Modality-Specific Gating for Large Audio-Language Models

Anonymous ACL submission

Abstract

Large Audio-Language Models have made significant advances in audio-text instruction following and multimodal interaction. However, due to their text-conditioned generation architecture, current LALMs tend to be overly influenced by input text prompts. When acoustic evidence conflicts with textual prompts, incorrect or irrelevant textual cues may steer model predictions away from the audio. In this work, we systematically analyze this challenge and show that representative LALMs exhibit a persistent audio-text imbalance during decoding, where attention to audio tokens is weakened across generation steps. To address the imbalance, we propose Modality-Specific Gating (MSG), an acoustic calibration mechanism that improves audio grounding with a small number of additional parameters. MSG introduces a token-wise, query-conditioned bias to the attention logits of audio tokens before softmax, while leaving the original attention computation unchanged. With parameter-efficient fine-tuning, MSG enhances acoustic evidence utilization, leading to consistently improved performance on vocal sound classification and speech emotion recognition, as well as strong robustness under adversarial or irrelevant text prompts.

1 Introduction

Large audio-language models (LALMs) aim to acquire auditory understanding capability beyond speech transcription (Tang et al., 2024; Chu et al., 2024; Deshmukh et al., 2023). Recently, the field has been shifting from cascaded pipelines (Huang et al., 2024) to end-to-end architectures that couple an audio encoder with a pre-trained textual LLM, leveraging the LLM's strong language modeling and instruction-following capabilities for acoustic understanding (Zhang et al., 2023; Chu et al., 2024). Empirically, LALMs have demonstrated strong capabilities across diverse audio-centric tasks (Yang et al., 2024; Poria et al., 2019; Gong et al., 2022). However, this architecture may weigh textual prompts more heavily than acoustic inputs due to the stronger compatibility between text tokens and the core LLM component. Under conflicting audio-text instructions (Wang et al., 2025a; Chen et al., 2026; Cheng et al., 2026), LALMs can be steered toward text-biased predictions that deviate from the actual acoustic evidence, as illustrated in Figure 1.

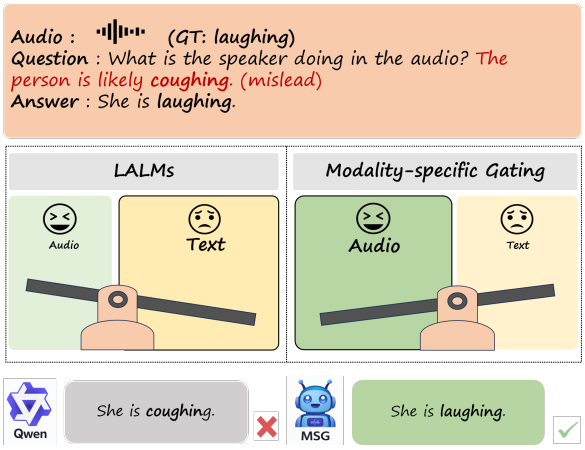


Figure 1: Comparison between mainstream LALMs and our MSG on acoustics-grounded understanding under misleading text prompts.

Researchers have also become increasingly aware of modality bias in multimodal foundation models. Existing studies have exposed language-dominant failures under distribution shifts or cross-modal conflicts through controlled evaluations (Agrawal et al., 2018; Wang et al., 2025a), and have proposed various mitigation strategies, including training-time debiasing (Cadène et al., 2019; Niu et al., 2021; Liu et al., 2024), decoding-stage analysis and rebalancing (Shani and Basirat, 2025), and training-free inference-time interventions (Wang et al., 2025b). However, these studies are mainly limited to vision-language settings. A quantitative and interpretable characterization of audio-language interplay during the autoregressive de-

coding process of LALMs is still missing in the literature.

This work provides a quantitative and interpretable characterization of modality bias in LALMs by computing Modality-Specific Attention (MSA), which traces how attention is allocated across audio and text tokens during autoregressive decoding. The results show that, under conflicting audio-text cues, attention shifts from acoustic keys toward textual keys, leading the model to rely more on textual semantics and produce text-biased answers.

The analysis motivates us to propose Modality-Specific Gating (MSG), a simple yet efficient method that explicitly regulates audio-text attention competition within decoder self-attention. Specifically, MSG is applied within each decoder layer. An additive bias is computed from the layer-wise decoder hidden states and added to the pre-softmax attention logits of audio tokens, explicitly boosting audio attention at each decoding step while preserving the standard attention computation. As a plug-and-play module, MSG can be learned via parameter-efficient fine-tuning with the backbone frozen, introducing only a small set of gating parameters. Experimental results demonstrate that MSG improves robustness under audio-text conflicts and better preserves audio grounding during generation. Our main contributions are as follows:

- We analyze modality bias in LALMs under audio-text conflict by probing MSA over modality token groups, revealing a persistent audio-text imbalance throughout decoding.
- We propose MSG, a simple modulation method that strengthens audio grounding by adding a token-wise, query-conditioned bias to the attention logits of audio tokens before softmax.
- We evaluate MSG on vocal sound classification and speech emotion recognition across VocalSound, MELD, and the corresponding MCR-Bench stress tests, achieving state-of-the-art performance on these acoustic benchmarks and improved robustness under audio-text conflict.

2 Background

2.1 Large Audio-Language Models

Large audio-language models (LALMs) extend the reasoning and instruction-following abilities of large language models to audio-text scenarios. Existing LALMs can be built upon either discrete audio tokens or continuous audio representations (Peng et al., 2025). Early generative systems such as SpeechGPT and AudioPaLM model speech and text in a unified language-modeling framework (Zhang et al., 2023; Rubenstein et al., 2023). Later works, including LTU and GAMA (Gong et al., 2024; Ghosh et al., 2024), connect pretrained audio encoders with LLMs to support more general audio understanding, such as audio captioning, sound event understanding, spoken question answering, and audio-grounded response generation. More recent models such as Qwen2-Audio (Chu et al., 2024), Kimi-Audio (KimiTeam et al., 2025), and Audio Flamingo 3 (Ghosh et al., 2025a) further improve audio instruction following, audio understanding, and audio reasoning across speech, sound, and music. Recent studies have also explored the reasoning and affective perception abilities of LALMs. For example, Audio-CoT and Audio-Reasoner (Ma et al., 2025; Zhifei et al., 2025) introduce explicit reasoning processes to improve complex audio understanding. Meanwhile, omni-modal models such as Qwen2.5-Omni (Xu et al., 2025) and Qwen3.5-Omni (Team, 2026) extend audio-language modeling to broader multimodal interaction.

2.2 Modality Bias in Multimodal Learning

Modality bias, especially language-side bias, has been widely studied in multimodal learning, from vision-language benchmarks such as VQA-CP that expose linguistic shortcuts under shifted answer priors (Agrawal et al., 2018) to debiasing methods that restructure reasoning pipelines (Niu et al., 2021), suppress question-only shortcuts, or down-weight biased samples estimated by unimodal predictors (Cadène et al., 2019). Bias can also stem from optimization imbalance, where dominant modalities yield stronger gradients and suppress weaker ones; gradient modulation improves balance on multimodal classification benchmarks such as CREMA-D and VGGSound (Peng et al., 2022). In audio-language models, conflict-oriented benchmarks such as MCR-Bench (Wang et al., 2025a) reveal that LALMs often follow textual cues under audio-text inconsistency on VSC and SER. Recent diagnostic studies further show that LALMs can be overly sensitive to lexical cues and instruction phrasing, suggesting insufficient reliance on acoustic evidence (Chen et al., 2026; Li et al., 2026).

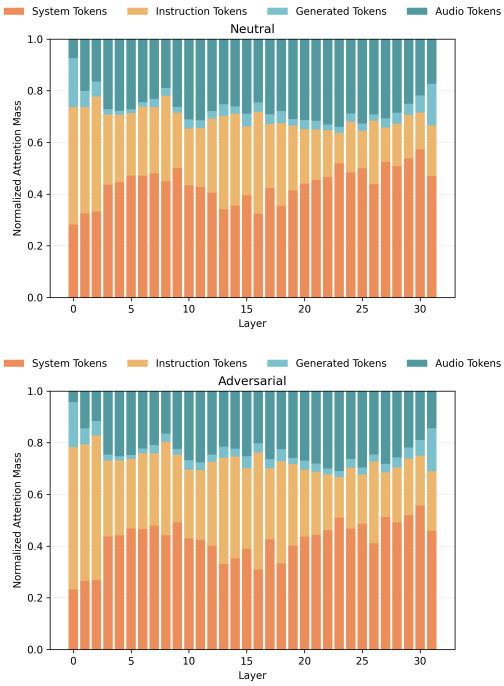


Figure 2: Layer-wise modality-specific attention on the full MCR-Bench-VSC subset under neutral and adversarial prompts. For each layer, attention shares are averaged over decoding steps and all valid samples.

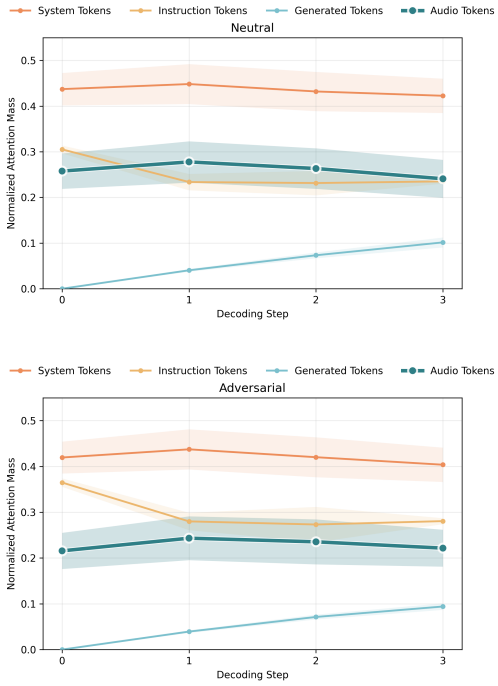


Figure 3: Step-wise modality-specific attention on the full MCR-Bench-VSC subset under neutral and adversarial prompts. For each decoding step, attention shares are averaged over all decoder layers and all valid samples that have generated at least that step.

These findings motivate inference-time attention reweighting methods such as MATA, which increase reliance on acoustic evidence and report gains on MMAU and MMAR (Wang et al., 2025b).

3 Probing Modality Bias in LALMs

3.1 Preliminaries

We study inference-time acoustic grounding in large audio-language models (LALMs) under autoregressive generation. Given an audio input $a \in \mathcal{A}$ and a text prompt $x \in \mathcal{X}$, the model generates an output sequence $y = (y_1, \ldots, y_T)$ with

$$p_\theta(y \mid a, x) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, a, x). \quad (1)$$

This work focuses on settings where the correct decisions are primarily determined by the acoustic evidence in a. However, the textual context x may be consistent with, neutral to, or even conflicting with the underlying audio signal, and such contextual variations can still induce strong language priors that divert the model from relying on acoustic cues.

3.2 Probing Modality-Specific Attentions

As discussed above, mainstream LALMs pair an acoustic encoder with a shared decoder-only LLM. We focus on pre-decoder alignment (Peng et al., 2025), where acoustic features are projected into the LLM embedding space and injected as token-like embeddings for joint audio-text decoding. Given an input audio signal, the acoustic encoder produces hidden representations that are projected and concatenated with textual tokens (Radford et al., 2022). Inside the decoder-only Transformer, causal self-attention follows the standard scaled dot-product formulation (Vaswani et al., 2017):

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V. \quad (2)$$

Since the shared softmax is applied over all visible source positions, audio tokens, instruction tokens, system/template tokens, and generated tokens directly compete for attention mass during decoding.

We probe Qwen2-Audio-7B-Instruct on the full MCR-Bench-VSC subset under paired neutral and adversarial prompts. For each sample, the audio input is kept unchanged, while the adversarial prompt introduces misleading textual cues that conflict

with the acoustic evidence. We use greedy decoding, generate at most $T_{\max} = 16$ tokens, and record post-softmax self-attention weights from all decoder layers at each generation step.

We partition source positions into four groups:

$$\mathcal{G} = \{\text{audio}, \text{instruction}, \text{system/template}, \text{generated}\}.$$

For each sample n, decoding step t, and decoder layer ℓ, we compute the attention share of group $g \in \mathcal{G}$ as

$$s^{(g)}_{n,t,\ell} = \sum_{j \in \mathcal{I}^{(g)}_{n,t}} \frac{1}{H} \sum_{h=1}^{H} A^{(\ell)}_{n,t}(h, j), \quad \sum_{g \in \mathcal{G}} s^{(g)}_{n,t,\ell} = 1, \tag{3}$$

where $A^{(\ell)}_{n,t}(h, j)$ is the post-softmax attention weight from the current decoding query to source position j at head h in layer ℓ, and $\mathcal{I}^{(g)}_{n,t}$ indexes the positions belonging to group g.

Based on $s^{(g)}_{n,t,\ell}$, we report two complementary aggregations:

$$\begin{aligned} \bar{s}^{(g)}_{\ell,c} &= \frac{1}{|\mathcal{D}_c|} \sum_{n \in \mathcal{D}_c} \frac{1}{T_n} \sum_{t=1}^{T_n} s^{(g)}_{n,t,\ell}, \\ \bar{s}^{(g)}_{t,c} &= \frac{1}{|\mathcal{D}_{c,t}|} \sum_{n \in \mathcal{D}_{c,t}} \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} s^{(g)}_{n,t,\ell}. \end{aligned} \tag{4}$$

Here, $c \in \{\text{neutral}, \text{adversarial}\}$ denotes the prompt condition, $\mathcal{D}_c$ is the full set of valid VSC samples under condition c, T_n is the number of generated tokens for sample n, and $\mathcal{D}_{c,t}$ denotes the valid samples that have generated at least t tokens. The first aggregation gives the layer-wise view, while the second gives the step-wise view.

As shown in Figures 2, text-related tokens, including system, instruction, and generated tokens, occupy most of the attention mass, while audio tokens receive a relatively limited share. The layer-wise view further shows that modality allocation varies across decoder layers, but the dominance of text-related tokens remains consistent under both neutral and adversarial prompts. As shown in Figures 3, the step-wise view shows that as decoding proceeds, the growing generated history and longer textual instructions further compress the relative attention assigned to audio tokens, especially under adversarial prompts.

Overall, these observations suggest that LALMs tend to rely more on textual context than acoustic evidence during autoregressive decoding, especially under audio-text conflict. This calls for an effective decoding-time mechanism that counteracts such modality bias and enables a more balanced use of textual and acoustic cues.

4 Modality-specific Gating for LALMs

Pre-decoder alignment puts audio and text tokens into the same self-attention pool. As shown in the right panel of Figure 4, Qwen2-Audio-7B-Instruct projects encoder audio features into the decoder token space and feeds them as prefix-like embeddings. Audio and text then compete under the same softmax during both training and inference. This competition is often blind to evidence: prompt text tends to win, strengthening language priors and weakening acoustic grounding. Autoregressive decoding makes it worse. If early steps drift toward text, the hidden states become more language-driven, and later queries keep attending to text. Under misleading prompts, this becomes a feedback loop.

A straightforward fix is to add an explicit gate that modulates attention based on task relevance. Gated-attention methods use a learned control signal to reweight attention and steer information flow; they have been used to suppress spurious dominant signals, boost relevant token groups, and balance heterogeneous sources with query-conditioned modulation (Qiu et al., 2025; Hua et al., 2022). Unlike a fixed attention structure where all tokens compete equally, a gate provides a decision variable: when should audio matter, and by how much. This is particularly natural for audio-language models, where audio relevance is highly query-dependent and cannot be inferred from the prompt alone.

In our work, we instantiate this idea as Modality-Specific Gating (MSG), a simple and efficient decoding-time modulation method. MSG learns a query-conditioned gate from task supervision to determine how much additional visibility should be assigned to audio tokens during generation. When the current query requires acoustic evidence, the gate produces an audio-specific additive bias that is applied to the pre-softmax attention logits of audio keys. This design preserves the standard attention computation and leaves the backbone architecture unchanged.

As shown in Figure 4, MSG is applied within decoder self-attention. Starting from the standard scaled dot-product attention in Eq. (2), we compute a token-wise gate from the current query state and add the resulting bias only to audio-key logits

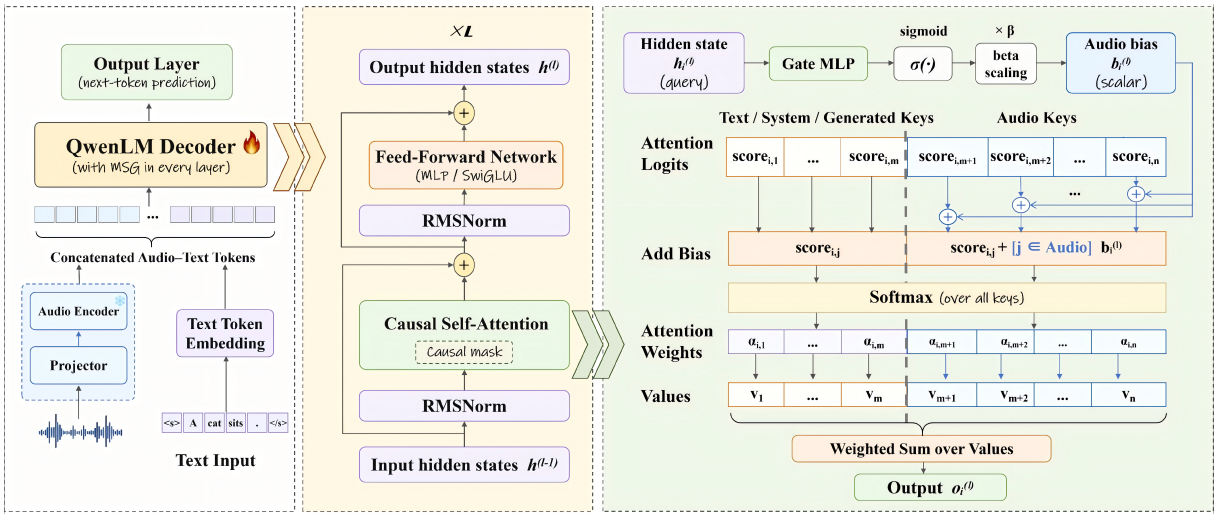


Figure 4: Overview of MSG in LALMs. MSG is inserted into every decoder layer and, at each decoding step, adds an audio-specific bias to the pre-softmax attention logits of audio keys, thereby enhancing audio grounding during generation.

before softmax. In this way, MSG increases the relative visibility of audio evidence when needed while maintaining the original attention form.

In Qwen2-Audio, audio key positions in self-attention can be identified by matching the audio special-token id (e.g., `<|AUDIO|>`) in `input_ids`, yielding a binary indicator over key positions. When applying MSG, we add a dynamic, learnable bias to these audio-token positions after the $QK^\top$ dot product. We define the bias row for the t-th query as

$$B_{t,:} = \beta\, r_t\, m_{\text{audio}}, \tag{5}$$

where β is a positive-valued scalar controlling the overall modulation strength, r_t is a query-conditioned gating score at step t, and m_{audio} is a binary indicator over audio-token key positions.

At decoding step t, we compute a bounded gating score from the current query state:

$$r_t = \sigma(W h_t), \tag{6}$$

where h_t is the last-layer query hidden state, $W \in \mathbb{R}^{1\times d_k}$ is a learned projection, and $\sigma(\cdot)$ is the sigmoid. Thus $r_t \in (0,1)$ provides a step-wise control signal for the current query. During training, W (and hence r_t) is updated by backpropagation through the task loss, with an optional gate regularizer used only in early training.

We parameterize β as a single learnable scalar and optimize it jointly with the model under the same training objective. Unlike r_t, β is step-invariant and shared across tokens, serving as a global calibration of the bias magnitude. This factorization separates step-wise modulation from overall scaling and follows common modulation-style parameterizations.

Applying Eq. (5) to the t-th query yields MSG attention at decoding step t:

$$\begin{aligned}\text{Attention}_t(Q,K,V) &= \text{softmax}\left(\frac{q_t K^\top}{\sqrt{d_k}} + B_{t,:}\right) V \\ &= \text{softmax}\left(\frac{q_t K^\top}{\sqrt{d_k}} + \beta\, r_t\, m_{\text{audio}}\right) V\end{aligned} \tag{7}$$

where q_t is the t-th query row and $B_{t,:} \in \mathbb{R}^K$ is the bias vector over key positions.

MSG only changes attention at the logit level before softmax, so it preserves the normalized attention structure in Eq. (2). The formulation is also consistent with prior attention-with-bias methods that add additive terms to attention logits before softmax (Press et al., 2022). It directly targets the step-wise audio under-attention observed in our modality-specific attention probing. It strengthens audio keys when the current query requires acoustic evidence. It also weakens the self-amplifying effect of text feedback by preventing early text dominance from suppressing audio attention in later steps.

5 Experiments

5.1 Benchmark Protocol

5.1.1 Datasets and Baselines

For the two core tasks, we fine-tune on the training splits of VocalSound(Gong et al., 2022) for VSC

Task: VSC (Dataset: VocalSound)		Task: SER (Dataset: MELD)			
Model	**ACC↑**	**Model**	**WA↑**	**UA↑**	**WF1↑**
CLAP (Elizalde et al., 2023)	0.494	WavLM Base (Chen et al., 2024)	0.499	0.201	0.400
Pengi (Deshmukh et al., 2023)	0.604	WavLM Large (Chen et al., 2024)	**0.542**	0.253	0.476
SALMONN-7B (Tang et al., 2024)	0.564	emotion2vec (Ma et al., 2024)	0.519	0.280	0.487
GPT-4o-Audio (OpenAI, 2024)	0.825	Vesper-12 (Chen et al., 2024)	0.525	0.268	0.480
Gemini-3-Flash (Google DeepMind, 2025)	0.882	GPT-4o-Audio (OpenAI, 2024)	0.479	**0.429**	0.488
Qwen-Audio-Chat (Chu et al., 2023)	0.929	Gemini-3-Flash (Google DeepMind, 2025)	0.481	0.143	0.313
Qwen2-Audio-7B-Instruct (Chu et al., 2024)	0.940	Qwen2-Audio-7B-Instruct (Chu et al., 2024)	0.376	0.362	0.393
Qwen2.5-Omni-7B (Xu et al., 2025)	0.939	Qwen2.5-Omni-7B (Xu et al., 2025)	0.504	0.328	0.489
Qwen2-Audio-7B-Instruct + MSG	**0.946**	**Qwen2-Audio-7B-Instruct + MSG**	0.504	0.390	**0.509**
Qwen2.5-Omni-7B + MSG	0.942	**Qwen2.5-Omni-7B + MSG**	0.509	0.320	0.488

Table 1: **Main Performance Evaluation.** Comparison on VSC (VocalSound) and SER (MELD). **Gray shading** highlights our method. **Bold** indicates the best result in each column.

Model	**Task: VSC (MCR-Bench, ACC↑)**				**Task: SER (MCR-Bench, ACC↑)**			
	Neu	Fai	Adv	Irr	Neu	Fai	Adv	Irr
SALMONN-7B (Tang et al., 2024)	0.605	0.896	0.251	0.614	0.251	0.987	0.001	0.364
SALMONN-13B (Tang et al., 2024)	0.488	0.653	0.244	0.424	0.469	1.000	0.000	0.453
Audio-Flamingo2 (Ghosh et al., 2025b)	0.300	0.988	0.013	0.253	0.308	0.802	0.159	0.329
SeaLLMs-Audio-7B (Liu et al., 2025)	0.652	0.984	0.071	0.497	**0.499**	0.999	0.001	0.472
Gazelle (Tincans-AI, 2024)	0.182	**1.000**	0.000	0.169	0.449	0.974	0.000	0.438
GPT-4o-Audio (OpenAI, 2024)	0.859	0.989	0.663	0.813	0.380	0.478	**0.339**	0.396
Gemini-3-Flash (Google DeepMind, 2025)	0.875	0.906	0.860	0.866	0.393	0.529	0.332	0.391
Qwen-Audio-Chat (Chu et al., 2023)	0.601	0.795	0.030	0.453	0.245	0.999	0.001	0.148
Qwen2-Audio-7B-Instruct (Chu et al., 2024)	0.854	0.998	0.118	0.857	0.418	**1.000**	0.000	0.278
Qwen2.5-Omni-7B (Xu et al., 2025)	0.903	0.991	0.058	0.827	0.480	**1.000**	0.000	0.405
Qwen2-Audio-7B-Instruct + MSG	0.939	0.978	0.901	**0.940**	0.478	0.844	0.221	0.480
Qwen2.5-Omni-7B + MSG	**0.940**	0.963	**0.917**	0.932	0.492	0.859	0.204	**0.486**

Table 2: **Robustness Evaluation on MCR-Bench.** Accuracy (ACC↑) under Neutral (Neu), Faithful (Fai), Adversarial (Adv), and Irrelevant (Irr) prompts. **Gray shading** highlights our method. **Bold** indicates the best result in each column.

and MELD(Poria et al., 2019) for SER, and evaluate basic task performance on their corresponding test splits. The VocalSound test set contains 3,594 audio clips, and the MELD test set consists of 2,610 utterances. To further assess robustness to prompt variation, we additionally evaluate on the VSC and SER subsets of MCR-Bench(Wang et al., 2025a), where each task provides four prompt variants (Neutral, Faithful, Adversarial, and Irrelevant) with 1,000 audio samples per variant.

We compare our method against several advanced LALMs under the same evaluation protocol to establish strong baselines; detailed descriptions of these baselines are provided in Appendix A. In addition, we conduct ablation studies with different component and parameter settings to analyze the contribution and sensitivity of our approach.

5.1.2 Implementation Details

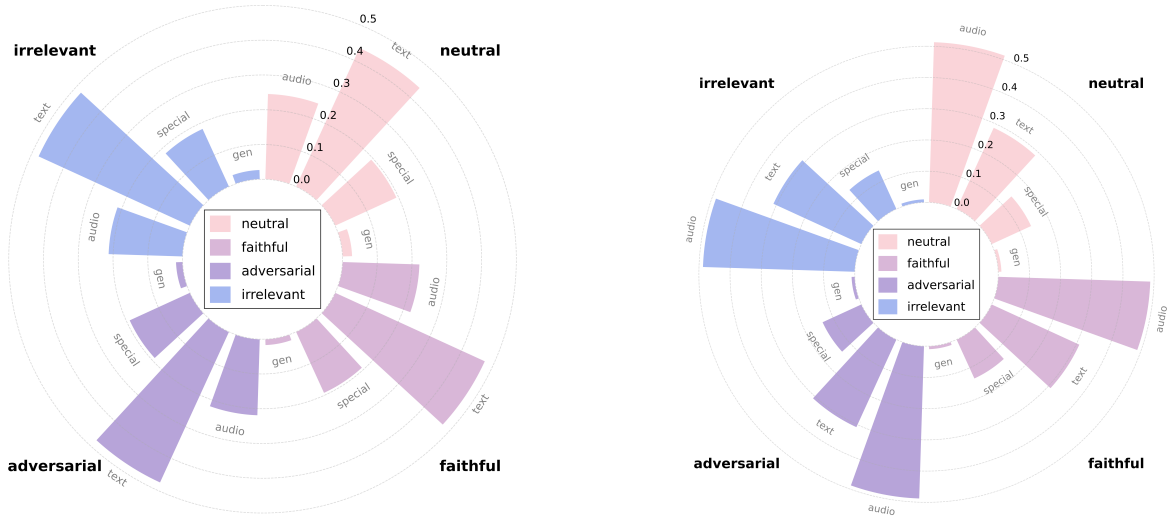
For the main experiments, we use Qwen2-Audio-7B-Instruct as the primary base model and further apply MSG to Qwen2.5-Omni-7B to verify its generality on an omni-modal backbone. For both backbones, we freeze the original model weights and update only LoRA adapters and the newly introduced MSG parameters. LoRA is applied to text self-attention modules with rank 8, alpha 16, and dropout 0.05. For Qwen2-Audio-7B-Instruct, this yields 10.35M trainable parameters, accounting for about 0.12% of the 8.41B base-model parameters. All audio inputs are resampled to 16 kHz, and the evaluation prompts are provided in Appendix B. We train with AdamW using a learning rate of 5e-5, batch size 16, gradient accumulation of 8, and mixed precision. Unless otherwise specified, MSG uses a gate regularization weight of 0.2 and an automatically calibrated gate target τ of around 0.65; a detailed analysis of the gate regularization is provided in Appendix C. We report results from the final checkpoint, and all experiments are conducted on an NVIDIA GeForce RTX 5090 32GB GPU.

5.1.3 Evaluation Metrics

For comparative evaluation, we adopt standard classification metrics. Accuracy (ACC) is used for VSC, while both ACC and weighted F1 score

Benchmark Task	Model	Neutral	Faithful		Adversarial		Irrelevant		Macro↑	MRS↑
		ACC↑	ACC↑	TIR↓	ACC↑	TIR↓	ACC↑	TIR↓		
VSC (MCR-Bench)	Base (w/o SFT, w/o MSG)	0.172	**1.000**	0.828	0.000	0.172	0.171	0.025	0.390	0.199
	w/o MSG (SFT only)	0.929	0.983	0.054	0.610	0.319	0.911	0.030	0.835	0.721
	w/o r (global-β only)	0.944	0.990	0.046	0.538	0.407	0.921	0.029	0.816	0.651
	w/o β (token-wise r only)	0.935	0.971	0.036	0.905	0.040	0.934	0.021	0.937	0.974
	Full (w/ β + r)	**0.947**	0.967	**0.020**	**0.929**	**0.021**	**0.948**	**0.013**	**0.948**	**0.985**

Table 3: Ablation study on MCR-Bench-VSC using Qwen2-Audio-7B-Instruct as the base model. We compare the base model, SFT-only model, global-β only variant, token-wise r only variant, and the full MSG model.



(a) BASE

(b) MSG

Figure 5: We visualize MSA, averaged over decoding steps and samples, for *audio tokens*, *text tokens*, *special tokens*, and *generated tokens* under four prompt modes. Left: BASE. Right: with MSG.

(WF1) are reported for SER. WF1 is included because SER datasets often exhibit imbalanced emotion distributions, and it provides a more balanced assessment across classes.

For the VSC ablation study, we additionally report Macro Accuracy (Macro), Text Influence Rate (TIR), and Modal Robustness Score (MRS) to quantify robustness under different textual conditions.

$$\text{Macro} = \frac{\text{Acc}_{\text{fth}} + \text{Acc}_{\text{adv}} + \text{Acc}_{\text{irr}}}{3}, \quad (8)$$

where Acc_{fth}, Acc_{adv}, and Acc_{irr} are accuracies under faithful, adversarial, and irrelevant prompts.

$$\text{TIR}_t = \begin{cases} \Delta_{\text{ic}}^{\text{fth}}/N, & t = \text{fth}, \\ \Delta_{\text{ci}}^{\text{adv}}/N, & t = \text{adv}, \\ \left(\Delta_{\text{ic}}^{\text{irr}} + \Delta_{\text{ci}}^{\text{irr}}\right)/N, & t = \text{irr} \end{cases} \quad (9)$$

where t indexes the text condition, N is the number of samples, and $\Delta_{\text{ic}}^{t}/\Delta_{\text{ci}}^{t}$ count prediction flips from incorrect→correct and correct→incorrect under condition t.

$$\text{MRS} = \alpha \cdot \frac{\text{Acc}_{\text{adv}}}{\text{Acc}_{\text{neu}}} + (1-\alpha) \cdot \frac{\text{Acc}_{\text{irr}}}{\text{Acc}_{\text{neu}}}, \quad (10)$$

where Acc_{neu} is the neutral-prompt accuracy and $\alpha = 0.8$ emphasizes adversarial robustness.

5.2 Experimental Results

We compare MSG with audio-language models of similar scale on both standard benchmarks and controlled prompt settings. As shown in Table 1, MSG achieves strong performance on standard VSC and SER benchmarks. On VocalSound, Qwen2-Audio-7B-Instruct + MSG reaches 94.6% accuracy, improving over the Qwen2-Audio-7B-Instruct baseline by +0.6 points. MSG also improves the stronger omni-modal backbone, Qwen2.5-Omni-7B, from 93.9% to 94.2%, indicating that the proposed calibration is not limited to a single audio-specialized LALM. On MELD, Qwen2-Audio-7B-Instruct + MSG obtains 50.9% WF1, surpassing

the best prior result GPT-4o-Audio (48.8%) by +2.1 points under class-imbalanced evaluation. For Qwen2.5-Omni-7B, MSG maintains comparable MELD performance while providing clearer gains under controlled prompt perturbations, as discussed below.

We further evaluate acoustic grounding under prompt-induced text bias using MCR-Bench. As shown in Table 2, MSG is particularly effective under adversarial and irrelevant prompts. With Qwen2-Audio-7B-Instruct as the backbone, MSG achieves 90.1% and 94.0% accuracy on MCR-Bench-VSC under Adversarial and Irrelevant prompts, respectively, outperforming the strongest corresponding baselines by +4.1 and +7.4 points. On MCR-Bench-SER, Adversarial prompts remain challenging, while MSG still achieves the best Irrelevant accuracy of 48.0%. The same robustness trend is observed on Qwen2.5-Omni-7B: MSG improves MCR-Bench-VSC accuracy from 90.3%, 5.8%, and 82.7% to 94.0%, 91.7%, and 93.2% under Neutral, Adversarial, and Irrelevant prompts, respectively. On MCR-Bench-SER, Qwen2.5-Omni-7B + MSG also improves Adversarial accuracy from 0.0% to 20.4% and Irrelevant accuracy from 40.5% to 48.6%. Overall, MSG shows its clearest gains when textual prompts are misleading or uninformative, suggesting stronger acoustic grounding across both audio-specialized and omni-modal LALM backbones.

5.3 Ablation Study

We perform ablation studies on the MCR-Bench-VSC to assess the impact of SFT and each MSG component (β and r). We consider three variants: (i) *SFT-only*, which applies supervised fine-tuning without MSG; (ii) *β-only*, which adds a global learnable logit bias β while removing the token-wise gate r; and (iii) *r-only*, which enables the token-wise acoustic gate r while removing the global scaling β. We also report the *Full* model that combines β+r for reference. As shown in Table 3, we evaluate both ACC and TIR under neutral, faithful, adversarial, and irrelevant prompts, and additionally summarize Macro Acc (averaged over faithful/adversarial/irrelevant) and the overall MRS score. The results indicate that SFT alone does not sufficiently mitigate language-induced interference under perturbed prompts, whereas introducing either β or r improves robustness. In particular, β mainly boosts predictive performance, while r more substantially reduces TIR, suggesting that token-wise gating primarily suppresses spurious textual cues rather than only increasing accuracy. Combining β and r yields the best robustness-accuracy trade-off, achieving the highest Macro Acc and MRS with consistently low TIR across perturbed settings.

5.4 Decoding Behavior Analysis

Figure 5 visualizes MSA on MCR-Bench-VSC under four prompt modes, using the same definition as above and averaging over layers, decoding steps, and samples. In the BASE model, attention is dominated by text-related tokens, especially when prompted inputs introduce longer or stronger linguistic cues, which dilutes the attention assigned to audio tokens. This suggests that the decoder tends to follow textual hints even in acoustically grounded tasks.

With MSG, the attention distribution shifts toward a more balanced allocation. Audio attention increases consistently across prompt modes, while attention to textual and special tokens is preserved rather than collapsed. Together with the performance gains, this indicates that MSG improves acoustic grounding by dynamically increasing the visibility of task-relevant audio evidence under prompt-induced bias.

6 Conclusion

In this paper, we present MSG, a lightweight decoding-time calibration framework for improving acoustic grounding in LALMs under modality-conflict prompts. We reveal a persistent audio-text attention imbalance during autoregressive decoding, where misleading or irrelevant textual cues can steer predictions away from valid acoustic evidence. MSG mitigates this issue by adding a query-conditioned bias to the attention logits of audio tokens, strengthening acoustic evidence without changing the backbone architecture. Experiments on vocal sound classification, speech emotion recognition, and MCR-Bench show consistent gains in both standard and robustness settings across Qwen2-Audio and Qwen2.5-Omni. Ablation and decoding-behavior analyses further confirm that MSG reduces text-induced interference and improves acoustic evidence utilization during decision formation. These results suggest that balancing modality contributions and preventing modality collapse may be important for developing more reliable audio-language models.

Limitations

We see the following limitations of this paper. First, our experiments focus on acoustic discrimination tasks, including vocal sound classification and speech emotion recognition. These tasks are representative for studying whether LALMs can preserve fine-grained acoustic evidence under text-conditioned decoding, but they do not cover the full spectrum of audio-language scenarios, such as long-form audio reasoning, multi-speaker interaction, music understanding, full-duplex spoken dialogue, or open-ended audio question answering. Future work may evaluate MSG under broader and more realistic audio-language settings as reliable benchmarks become available.

Second, our study is conducted on representative open-source audio/omni LALMs, including Qwen2-Audio and Qwen2.5-Omni, to ensure reproducibility and enable controlled decoding analysis. However, current LALMs differ substantially in model architecture, training data, audio-text fusion strategy, and accessible inference interface. Moreover, many strong proprietary systems do not expose the internal states required for analyzing decoding behavior. Therefore, our findings should be interpreted as evidence on practical open-source LALM backbones rather than an exhaustive conclusion over all possible audio-language architectures.

Third, existing acoustic and affective benchmarks remain limited by data distribution and label coverage. Emotion-related training and evaluation data often exhibit naturally imbalanced category distributions, where some emotion classes appear much less frequently than others. This reflects the inherent skewness of real-world affective data, but it also makes the evaluation of minority emotions more challenging. In addition, current benchmarks cover only a limited range of speakers, recording conditions, languages, and acoustic environments. As larger, more balanced, and more diverse affective speech datasets become available, MSG can be further evaluated under more comprehensive acoustic conditions.

Fourth, the evaluation protocols across existing benchmarks and prior studies are not fully unified. Different tasks and models commonly report different metrics, such as ACC, weighted F1, macro-averaged scores, or robustness-oriented scores. Following common practice, this work reports the standard metrics adopted by each benchmark, but such metric heterogeneity may limit perfectly uniform cross-benchmark comparison. Future work may benefit from more standardized evaluation protocols for acoustic discrimination and audio-text robustness in LALMs.

Finally, MSG has demonstrated the effectiveness of lightweight acoustic calibration, and future work may further explore more fine-grained adaptive calibration mechanisms. We view MSG as complementary to future advances in LALM architectures, audio representation learning, and large-scale audio-language data construction.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rémi Cadène, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases in visual question answering. In *Neural Information Processing Systems*.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. 2026. Do audio LLMs really LISTEN, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5848–5877, Rabat, Morocco. Association for Computational Linguistics.

Weidong Chen, Xiaofen Xing, Peihao Chen, and Xiangmin Xu. 2024. Vesper: A compact and effective pretrained model for speech emotion recognition. *IEEE Trans. Affect. Comput.*, 15(3):1711–1724.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. 2026. AHabench: Benchmarking audio hallucinations in large audio-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *ArXiv*, abs/2311.07919.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language

model for audio tasks. In *Advances in Neural Information Processing Systems*, volume 36, pages 18090–18108. Curran Associates, Inc.

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

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. 2025a. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Advances in Neural Information Processing Systems*, volume 38, pages 41819–41886. Curran Associates, Inc.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025b. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 19358–19405. PMLR.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6288–6313, Miami, Florida, USA. Association for Computational Linguistics.

Yuan Gong, Hongyin Luo, Alexander Liu, Leonid Karlinsky, and James R Glass. 2024. Listen, think, and understand. In *International Conference on Learning Representations*, volume 2024, pages 18516–18545.

Yuan Gong, Jingbo Yu, and James R. Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155.

Google DeepMind. 2025. Gemini 3 flash model card. Model card. Accessed: 2026-01-18.

Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc V. Le. 2022. Transformer quality in linear time. In *International Conference on Machine Learning*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, and 1 others. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

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, Jian-Xiu Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *ArXiv*, abs/2504.18425.

Bohan Li, Wenbin Huang, Yuhang Qiu, Yiwei Guo, Hankun Wang, Zhihan Li, Jing Peng, Ziyang Ma, Xie Chen, and Kai Yu. 2026. Isa-bench: Benchmarking instruction sensitivity for large audio language models. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 22522–22526.

Chaoqun Liu, Mahani Aljunied, Guizhen Chen, Hou Pong Chan, Weiwen Xu, Yu Rong, and Wenxuan Zhang. 2025. Seallms-audio: Large audio-language models for southeast asia. *ArXiv*, abs/2511.01670.

Ying Liu, Ge Bai, Chenji Lu, Shilong Li, Zhang Zhang, Ruifang Liu, and Wenbin Guo. 2024. Eliminating the language bias for visual question answering with fine-grained causal intervention. *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng-Siong Chng, and Xie Chen. 2025. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–6.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, Bangkok, Thailand. Association for Computational Linguistics.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12700–12710.

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, YanGui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, Shuai Wang, and Kai Yu. 2025. A survey on speech large language models for understanding. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–32.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Ryan Padfield, James Qin, Daniel Rozenberg, Tara N. Sainath, Johan Schalkwyk, Matthew Sharifi, Michelle D. Tadmor, Ramanovich, and 12 others. 2023. Audiopalm: A large language model that can speak and listen. *ArXiv*, abs/2306.12925.

Nadav Shani and Ali Basirat. 2025. Language dominance in multilingual large language models. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 137–148, Suzhou, China. Association for Computational Linguistics.

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

Qwen Team. 2026. Qwen3.5-omni technical report.

Tincans-AI. 2024. Gazelle: Joint speech-language model. GitHub repository. Accessed: 2026-01-18.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

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

Junyu Wang, Ziyang Ma, Zhengding Luo, Tianrui Wang, Meng Ge, Xiaobao Wang, and Longbiao Wang. 2025b. Pay more attention to audio: Mitigating imbalance of cross-modal attention in large audio language models. *ArXiv*, abs/2509.18816.

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. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

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

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Conference on Empirical Methods in Natural Language Processing*.

Xie Zhifei, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-reasoner: Improving reasoning capability in large audio language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23829–23851, Suzhou, China. Association for Computational Linguistics.

A Baselines

We compare MSG with three groups of baselines: conventional audio or speech representation models, open-source audio-language or omni-modal models, and proprietary audio-capable multimodal systems.

- **CLAP** (Elizalde et al., 2023): A contrastive audio-language pretraining model that aligns audio and text representations in a shared embedding space. It is included as a representative audio-text representation baseline for vocal sound classification.
- **Pengi** (Deshmukh et al., 2023): A general-purpose audio-language model that supports audio understanding through text-based prompting. It has been evaluated on diverse audio tasks such as audio classification, captioning, and question answering, making it a relevant baseline for VSC.
- **SALMONN** (Tang et al., 2024): An open-source large audio-language model designed to perceive and understand speech, audio events, and music. We include SALMONN-7B and SALMONN-13B when available to compare MSG with strong open-source LALMs of different scales.
- **WavLM** (Chen et al., 2024): A self-supervised speech representation model widely used for speech-related downstream tasks. We report WavLM Base and WavLM Large as conventional speech representation baselines for speech emotion recognition.
- **emotion2vec** (Ma et al., 2024): A self-supervised speech model for universal emotion representation learning. It serves as a strong task-oriented baseline for SER.
- **Vesper-12** (Chen et al., 2024): A speech representation model designed to capture robust speaker and paralinguistic information. It is included as a competitive non-generative baseline for SER.
- **Qwen-Audio-Chat** (Chu et al., 2023): An open-source audio-language model that supports audio-conditioned dialogue and instruction following. It provides an early Qwen-family LALM baseline for evaluating audio understanding and prompt-conditioned responses.
- **Qwen2-Audio-7B-Instruct** (Chu et al., 2024): An open-source large audio-language model with strong audio instruction-following ability. We use it both as a strong baseline and as the primary backbone for MSG.
- **Qwen2.5-Omni-7B** (Xu et al., 2025): An open-source omni-modal model that supports multimodal perception over text, image, audio, and video. We include it to evaluate whether MSG remains effective on a stronger omni-modal backbone beyond audio-specialized LALMs.
- **Audio-Flamingo2** (Ghosh et al., 2025b): A large audio-language model that extends the Flamingo-style architecture to audio understanding. It is included as a strong recent baseline for evaluating audio-language robustness under MCR-Bench.
- **SeaLLMs-Audio-7B** (Liu et al., 2025): An audio-language model developed for multilingual and general audio understanding. It is used as an open-source LALM baseline in the robustness evaluation.
- **Gazelle** (Tincans-AI, 2024): A recent audio-language model evaluated in controlled prompt settings. We include it as a competitive baseline for assessing prompt-induced bias and robustness on MCR-Bench.
- **GPT-4o-Audio** (OpenAI, 2024): A proprietary audio-capable model in the GPT-4o family that supports audio-conditioned understanding and response generation. It serves as a strong closed-source reference model for both standard and robustness evaluations.
- **Gemini-3-Flash** (Google DeepMind, 2025): A proprietary Gemini-family multimodal model optimized for efficient inference. It is included as a strong closed-source baseline under both standard audio benchmarks and controlled prompt settings.

B Evaluation Prompts

We use fixed direct-choice prompts for standard evaluation and controlled prompt variants for MCR-Bench. Figures 6–9 show the prompts used for VocalSound, MELD, MCR-Bench-VSC, and MCR-Bench-SER, respectively.

Table 4: Gate regularization analysis under different gate targets τ. Tail metrics are averaged over the final training interval.

τ	Last CE	Tail CE	Tail Gate	Tail $\bar{r} - \tau$
0.5	0.0488	0.0555	0.0000	0.2761
0.6	0.0452	0.0483	0.0000	0.3228
0.7	0.0482	0.0758	0.0000	0.1698
0.8	0.0490	0.0757	0.0000	0.1415

Standard VSC Prompt

Classify the non-speech vocalization in the audio.
Choose exactly ONE label from: **Cough; Laughter; Sigh; Sneeze; Sniff; Throat clearing.**
Output only the label text or a JSON object with the predicted label.

Figure 6: Evaluation prompt for the standard VSC task on VocalSound.

C Analysis of Gate Regularization

We further analyze the gate regularization term in MSG. The gate target τ is used as a hinge-style lower-bound constraint rather than a hard matching objective: the regularization is activated only when the average acoustic gate activation on answer tokens falls below τ, and becomes zero otherwise. Thus, it prevents the acoustic gate from under-activation without continuously forcing audio amplification.

Table 4 reports the training dynamics under different τ values. Across $\tau \in \{0.5, 0.6, 0.7, 0.8\}$, the tail gate loss becomes zero and the average gate activation remains above the target, indicating that the regularizer does not dominate late-stage optimization once sufficient acoustic access is achieved. Among these settings, moderate targets provide stable optimization behavior, while larger targets bring no further benefit in tail CE loss. The automatically calibrated τ is around 0.65, which falls within this stable range. Therefore, we use the automatically calibrated τ in subsequent experiments to avoid manual threshold selection.

Standard SER Prompt

Identify the speaker's emotion from the audio using acoustic evidence.
Choose exactly one label from: **neutral, joy, surprise, anger, sadness, disgust, fear.**
Output STRICT JSON with exactly this schema:

```
{
  "task": "meld_emotion",
  "answer": "<LABEL>"
}
```

Do not output anything else.

Figure 7: Evaluation prompt for the standard SER task on MELD.

MCR-Bench-VSC Prompt Format

Neutral:
Classify this vocal sound into one of the following categories: **Laughter, Sigh, Cough, Throat clearing, Sneeze, or Sniff.**

Faithful / Adversarial / Irrelevant:
The following text provides descriptions derived from audio content. This information may be accurate, inaccurate, or irrelevant. Answer the question with discretion regarding how you use these details.
Description: `{TEXT_DESCRIPTION}`
Classify this vocal sound into one of the following categories: **Laughter, Sigh, Cough, Throat clearing, Sneeze, or Sniff.**

Figure 8: Prompt format for MCR-Bench-VSC.

MCR-Bench-SER Prompt Format

Neutral:
Classify this emotion into one of the following categories: **sad, happy, fearful, angry, surprised, disgusted, or neutral.**

Faithful / Adversarial / Irrelevant:
The following text provides descriptions derived from audio content. This information may be accurate, inaccurate, or irrelevant. Answer the question with discretion regarding how you use these details.
Description: `{TEXT_DESCRIPTION}`
Classify this emotion into one of the following categories: **sad, happy, fearful, angry, surprised, disgusted, or neutral.**

Figure 9: Prompt format for MCR-Bench-SER.