed while assigning opposite correct answers to contrasting audio conditions. PCT adopts this relation as its training signal.

3 A Lightweight Intervention: Paralinguistic Counterfactual Training

PCT combines the counterfactual judgments defined in Section 2.1 with human quality preferences through supervised low-rank adaptation. Figure 4 summarizes the framework.

3.1 Training Set Construction

We construct $\mathcal{D}_C$ by including the four labeled inputs in Equation 1 for each ContraProST training example. Each input is trained independently.

We construct $\mathcal{D}_Q$ from human scores in the IWSLT speech translation evaluation development set, retaining candidate pairs with sufficiently different scores and presenting each in both orders. The higher-scoring translation is labeled as preferred. Both sources use A/B targets and contribute 5,356 views each, yielding 10,712 equally mixed training views.

3.2 Parameter-Efficient Training

For Qwen2.5-Omni (Xu et al., 2025), we apply LoRA (Hu et al., 2021) to selected attention and feed-forward projections in the Thinker backbone, freezing all pretrained weights. Only the A/B choice-value token in the target JSON contributes to the loss:

$$\ell(x, y; \phi) = -\log p_{\theta_0,\phi}\left(v(y) \mid x, s_{\text{prefix}}\right), \tag{3}$$

where θ_0 and ϕ are frozen and trainable parameters, $v(y)$ is the label token, and s_{prefix} is the preceding JSON prefix. Probabilities use the original full-vocabulary output head. With equal source sizes, the objective is

$$\mathcal{L}(\phi) = \tfrac{1}{2}\mathcal{L}_Q(\phi) + \tfrac{1}{2}\mathcal{L}_C(\phi), \tag{4}$$

where each term is the mean choice loss for its source. No additional contrastive or consistency loss is used. Construction and implementation details are provided in Appendix B.

4 Experiment

4.1 Experimental Setup

Models. Our experiments cover Qwen2.5-Omni-3B and Qwen2.5-Omni-7B (Xu et al., 2025), together with four post-trained omni variants, and further include Gemma-4-E4B (Team et al., 2026) as a different backbone,providing broad model coverage for our evaluation.

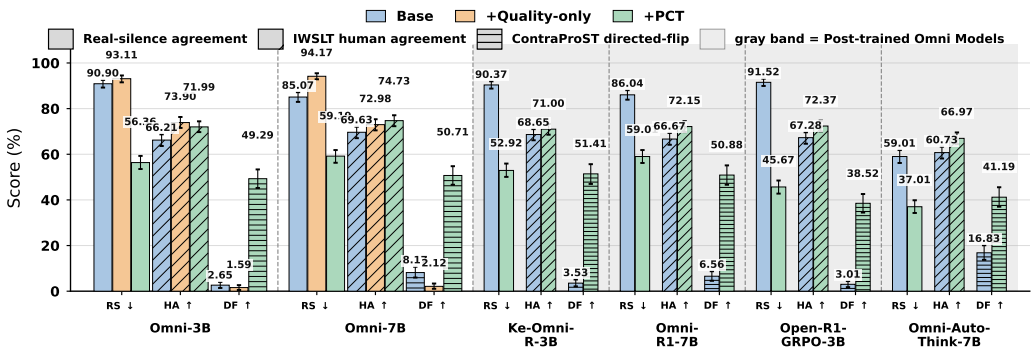


Figure 5: Audio-grounded judgment and translation-quality evaluation. For Omni-3B/7B, bars show Base, +Quality-only, and +PCT; for post-trained omni models, bars show Base and +PCT only. Error bars are 95% bootstrap confidence intervals. The gray band marks post-trained omni models.

Evaluation Datasets. Our evaluation consists of two parts: targeted audio-grounded judgment and transfer to general audio understanding. For targeted evaluation, ContraProST (Tsiamas et al., 2024) contains 1,132 real-audio trials derived from 283 examples, while the IWSLT evaluation contains 1,314 pairwise trials. For the real–silence control, we replace only the audio in the 1,132 ContraProST trials with a fixed silent waveform, yielding 1,132 matched silence trials. For general audio transfer, we evaluate on MMAU test-mini (Sakshi et al., 2025), AIR-Bench Foundation (Yang et al., 2024), and MMSU (WANG et al., 2026), using 1,000, 24,683, and 4,994 examples, respectively.

Metrics and Evaluation Protocol. For targeted translation judgment, we report three complementary metrics. IWSLT Human Agreement measures agreement between model choices and human translation-quality preferences over 1,314 pairwise trials. ContraProST Directed Flip counts an order-pair as correct only when both same-transcript recordings with contrasting prosody are judged correctly under the same candidate order. The definition of DF is provided in Section 2. The full evaluation contains 566 DF pairs (283 examples $\times$ two orders). Real–silence choice agreement measures the proportion of unchanged choices between the 1,132 real-audio trials and their matched silence counterparts. Further implementation and scoring details are provided in the appendix C.1.

4.2 Main Results on Counterfactual Grounding and Translation Quality

Figure 5 shows that PCT improves both agreement with human preferences and counterfactual audio judgment. For Qwen2.5-Omni-3B and 7B, PCT increases IWSLT human agreement from 66.21% to 71.99% and from 69.63% to 74.73%, respectively; ContraProST DF increases from 2.65% to 49.29% and from 8.17% to 50.71%. This improvement also extends to post-trained omni models, where PCT raises DF from a Base range of 3.01–16.83% to 38.52–51.41% while improving IWSLT human agreement in all cases. These gains reflect not only better agreement with human quality preferences but also more accurate responses to changes in vocal meaning.

The Quality-only control further shows that higher quality-judgment scores do not necessarily lead to more reliable use of vocal information. Training solely on human quality preferences increases IWSLT human agreement to 73.90% for 3B and 72.98% for 7B, while DF decreases to 1.59% and 2.12%, respectively. Notably, at 3B, Quality-only even exceeds PCT's human agreement of 71.99%, yet performs substantially worse on counterfactual judgments. Within the same backbone, this comparison reproduces the central observation from the preliminary analysis: better agreement with human translation-quality preferences does not imply an ability to correctly distinguish recordings that share a transcript but convey different meanings. These results support the value of explicitly supervising the correspondence between vocal expression and candidate preference.

Real–silence comparisons provide complementary behavioral evidence. Choice agreement increases to 93.11% and 94.17% under Quality-only, but decreases to 56.36% and 59.19% under PCT; the same reduction is observed across all four post-trained models. Lower choice agreement alone does not imply more accurate judgments; together with the higher DF scores, however, these results show that PCT changes its choices more often when audio is replaced and more frequently makes

Table 1: Results on general audio understanding. We report accuracy (%). Avg. is the arithmetic mean of the three benchmark scores. Bold indicates the higher score within each Base/+PCT pair.

Model	MMAU ↑		AIR-Bench ↑		MMSU ↑		Avg. ↑	
	Base	+PCT	Base	+PCT	Base	+PCT	Base	+PCT
Omni Models								
Omni-3B	69.50	**72.10**	67.38	**68.99**	60.43	**65.92**	65.77	**69.00**
Omni-7B	76.50	**76.80**	70.10	**70.98**	68.88	**70.30**	71.83	**72.69**
Gemma-4-E4B	**63.30**	63.10	40.02	**41.14**	62.74	**65.14**	55.35	**56.46**
Post-trained Omni Models								
Ke-Omni-R-3B	57.50	**57.60**	**69.54**	69.11	67.88	**67.94**	**64.97**	64.88
Omni-R1-7B	**76.70**	75.90	69.99	**70.86**	68.96	**71.83**	71.88	**72.86**
Open-R1-GRPO-3B	**69.40**	69.00	66.78	**67.05**	**62.43**	62.05	**66.20**	66.03
Omni-AutoThink-7B	73.20	**75.30**	60.08	**60.98**	63.94	**66.36**	65.74	**67.55**

Table 2: Ablation of supervision sources on Qwen2.5-Omni-3B. Avg. is the arithmetic mean of the three benchmark scores. Bold indicates the best result in each column.

Training variant	MMAU ↑	AIR-Bench ↑	MMSU ↑	Avg. ↑
Base	69.50	67.38	60.43	65.77
+ Quality-only	69.00	66.98	55.87	63.95
+ Counter-only	63.40	66.45	54.06	61.30
+ PCT	**72.10**	**68.99**	**65.92**	**69.00**

correct adjustments when vocal meaning changes. Overall, PCT improves counterfactual audio judgment and this effect extends to models that have already undergone additional post-training. These findings indicate that PCT not only improves evaluation scores but also yields behavioral improvements in using vocal information.

4.3 Transfer to General Audio Understanding

We further examine whether PCT benefits general audio task performance in omni models. Table 1 presents transfer results on MMAU, AIR-Bench, and MMSU for omni models and their post-trained variants. Although PCT draws supervision from counterfactual translation selection and human quality preferences, its lightweight adaptation yields general-task gains across multiple models, indicating benefits beyond candidate translation evaluation.

Omni-3B and Omni-7B improve on all three benchmarks, with average scores increasing from 65.77 to 69.00 and from 71.83 to 72.69, respectively. In particular, Omni-3B gains 5.49 percentage points on MMSU, showing transfer from translation judgment to broader spoken language understanding. Gemma-4-E4B also gains 1.12 and 2.40 percentage points on AIR-Bench and MMSU, respectively; despite a slight decline on MMAU, its average score increases from 55.35 to 56.46.

PCT also provides additional gains for some omni models that have already undergone post-training. Omni-R1-7B improves its average score from 71.88 to 72.86, while Omni-AutoThink-7B gains 2.10 and 2.42 percentage points on MMAU and MMSU, respectively. However, Ke-Omni-R-3B remains nearly unchanged on average, and Open-R1-GRPO-3B declines slightly on both reported benchmarks, indicating that transfer still depends on the model and task. These results show that lightweight training centered on counterfactual translation selection can improve the diagnosed audio-conditioned judgments while benefiting broader tasks in some omni models. Additional results on audio-text-to-text models and a related post-trained model are provided in Appendix E.

4.4 Ablation of Supervision Sources

We compare Quality-only, which uses human translation-quality preferences; Counter-only, which uses counterfactual audio supervision and PCT, which combines both data on Qwen2.5-Omni-3B.

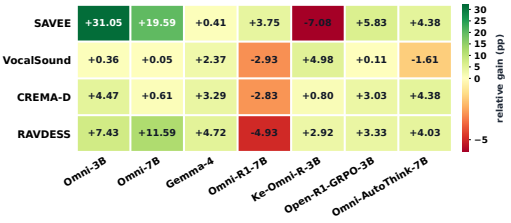


Figure 6: Transfer to paralinguistic tasks after PCT. Cells show accuracy changes relative to Base. Positive values indicate improvements.

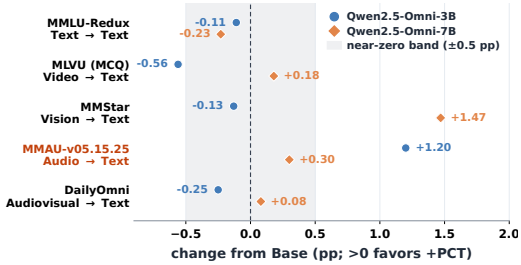


Figure 7: Cross-modal performance changes after PCT adaptation. Points show accuracy changes relative to Base in percentage points.

Table 2 shows that PCT achieves the highest score on all three benchmarks, with an average of 72.10, exceeding Base by 2.6 points and Quality-only and Counter-only by 3.1 and 8.7 points, respectively. The largest differences occur on MMSU, where PCT reaches 65.92, outperforming the two single-source variants by 10.05 and 11.86 percentage points. In contrast, both Quality-only and Counter-only obtain lower average scores than Base. These results support the value of PCT: combining quality preferences with counterfactual audio supervision yields better transfer than using either source alone.

5 Analysis

5.1 Transfer to Paralinguistic Tasks

We evaluate transfer to speech emotion recognition on SAVEE (Haq & Jackson, 2011), CREMA-D (Cao et al., 2014), and RAVDESS (Livingstone & Russo, 2018), and to sound recognition on VocalSound (Gong et al., 2022). Figure 6 shows the accuracy changes after PCT adaptation relative to Base. Omni-3B, Omni-7B, and Gemma-4 improve on all benchmarks, indicating that lightweight adaptation based on translation judgments can transfer to other paralinguistic tasks.

The magnitude of the gains varies across tasks. Omni-3B and Omni-7B gain 31.05 and 19.59 percentage points on SAVEE and 7.43 and 11.59 points on RAVDESS, respectively, while also improving on CREMA-D; their gains on VocalSound are smaller, at 0.36 and 0.05 percentage points, respectively. Gemma-4 likewise exhibits cross-task gains, including increases of 3.29 and 4.72 percentage points on CREMA-D and RAVDESS, respectively, showing that these transfer benefits are not restricted to the Qwen2.5-Omni backbones.

Omni-R1-7B, which has undergone further post-training, exhibits mixed results: a 3.75-percentage-point gain on SAVEE but decreases of 2.93, 2.83, and 4.93 percentage points on VocalSound, CREMA-D, and RAVDESS, respectively. Overall, supervision based on counterfactual translation judgments can yield cross-task gains for some models, but the direction and magnitude of transfer remain model- and task-dependent. Full Base/PCT scores for all configurations, additional results for R1-AQA, and evaluation details are provided in Appendix F.

5.2 Preservation of Cross-Modal Capabilities

Figure 7 shows performance changes across input modalities after PCT. We evaluate Qwen2.5-Omni-3B and 7B with text, video, image, audio, and audiovisual inputs, with all tasks producing text outputs. The results indicate that lightweight training focused on vocal expression largely preserves performance in the other evaluated modalities while improving performance on the audio task.

Omni-3B exhibits decreases of 0.11–0.56 points on the four non-audio tasks; Omni-7B declines by only 0.23 points on MMLU-Redux and improves on the other three tasks, including a 1.47-point gain on MMStar. Performance on the audio benchmark also improves by 1.20 and 0.30 percentage points at the two model scales, respectively. These results suggest that, within the evaluated model scales and tasks, the audio-task gains of PCT are not accompanied by substantial declines in other

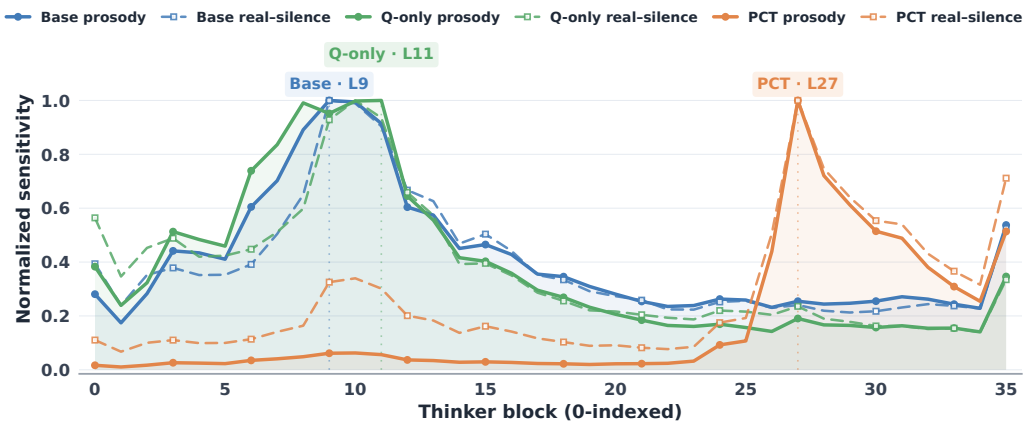
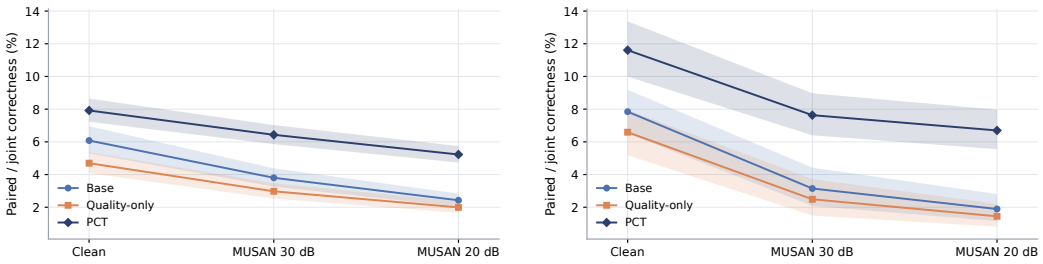


Figure 8: Sensitivity to prosodic variation and audio removal on ContraProST. Solid curves show normalized mean cosine distances between same-transcript recordings with contrasting prosody; dashed curves show those between real audio and matched silence.



(a) Robustness of Omni-3B on CREMA-D.

(b) Robustness of Omni-3B on RAVDESS.

Figure 9: Paired emotion judgments under MUSAN noise on Omni-3B. Shaded bands denote actor-clustered 95% bootstrap intervals. Noisy inputs retain the original intended-emotion labels.

modalities, supporting its use as a lightweight supplementary training approach for existing omni models. Full Base/PCT scores and evaluation sample counts are provided in Appendix G.

5.3 Layer-wise Changes in Audio Sensitivity

To examine how PCT changes internal responses to vocal information, we compare Base, Quality-only, and PCT on same data from ContraProST. At each Thinker block, we compute the $1 - \cos$ between the two prosodic variants and between real audio and matched silence, yielding profiles of prosody sensitivity and sensitivity to audio removal, respectively. Figure 8 presents the normalized profiles to compare the relative distribution of responses across layers and training configurations.

The prosody-sensitivity peaks of Base and Quality-only occur around the relatively shallow blocks 9 and 11, respectively, with their audio-removal responses also concentrated in this shallow region. In contrast, both PCT profiles peak around the deeper block 27, indicating a shift in the location of the dominant relative response. This distinction complements the behavioral results in Figure 5: Quality-only improves agreement with human quality preferences without improving counterfactual audio judgment, whereas PCT achieves higher DF alongside a different deep-layer response pattern.

5.4 Robustness of Paralinguistic Judgments

We further examine whether the paired emotion-judgment gains of PCT persist under background noise. Figure 9 compares Base, Quality-only, and PCT on Omni-3B using CREMA-D (Cao et al., 2014) and RAVDESS (Livingstone & Russo, 2018). J_{pair} requires both recordings to be classi-

fied correctly; at MUSAN (Snyder et al., 2015) SNRs of 30 and 20 dB, J_{robust} requires all four judgments on the two original recordings and their noisy counterparts to be correct.

PCT outperforms Base and Quality-only in every evaluated condition on both datasets. PCT improves paired correctness over Base by 1.84 and 3.76 percentage points on CREMA-D and RAVDESS, respectively, and over Quality-only by 3.22 and 5.02 points. At 20 dB, its joint-correctness gains over Base are 2.80 and 4.80 percentage points, with paired 95% confidence intervals of $[2.43, 3.17]$ and $[3.89, 5.91]$, respectively. Although joint correctness decreases under noise, PCT retains its relative advantage under the stricter requirement that both original and noisy recordings be classified correctly. Details are provided in Appendix H.

6 Related Work

Omni Models and Audio Grounding Recent omni models integrate multi-modal within unified architectures, supporting tasks ranging from speech understanding to multimodal reasoning. Qwen2.5-Omni (Xu et al., 2025) adopts a Thinker–Talker architecture to jointly process multiple input modalities. Gemma 4 (Team et al., 2026) further represents the emerging generation of natively multimodal models supporting audio and visual inputs. Despite their strong performance across audio and multimodal tasks, aggregate benchmark scores alone do not establish whether a model's final decisions actually depend on the task-relevant evidence contained in the audio. Recent work has increasingly targeted audio grounding through post-training. He et al. (2026) partitions training examples according to the contribution of audio evidence to the correct answer and applies different training strategies to examples with different levels of audio contribution. Chen et al. (2026) instead uses response-side counterfactual hard negatives to improve audio reasoning and reduce errors in event identity, omission, temporal relations, and counting. Omni-R1 (Rouditchenko et al., 2025) and Ke-Omni-R (Zhao et al., 2025) further apply RL to post-train omni models for audio reasoning. These approaches show that data selection, preference construction, and reinforcement learning can improve performance on audio tasks.

Paralinguistic Information Modeling in Speech Existing work has modeled paralinguistic information at the levels of speech representation, spoken interaction, and speech translation. Emotion2vec (Ma et al., 2024) combines utterance- and frame-level objectives to learn speech emotion representations transferable across languages and emotion-related tasks. ParalinGPT (Lin et al., 2024b) combines text, speech embeddings, and paralinguistic attributes in a serialized multitask framework that predicts current-turn sentiment, response sentiment, and response text. Spoken-LLM (Lin et al., 2024a) uses two-stage training on StyleTalk to produce style-sensitive responses when identical lexical content is expressed in different speaking styles. GOAT-SLM (Chen et al., 2025) combines linguistic and acoustic output heads with staged training that progressively incorporates linguistic, paralinguistic, and speaker information. ContraProST study (Tsiamas et al., 2024) finds that models encode some prosodic information, but its influence is often insufficient to change their translations. SpeechCOMET (Züfle et al., 2026) incorporates speech encoders into translation-quality estimation, yet improvements on standard evaluation do not consistently translate into correct judgments of speech-specific phenomena. Our work diagnoses and improves how models use paralinguistic information in speech understanding.

7 Conclusion

Through text-controlled translation selection, we have identified a gap between general quality judgment and reliable use of vocal expression in omni models, and introduced PCT as a lightweight intervention. PCT combines same-transcript counterfactual audio pairs, human translation-quality preferences, and candidate-order augmentation using supervised low-rank adaptation. Experiments across multiple omni models show improvements in both counterfactual audio judgment and agreement with human preferences. Evaluations demonstrate transfer to general audio and paralinguistic tasks across multiple models, alongside largely preserved performance on other evaluated modalities. The Quality-only control shows that improving quality-judgment scores alone is insufficient to improve audio-conditioned judgments, supporting the value of explicitly linking vocal expression to the correct choice in supervision. Overall, these findings show that targeted supervision can improve how models use vocal information, with benefits extending beyond the supervision task.

AI USE STATEMENT

In this work, we primarily used generative AI tools to assist with manuscript translation, code debugging, and data visualization. We designed the entire experimental workflow ourselves. We have manually reviewed all AI-assisted content and take full responsibility for the final content of this work, including the text, code, figures and experimental results.

ETHICS STATEMENT

This work does not involve new human-subject recruitment or data collection. All experiments use publicly released speech, translation, and multimodal datasets under their respective licenses and intended research uses. Our study focuses on how models use paralinguistic information in speech and translation judgments; we do not use model predictions to infer sensitive attributes of real individuals or for high-stakes decision making. We acknowledge that existing emotion and paralinguistic datasets may reflect biases in speaker demographics, acted expressions, languages, and annotation protocols. Accordingly, our findings should be interpreted within the scope of the evaluated datasets and settings, rather than as general judgments about real-world emotional or personal states.

REPRODUCIBILITY STATEMENT

To support reproducibility, we report dataset construction and splits, training-data composition, prompt templates, model adaptation details, training objectives, evaluation protocols, and statistical procedures throughout the main text and appendices. The grouping and train/test splits of ContraProST, the mixed training construction used by PCT, and the LoRA configuration are documented in the method and supplementary material. We also provide the sample counts, scoring rules, and perturbation settings used for the general audio, paralinguistic transfer, cross-modal retention, and MUSAN robustness evaluations. All codes and models will be available after accepted.

REFERENCES

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

Hongjie Chen, Zehan Li, Yaodong Song, Wenming Deng, Yitong Yao, Yuxin Zhang, Hang Lv, Xuechao Zhu, Jian Kang, Jie Lian, Jie Li, Chao Wang, Shuangyong Song, Yongxiang Li, Zhongjiang He, and Xuelong Li. Goat-slm: A spoken language model with paralinguistic and speaker characteristic awareness, 2025. URL https://arxiv.org/abs/2507.18119.

Yanxi Chen, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Xin Li, Peijie Qiu, Hao Wang, Xuanzhao Dong, Yujian Xiong, Anderson Schneider, Yuriy Nevmyvaka, and Yalin Wang. AHA: Aligning large audio-language models for reasoning hallucinations via counterfactual hard negatives. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29294–29306, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1464. URL https://aclanthology.org/2026.findings-acl.1464/.

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. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Hiroya Fujisaki. Information, prosody, and modeling - with emphasis on tonal features of speech -. In *Speech Prosody 2004*, pp. 1–10, 2004. doi: 10.21437/SpeechProsody.2004-1.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume

38, Main Conference, pp. 41819–41886. Curran Associates, Inc., 2025. doi: 10.52202/085713-1396. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3babb6b453cb59d87cb58a1219ef914b-Paper-Conference.pdf.

Yuan Gong, Jin Yu, and James Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 151–155, 2022. doi: 10.1109/ICASSP43922.2022.9746828.

Sanaul Haq and Philip JB Jackson. Multimodal emotion recognition. In *Machine audition: principles, algorithms and systems*, pp. 398–423. IGI Global Scientific Publishing, 2011.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, Chunyat Wu, Qianhua He, Tan Lee, Xie Chen, Wei-Long Zheng, Weiqiang Wang, Mark D. Plumbley, Jian Liu, and Qiuqiang Kong. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 110411–110432, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b2fbf1c9bc92e7ef2f6cab2e8a3e09af-Paper-Conference.pdf.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report, 2025. URL https://arxiv.org/abs/2504.18425.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*, 2025. URL https://github.com/xiaomi-research/r1-aqa;https://huggingface.co/mispeech/r1-aqa.

Guan-Ting Lin, Cheng-Han Chiang, and Hung-yi Lee. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6626–6642, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.358. URL https://aclanthology.org/2024.acl-long.358/.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Ankur Gandhe, Chao-Han Huck Yang, Yile Gu, Shalini Ghosh, Andreas Stolcke, Hung-Yi Lee, and Ivan Bulyko. Paralinguistics-enhanced large language modeling of spoken dialogue. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10316–10320, 2024b. doi: 10.1109/ICASSP48485.2024.10446933.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, Amélie Martin, Anmol Agarwal, Antoine Roux, Arthur Darcet, Arthur Mensch, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Chris Bamford, Christian Wallenwein, Christophe Renaudin, Clémence Lanfranchi, Darius Dabert, Devendra Singh Chaplot, Devon Mizelle, Diego de las Casas, Elliot Chane-Sane, Emilien Fugier, Emma Bou Hanna, Gabrielle Berrada, Gauthier Delerce, Gauthier Guinet, Georgii Novikov, Guillaume Martin, Himanshu Jaju, Jan Ludziejewski, Jason Rute, Jean-Hadrien Chabran, Jessica Chudnovsky, Joachim Studnia, Joep Barmentlo, Jonas Amar, Josselin Somerville Roberts, Julien Denize, Karan Saxena, Karmesh Yadav, Kartik Khandelwal, Kush Jain, Lélio Renard Lavaud, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier,

Luyu Gao, Marie Pellat, Mathilde Guillaumin, Mathis Felardos, Matthieu Dinot, Maxime Darrin, Maximilian Augustin, Mickaël Seznec, Neha Gupta, Nikhil Raghuraman, Olivier Duchenne, Patricia Wang, Patryk Saffer, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Rémi Delacourt, Romain Sauvestre, Roman Soletskyi, Sagar Vaze, Sandeep Subramanian, Saurabh Garg, Shashwat Dalal, Siddharth Gandhi, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Thibault Schueller, Thibaut Lavril, Thomas Robert, Thomas Wang, Timothée Lacroix, Tom Bewley, Valeriia Nemychnikova, Victor Paltz, Virgile Richard, Wen-Ding Li, William Marshall, Xuanyu Zhang, Yihan Wan, and Yunhao Tang. Voxtral, 2025. URL https://arxiv.org/abs/2507.13264.

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

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.931. URL https://aclanthology.org/2024.findings-acl.931/.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Omni-r1: Do you really need audio to fine-tune your audio llm?, 2025. URL https://arxiv.org/abs/2505.09439.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d36f208919582785db965fe648b9fe59-Paper-Conference.pdf.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39, 2013. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2012.02.005. URL https://www.sciencedirect.com/science/article/pii/S0885230812000162. Special issue on Paralinguistics in Naturalistic Speech and Language.

David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus, 2015. URL https://arxiv.org/abs/1510.08484.

Gemma Team. Gemma 3n. 2025. URL https://ai.google.dev/gemma/docs/gemma-3n.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier

Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Ioannis Tsiamas, Matthias Sperber, Andrew Finch, and Sarthak Garg. Speech is more than words: Do speech-to-text translation systems leverage prosody? In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1235–1257, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.119. URL https://aclanthology.org/2024.wmt-1.119/.

Dingdong WANG, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 31374–31410, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/34e278fbbd7d6d7d788c98065988e1a9-Paper-Conference.pdf.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Dongchao Yang, Songxiang Liu, Disong Wang, Yuanyuan Wang, Guanglu Wan, and Helen Meng. Omni-autothink: Adaptive multimodal reasoning via reinforcement learning, 2025. URL https://arxiv.org/abs/2512.03783.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-bench: Benchmarking large audio-language models via generative comprehension. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.109. URL https://aclanthology.org/2024.acl-long.109/.

Shuaijiang Zhao, Tingwei Guo, Cheng Wen, Bajian Xiang, and Wei Zou. Ke-omni-r: Achieving advanced audio reasoning with a concise 50-words think process. https://github.com/shuaijiang/Ke-Omni-R, 2025.

Maike Züfle, Danni Liu, Vilém Zouhar, and Jan Niehues. Why we need speech to evaluate speech translation, 2026. URL https://arxiv.org/abs/2605.28227.

Table 3: Prompt template for reference-free speech-to-text translation evaluation. Fields in braces are placeholders.

System Prompt
You are a strict reference-free evaluator of speech-to-text translation.
Judge whether the candidate translation preserves the meaning intended by the source speech.
Speech-dependent cues such as contrastive stress, pauses, intonation, emotion, politeness, speaker gender, names, and numbers matter only when they affect the translated meaning. Do not reward verbosity. Do not invent a reference translation. Return exactly one JSON object and no markdown.
Audio Evidence Instruction
Listen to the attached source audio before deciding. Treat audio as the primary evidence for speech-only meaning. Attend to which word is emphasized, pause structure, final intonation, emotion, urgency, and politeness. If one candidate preserves that intended meaning better, choose it even when both preserve the literal words.
User Prompt Template
Compare two {target} translations of the same source utterance.
{evidence_instruction}
{transcript_section}Candidate A:
{candidate_a}
Candidate B:
{candidate_b}
{output_instruction}
Choose TIE only when the available source evidence genuinely does not favor either candidate.

A Prompt Template

Table 3 shows template of our method.

B Supplementary Method Details

B.1 Judgment Relations

Audio-conditioned evaluation. The evaluator compares the content and intent conveyed by two candidate translations using the source audio and lexical transcript, without a reference translation at inference time. The label TIE indicates that the available evidence does not support a preference for either candidate. The task interface permits this response, while both training sources contain only clear A/B preferences. Candidate evaluation involves overall translation-quality judgment and audio-semantic discrimination. The former compares candidate quality using the source speech and its lexical content; the latter concerns cases in which both candidates are compatible with the transcript but interpret vocal cues such as stress or intonation differently.

Audio-semantic discrimination. We consider paired inputs with identical textual conditions but opposite correct labels:

$$x_1 = (a_1, t, c_A, c_B), \qquad x_2 = (a_2, t, c_A, c_B), \qquad y_1 \neq y_2. \tag{5}$$

The two recordings support different candidates, while the transcript, candidate content, and presentation order remain unchanged. The desired judgments satisfy $f_\theta(x_1) = y_1$ and $f_\theta(x_2) = y_2$. A deterministic evaluator that relies only on text cannot answer both conditions correctly. This condition concerns changes in audio that alter the preferred candidate; it does not require every audio perturbation to change the decision. A change in the output label alone is insufficient: both judgments must match their respective correct labels.

Candidate-order consistency. Let π swap candidate order and the associated labels, with $\pi(A) = B$, $\pi(B) = A$, and $\pi(\text{TIE}) = \text{TIE}$. The desired consistency is:

$$f_\theta(a, t, c_B, c_A) = \pi(f_\theta(a, t, c_A, c_B)) \,. \tag{6}$$

Swapping candidate positions should change the output label without changing the preferred translation. These relations guide the construction of supervision; their empirical satisfaction is evaluated separately.

B.2 Data Construction and Splits

Human quality-preference supervision. The quality data come from the publicly available IWSLT speech translation evaluation development set, containing source speech, transcripts, candidate translations from different systems, and human quality scores. We group records by source utterance identifier and language pair, and average repeated ratings of the same translation from the same system to obtain the quality score $\bar{s}_i$ for candidate c_i. We enumerate pairs of distinct translations for each source utterance and retain those satisfying:

$$|\bar{s}_i - \bar{s}_j| \geq \tau. \tag{7}$$

We use $\tau = 15$ on the original rating scale and label the higher-scoring translation as preferred. Training candidates are also filtered using a predefined length limit. The supervision is derived directly from human scores rather than automatic translation metrics.

Before constructing candidate pairs, we exclude records sharing a source identifier or audio hash with the held-out evaluation set. Eligible pairs are selected using a hash-based ordering with a fixed seed. Each retained pair is presented in both AB and BA order, with the preference labels swapped accordingly. This yields 2,678 quality-preference pairs and 5,356 training views.

Audio-semantic contrast supervision. We use the paired recordings and candidate translations already provided by ContraProST (Tsiamas et al., 2024), without synthesizing additional counterfactual audio. Audio and translation indices in Equation 1 specify correspondences within each example, rather than semantic categories shared across examples. Each training view contains only one recording.

We group examples by normalized source sentence, keeping cross-language variants and duplicate records of the same sentence together. Groups are assigned to training, development, and test splits with stratification by category, and all four views of an example remain in the same split. The full training split contains 1,356 language-specific examples and 5,424 views. Mixed training uses a predefined subset of 1,339 examples, retaining all four views of each to obtain 5,356 training records. The number of language-specific examples is distinguished from the number of distinct source-sentence groups.

Mixture and ordering. The resulting mixed dataset is:

$$\mathcal{D} = \mathcal{D}_Q \cup \mathcal{D}_C, \qquad |\mathcal{D}_Q| = |\mathcal{D}_C| = 5{,}356, \qquad |\mathcal{D}| = 10{,}712. \tag{8}$$

The $1:1$ ratio is defined by training-view count, not by the number of distinct recordings or candidate pairs. The two views of each quality-preference pair and the four views of each semantic contrast example are kept in bundles, which are shuffled and organized into batches at the bundle level. An equal source mixture is not required in every batch.

B.3 Input Interface and Loss Masking

Both data sources use the same system instruction and user-prompt structure, comprising one source recording, the target language, the lexical transcript, and candidate translations A/B. The prompt states that the lexical transcript may omit meaning conveyed only by speech and asks the model to use the audio when such information is relevant. Human scores, candidate system identities, reference translations, correct labels, and meaning annotations used to construct supervision are excluded from the input prompt.

Outputs use compact JSON, such as {"choice":"A"} or {"choice":"B"}. The prompt also permits {"choice":"TIE"}, although both training sources contain only clear A/B preferences.

During training, the target JSON is appended as the supervised assistant sequence; at inference time, the model generates the choice without receiving the target.

For A/B supervision, the choice value corresponds to a single token. System instructions, user text, audio placeholders, JSON syntax, end-of-sequence tokens, and padding positions are all masked out of the loss. For the templated sequence $z = (z_1, \ldots, z_T)$, position k is the choice-value position in Equation 3. Its prefix includes the audio conditioning, task prompt, and fixed JSON prefix, but not the correct choice value itself. The full target sequence is processed with a causal mask, so tokens following the choice position do not contribute to predicting that choice. Only the designated choice position contributes to the training loss. The probability is produced by the original language modeling head's full-vocabulary softmax, rather than a separate binary classifier.

Both supervision sources use the mean choice loss in Equation 4 and share the same adapters and output interface. Their complementarity is encoded in the examples' input–label relations; no separate task heads, contrastive loss, ranking loss, or pairwise consistency regularizer are added.

B.4 Low-rank Parameterization

We use LoRA (Hu et al., 2021) to adapt selected linear transformations while freezing their pretrained weights. For $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, the effective weight is:

$$W = W_0 + \Delta W, \qquad \Delta W = \frac{\alpha}{r} UV. \tag{9}$$

Here, $U \in \mathbb{R}^{d_{\text{out}} \times r}$ and $V \in \mathbb{R}^{r \times d_{\text{in}}}$ are trainable low-rank matrices, r is the rank, and α is the scaling factor. Only the adapter parameters ϕ are optimized; all pretrained parameters θ_0 remain frozen.

In Qwen2.5-Omni, adapters are inserted into selected q/k/v/o attention projections and gate/up/down feed-forward projections in the Thinker language backbone. The audio encoder and its projection, vision encoder, and remaining pretrained weights stay frozen. The model retains its original language modeling head for choice generation and does not introduce a quality-regression or discriminative classification head. The speech generation branch is not used during training or evaluation. Model-specific adaptation configurations, optimization schedules, resource costs, and comparisons with full fine-tuning are reported in the experimental setup and analysis.

C Details of the Preliminary Analysis

C.1 Dataset Size and Splits

Our local ContraProST collection contains 1,921 language-specific examples spanning three translation directions: English-to-German (607), English-to-Spanish (641), and English-to-Japanese (673). The examples cover five phenomena: prosodic breaks, emotional prosody, intonation, sentence stress, and pragmatic prosody. Each example corresponds to a source-utterance–target-language combination and contains a shared transcript, two audio variants, and two candidate translations. Source recordings can be reused across target languages, so the number of language-specific examples is distinct from the number of unique audio pairs.

We group examples by normalized source sentence, keeping cross-language variants and records with the same normalized source sentence together. This produces 1,033 source-sentence groups. Using a fixed seed of 20260803, we partition these groups into training, development, and test splits, with no source-sentence-group overlap between splits. Table 4 summarizes the complete partition.

The test split contains 79 English-to-German, 96 English-to-Spanish, and 108 English-to-Japanese examples. Its phenomenon counts are 87 for prosodic breaks, 75 for emotional prosody, 48 for intonation, 37 for sentence stress, and 36 for pragmatic prosody.

For each test example, crossing two audio conditions with two candidate orders gives four individual judgment views. At each candidate order, the two audio conditions form one directed-flip pair. The complete test protocol therefore contains 1,132 judgment views and 566 directed-flip pairs per model. These are evaluation pairs rather than counts of unique source audio pairs. All counts in Table 4 refer to one four-view audio-text setting, not the sum of different input modes.

Table 4: ContraProST split statistics after grouping by normalized source sentence. Each language-specific example yields four audio-text evaluation views and two directed-flip pairs. The preliminary diagnosis uses the test split.

Split	Source-sentence groups	Language-specific examples	Evaluation views	Directed-flip pairs
Train	723	1,356	5,424	2,712
Development	156	282	1,128	564
Test	**154**	**283**	**1,132**	**566**
Total	1,033	1,921	7,684	3,842

The training row describes the complete training split before selecting the subset used for PCT. Likewise, the full-collection zero-shot evaluation is separate from the held-out preliminary diagnosis and is not used to determine its sample counts.

C.2 Counterfactual Evaluation Inputs

Following the task definition in the main text, we denote counterfactual group i as

$$g_i = (a_{i1}, a_{i2}, t_i, c_{i1}, c_{i2}), \tag{10}$$

where the two recordings share the lexical transcript t_i and support candidate translations c_{i1} and c_{i2}, respectively. Audio and translation indices indicate correspondences within a group, rather than semantic categories shared across examples.

Let $\mathcal{O} = \{\mathrm{AB}, \mathrm{BA}\}$ denote the two candidate orders. Under order AB, candidates A and B are c_{i1} and c_{i2}, respectively; their positions are reversed under order BA. For audio condition $r \in \{1, 2\}$, the corresponding inputs are

$$\begin{aligned} x_{i,r,\mathrm{AB}} &= (a_{ir}, t_i, c_{i1}, c_{i2}), \\ x_{i,r,\mathrm{BA}} &= (a_{ir}, t_i, c_{i2}, c_{i1}). \end{aligned} \tag{11}$$

Ordering the views by the two audio conditions under AB and then the two under BA gives the gold-label sequence

$$\left(y_{i,1,\mathrm{AB}}, y_{i,2,\mathrm{AB}}, y_{i,1,\mathrm{BA}}, y_{i,2,\mathrm{BA}}\right) = (A, B, B, A). \tag{12}$$

Here, A/B identify candidate positions rather than fixed translation identities. Each model call receives only one recording; relations among the four views are used to construct evaluation contrasts and compute metrics, rather than to present both recordings jointly.

C.3 IWSLT Human Agreement

IWSLT Human Agreement measures whether model choices agree with translation preferences derived from human quality scores. Each evaluation view contains source audio, its transcript, and two candidate translations, with the higher-scoring candidate serving as the preferred choice. Let M denote the number of evaluation views, y_j^{H} the human-preferred label for view j, and $\hat{y}_j$ the extracted model choice:

$$\mathrm{HA} = \frac{100}{M} \sum_{j=1}^{M} \mathbf{1}\left[\hat{y}_j = y_j^{\mathrm{H}}\right]. \tag{13}$$

HA evaluates the correspondence between model choices and human quality preferences on individual views, whereas DF evaluates joint correctness over the two audio conditions at a fixed candidate order. The metrics differ in both datasets and units of correctness, so their numerical values should not be treated as directly equivalent measurements of a single capability.

C.4 Real–Silence Choice Agreement

For each original input

$$x_j = (a_j, t_j, c_{jA}, c_{jB}),$$

we replace only the source audio to construct

$$x_j^{\text{sil}} = (a_{\text{sil}}, t_j, c_{jA}, c_{jB}). \tag{14}$$

The instruction, target-language specification, lexical transcript, candidate translations, and candidate order remain unchanged. The silence control uses a predefined 0.5-second all-zero waveform, replacing different recordings with the same silent clip rather than a duration-matched silent version of each recording.

Let $\hat{y}_j^{\text{real}}$ and $\hat{y}_j^{\text{sil}}$ denote the extracted choices under real and silent audio, respectively, and let N_S be the number of matched comparisons:

$$\text{Agreement}_{\text{real,sil}} = \frac{100}{N_S} \sum_{j=1}^{N_S} \mathbf{1}\left[\hat{y}_j^{\text{real}} = \hat{y}_j^{\text{sil}}\right]. \tag{15}$$

Because candidate order is unchanged, equal A/B labels within a comparison refer to the same translation. This metric measures the stability of discrete choices, not their correctness, and does not imply that the full predictive distribution or confidence remains unchanged. High agreement therefore does not by itself establish that a model entirely ignores audio; we interpret it together with DF to examine correct responses to changes in vocal meaning and choice stability after real speech is removed.

C.5 Reading the Diagnostic Plots

In Figure 3(a), the horizontal axis is IWSLT Human Agreement and the vertical axis is ContraProST Directed Flip. The dashed $y = x$ line marks numerical equality between the two percentage scores, rather than a theoretical relationship that an audio-grounded model should satisfy. Our analysis concerns observed performance across tasks and comparisons between models, rather than interpreting distance from the diagonal as a measure of audio information use.

Figure 3(b) additionally reports real–silence choice agreement. Together, the two panels characterize the evaluated models' behavior: general quality-preference judgment and correct audio-dependent decision-making can exhibit different performance, and many discrete choices remain unchanged after real speech is removed. These observations reveal limitations in reliable use of vocal information in this setting, but do not identify whether their source lies in acoustic perception, meaning interpretation, or the integration of audio evidence into the final decision.

D Additional Experimental Settings

D.1 Benchmarks

Targeted judgment evaluation. Our primary evaluation uses ContraProST and IWSLT to examine counterfactual audio judgment and agreement with human translation-quality preferences, respectively. We reformulate ContraProST (Tsiamas et al., 2024) as a text-controlled translation-selection task. The held-out set contains 283 language-specific examples from 154 normalized-source-sentence groups across English-to-German, English-to-Spanish, and English-to-Japanese translation. The complete four-view protocol yields 566 directed-flip pairs per model. IWSLT evaluation compares model choices with preferences derived from human translation-quality scores. Real–silence comparisons provide an additional diagnostic of choice stability under audio replacement.

MMAU. MMAU (Sakshi et al., 2025) evaluates understanding and reasoning across speech, environmental sounds, and music, covering information extraction and reasoning questions that require domain knowledge. We use the public test-mini split containing 1,000 examples.

AIR-Bench. AIR-Bench (Yang et al., 2024) evaluates audio-language models through audio inputs and text responses across speech, natural sounds, and music, with Foundation and Chat components. We use the Foundation component, with 24,683 records in our evaluation manifest.

MMSU. MMSU (WANG et al., 2026) evaluates fine-grained spoken language understanding and reasoning across semantic, phonological, and paralinguistic phenomena. The released benchmark contains 5,000 examples across 47 tasks. We evaluate a fixed eligible subset of 4,994 examples after excluding six recordings shorter than 0.5 seconds. This subset contains 2,575 perception examples and 2,419 reasoning examples, all presented as four-option multiple-choice questions. The same eligible set is used across model configurations.

D.2 MODELS

The general audio transfer evaluation includes twelve models organized into three reporting groups: three omni models, four audio-text-to-text models, and five further post-trained models. The targeted judgment comparison uses Qwen2.5-Omni-3B and 7B, including their Base, Quality-only, and PCT configurations.

Omni models. This group comprises Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, and Gemma-4-E4B. Qwen2.5-Omni (Xu et al., 2025) uses a Thinker–Talker architecture to process text, images, audio, and video in a unified system and supports text and speech generation. Gemma 4 (Team et al., 2026) is a family of natively multimodal open-weight models supporting audio and visual inputs; we evaluate its E4B variant.

Audio-text-to-text models. Qwen2-Audio-7B (Chu et al., 2024) supports audio analysis and speech instruction following, producing corresponding text responses. Gemma-3n-E4B (Team, 2025) is designed for resource-constrained devices and supports audio, visual, and text inputs, including speech recognition, translation, and audio analysis. Kimi-Audio-7B (KimiTeam et al., 2025) is an audio foundation model covering speech, environmental sounds, and music, with capabilities in audio understanding, generation, and conversation. Voxtral-Mini-3B (Liu et al., 2025) is an audio chat model trained to understand spoken audio and text. We evaluate these models through their audio-conditioned text responses.

Post-trained audio-language models. This group comprises Ke-Omni-R-3B, Omni-R1-7B, R1-AQA, Open-R1-GRPO-3B, and Omni-AutoThink-7B. Ke-Omni-R-3B (Zhao et al., 2025) builds on Qwen2.5-Omni and uses GRPO with concise reasoning traces for audio reasoning post-training. Omni-R1-7B (Rouditchenko et al., 2025) also builds on Qwen2.5-Omni and applies GRPO to audio question answering data. R1-AQA (Li et al., 2025) applies GRPO-based audio question answering post-training to Qwen2-Audio-7B-Instruct. Open-R1-GRPO-3B is included as an additional post-trained baseline. Omni-AutoThink-7B (Yang et al., 2025) builds on Qwen2.5-Omni-7B and combines adaptive supervised fine-tuning with adaptive GRPO for multimodal reasoning. This group tests whether PCT provides additional benefits after prior post-training.

D.3 EVALUATION PROTOCOL

Comparison configurations. Base denotes the corresponding model before additional PCT adaptation, and +PCT denotes the model after applying our training method. For the post-trained group, Base retains the model's original post-training but does not include additional PCT adaptation. Quality-only and Counter-only use human translation-quality preferences and counterfactual audio supervision, respectively, whereas PCT combines the two sources. The targeted judgment table compares Base, Quality-only, and PCT at both Qwen2.5-Omni scales; the general-audio supervision ablation compares Base, Quality-only, Counter-only, and PCT on Omni-3B.

Inputs and scoring. For general audio benchmarks, models receive the audio and task instruction, together with candidate options when applicable, and produce text responses. MMAU accuracy is computed using the string-matching rule in its official evaluation script. AIR-Bench Foundation is micro-averaged over the 24,683 records in our evaluation manifest. For MMSU, we use the MMAU string-matching implementation under our fixed local evaluation protocol, retaining the original option order and using greedy decoding with a maximum of 128 new tokens. MMSU accuracy is aggregated over all 4,994 eligible examples. Runtime failures and empty predictions count as incorrect within each benchmark's fixed eligible set. Unavailable benchmark results are reported as – rather than estimated from incomplete evaluations.

Table 5: Additional transfer results on audio-text-to-text models and R1-AQA. Scores are reported on a 0–100 scale. Avg. is the arithmetic mean of the three benchmark scores. Bold indicates the higher score within each Base/+PCT pair.

Model	MMAU ↑		AIR-Bench ↑		MMSU ↑		Avg. ↑	
	Base	+PCT	Base	+PCT	Base	+PCT	Base	+PCT
Audio-Text-to-Text Models								
Qwen2-Audio-7B	56.10	**58.10**	56.70	**58.60**	**52.94**	52.84	55.25	**56.51**
Gemma-3n-E4B	**64.30**	62.70	44.41	**44.80**	**56.21**	53.86	**54.97**	53.79
Kimi-Audio-7B	20.00	**30.20**	49.42	**54.09**	3.30	**5.13**	24.24	**29.81**
Voxtral-Mini-3B	58.10	**59.70**	48.89	**52.19**	60.33	**61.83**	55.77	**57.91**
Post-trained Audio-Text-to-Text Model								
R1-AQA (Qwen2-Audio-7B)	30.60	**43.30**	60.31	**63.73**	31.50	**39.11**	40.80	**48.71**

Aggregation. All scores are reported on a 0–100 scale. Avg. is the arithmetic mean over benchmark scores available for both members of a Base/+PCT comparison, using the same benchmark set for the two entries. It is a benchmark-level average rather than a pooled accuracy over all examples from different benchmarks.

E Additional Transfer Results on Audio-Text-to-Text Models

To examine whether the transfer benefits of PCT extend beyond the omni models discussed in the main text, we additionally evaluate Qwen2-Audio-7B, Gemma-3n-E4B, Kimi-Audio-7B, and Voxtral-Mini-3B, together with R1-AQA, a post-trained model based on Qwen2-Audio-7B. Table 5 reports Base/PCT comparisons on the same three general audio benchmarks, supplementing the results in the main text.

Voxtral-Mini-3B improves on all three benchmarks, increasing its average score from 55.77 to 57.91, including a 3.30-percentage-point gain on AIR-Bench. Qwen2-Audio-7B gains 2.00 and 1.90 percentage points on MMAU and AIR-Bench, respectively, while declining by 0.10 points on MMSU; its average score nevertheless increases from 55.25 to 56.51. Kimi-Audio-7B increases its MMAU score from 20.00 to 30.20 and its AIR-Bench score from 49.42 to 54.09. In contrast, Gemma-3n-E4B improves slightly only on AIR-Bench, while declining by 1.60 and 2.35 percentage points on MMAU and MMSU, respectively, reducing its average score from 54.97 to 53.79.

For R1-AQA, which has already undergone audio question answering post-training, PCT increases scores on all three benchmarks. Its MMAU, AIR-Bench, and MMSU scores increase by 12.70, 3.42, and 7.61 percentage points, respectively, raising the average from 40.80 to 48.71. These additional results show that the cross-task benefits of counterfactual translation-judgment supervision also occur in some audio-text-to-text models and a post-trained variant, rather than being confined to the omni models in the main text. However, the declines in Gemma-3n-E4B and the task-dependent changes in other models indicate that this transfer is not uniform across backbones and tasks.

F Details of Transfer to Paralinguistic Tasks

F.1 Datasets and Evaluation Sizes

SAVEE. SAVEE (Haq & Jackson, 2011) contains 480 utterances from four male native English speakers, covering seven emotion categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. We use its audio recordings for speech emotion recognition, with the evaluation count denoted by N_{SAVEE}.

Table 6: Cross-benchmark transfer on SAVEE, VocalSound, CREMA-D, and RAVDESS. Scores are reported on a 0–100 scale. Bold indicates the higher score within each Base/+PCT pair. For Gemma-4, +PCT refers to the +PCT configuration.

Model	SAVEE ↑		VocalSound ↑		CREMA-D ↑		RAVDESS ↑	
	Base	+PCT	Base	+PCT	Base	+PCT	Base	+PCT
Omni-3B	28.33	**59.38**	90.98	**91.34**	28.61	**33.08**	29.86	**37.29**
Omni-7B	43.33	**62.92**	91.48	**91.53**	64.20	**64.81**	56.53	**68.12**
Gemma-4	22.92	**23.33**	30.74	**33.11**	16.02	**19.31**	15.97	**20.69**
R1-AQA	**47.92**	45.42	88.67	**90.34**	**77.37**	69.00	**59.44**	34.31
Omni-R1-7B	46.46	**50.21**	**91.90**	88.97	**61.82**	58.99	**58.54**	53.61

Table 7: Dataset tasks and evaluation sizes for the transfer experiments. N counts individual audio recordings rather than audio pairs.

Dataset	Task	Classes	Evaluation N
SAVEE	Speech emotion	7	480
VocalSound	Vocal sound	6	21,024
CREMA-D	Speech emotion	6	7,442
RAVDESS	Speech emotion	8	1,440

VocalSound. VocalSound (Gong et al., 2022) contains 21,024 crowdsourced recordings from 3,365 participants across six non-speech vocal sound categories: laughter, sighs, coughs, throat clearing, sneezes, and sniffs. This task evaluates vocal sound type recognition rather than emotion classification; we denote the number of evaluated recordings by N_{VS}.

CREMA-D. CREMA-D (Cao et al., 2014) contains 7,442 recordings from 91 actors performing 12 sentences with six emotions: anger, disgust, fear, happiness, neutral, and sadness. The dataset includes emotional expressions at different intensity levels, and our per-recording emotion recognition evaluation uses 7,442 audio clips.

RAVDESS. RAVDESS (Livingstone & Russo, 2018) contains emotional speech and song in audio-only, video-only, and audiovisual formats. We use its audio-only speech subset of 1,440 recordings from 24 actors speaking two fixed statements, excluding song and visual inputs. This subset covers eight emotion categories: neutral, calm, happiness, sadness, anger, fear, disgust, and surprise.

F.2 Evaluation Protocol

We classify each recording independently, using the emotion or vocal sound label in the evaluation manifest as the reference. We report percentage classification accuracy, where N_D is the number of evaluated recordings in dataset D, y_i is the reference label, and $\hat{y}_i$ is the category extracted from the model response.

$$\mathrm{Acc}_D = \frac{100}{N_D} \sum_{i=1}^{N_D} \mathbf{1}[\hat{y}_i = y_i] \,. \tag{16}$$

This metric is computed per recording and differs from the paired-correctness metrics used in the subsequent robustness analysis. This evaluation measures cross-task transfer from translation-judgment supervision without additional fine-tuning on these task labels.

F.3 Detailed Results

Detailed results of each modal are shown in 6.

Table 8: Cross-modal performance before and after PCT adaptation. All tasks produce text outputs. Scores are percentages; values in parentheses indicate changes from Base in percentage points.

Modality	Benchmark	N	Qwen2.5-Omni-3B		Qwen2.5-Omni-7B	
			Base	+PCT (Δ)	Base	+PCT (Δ)
Text $\rightarrow$ Text	MMLU-Redux	2,722	55.84	55.73 (-0.11)	66.50	66.27 (-0.23)
Video $\rightarrow$ Text	MLVU (MCQ)	2,171	63.80	63.24 (-0.56)	66.01	66.19 ($+0.18$)
Vision $\rightarrow$ Text	MMStar	1,500	54.40	54.27 (-0.13)	59.20	60.67 ($+1.47$)
Audio $\rightarrow$ Text	MMAU-v05.15.25	1,000	74.00	**75.20** ($+1.20$)	70.90	**71.20** ($+0.30$)
Audiovisual $\rightarrow$ Text	DailyOmni	1,196	58.19	57.94 (-0.25)	61.54	61.62 ($+0.08$)

G Supplementary Results on Cross-Modal Capability Retention

G.1 Evaluation Scope and Metrics

We compare cross-modal performance before and after PCT adaptation on Qwen2.5-Omni-3B and 7B, using MMLU-Redux for text inputs, MLVU (MCQ) for video inputs, MMStar for image inputs, MMAU-v05.15.25 for audio inputs, and DailyOmni for audiovisual inputs. All tasks produce text outputs. Table 8 reports the evaluation sample count, Base and PCT scores, and corresponding changes for each task.

Let $S_D^{(m)}$ denote the percentage score of model m on benchmark D. The change after PCT adaptation is defined as

$$\Delta_D^{(m)} = S_D^{(m,\mathrm{PCT})} - S_D^{(m,\mathrm{Base})}. \tag{17}$$

Δ is measured in percentage points, with positive values indicating improvements and negative values indicating declines. The main-text dot plot visualizes these differences, while the appendix table retains the corresponding absolute scores.

G.2 Complete Evaluation Results

These results correspond directly to the main-text dot plot; scores from different benchmarks are neither pooled nor reweighted. The complete scores show small declines for Omni-3B on all four text, video, image, and audiovisual tasks, with the largest decrease of 0.56 percentage points on MLVU. Omni-7B shows a slight decline on the text task, while improving by 0.18, 1.47, and 0.08 percentage points on the video, image, and audiovisual tasks, respectively. Both model scales improve on the audio benchmark. These results therefore support largely preserved performance in other modalities on the evaluated tasks, rather than unchanged scores on every task.

G.3 Plot Interpretation and Scope

In Figure 7, circles and diamonds represent Qwen2.5-Omni-3B and 7B, respectively; the horizontal axis shows score changes after PCT relative to Base, and the vertical dashed line marks zero change. The shaded ± 0.5-percentage-point band serves only as a visual reference and does not represent a confidence interval, statistical significance, or a predefined non-inferiority criterion.

This analysis describes performance changes using point estimates, with conclusions limited to the two model scales and the listed benchmarks. Because all evaluations use text outputs, these results concern performance retention across input modalities and do not establish whether speech generation or capabilities in other output modalities are preserved.

H Details of Acoustic Robustness Evaluation

H.1 Models and Evaluation Pools

The noise experiment uses five fixed configurations: Base, Quality-only, and PCT at Omni-3B, and Base and PCT at Omni-7B.

The common evaluation pool contains 7,349 CREMA-D recordings from 91 actors and 1,440 RAVDESS recordings from 24 actors. Strict pairs connect recordings with different emotions within the same actor, sentence, and intensity group; RAVDESS additionally requires the same repetition index. Before perturbation, 93 recordings are excluded: 92 have no contrasting emotion within their strict group, and one has zero RMS. The final CREMA-D pair pool decreases from 17,685 to 17,680 pairs, while RAVDESS contains 4,704 pairs; clean and noisy metrics use the same common subset.

References are the intended-emotion labels derived from filenames in the evaluation manifests, not the VoiceVote perceptual labels used in the representation analysis.

H.2 MUSAN Noise Generation

Perturbations use only the noise partition of MUSAN (Snyder et al., 2015), excluding its music and speech partitions. Clean and noise audio are loaded as 16-kHz mono signals; the fixed noise inventory contains 879 files lasting at least one second, comprising 796 free-sound and 83 sound-bible files. Recordings are processed in a fixed dataset–ID order using NumPy PCG64 with seed 20260921, first sampling uniformly from sufficiently long noise files and then uniformly sampling a crop offset. Noise is assigned per recording; the same file and crop are reused at 30 and 20 dB with only the gain changed, and each audio–SNR mixture is shared across all configurations.

Let a be the clean waveform, n an equal-length noise crop, and $r(\cdot)$ the full-clip RMS; the mixture at SNR s is constructed as follows:

$$r(x) = \sqrt{\frac{1}{T}\sum_{t=1}^{T} x_t^2}, \qquad g_s = \frac{r(a)}{r(n)10^{s/20}}, \qquad \widetilde{a}^{(s)} = a + g_s n, \tag{18}$$

where $s \in \{30, 20\}$ dB and RMS is computed over the entire clip.

Mixtures are saved as 16-kHz mono float32 WAV files without additional scaling or clipping; peak amplitudes are logged by the mixer. Fallback rules repeat the longest noise file when necessary or advance the offset when crop RMS is below 10^{-8}; neither fallback is triggered in this batch. The 8,789 recordings yield 17,578 mixtures at the two SNRs and 87,890 noisy-input generations across the five configurations.

H.3 Prompts, Decoding, and Scoring

All configurations use the same system prompt and dataset-specific fixed user prompts, with emotion labels presented in manifest order.

System prompt.

```
You are an audio understanding assistant.
```

CREMA-D user prompt.

```
Identify the emotion expressed by the speaker.  Choose
exactly one label from:  angry, disgust, fearful, happy,
neutral, sad.  Output only that label.
```

RAVDESS user prompt.

```
Identify the emotion expressed by the speaker.  Choose
exactly one label from:  neutral, calm, happy, sad, angry,
fearful, disgust, surprised.  Output only that label.
```

We use greedy decoding with a maximum of 32 new tokens (`do_sample=false`), with `return_audio=false` and `use_audio_in_video=false`.

Correctness is determined by `raw.strip() == reference`: only leading and trailing whitespace is removed, matching is case-sensitive, and no alias mapping or substring fallback is used. The output must be the label string itself rather than an option letter; this single-label task has no TIE path. Completed attempts with runtime failures, empty outputs, or invalid responses count as incorrect and remain in the fixed denominator; raw outputs are preserved. Unrun records or missing artifacts are separately marked as missing, rather than imputed as zero or silently removed.

H.4 Metrics and Statistical Analysis

For the fixed pair set $\mathcal{P}_D$ of dataset D, let $G_D = |\mathcal{P}_D|$ and let $C_{ir}^{(m,s)}$ indicate whether model m correctly classifies endpoint r of pair i under condition s, where $s = 0$ denotes clean audio. J_{pair} measures correctness of both endpoints, whereas $J_{\text{robust}}(s)$ requires both clean endpoints and both noisy counterparts at SNR s to be correct:

$$J_{\text{pair}}^{(m)} = \frac{100}{G_D} \sum_{i \in \mathcal{P}_D} C_{i1}^{(m,0)} C_{i2}^{(m,0)}, \tag{19}$$

$$J_{\text{robust}}^{(m)}(s) = \frac{100}{G_D} \sum_{i \in \mathcal{P}_D} C_{i1}^{(m,0)} C_{i2}^{(m,0)} C_{i1}^{(m,s)} C_{i2}^{(m,s)}. \tag{20}$$

Each noise condition is evaluated jointly with clean inputs; success at 20 dB does not additionally require success at 30 dB. All metrics use the full common pair pool as the denominator, rather than measuring retention only on a model-specific clean-correct subset.

As a consistency check, we compute pair-weighted endpoint accuracy A_{pair} and the proportion D_{one} with exactly one correct endpoint, and verify $J_{\text{pair}} = A_{\text{pair}} - D_{\text{one}}/2$:

$$A_{\text{pair}}^{(m)} = \frac{100}{2G_D} \sum_{i \in \mathcal{P}_D} \left(C_{i1}^{(m,0)} + C_{i2}^{(m,0)} \right), \tag{21}$$

$$D_{\text{one}}^{(m)} = \frac{100}{G_D} \sum_{i \in \mathcal{P}_D} \mathbf{1}\left[C_{i1}^{(m,0)} + C_{i2}^{(m,0)} = 1 \right]. \tag{22}$$

The numerical residual of this identity is approximately 10^{-17} across configurations; supplementary summaries also average per-actor J scores with equal actor weights.

Confidence intervals use 2,000 actor-cluster bootstrap, sharing each resample across configurations and audio conditions. We report the 2.5th and 97.5th percentiles of the bootstrap distribution, computing model contrasts directly within each shared resample. These intervals quantify uncertainty over evaluation actors for fixed models, not stability across model-training seeds.

H.5 Complete Results and Limitations

Tables 9 and 10 report complete scores and paired contrasts; contrasts are taken from the reported full-precision results before rounding to two decimals. Omni-3B exceeds Base and Quality-only under all conditions on both datasets, with all corresponding 95% contrast intervals above zero.

Noisy-input scoring retains the original intended-emotion labels without a human reassessment of perceived emotion or meaning after perturbation, making this a label-based noise evaluation.

Table 9: Complete paired-judgment results under the fixed robustness protocol (%). Clean reports J_{pair}; 30 and 20 dB report J_{robust} for that SNR jointly with clean inputs. The fixed pair counts are 17,680 for CREMA-D and 4,704 for RAVDESS.

Dataset	Configuration	Clean J_{pair}	30 dB J_{robust}	20 dB J_{robust}
CREMA-D	3B Base	6.08	3.80	2.43
	3B Quality-only	4.69	2.98	1.99
	3B PCT	7.92	6.43	5.23
RAVDESS	3B Base	7.84	3.15	1.89
	3B Quality-only	6.59	2.49	1.45
	3B PCT	11.61	7.63	6.70

Table 10: Paired PCT contrasts in percentage points. Intervals use 2,000 actor-cluster bootstrap draws, shared across configurations, and the 2.5th and 97.5th percentiles. Q-only denotes Quality-only. Differences are taken from the reported full-precision contrasts before rounding.

Dataset	Comparison	Condition	Δ	95% CI
CREMA-D	3B: PCT−Base	Clean	$+1.84$	$[1.30, 2.39]$
		30 dB	$+2.63$	$[2.15, 3.08]$
		20 dB	$+2.80$	$[2.43, 3.17]$
	3B: PCT−Q-only	Clean	$+3.22$	$[2.75, 3.68]$
		30 dB	$+3.46$	$[3.01, 3.88]$
		20 dB	$+3.24$	$[2.85, 3.63]$
RAVDESS	3B: PCT−Base	Clean	$+3.76$	$[2.30, 5.36]$
		30 dB	$+4.49$	$[3.44, 5.65]$
		20 dB	$+4.80$	$[3.89, 5.91]$
	3B: PCT−Q-only	Clean	$+5.02$	$[3.49, 6.59]$
		30 dB	$+5.14$	$[3.95, 6.42]$
		20 dB	$+5.25$	$[4.29, 6.40]$