RETHINKING, REVEALING, AND REDUCING REAL MODALITY GAP OF AUDIO LANGUAGE MODELS UNDER LINGUISTIC-PARALINGUISTIC TRADE-OFF

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Audio Language Models (LALMs) perform worse on audio than on semantically equivalent text, which is known as the modality gap. Recent studies claim that the gap can be reduced without degrading the pre-trained paralinguistic capability. However, we argue that the current protocol, which measures the two sides on disjoint data respectively built for linguistic- and paralinguistic-centered tasks, ignores their real-world tension and thus cannot actually support the claim. In response, we propose ReMG, a new protocol that crosses two linguistic contents with two paralinguistic states into a matched 4-tuple, holding the remaining acoustic covariates fixed. Each audio clip then has a variant that changes along only one axis and is also queried on both axes. This leaves no shortcut for LALMs but to directly trade off linguistic against paralinguistic capability on the same audio, only under which we believe a real modality gap can be meaningfully claimed. The result under ReMG challenges the previous view: paralinguistic collapse is widespread, and the only released checkpoint shows no substantial gap reduction. ReMG helps address this situation by guiding LALM fine-tuning. We use simple on-policy distillation without stronger teachers or ground-truth labels, so as not to mix the protocol contribution with that from new knowledge learning, as existing methods do. Across nine LALMs, four baseline methods, and seven benchmarks, the proposed ReMG is demonstrated to effectively reveal and reduce the real modality gap under the linguistic-paralinguistic trade-off.

1 INTRODUCTION

In recent years, end-to-end Large Audio Language Models (LALMs) have emerged as a prominent paradigm in multimodal intelligence, enabling natural, expressive, and human-like interaction (OpenAI, 2024; Tang et al., 2024; Chu et al., 2024; Abouelenin et al., 2025; Xu et al., 2025a;b). Typically built upon pretrained Large Language Model (LLM) backbones equipped with an acoustic encoder and cross-modal adapter, LALMs are expected to inherit the knowledge and reasoning capability from LLMs while simultaneously perceiving rich acoustic and paralinguistic cues (Peng et al., 2025; Abouelenin et al., 2025; Xu et al., 2025a). However, recent investigations have uncovered a critical challenge known as the **modality gap**, demonstrating that when given semantically identical queries, state-of-the-art (SOTA) LALMs consistently exhibit degraded performance on audio inputs compared to their corresponding text content (Xiang et al., 2025; Hsu et al., 2026).

Although recent works have reported considerable reduction in the LALM modality gap without significantly degrading the paralinguistic capability, they typically achieve this result by employing two separate categories of datasets (Wang et al., 2026a; Hu et al., 2026; Wang et al., 2026b; Cao et al., 2026). One relies only on the linguistic content within the input audio (e.g., math questions), and the other only needs the paralinguistic information (e.g., emotion recognition). In these works, the modality gap is **only** defined on the linguistic data and then claimed to be reduced by their proposed methods, while the paralinguistic data is **only** used for evaluation to show that such methods cause no catastrophic forgetting of the pre-trained acoustic capability (Wang et al., 2026a;b).

However, this current protocol leads to three issues: **First**, as a high-dimensional signal, audio contains uninterpretable, unstructured residuals derived from the audio sources, synthesis artifacts,

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

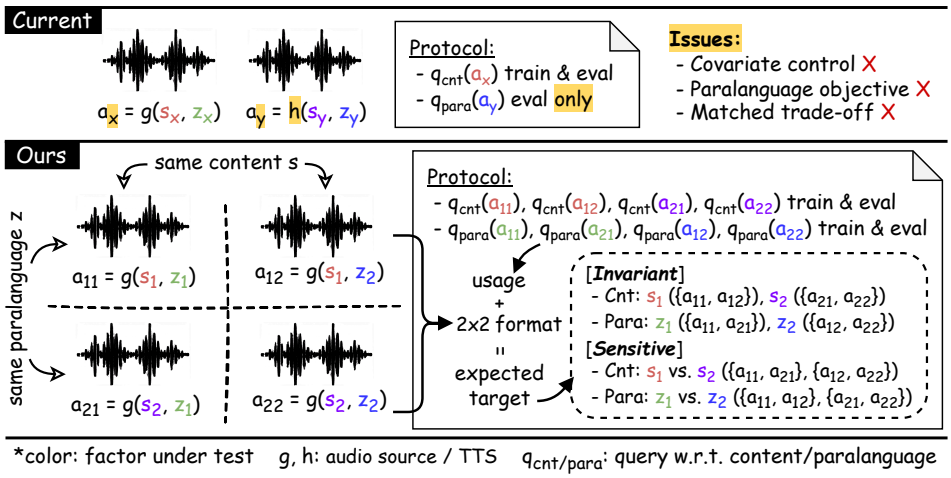

Figure 1: Protocol-level comparison of LALM modality gap between current knowledge and ours. The colored $\boldsymbol{s}$ and $\boldsymbol{z}$ respectively denote different linguistic and paralinguistic features, while the colors for audio $\boldsymbol{a}$ in query $\boldsymbol{q}$ indicate the intended factor (i.e., specific $\boldsymbol{s}$ or $\boldsymbol{z}$ with the corresponding color) under test by that query. The components with yellow background highlight the three issues with the current protocol, as detailed in the third paragraph of §1. The expected model behavior of our protocol is illustrated in the dashed box. For instance, $\boldsymbol{a}_{11}$ and $\boldsymbol{a}_{12}$ contain the same linguistic content s_1 but different paralinguistic features z_1 *vs.* z_2, thus the LALM responses to them are expected to be invariant given linguistic query $\boldsymbol{q}_{cnt}$ while sensitive given paralinguistic query $\boldsymbol{q}_{para}$.

channel effects, and so on (Cui et al., 2025; Chen et al., 2026; Gong et al., 2026). Thus, without controlling these covariates, it is not rigorous for the current protocol to uniquely attribute the observed results to either linguistic or paralinguistic features. **Second**, the current philosophy is to focus on developing algorithms to reduce the gap on the linguistic side, meanwhile just "expect" them to be mild enough not to significantly degrade the paralinguistic score (Wang et al., 2026a; Hu et al., 2026; Wang et al., 2026b; Cao et al., 2026). In other words, paralinguistic capability is just a passive objective. However, we believe explicitly maintaining the pre-trained acoustic capability is vital for LALMs, as it is the core competitiveness that distinguishes an LALM from its LLM backbone. **Finally**, even if both the gap reduction and paralinguistic maintenance are observed under the current protocol, it does not mean an improved Pareto front. As their data for the two sides are completely separate, the joint explanation of the two results can only provide some high-level intuitions, rather than indicate any linguistic-paralinguistic trade-off in real-world practices.

In response to these issues, we **Re**think how a **Re**al **M**odality **G**ap of LALMs should be claimed under the linguistic-paralinguistic trade-off, and propose **ReMG**, a new protocol to **Re**veal and **Re**duce that gap. As shown in Figure 1, we construct 4-tuple data in a 2×2 format, crossing two linguistic contents and two paralinguistic states, while keeping all other audio covariates constant. That means every audio has its variant with only linguistic or paralinguistic changes. During both training and evaluation, each audio receives two queries, one on its linguistic content and another on its paralinguistic state. As a result, we inject two kinds of structured priors: If two audio clips share a factor (whether linguistic or paralinguistic) being queried, the response correctness should be the same (***Invariant***), while if they differ in that factor, the result is expected to follow the change (***Sensitive***). Two clips with the same content but different emotions, for example, should get similar responses under content queries but not under emotion ones. The philosophy behind our protocol is that it *explicitly encourages the target LALMs to perceive and distinguish linguistic and paralinguistic features within a* ***single*** *audio, and to autonomously learn how to* ***selectively*** *utilize (i.e., directly* ***trade off****) them based on different cases*, rather than, as in the current modality gap works, treating the paralinguistic features as inherent noise that should be discarded.

We instantiate our ReMG protocol upon two tiers of data sources: 1) reasoning data that is originally challenging linguistic content in text and then rendered to audio via controlled Text-to-Speech (TTS) synthesis, and 2) perception data that is originally real human audio carrying natural paralinguistic features and then supplemented with content queries. We realize the full 4-tuple protocol in each tier independently and construct both queries for each audio clip, rather than limit it to reasoning data

with linguistic queries while human audio with paralinguistic ones as the existing protocol. Upon this design, we are able to define a series of new metrics to quantify the real modality gap along with the linguistic-paralinguistic trade-off on the same data. We also demonstrate how the ReMG protocol helps reduce that gap under the trade-off by fine-tuning LALMs. Recent works typically follow On-Policy Distillation (OPD) (Agarwal et al., 2024; Lu & Lab, 2025) but append instance-specific external knowledge from stronger teachers, ground-truth answers/labels, or reinforcement rewards (Wang et al., 2026b; Hu et al., 2026; Cao et al., 2026). As a consequence, their reported gap reduction actually mixes the contributions from their own design with those from the new knowledge. In contrast, we simply adopt OPD without any external instance-specific information, so that our observed improvement can be uniquely attributed to the proposed ReMG itself. Extensive experiments across six commercial plus three open-source LALMs compared with four SOTA existing works on seven benchmarks demonstrate the effectiveness of ReMG in both revealing and reducing the real modality gap. Our contributions can be summarized as follows:

- We identify **three** protocol-level issues: uncontrolled covariates, neglected paralinguistic capability, and unmatched linguistic-paralinguistic tension, along with **one** methodology-level issue: mixed contribution from new knowledge, for the current modality gap studies.
- A novel ReMG protocol is proposed to address the issues, revealing the real modality gap under linguistic-paralinguistic trade-off. Our result **challenges** the current view: The paralinguistic collapse across existing works is exposed once both query types are placed on the same data, and the only released checkpoint reporting a substantial gap reduction also does not survive ReMG.
- We verify the effectiveness of ReMG in fine-tuning LALMs to reduce the real gap while keeping a considerable trade-off, without mixing any contributions from external sources. Our code, new data, and checkpoints are available at `https://anonymous.4open.science/r/ReMG`.

2 REMG: REVEALING REAL MODALITY GAP OF LALMS

In this section, we reveal the real modality gap of LALMs by turning the protocol in Figure 1 into a concrete evaluation framework and redefining how a modality-gap reduction should be measured and claimed under the linguistic-paralinguistic trade-off.

2.1 THE 4-TUPLE UNIT

We model an audio clip as the output of a rendering process $\boldsymbol{a} = g(\boldsymbol{s}, \boldsymbol{z}; c)$, where $\boldsymbol{s}$ is linguistic content, $\boldsymbol{z}$ is paralinguistic state, and c collects the remaining acoustic covariates, such as reference voice, recording or synthesis channel, and unstructured residuals of the audio source. The atomic unit of ReMG crosses two contents with two paralinguistic states under one covariate condition:

$$\boldsymbol{u}(c) = \big\{ \boldsymbol{a}_{ij} = g(\boldsymbol{s}_i, \boldsymbol{z}_j; c) \,\big|\, i, j \in \{1, 2\} \big\}. \tag{1}$$

Because c is held fixed inside a unit, any behavioral difference within $\boldsymbol{u}$ is attributable to $\boldsymbol{s}$ or $\boldsymbol{z}$ alone, which is exactly the attribution that uncontrolled audio in previous works cannot provide, as the first issue mentioned in §1.

For every audio clip, there are two queries: A content query $\boldsymbol{q}_{cnt}$ targets $\boldsymbol{s}_i$ and treats $\boldsymbol{z}_j$ as a nuisance factor, and a paralinguistic query $\boldsymbol{q}_{para}$ targets $\boldsymbol{z}_j$ and treats $\boldsymbol{s}_i$ as a nuisance factor. This gives eight scored responses per unit, inducing four pairwise interventions as demonstrated in Figure 1. Along a row, where $\boldsymbol{s}$ is fixed and $\boldsymbol{z}$ changes, content correctness should keep its original state, while paralinguistic predictions should track the changed gold label. On the contrary, along a column, where $\boldsymbol{z}$ is fixed and $\boldsymbol{s}$ changes, the mirror image holds. We call the preservation requirement *invariability* and the tracking one *sensitivity*, both of which are scored on correctness states rather than surface strings, as to be detailed in §2.3.

2.2 TWO-SOURCED INSTANTIATIONS OF REMG PROTOCOL

There is no single existing audio corpus that can host both sides of the ReMG protocol. The linguistic test requires challenging content such as Question-Answering (QA) rather than daily conversations, which acted speech datasets usually fail to provide. The genuine paralinguistics, on the other hand,

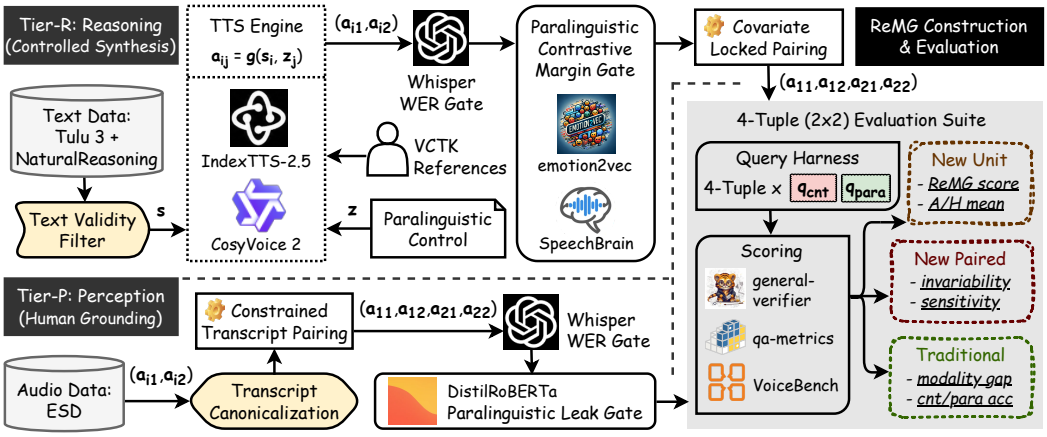

Figure 2: The ReMG data construction and evaluation framework. **Top/Bottom-left:** Text- and audio-sourced data are adopted in two tiers to realize the 4-tuple (2×2) unit as our proposed protocol in Figure 1 respectively (§2.2). **Bottom-right:** Every audio clip within each unit is queried twice for linguistic and paralinguistic features respectively, to support our newly defined metrics (§2.3).

requires real human recordings, which TTS synthesis is not sufficiently qualified. That is one reason why prior evaluations also typically combine two sources. However, they eventually place the two query types respectively on these unrelated audios, leading to the third issue mentioned in §1. Instead, ReMG realizes the proposed 4-tuple protocol twice: **Tier-R** (Reasoning) under controlled synthesis guarantees valid linguistic content, and **Tier-P** (Perception) on human speech grounds paralinguistics in real voices. Note that the design of tiers is intended to take into account the natural characteristics and distribution of both sources of data, rather than using them as representatives for linguistic and paralinguistic tasks as was done in previous works. Actually, both tiers realize the whole 4-tuple protocol independently, which means every audio clip in both tiers is tested under both linguistic and paralinguistic tasks by answering both queries.

Tier-R. As shown in Figure 2 (Top-left), the construction of Tier-R begins from the reasoning questions in text drawn from Tulu 3 (Lambert et al., 2025) and NaturalReasoning (Yuan et al., 2025), which is filtered to 8 to 50 words and stripped of formatting noise to cater to better TTS quality, as detailed in Appendix B. Next, each linguistic content $\boldsymbol{s}_i$ is rendered under two different paralinguistic states $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ by two zero-shot TTS synthesizers, IndexTTS-2.5 (Li et al., 2026) and CosyVoice 2 (Du et al., 2024), conditioned on reference voices from VCTK (Yamagishi et al., 2019). The two TTS engines are intended to prevent potential overfitting to synthesizer-specific artifacts. For simplicity, we mainly adopt emotion for $\boldsymbol{z}_j$ in this work, while this is just a practical option rather than a restricted condition of our protocol. The candidate pairs $(\boldsymbol{a}_{i1}, \boldsymbol{a}_{i2})$ should then survive a Word Error Rate (WER) gate by Whisper-medium (Radford et al., 2023), which requires WER $\leq 20\%$ for single audio clips and pairwise ΔWER $\leq 10/10/5\%$ between $\boldsymbol{a}_{i1}$ and $\boldsymbol{a}_{i2}$, respectively on the training/validation/evaluation split, so that the speech recognition itself would not be a confounding factor. Also built upon two backbones, emotion2vec (Ma et al., 2024) and SpeechBrain (Ravanelli et al., 2024), the following gate on the contrastive margin requires that the paralinguistic intervention within each $(\boldsymbol{a}_{i1}, \boldsymbol{a}_{i2})$ (i.e., between $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$) is indeed perceptible and distinguishable, as detailed in Appendix B. Finally, surviving pairs are further paired into 4-tuple units $\boldsymbol{u}$, based on the definition of our 2×2 protocol and a principle that all the audio clips within one unit share the same covariates c, including the specific synthesizer and reference voice. In evaluation, as in Figure 2 (Bottom-right), each audio clip has two query types, both of which are aligned with the format in VoiceBench (Chen et al., 2026). Specifically, the linguistic one adopts the audio itself as the reasoning question, and the free-form LALM responses are then scored by routed automatic verifiers, including general-verifier (Ma et al., 2025), qa-metrics (Li et al., 2024), and VoiceBench (Chen et al., 2026). The paralinguistic query uses the MMSU emotion recognition surface from VoiceBench (Chen et al., 2026) with a closed label set.

Tier-P. As shown in Figure 2 (Bottom-left), Tier-P is built from the English partition of the Emotional Speech Database (ESD) (Zhou et al., 2022), in which 10 actors read 350 fixed scripts in 5 emotions, giving naturally paired real-human audio clips $(\boldsymbol{a}_{i1}, \boldsymbol{a}_{i2})$. As there are slight differences

between the official transcriptions of 5 clips with the same script, such as different word order or punctuation, we use the neutral-emotion transcript as the canonical transcript for all 5 clips to avoid introducing potential shortcuts. Then, we pair two scripts into 4-tuples under the constraints of the same actor and emotions, followed by a similar Whisper-medium WER gate for basic recognizability. Finally, a paralinguistic leak gate based on DistilRoBERTa (Sanh et al., 2019) is adopted to filter out those scripts whose emotion can be guessed from the linguistic content alone. More details are deferred to Appendix C. In evaluation, as the contents in ESD are mainly daily conversations rather than QA, the linguistic query is a choice between the two canonical transcripts of the 4-tuple, asking which one matches the audio. It is simple but complementary to Tier-R, where the original modality gap is defined and calculated, by contributing real human paralinguistics. The paralinguistic query, on the other hand, uses the same MMSU surface as Tier-R with the 5 ESD labels.

2.3 Modality Gap under Linguistic-Paralinguistic Trade-Off

Now it is time to rethink how to claim modality gap in a manner that is significant and responsible for real-world practices. The reduction of modality gap implies an inherent philosophy: To ignore all features other than linguistic content given linguistic queries. The problem here is how such ignore happens. If it is realized by forcing the LALM to learn all other features as noise (even if only in linguistic tasks), the hurt to the model capability of paralinguistic perception and processing is basically unavoidable. Thus, we argue that **modality gap should always be claimed and measured together with linguistic-paralinguistic trade-off**. Unfortunately, existing methods did not account for this point, nor do their protocol and metrics possess the capability to achieve it. In this section, we propose new pair- and unit-based metrics upon our ReMG protocol, as illustrated in Figure 2 (Bottom-right), to achieve more real and comprehensive measurement.

Traditional metrics. First, we retain the original definition of modality gap (Wang et al., 2026a;b; Cao et al., 2026), as well as some traditional metrics to remain comparable with prior studies. Given the target LALM f, data split $\mathcal{D}$, units $\boldsymbol{u} \sim \mathcal{D}$ with audio $\boldsymbol{a}_{ij}$ and their content $\boldsymbol{s}_i$ in text, the *modality gap* G is formulated in Tier-R as the difference in mean accuracy under the scoring h between their responses:

$$G = \mathbb{E}_{\boldsymbol{u},i,j}\left[\, h(f(\boldsymbol{s}_i)) - h(f(\boldsymbol{a}_{ij}))\,\right]. \tag{2}$$

We also report the mean accuracies over the content and paralinguistic queries of each tier, respectively as *cnt accuracy* A_{cnt} and *para accuracy* A_{para}. While aligned with existing works, these metrics only describe how the LALM behaves on linguistic or paralinguistic tasks in isolation, failing to model their real-world trade-off.

New pair metrics. Under our ReMG protocol, any pairs of adjacent audio clips within one unit share either the same linguistic content $\boldsymbol{s}_i$ or the same paralinguistic state $\boldsymbol{z}_j$, which makes it possible to conduct controlled comparison along each axis. This first leads to two categories of paired metrics, namely invariability and sensitivity. Specifically, given different $\boldsymbol{s}_i$ yet the same $\boldsymbol{z}_j$, $\boldsymbol{q}_{para}$ expects to get correct and equivalent responses, while $\boldsymbol{q}_{cnt}$ expects to see correct changes according to $\boldsymbol{s}_i$, and vice versa. The achievement rates of such expectations are formulated as *cnt invariability* I_{cnt} and *para invariability* I_{para}:

$$I_{cnt} = \mathbb{E}_{\boldsymbol{u},i}\left[h(f(\boldsymbol{a}_{i1})) \cdot h(f(\boldsymbol{a}_{i2}))\right], \quad I_{para} = \mathbb{E}_{\boldsymbol{u},j}\left[h(f(\boldsymbol{a}_{1j}, \boldsymbol{q}_{para})) \cdot h(f(\boldsymbol{a}_{2j}, \boldsymbol{q}_{para}))\right], \tag{3}$$

as well as *cnt sensitivity* S_{cnt} and *para sensitivity* S_{para}:

$$S_{cnt} = \mathbb{E}_{\boldsymbol{u},j}\left[h(f(\boldsymbol{a}_{1j})) \cdot h(f(\boldsymbol{a}_{2j}))\right], \quad S_{para} = \mathbb{E}_{\boldsymbol{u},i}\left[h(f(\boldsymbol{a}_{i1}, \boldsymbol{q}_{para})) \cdot h(f(\boldsymbol{a}_{i2}, \boldsymbol{q}_{para}))\right]. \tag{4}$$

Careful readers may notice that these metrics do not directly trade off the linguistic and paralinguistic axis, as they are calculated for either "cnt" or "para". However, as introduced in the fourth paragraph of §1, they are expected to reveal the model capability in distinguishing linguistic and paralinguistic features in single audio clips, which we believe is important for reducing the modality gap. This would be empirically validated in §4.

New unit metrics. Finally, based on the entire 4-tuple unit in our new protocol, we redefine how a modality gap of real-world significance should be claimed under the linguistic-paralinguistic trade-off. Writing unit-wise scores regarding linguistic and paralinguistic tasks respectively as:

$$U_{cnt} = \mathbb{E}_{\boldsymbol{u},i,j}[\, h(f(\boldsymbol{a}_{ij}))\,], \quad U_{para} = \mathbb{E}_{\boldsymbol{u},i,j}[\, h(f(\boldsymbol{a}_{ij}, \boldsymbol{q}_{para}))\,], \tag{5}$$

Table 1: Evaluation of three base LALMs as reported by previous works on modality gap and as re-measured under ReMG. All values are percentages. The AM and HM are standard arithmetic/harmonic means for previous works, but are U_a and U_h over the unit for ReMG, calculated on Tier-R where G is defined. Collected on different data and protocols, this table supports not point-wise comparison but how the intuition about the same base model changes under ReMG.

Model	**Result Source**	$G \downarrow$	$A_{cnt} \uparrow$	$A_{para} \uparrow$	AM $\uparrow$	HM $\uparrow$
Qwen2.5-Omni-7B	CKD (Wang et al., 2026b)	3.52	75.08	74.20	74.64	74.64
	TARS (Wang et al., 2026a)	4.87	71.30	—	—	—
	X-OPD (Cao et al., 2026)	1.93	49.38	—	—	—
	ReMG (Ours)	**3.41**	**52.27**	**39.49**	**45.88**	**44.99**
Qwen2-Audio-7B	Xiang et al. (Xiang et al., 2025)	13.30	47.89	—	—	—
	CKD (Wang et al., 2026b)	9.19	55.35	—	—	—
	CORD (Hu et al., 2026)	15.25	35.99	60.81	48.40	45.22
	ReMG (Ours)	**3.98**	**40.62**	**0.57**	**20.60**	**1.12**
Phi-4-Multimodal	TARS (Wang et al., 2026a)	15.16	63.23	56.70	59.97	59.79
	ReMG (Ours)	**19.32**	**48.30**	**41.19**	**44.74**	**44.46**

the **real modality gap** should be claimed together with the *unit arithmetic means* U_a and *unit harmonic means* U_h between U_{cnt} and U_{para}, respectively formulated as:

$$U_a = (U_{cnt} + U_{para}) \,/\, 2, \quad U_h = 2 \cdot U_{cnt} \cdot U_{para} \,/\, (U_{cnt} + U_{para}). \tag{6}$$

Note that different from previous, due to the existence of our new protocol, such originally simple metrics turn to directly reflect the tension between linguistic and paralinguistic capability on exactly the same controlled data within each unit. When a base model B exists (e.g., tuning-based methods, which is the mainstream solution for reducing modality gap), such a philosophy can be simplified into one *ReMG score* U:

$$U = \max(0,\, (G^B - G)) - \max(0,\, (U^B_{para} - U_{para})), \tag{7}$$

which prices the linguistic gap reduction against the paralinguistic damage paid for it, where a positive value means that the gap improved by more than the paralinguistic capability dropped.

2.4 Revisiting Reported Modality Gaps

We close the ReMG framework by asking what changes once the evaluation controls the linguistic content, paralinguistic state, and covariates. Table 1 places the numbers reported by recent studies on modality gap for three mainstream base LALMs they adopt with our re-evaluation results under ReMG. Collected on different data with different query formats, point-wise numerical comparison is not meaningful here. What matters is how the conclusions about the same model change.

The blanks in the table first reveal a disturbing reality that over half of previous records did not even report a paralinguistic score. While for those reported, there is a huge drop under our ReMG test, which can be attributed to two main reasons: 1) ReMG eliminates the shortcut to inflate scores over two disjoint datasets, thereby directly exposing the linguistic-paralinguistic tension of the model, which turns out to mainly sacrifice the paralinguistic side in previous works, 2) For those ReMG audios whose linguistic content is a challenging question but actually queried regarding paralinguistic features, LALM can be wrongly captured into answering the question. The collapse of Qwen2-Audio is such a demonstration, 318 of whose 352 paralinguistic responses are actually free-form answers to the linguistic content. Both of these failure patterns are invisible to prior protocols, where the two query types are placed on unrelated audio, varying more than one factor. This further enhances the motivation of ReMG: If the conclusions about even a base model depend heavily on whether the evaluation controls content, paralinguistic, and covariates axis, then a real reduction of modality gap should even more be claimed and measured under such control.

3 Reducing Real Modality Gap via ReMG

This section describes how the proposed ReMG protocol helps reduce the modality gap by fine-tuning LALMs. Following recent works (Wang et al., 2026a;b; Hu et al., 2026; Cao et al., 2026), we

Table 2: Main results on Tier-R under ReMG. All models are evaluated with greedy decoding, and all values are percentages. U is defined only for fine-tuned models against their own base.

Model	$G\downarrow$	$A_{cnt}\uparrow$	$A_{para}\uparrow$	$I_{cnt}\uparrow$	$S_{cnt}\uparrow$	$I_{para}\uparrow$	$S_{para}\uparrow$	$U_a\uparrow$	$U_h\uparrow$	$U\uparrow$
Commercial LALMs										
Qwen3.5-Omni-Flash	7.67	52.56	60.23	62.26	53.33	56.82	20.45	56.39	56.13	—
Qwen3-Omni-Flash	6.82	47.73	55.11	60.42	48.21	51.14	10.80	51.42	51.16	—
Doubao Seed2.0 Lite	9.38	71.31	69.60	78.87	71.43	61.93	39.20	70.45	70.44	—
GPT-Audio	9.66	68.18	51.70	72.99	65.74	40.34	14.20	59.94	58.81	—
Gemini 3.7 Flash	6.82	64.20	66.48	76.00	70.21	57.95	35.80	65.34	65.32	—
Gemini 3.1 Pro	7.10	62.22	56.25	76.23	67.78	45.45	21.59	59.23	59.08	—
Publicly Released Checkpoint										
TARS-Qwen2.5 (Wang et al., 2026a)	8.52	54.55	48.01	63.96	52.86	35.80	11.93	51.28	51.07	0.00
Open-Source LALMs										
Qwen2.5-Omni-7B	3.41	52.27	39.49	68.37	**62.96**	27.84	5.11	45.88	44.99	—
+ CKD (Wang et al., 2026b)	3.41	**53.98**	37.22	**71.29**	61.67	26.70	3.41	45.60	44.06	-2.27
+ TARS (Wang et al., 2026a)	4.55	52.84	48.30	64.36	54.84	39.77	7.39	50.57	50.47	0.00
+ CORD (Hu et al., 2026)	4.26	53.69	26.42	68.63	55.36	11.36	2.27	40.06	35.41	-13.07
+ X-OPD (Cao et al., 2026)	3.69	51.99	28.69	66.33	53.70	14.20	2.27	40.34	36.98	-10.80
+ **ReMG (Ours)**	**0.57**	53.41	**52.56**	67.37	59.62	**48.30**	**7.39**	**52.98**	**52.98**	**+2.84**
Phi-4-Multimodal	19.32	48.30	41.19	47.06	39.47	**32.95**	1.14	44.74	44.46	—
+ **ReMG (Ours)**	**8.52**	**52.84**	**49.15**	**68.52**	**55.00**	27.27	**17.61**	**50.99**	**50.93**	**+10.80**

adopt OPD as the tuning framework. However, different from recent works, we avoid introducing any instance-specific knowledge from external sources, such as stronger teachers or ground-truth answers/labels, as argued in the fifth paragraph of §1. In contrast, our fine-tuning adopts a frozen copy of the target LALM at step 0 as the OPD teacher and incorporates no external rewards. Thus, unlike previous results mixing contributions from protocol and additional knowledge, any observed improvement in our work can be uniquely attributed to the ReMG protocol itself.

Following the two query types $\boldsymbol{q}_{cnt}$ and $\boldsymbol{q}_{para}$ of the ReMG protocol, each training audio clip also carries two losses $\mathcal{L}_{cnt}$ and $\mathcal{L}_{para}$, respectively regarding linguistic and paralinguistic features. For the former aiming at reducing the modality gap, the objective is an on-policy sampled-token reverse Kullback–Leibler (KL) divergence with a clip in the style of Proximal policy optimization (PPO) (Schulman et al., 2017). Let $y \sim \pi_\theta(\cdot \,|\, \boldsymbol{a},\, \boldsymbol{q}_{cnt})$ denote a response sampled from the target LALM θ, $r_t = \pi_\theta(y_t \,|\, \boldsymbol{a},\, y_{<t}) \,/\, \pi_{\theta_{old}}(y_t \,|\, \boldsymbol{a},\, y_{<t})$ denote the importance ratio against the behavior policy, and $V_t = \log \pi_B(y_t \,|\, \boldsymbol{s},\, y_{<t}) - \log \pi_{\theta_{old}}(y_t \,|\, \boldsymbol{a},\, y_{<t})$ denote the per-token advantage of the teacher over the behavior policy given text input $\boldsymbol{s}$, the loss for linguistic content is formulated as:

$$\mathcal{L}_{cnt} = -\frac{1}{T}\sum_{t=1}^{T} \min\left(r_t V_t,\ \mathrm{clip}\left(r_t,\, 1-\epsilon,\, 1+\epsilon\right) V_t\right), \tag{8}$$

where T is the number of response tokens over which the loss is averaged, and ϵ is the clipping threshold. While on the paralinguistic side, the objective is an offline forward KL along the greedy teacher trajectory y^* given the same audio input $\boldsymbol{a}$ and query $\boldsymbol{q}_{para}$ as the target LALM, as:

$$\mathcal{L}_{para} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{KL}\left(\pi_B(\cdot \,|\, \boldsymbol{a},\, \boldsymbol{q}_{para},\, y^*_{<t}) \,\|\, \pi_\theta(\cdot \,|\, \boldsymbol{a},\, \boldsymbol{q}_{para},\, y^*_{<t})\right), \tag{9}$$

which is expected to start near zero and act against the drift of pre-trained audio capability.

Inspired by On-Policy Self-Distillation (OPSD) (Zhao et al., 2026), a recent variant of OPD, we introduce the concept of privileged context, which enables providing the teacher with more information than the student, to both $\mathcal{L}_{cnt}$ and $\mathcal{L}_{para}$. However, different from its original definition, which directly provides instance-specific answers to the teacher as privilege, we narrow it to only supplement task-level instructions. Specifically, for $\mathcal{L}_{para}$, the privileged context is fixed to "*Ignore the linguistic content of the speech, including any problem being read aloud. Classify only the speaker's affect from speaking style*", to prompt the teacher to follow the design of our ReMG protocol better. While for $\mathcal{L}_{cnt}$, the privileged context is "*Answer concisely and accurately. Focus on the question itself. Avoid redundancy, boilerplate, and lengthy tutorials*", which is mainly to raise

Table 3: Evaluation of generalization ability to six existing benchmarks on Qwen2.5-Omni-7B.

Model	VoiceBench				MMAU				Overall	
	MMSU	OBQA	SDQA	Mean	Speech	Sound	Music	Mean	AM	HM
Base model	49.28	79.78	56.96	62.01	39.94	64.26	50.90	51.70	56.86	56.39
+ CKD	52.86	83.08	61.30	**65.75**	33.93	64.56	46.41	48.30	57.03	55.69
+ TARS	51.82	80.00	58.05	63.29	37.84	66.07	49.70	51.20	57.25	56.61
+ CORD	54.42	83.52	59.13	65.69	31.83	60.96	39.22	44.00	54.85	52.70
+ X-OPD	53.45	83.08	60.40	65.64	34.83	63.66	43.71	47.40	56.52	55.05
+ **ReMG (Ours)**	54.46	81.70	58.59	64.92	40.54	64.26	49.70	**51.50**	**58.21**	**57.44**

the distillation efficiency. Besides, as $\mathcal{L}_{cnt}$ and $\mathcal{L}_{para}$ are both on audio capability, to limit potential text drift, we also provide an optional loss that distills the Tier-R text trajectory with a similar offline forward KL as Eq. 9, which is enabled only when text degradation occurs in practice.

4 EXPERIMENTS

In this section, we demonstrate how the proposed ReMG reveals and reduces the real modality gap with comprehensive empirical evidence across six commercial LALMs and two open-source backbones, Qwen2.5-Omni-7B (Xu et al., 2025a) and Phi-4-Multimodal (Abouelenin et al., 2025). The detailed experimental setup is deferred to Appendix D.

Main results. The main comparison involves four recent SOTA baselines, CKD (Wang et al., 2026b), TARS (Wang et al., 2026a), CORD (Hu et al., 2026), and X-OPD (Cao et al., 2026), through a knowledge-controlled reproduction. As argued in §3, we remove all their external instance-specific knowledge sources to keep them at the same knowledge level as ours for fair comparison, and to distinguish their real protocol-level contributions with those from new knowledge learning. The fine-tuning of the baselines shares the same data with ours, while excluding access to the ReMG protocol upon it. Besides, as TARS is the only one publicly releasing its fine-tuned checkpoint, we also involve it for reference. Table 2 shows the results on Tier-R, while Table 8 in Appendix E supplements those on Tier-P, which reveals that, **unlike the current view, most existing methods significantly degrade the pre-trained acoustic capability, while none of them substantially reduce the gap**. Compared with the base model, CKD retains the same gap level but results in even a worse linguistic-paralinguistic trade-off. CORD and X-OPD witness a collapse in paralinguistic scores, which makes them inapplicable in real-world practices. TARS is the only one to successfully raise the Pareto front upon the trade-off, while both sides are still behind ours, and the modality gap also becomes worse. In contrast, ours is the only method with a positive U, which means the gap reduction costs relatively less paralinguistic degradation as its price. At the same time, we achieve the best U_a and U_h for the trade-off, which even exceeds the released TARS checkpoint with external knowledge and a commercial LALM, Qwen3-Omni-Flash. Besides, the four paired metrics help to locate the specific pain point of the baselines to I_{para} and S_{para}, which is especially clear when compared with the commercial LALMs. Then it turns out that we significantly improve I_{para} to a considerable level, while still leaving room for S_{para}, which might be a promising future direction.

Generalization to previous benchmarks. As the above main experiment involves new, task-specific data and metrics, there could be two certain questions: 1) whether we are just creating a new benchmark to chase the score, whose results are not of general significance, and 2) whether the reduced modality gap and improved trade-off reported on ReMG price other general acoustic abilities of LALMs. In response, in this section, we extend our evaluation to six previously recognized benchmarks, namely MMSU, OBQA, and SDQA from VoiceBench (Chen et al., 2026) for the linguistic capability, as well as the speech, sound, and music splits from MMAU (Sakshi et al., 2025) for the paralinguistic one. The results on Qwen2.5-Omni-7B are shown in Table 3. It can be found that all the existing methods show slight improvements in the mean linguistic score on VoiceBench, which is a similar tendency to the main results in Table 2 that might benefit from the self-regularization by the shared tuning framework, OPD. On the other hand, all methods witness a drop on the MMAU side, while ours turns out to be the best one with a degradation of only 0.2%. As a result, we finally achieve improved and the best overall AM and HM scores beyond all base-

Table 4: Ablations w.r.t. the components of the ReMG fine-tuning on Qwen2.5-Omni-7B. All values are percentages. The results verify the independent contributions from the ReMG protocol and other practical parts, and double-confirm the current issues of ignoring the paralinguistic capability.

Checkpoint	Tier-R						Tier-P	
	$G \downarrow$	$A_{cnt} \uparrow$	$A_{para} \uparrow$	$U_a \uparrow$	$U_h \uparrow$	$U \uparrow$	$U_a \uparrow$	$U_h \uparrow$
Base model	3.41	52.27	39.49	45.88	44.99	—	71.65	60.43
+ Vanilla OPD	6.82	51.14	35.80	43.47	42.11	-3.69	71.53	60.21
+ ReMG protocol	6.53	52.56	40.62	46.59	45.83	0.00	71.65	60.43
+ Privilege (**Ours**)	**0.57**	53.41	**52.56**	**52.98**	**52.98**	**+2.84**	**71.65**	**60.43**
− $\mathcal{L}_{para}$ (Current view)	4.55	**56.25**	39.77	48.01	46.60	0.00	71.42	59.98

lines. This result verifies that 1) the improvements observed on ReMG evaluation are generalizable to existing benchmarks, especially regarding the trade-off, and 2) the ReMG fine-tuning would not significantly hurt the pre-trained acoustic ability, which even brings slight advantages in most (i.e., 4 of 6) benchmarks. Besides, an interesting phenomenon is that CKD and TARS also achieved improved overall scores across the six benchmarks, both of which are also ahead of the other two baselines in our main results. Actually, the specific ranking within the four baselines is exactly the same between Table 2 (the overall AM and HM) and Table 3 (the three U series metrics), which double-confirms the effectiveness and generalization ability of the proposed ReMG. Nevertheless, ReMG still provides new insights into CKD and TARS, such as their non-positive U indicating that the Pareto front is not successfully pushed forward. This is exactly where the unique philosophy and value of our ReMG lie: scores on disjoint benchmarks do not capture certain behaviors exclusive to our controlled, matched new protocol with pairs and units.

Ablation study. To illustrate the new contributions of each component of our ReMG compared with the current knowledge, we conduct comprehensive ablations on Qwen2.5-Omni-7B, as shown in Table 4. We first try vanilla OPD, where both $\mathcal{L}_{cnt}$ and $\mathcal{L}_{para}$ exist but are optimized on disjoint data (i.e., the information is the same as ReMG data, but the protocol is not), which gets even worse results than the base model. Next, by simply adding the ReMG protocol upon vanilla OPD, we can observe a reduction of modality gap and an improvement of both linguistic and paralinguistic capabilities, which directly confirms its effectiveness. However, the absolute gap is still larger than the base model at this time. Through detailed analysis of the LALM dynamics and responses during fine-tuning, we found that this is because the vanilla OPD framework is not capable of fully following our ReMG protocol without specific enhancement. Therefore, as detailed in §3, we then add task-level instructions as the privilege along with both text and audio queries to the teacher. Note that this is a practical- rather than protocol-level design, as our ReMG protocol is itself an independent contribution not tied to any specific optimization algorithms. Finally, to empirically support the second issue in §1 that the current view excludes the paralinguistic capability as an optimization objective, we remove the $\mathcal{L}_{para}$ update from ours while keeping all other unchanged. It is immediately clear that the gap re-appears and the trade-off collapses with the paralinguistic side, which double-confirms our argument. More ablations on Phi-4-Multimodal are deferred to Appendix E.

5 Conclusion

In this work, we revisit the current protocol of the LALM modality gap and find it incapable of supporting the current claim: reducing the gap while retaining the pre-trained paralinguistic capability. As the two parts are considered on disjoint data without any controls, such a claim can be a false prosperity that collapses once facing the real-world tension between linguistic and paralinguistic capabilities. We fix this flaw with a new protocol, ReMG, forcing LALM to face the trade-off on the same data and to price gap reduction against paralinguistic degradation, instead of finding shortcuts to chase two disjoint scores. Previous works then show widespread paralinguistic collapses without substantial gap improvement under ReMG, while simple OPD fine-tuning with the ReMG protocol is demonstrated to be effective and generalizable in addressing this problem. As the new ReMG protocol itself is an independent contribution not specific to our data and method, further constructing benchmarks upon ReMG with larger scale or more real-world data, or developing more advanced optimization algorithms to better facilitate its potential are promising future directions.

AI USE STATEMENT

In this work, we used generative AI tools for [generate synthetic data sets], and [clean and reformat dataset]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks], [propose or refine hypotheses], [design or provide feedback on research methodology or experiments], [implement methods], [assist with translation], [support qualitative and thematic data analysis], and [interpret results], and [formulate mathematical claims], [provide critical ingredients for proving mathematical claims], and [assist in the writing of proofs] are not applicable to this work. Additionally, we used generative AI tools for [create or edit software code]. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We fully support the reproducibility policy of ICLR 2027 and have made best efforts to ensure that our results are reproducible. The construction of the ReMG benchmark, including the data sources, the pair and unit organization, and the quality control pipeline, is described in §2.1 and §2.2, as well as Appendix B and C. The definitions of all metrics are given in §2.3, and all prompt templates are listed in Appendix F. The complete training and evaluation details are provided in Appendix D. A downloadable repository containing our source code, constructed data, and fine-tuned checkpoints is provided via an anonymous link in §1. We also promise to open it to the public upon acceptance.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024.

Di Cao, Dongjie Fu, Hai Yu, Siqi Zheng, Xu Tan, and Tao Jin. X-opd: Cross-modal on-policy distillation for capability alignment in speech llms. *arXiv preprint arXiv:2603.24596*, 2026.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *Transactions of the Association for Computational Linguistics (TACL)*, 2026.

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

Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, and Irwin King. Voxeval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

Yitian Gong, Luozhijie Jin, Kuangwei Chen, Dong Zhang, Ruifan Deng, Xiaogui Yang, Xin Zhang, Zhaoye Fei, Qinyuan Cheng, Shimin Li, et al. Xy-tokenizer: Mitigating the semantic-acoustic conflict in low-bitrate speech codecs. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

Ming-Hao Hsu, Xueyao Zhang, Xiaohai Tian, Jun Zhang, and Zhizheng Wu. Anatomy of the modality gap: Dissecting the internal states of end-to-end speech llms. *arXiv preprint arXiv:2603.01502*, 2026.

Jing Hu, Danxiang Zhu, Xianlong Luo, Dan Zhang, Shuwei He, Yishu Lei, Shikun Feng, Hai-Tao Zheng, Jingzhou He, Yu Sun, Hua Wu, and Haifeng Wang. Cord: Bridging the audio–text reasoning gap via weighted on-policy cross-modal distillation. In *Findings @ Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, et al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *International Conference on Learning Representations (ICLR)*, 2025.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. In *Conference on Language Modeling (CoLM)*, 2025.

Yunpei Li, Xun Zhou, Jinchao Wang, Lu Wang, Yong Wu, Siyi Zhou, Yiquan Zhou, Yining Wang, Yaogen Yang, Zhetao Hu, et al. Indextts 2.5 technical report. *arXiv preprint arXiv:2601.03888*, 2026.

Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan Lee Boyd-Graber. Pedants: Cheap but effective and interpretable answer equivalence. In *Findings @ Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Kevin Lu and Thinking Machines Lab. On-policy distillation, 2025. URL `https://thinkingmachines.ai/blog/on-policy-distillation`.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. General-reasoner: Advancing llm reasoning across all domains. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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 *Findings @ Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

OpenAI. Gpt-4o-audio, 2024. URL `https://developers.openai.com/api/docs/models/gpt-4o-audio-preview`.

Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, Yangui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, et al. A survey on speech large language models for understanding. *IEEE Journal of Selected Topics in Signal Processing*, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, 2023.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain De Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, et al. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research (JMLR)*, 2024.

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 *International Conference on Learning Representations (ICLR)*, 2025.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Chaoren Wang, Heng Lu, Xueyao Zhang, Shujie Liu, Yan Lu, Jinyu Li, and Zhizheng Wu. Closing the modality reasoning gap for speech large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026a.

Enzhi Wang, Qicheng Li, Zhiyuan Tang, and Yuhang Jia. Cross-modal knowledge distillation for speech large language models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026b.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

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. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit, 2019. URL `https://doi.org/10.7488/ds/2645`.

Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Dong Wang, Ilia Kulikov, Kyunghyun Cho, Yuandong Tian, Jason Weston, et al. Naturalreasoning: Reasoning in the wild with 2.8m challenging questions. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *International Conference on Machine Learning (ICML)*, 2026.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 2022.

APPENDIX

A RELATED WORKS

The current view is to treat the modality gap mainly as a problem on the linguistic content side. Mainstream existing works therefore try to close this gap by transferring text-channel behavior onto the audio channel through distillation, on-policy alignment, or reinforcement learning. Cross-modal Knowledge Distillation (CKD) Wang et al. (2026b) establishes dual text-to-text and audio-to-text Knowledge Distillation (KD) channels, together with a sequence-level cross-entropy (CE) on teacher-generated pseudo-labels. Beyond distillation, Trajectory Alignment for Reasoning in Speech (TARS) Wang et al. (2026a) uses Group Relative Policy Optimization (GRPO) Shao et al. (2024) to enforce trajectory-level alignment on spoken reasoning rollouts and to optimize sequence-level policy behaviors. Cross-modal weighted On-policy Reward-guided Distillation (CORD) Hu et al. (2026) proposes a cross-modal OPD framework that takes the in-model text modality as the reference, without an independent frozen teacher copy. It aligns audio rollouts with an importance- and position-weighted reverse KL at the token level, and applies judge-guided sequence-level optimization through GRPO. Finally, Cross-modal OPD (X-OPD) Cao et al. (2026) optimizes LALMs from self-generated student trajectories on paired speech-text inputs, using policy gradients with token-level feedback from a text teacher. However, it is not hard to notice that all these methods do not involve explicit design for the retention of pre-trained paralinguistic capability. External instance-specific knowledge is also mixed for all to different extents, which makes it hard to distinguish the contributions from their specific designs and new knowledge learning.

Table 5: Comparison of our ReMG with the four SOTA baselines in terms of implementation details. The "A" and "T" in the row of paradigm respectively refer to "Audio" and "Text".

		CKD	TARS	CORD	X-OPD	Ours
Venue		ICASSP'26	ACL'26	ACL'26 Findings	INTERSPEECH'26	—
Training	**Data Source**	OpenOrca	UnifiedQA	NuminaMath	Tulu 3 NaturalReasoning	Tulu 3 NaturalReasoning ESD
	TTS	CosyVoice 2	CosyVoice 2 OpenAudio S1-mini	Kokoro	CosyVoice 3	CosyVoice 2 IndexTTS-2.5
Evaluation		VoiceBench MMAU	VoiceBench MMSU LibriSpeech	VoiceBench MMAU	VoiceBench BIG Bench Audio Audio Multi-Challenge MMAR	ReMG VoiceBench MMAU
Target Model		Qwen2.5-Omni-7B	Qwen2.5-Omni-7B Phi-4-Multimodal	Qwen2-Audio-7B Step-Audio-2-mini	Qwen3-Omni-A3B-Instruct	Qwen2.5-Omni-7B Phi-4-Multimodal
Distillation	**Teacher**	Qwen2.5-7B-Instruct	—	Self (In-model)	Qwen3-A3B-Instruct	Self (Frozen)
	Paradigm	KD (A→T + T→T)	—	OPD (A→T)	OPD (A→T + T→T)	OPD (A→T) KD (A→A)
Other Optimization		CE (Teacher pseudo-label)	GRPO (Trajectory alignment)	GRPO	—	—
Open-Source		×	✓	×	×	✓

B TIER-R CONSTRUCTION DETAILS

Corresponding to §2.2 and Figure 2, this section details each stage and component in the construction pipeline of Tier-R, with the specific counts summarized by Table 6.

Text source and filter. The Tier-R contents are drawn from Tulu 3 (Lambert et al., 2025) and NaturalReasoning (Yuan et al., 2025) datasets. Starting from 1,831,584 normalized candidates, we keep prompts of 8 to 50 words and at most 350 characters with answers of 1 to 20 words, remove code fences and formatting noise, keep English text with a langdetect probability of at least 0.85 together with an ASCII-letter heuristic, drop the multi-language `aya_100k` subset of

Table 6: The data construction funnel of Tier-R. The numbers in the table denote the count of samples or cases regarding the specific "Object". As the text filter runs before the dataset split, the first two rows have no count (i.e., −) specific to the training, validation, or evaluation split.

Tier-R	Object	Train	Valid	Eval	Total Size
Stage					
Normalized candidates	texts	—	—	—	1,831,584
Text filter	texts	—	—	—	315,359
Drawn texts	texts	8,250	1,000	750	10,000
Complete pairs	pairs	8,157	989	744	9,890
WER gate	pairs	4,023	498	326	4,847
Paralinguistic gate	pairs	2,397	296	190	2,883
Unit pairing	4-tuples	1,152	142	88	1,382
Component					
Reference voices	speakers	84	9	16	109

Tulu 3, and deduplicate prompts exactly. The filter keeps 315,359 texts, of which 296,887 are from NaturalReasoning, and 18,472 are from Tulu 3.

Reference voices. The adopted reference voices come from VCTK-Corpus-0.92 (Yamagishi et al., 2019). After excluding one speaker without text annotations, 109 speakers are split into 16 evaluation speakers (8 female and 8 male), 9 validation speakers, and 84 training speakers. For each speaker, we select one neutral 3 to 5 second clip recorded on the close-talk microphone whose Whisper-medium transcription has a WER of at most 5%, which succeeds for all 109 speakers. The same clip serves as the zero-shot reference for both synthesizers, so the reference voice belongs to the covariate c of a unit.

TTS synthesis. We draw 10,000 texts uniformly at random from the filtered pool, assign 5,000 to each engine, and render every text under both paralinguistic states, giving 20,000 jobs. IndexTTS-2.5 (Li et al., 2026) is controlled through its 8-dimensional emotion vector, with the specific coordinates set to 0.8 for the corresponding states. CosyVoice 2 (Du et al., 2024) is controlled through its instruction interface with the prompt "speak in a [emotion] tone". CosyVoice 2 completes all 10,000 jobs while IndexTTS-2.5 completes 9,780, and the 220 failed jobs are excluded from pairing. Before the full-scale run, a pilot of 20 texts $\times$ 2 engines $\times$ 2 states $\times$ 4 training-pool speakers is synthesized and passes a human listening check.

WER gate. Every candidate clip is transcribed by Whisper-medium (Radford et al., 2023) and scored against the spoken text after lowercasing and punctuation stripping. A clip survives when its WER is at most 20%, and a pair $(\boldsymbol{a}_{i1}, \boldsymbol{a}_{i2})$ survives when its two clips differ in WER by at most 10% on the training and validation pools while at most 5% on the evaluation pool. The single-clip cutoff is looser than the Tier-P gate because spoken mathematical content inflates recognition error. Of the 19,780 synthesized clips forming 9,890 complete pairs, 4,847 pairs survive, of which 326 belong to the evaluation pool.

Paralinguistic gate. Each surviving pair must additionally show that the paralinguistic state intervention is perceptible, using two off-the-shelf emotion recognizers, emotion2vec (Ma et al., 2024) and a SpeechBrain wav2vec2 classifier (Ravanelli et al., 2024). A unit passes the absolute rule when, on at least one backend, each of its two clips is classified as its constructed state with a top-1 minus top-2 margin of at least 0.2. Formally, with $P(\cdot)$ the posterior of one backend, a unit passes the contrastive rule when:

$$\big[\,P(\mathrm{e}_1) - P(\mathrm{e}_2)\,\big]_{\boldsymbol{a}_{i,\mathrm{e}_1}} - \big[\,P(\mathrm{e}_1) - P(\mathrm{e}_2)\,\big]_{\boldsymbol{a}_{i,\mathrm{e}_2}} \;\geq\; \delta \tag{10}$$

on either backend, where e_1, e_2 denote the two targeted paralinguistic states, and the threshold is $\delta = 0.10$. A unit survives when it passes either rule, giving 2,397 training, 296 validation, and 190 evaluation units, of which 973 come from CosyVoice 2 and 1,910 from IndexTTS-2.5 in total. The main paralinguistic contrast adopted is $\boldsymbol{z}_1$ = neutral versus $\boldsymbol{z}_2$ = happy, as this pair is the most natural and the most reliably realizable contrast for both content types. Human actors and zero-shot synthesizers can both produce it on demand, whereas rendering, for instance, an angry mathematical

Table 7: The data construction funnel of Tier-P. Different from Tier-R, it is in parallel with speaker- and script-hold-out constructions. The numbers in the table denote the sample counts.

Tier-P	Train	Valid	Eval	Total Size
Source				
Audio clips	—	—	—	17,500
Speaker-hold-out				
Unit pairing	1,218	174	348	1,740
WER gate	615	100	131	846
Paralinguistic gate	539	85	119	743
Script-hold-out				
Unit pairing	1,319	169	239	1,727
WER gate	658	78	116	852
Paralinguistic gate	564	66	116	746
Conflict-free split				
Units	400	146	220	766

problem is neither natural nor stable. However, this choice does not fix the protocol, which supports any pairing of paralinguistic states.

Unit pairing. Audio clips are paired into 4-tuple units within the same split, speaker, and engine by greedy shuffle with seed 42, so that all four clips within one unit share the same covariate c, which gives 1,152 training, 142 validation, and 88 evaluation units. The evaluation split therefore contains 88 4-tuples with 176 unique contents, 352 clips, and 704 queries, synthesized with the 16 held-out VCTK voices. Because each content is bound to its reference voice at synthesis time, this single speaker-disjoint cut yields unseen contents and unseen reference timbres at once.

Evaluation. The content query $\boldsymbol{q}_{cnt}$ presents the audio itself as the question. Free-form content responses are scored by routed automatic verifiers according to the source dataset: general-verifier (Ma et al., 2025) for NaturalReasoning, qa-metrics (Li et al., 2024) for Tulu 3 open QA, and the VoiceBench option extractor (Chen et al., 2026) for Tulu 3 multiple-choice. Paralinguistic query $\boldsymbol{q}_{para}$ targets the emotion within the same audio, with the responses parsed into the closed label set and scored by exact match.

C Tier-P Construction Details

Corresponding to §2.2 and Figure 2, this section details each stage and component in the construction pipeline of Tier-P, with the specific counts summarized by Table 7.

Audio source. Tier-P is built from the English partition of the Emotional Speech Database (ESD) (Zhou et al., 2022), in which 10 actors read 350 fixed scripts in 5 emotions, giving 17,500 clips that form 3,500 complete parallel groups.

Transcript canonicalization. The official release of ESD provides one transcript per audio clip, and in 150 of the 3,500 script groups the five transcripts of the same script differ, mostly through repeated or reordered words. Specifically, 268 clips differ from the neutral clip at the word level, and 14 differ by punctuation. Such differences may potentially break the 2×2 unit, since the two rows of a unit would no longer share exactly the same content $\boldsymbol{s}$, and may also leak the answer of the content query through surface hints. We therefore take the neutral transcript of each script as the canonical transcript of all five clips, while the WER gate below still scores each take against its own official transcript.

Unit pairing. Two scripts are paired into a 4-tuple only within the same actor and the same emotion pair, and a candidate pair is kept only when the two canonical transcripts differ by at most 3 words and at most 30% of the longer script, and when their Jaccard word-type overlap is at most 0.5, above which the pair is re-drawn. This gives 1,740 candidate 4-tuples under the speaker hold-out and then 1,727 under the script hold-out.

Table 8: Main results on Tier-P under the ReMG protocol. All values are percentages. As the content identification is basically a saturated control, the paralinguistic axis carries most of the signal.

Model	A_{cnt} ↑	A_{para} ↑	I_{para} ↑	S_{para} ↑	U_a ↑	U_h ↑
Commercial LALMs						
Qwen3.5-Omni-Flash	100.00	50.57	38.41	12.73	75.28	67.17
Qwen3-Omni-Flash	100.00	46.82	31.82	9.09	73.41	63.78
Doubao Seed2.0 Lite	100.00	60.00	42.73	32.50	80.00	75.00
GPT-Audio	90.23	43.30	26.82	7.05	66.76	58.51
Gemini 3.7 Flash	100.00	52.95	34.09	20.91	76.48	69.24
Gemini 3.1 Pro	99.20	40.00	25.23	7.05	69.60	57.01
Publicly Released Checkpoint						
TARS-Qwen2.5 (Wang et al., 2026a)	100.00	40.91	22.95	6.36	70.45	58.06
Open-Source LALMs						
Qwen2.5-Omni-7B	100.00	**43.30**	27.05	4.55	**71.65**	**60.43**
+ CKD (Wang et al., 2026b)	100.00	42.61	27.27	3.64	71.31	59.76
+ TARS (Wang et al., 2026a)	100.00	43.07	26.36	**5.23**	71.53	60.21
+ CORD (Hu et al., 2026)	100.00	42.61	26.82	3.41	71.31	59.76
+ X-OPD (Cao et al., 2026)	100.00	43.07	27.50	3.86	71.53	60.21
+ **ReMG (Ours)**	100.00	**43.30**	**28.18**	3.64	**71.65**	**60.43**
Phi-4-Multimodal	93.18	34.20	14.55	**5.45**	63.69	50.04
+ **ReMG (Ours)**	**99.20**	**48.75**	**41.59**	2.73	**73.98**	**65.37**

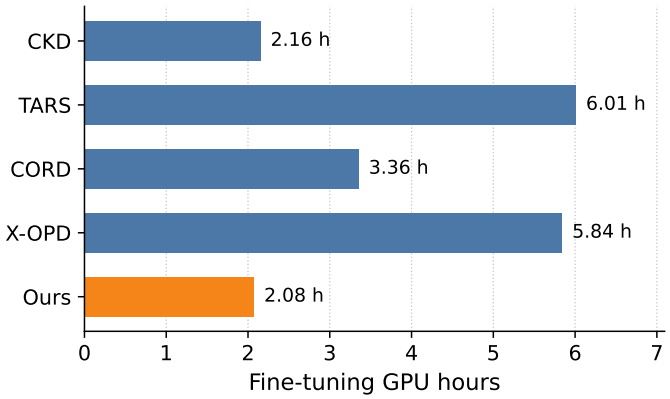

Figure 3: Fine-tuning GPU hour costs on Qwen2.5-Omni-7B and a single NVIDIA RTX PRO 6000.

WER gate. Similar to Tier-R, a Whisper-medium (Radford et al., 2023) is adopted to transcribe each audio clip. A clip is dropped when its WER against the official transcript exceeds 10%, and a script is dropped when its clips with the contrast paralinguistic states differ in WER by more than 3% for the evaluation split and 5% for the other two. This gate keeps 12,701 of the 17,500 clips.

paralinguistic gate. To make sure that the paralinguistic states cannot be leaked by words alone, we score the canonical transcript of every script with an off-the-shelf DistilRoBERTa emotion classifier (Sanh et al., 2019), and drop a script when the largest score over any candidate paralinguistic states reaches 0.9. This removes 24 of the 350 scripts (9 anger, 8 joy, 4 sadness, and 3 surprise), together with all 4-tuples that contain them.

Evaluation. As 10 actors can hardly support a simultaneous actor-and-script hold-out, the evaluation divides it into a speaker hold-out using actors 0011 and 0015 (one male and one female) and a script hold-out using 50 of the 350 scripts. After the gates above, the two constructions respectively contribute 119 and 116 evaluation units. When one audio appears in two units across the two constructions, we drop the whole script-hold-out unit, giving a conflict-free split of 220 units with

Table 9: The removed or replaced components with external instance-specific knowledge in our reproductions of the four current SOTA baselines. All variants share the same (but not paired as in the proposed ReMG) audio data and the fine-tuning setting as provided in §D.

Method	Removed / Replaced Component
CKD	External Qwen2.5-7B-Instruct teacher (replaced by the text channel of the same model as the student, Qwen2.5-Omni-7B), gold-label cross-entropy
TARS	Gold accuracy reward with xFinder, external Qwen3-Embedding (replaced by pooled hidden states of the text channel of the teacher)
CORD	Proprietary judge model, GRPO sequence loss
X-OPD	External Qwen3-A3B-Instruct teacher (replaced by the text channel of the same model as the student, Qwen2.5-Omni-7B)

Table 10: An additional ablation based on Phi-4-Multimodal to demonstrate the usage of the optional loss on the Tier-R text (rather than audio) as mentioned in §3. All values are percentages.

Checkpoint	Tier-R						Tier-P	
	$G \downarrow$	$A_{cnt} \uparrow$	$A_{para} \uparrow$	$U_a \uparrow$	$U_h \uparrow$	$U \uparrow$	$U_a \uparrow$	$U_h \uparrow$
Base model	19.32	48.30	41.19	44.74	44.46	—	63.69	50.04
Ours w/o optional loss	**7.39**	51.70	**49.43**	50.57	50.54	**+11.93**	**73.98**	**65.45**
Ours w/ optional loss	8.52	**52.84**	49.15	**50.99**	**50.93**	+10.80	**73.98**	65.37

880 audios and 1,760 queries. That means each unit has either an unseen actor or unseen scripts (or both). Training and validation splits come from the intersection of the two constructions, the remaining 8 actors reading the remaining 300 scripts. The content query $\boldsymbol{q}_{cnt}$ of Tier-P is a two-way identification between the two canonical transcripts of the unit. This is a simple task near saturation for the base models, which serves as a control averse to degradation after fine-tuning. The paralinguistic query $\boldsymbol{q}_{para}$ uses the same MMSU surface as Tier-R while presenting the ESD labels.

D Experimental Setup

All the checkpoints in this work share the same fine-tuning and evaluation settings. The random seed is 42. We train LoRA adapters of rank 16, scaling 32, and dropout 0.05 on the query, key, value, and output projections of the LLM self-attention, with the audio encoder and the audio adapter frozen. The teacher is a frozen copy of the student with the adapters disabled. We use AdamW with a learning rate of 1×10^{-4} and a cosine schedule for 1 epoch. The content loss is warmed up over the first 10% of steps. Rollouts are sampled at temperature 1.0 and top-p 1.0 with a PPO clip threshold of 0.2. Generation is capped at 512 tokens for content queries and 64 tokens for paralinguistic queries, where the answer is truncated to the first-line closed-set label. The Tier-R paralinguistic loss applies an offline top-16 forward KL along the teacher greedy trajectory. The experiments are conducted on NVIDIA RTX PRO 6000. To make full use of its 96GB memory, we set the mini-batch size to 6 with 21 accumulation steps, resulting in an effective batch size of 126. The $\mathcal{L}_{cnt}$ for Tier-P can be omitted in practice to improve the fine-tuning efficiency, as its task is originally near saturated. In that case, the masses over $\mathcal{L}_{cnt}$ for Tier-P and $\mathcal{L}_{para}$ for both tiers are assigned as 2:1:1 to keep the global 1:1 ratio between linguistic and paralinguistic losses. In evaluation, we use greedy decoding with at most 512 / 64 new tokens for $\boldsymbol{q}_{cnt}$ / $\boldsymbol{q}_{para}$, which matches their tuning-time generation budget.

E Additional Experimental Details

This section supplements more experimental details, including additional main results on Tier-P as Table 8 and the specific removed or replaced components when reproducing the baselines as

described in §4 as Table 9. Also, we conduct ablation on the optional text loss mentioned in §3. The experiment is based on Phi-4-Multimodal, as its optimized level is relatively low (Wang et al., 2026a), suffering more from text degradation, which is expected to fit the condition to enable the optional loss better and make the effect clearer. As the result shown in Table 10, with the optional loss enabled, a modality gap climb of 1.13% and a linguistic score improvement of 1.14% together restore 2.27% text degradation compared with the base model. Of course, there is a certain trade-off with the scores on the paralinguistic side. Therefore, its specific usage would be better left as an option depending on real-world practices. Finally, Figure 3 compares the fine-tuning GPU hour costs per epoch of ours against the four baseline reproductions on Qwen2.5-Omni-7B and a single NVIDIA RTX PRO 6000. The result also shows our advantage in efficiency.

F Query Surfaces

Content query. Tier-R presents the audio itself as the question, so there is no query in text adopted. Tier-P asks for a two-way identification between the two canonical transcripts of the unit, such that:

Tier-P content query

Which transcript matches this audio?

A. {transcript of the first content}

B. {transcript of the second content}

Answer with A or B only.

The responses are then mapped to A or B by exact option or normalized transcript matching.

Paralinguistic query. All evaluation paralinguistic queries use the Emotion Recognition surface of MMSU (Chen et al., 2026), such that:

Tier-R & Tier-P paralinguistic query

How does the speaker feel in the recording?

Choose one: {labels}.

Answer with one label only.

During training, the paralinguistic prompt is instead sampled uniformly from a closed pool of four published Dynamic-SUPERB (Huang et al., 2025) stems together with the MMSU stem above, each completed with the same closed answer space and the one-label instruction:

Additional training paralinguistic queries

Determine the emotion in the speech. The answer could be {labels}.

What is the emotion in the audio? The answer could be {labels}.

Can you tell me the emotion of the speaker? The answer could be {labels}.

Identify the emotion conveyed in the utterance. The answer could be {labels}.