AHA: Aligning Large Audio-Language Models for Reasoning Hallucinations via Counterfactual Hard Negatives

Anonymous ACL submission

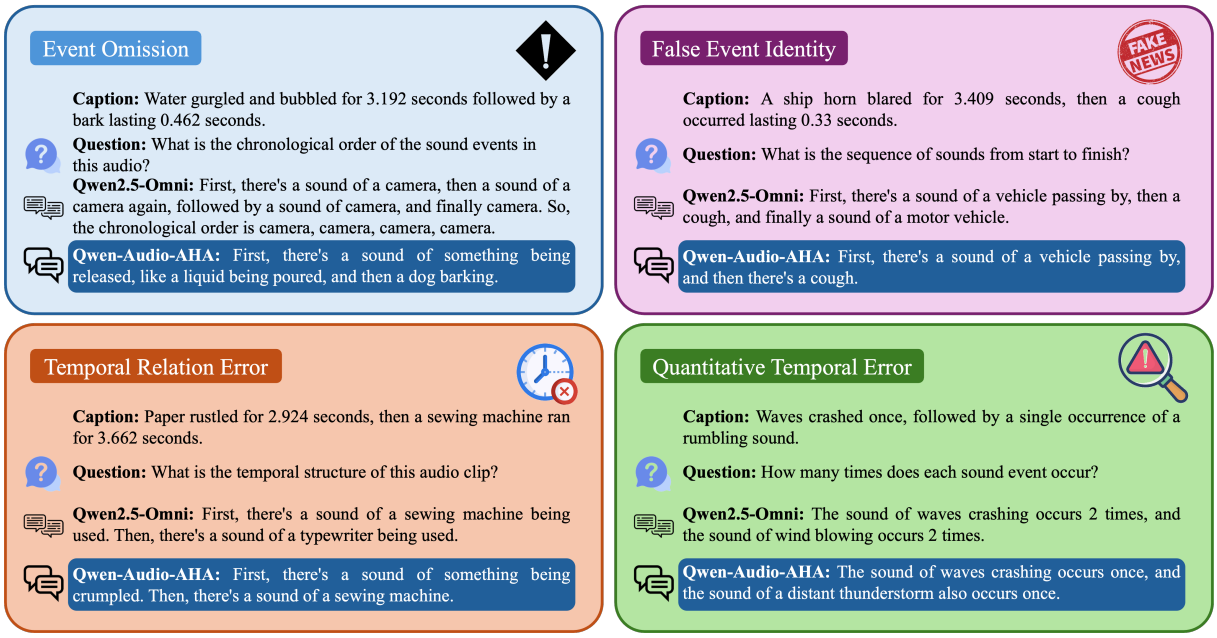

Figure 1: Illustrative examples showcasing the improvements after performing the proposed **AHA framework alignment**. Compared to the base model **Qwen2.5-Omni**, our **Qwen-Audio-AHA** effectively mitigates hallucinations and errors across four critical dimensions: (1) Event Omission, (2) False Event Identity, (3) Temporal Relation Error, and (4) Quantitative Temporal Error.

Abstract

Although Large Audio-Language Models (LALMs) deliver state-of-the-art (SOTA) performance, they frequently suffer from hallucinations, *e.g.,* generating text not grounded in the audio input. We analyze these grounding failures and identify a distinct taxonomy: **Event Omission**, **False Event Identity**, **Temporal Relation Error**, and **Quantitative Temporal Error**. To address this, we introduce the **AHA** (**A**udio **H**allucination **A**lignment) framework. By leveraging **counterfactual hard negative mining**, our pipeline constructs a high-quality preference dataset that forces models to distinguish strict acoustic evidence from linguistically plausible fabrications. Additionally, we establish **AHA-Eval**, a diagnostic benchmark designed to rigorously test these fine-grained reasoning capabilities. We apply this data to align Qwen2.5-Omni. The resulting model, **Qwen-Audio-AHA**, achieves a 13.7% improvement on AHA-Eval. Crucially, this benefit generalizes beyond our diagnostic set. Our model shows substantial gains on public benchmarks, including 1.3% on MMAU-Test and 1.6% on MMAR, outperforming latest SOTA methods.

1 Introduction

The development of Large Audio-Language Models (LALMs) has rapidly bridged the gap between acoustic perception and semantic reasoning. Foundational works like Whisper (Radford et al., 2023) established robust speech representations, paving the way for general audio understanding models such as Qwen-Audio (Chu et al., 2023) and

SALMONN (Tang et al., 2023), which integrate non-speech environmental cues through advanced instruction tuning. More recently, the field has shifted toward **omni-modal integration**, exemplified by GPT-4o (Hurst et al., 2024) and Qwen2.5-Omni (Xu et al., 2025). These models utilize simultaneous speech-text architectures to enable low-latency interaction. However, despite these architectural breakthroughs, relying solely on Supervised Fine-Tuning (SFT) leaves these models vulnerable to hallucinations (Ji et al., 2023) and misalignment with human preferences (Ouyang et al., 2022), often resulting in outputs that are linguistically fluent but factually detached from the acoustic reality. Audio hallucinations present distinct challenges because sound is inherently temporal. Unlike images, acoustic events unfold over time and often overlap in complex patterns. This requires LALMs to track temporal changes with high precision. While previous studies have analyzed hallucinations, they mostly limit their scope to specific subtasks. For example, ASR research highlights silence-induced loops (Koenecke et al., 2024; Radford et al., 2023), and audio captioning studies investigate fabricated objects (Mei et al., 2024). However, the field lacks a systematic taxonomy for the complex fine-grained reasoning failures found in general-purpose LALMs. When acoustic scenes become difficult to parse, these models often default to language priors. This reliance leads to critical inaccuracies in how they ground temporal information.

To address this, we deconstruct these grounding failures into a distinct taxonomy with four dimensions: **Event Omission** refers to cases where the model ignores perceptible events. **False Event Identity** occurs when the model fabricates or mislabels sounds. **Temporal Relation Error** involves distortions in chronological order or causality, such as reversing the sequence of two events. **Quantitative Temporal Error** captures incorrect event counts or duration estimates. This shift, from generic *hallucination* to specific breakdowns in temporal perception and reasoning, provides the rigorous foundation necessary for our targeted alignment strategy. Crucially, the SFT struggles to fix these errors as it lacks the necessary negative constraints (Ji et al., 2023), which potentially encourages models to prioritize linguistic plausibility over acoustic fidelity, leading to what we term *hallucinations of granularity*, where models fabricate precise details to sound coherent. We argue that effective alignment requires more than just positive examples; it demands *counterfactual hard negatives*. The model must learn to distinguish the true acoustic timeline from linguistically plausible but chronologically manipulated alternatives (e.g., reversed causal events), thereby forcing it to abandon language priors in favor of strict acoustic grounding.

To bridge this gap, we introduce the **AHA** (**A**udio **H**allucination **A**lignment) framework. This unified pipeline constructs a shared audio-question pool to generate two distinct resources. First, we synthesize a post-training alignment dataset by contrasting human-verified ground truths against "hard negative" hallucinations. These negatives specifically target the error types in our taxonomy. Second, we establish AHA-Eval, a dedicated benchmark designed to rigorously assess these reasoning capabilities. Training on our alignment data forces the model to distinguish factual acoustic evidence from plausible fabrications. Finally, we apply this framework to the Qwen2.5-Omni-7B base model by using Direct Preference Optimization (DPO), resulting in aligned model **Qwen-Audio-AHA**. Experimental results demonstrate that our framework significantly reduces error rates across all four hallucination dimensions on our diagnostic benchmark, as shown in Figure 1. Crucially, this improved grounding generalizes beyond the diagnostic set. **Qwen-Audio-AHA** achieves consistent accuracy gains on public benchmarks, including MMAU-Test and MMAR, demonstrating that mitigating specific hallucination behaviors contributes to broader improvements in multimodal reasoning.

In summary, our contributions are fivefold: **(1)** We propose a taxonomy that divides hallucinations into four dimensions, focusing on deficits in temporal logic rather than general errors. **(2)** We introduce a diagnostic benchmark based on this taxonomy to quantify hallucinations in LALMs. **(3)** We build a preference dataset serves as a model-agnostic resource for post-training alignment and the generation pipeline also can extends to other hallucinations in LALMs. **(4)** Empirical evaluations validate our framework, showing that our aligned **Qwen-Audio-AHA** significantly reduces hallucinations on diagnostic tasks with *zero alignment tax*, **even boosting** performance on public benchmarks. **(5)** Our work offers a novel perspective on LALMs hallucinations, encouraging the community to shift focus to fine-grained reasoning.

2 Related Work

2.1 Large Audio Language Models (LALMs)

Recent research has successfully extended Large Language Models (LLMs) into the audio domain. Whisper (Radford et al., 2023) serves as a foundational component, providing robust representations for many subsequent systems. Early models like AudioLM (Borsos et al., 2023), SpeechGPT (Zhang et al., 2023), and AudioPaLM (Rubenstein et al., 2023) treated audio as discrete tokens or utilized joint vocabularies for unified sequence modeling. Later work expanded this scope to general audio understanding. For instance, SALMONN (Tang et al., 2023) combined encoders to capture environmental cues, while Qwen-Audio (Chu et al., 2023) scaled up analysis through multi-task instruction tuning. Additionally, WavLLM (Hu et al., 2024) introduced dual-encoder architectures to improve context robustness. Current trends prioritize **omni-modal integration**, where models learn directly from mixed data streams (Hurst et al., 2024; Défossez et al., 2024; Xie and Wu, 2024). Qwen2.5-Omni (Xu et al., 2025) exemplifies this shift, utilizing a "Thinker-Talker" architecture for low-latency interaction. However, relying solely on Supervised Fine-Tuning (SFT) leaves these models prone to hallucinations (Ji et al., 2023) and misalignment with human preferences (Ouyang et al., 2022). To address this, we introduce AHA, a unified framework designed for post-training alignment that seamlessly integrates with existing LALMs. It mitigates temporal hallucinations through a targeted training dataset and provides a diagnostic benchmark to rigorously evaluate fine-grained reasoning capabilities.

2.2 Hallucination in LALMs

LALMs have made strides but remain prone to hallucinations ungrounded in acoustic signals. This issue is uniquely challenging in the audio domain because sound events are temporally dynamic and often overlapping. Previous research has primarily addressed isolated subtasks. For instance, ASR studies characterize silence-induced loops (Radford et al., 2023; Koenecke et al., 2024; Guerreiro et al., 2023; Frieske and Shi, 2024), while audio captioning work investigates object fabrication (Mei et al., 2024; Gong et al., 2023). However, the field lacks a systematic taxonomy for the complex fine-grained reasoning failures in general LALMs. To fill this gap, we categorize these errors into four dimensions: **Event Omission**, **False Event Identity**, **Temporal Relation Error**, and **Quantitative Temporal Error**. Standard SFT often fails to rectify such granular errors because it lacks negative constraints (Ji et al., 2023). We address this by introducing a specialized dataset and benchmark designed for effective post-training alignment.

3 Problem Statement

Formally, we define an LALM as a function that maps an acoustic sequence $\mathcal{A}$ and a textual instruction $\mathcal{I}$ to a response $\mathcal{R}$. Temporal audio reasoning, however, poses challenges that extend far beyond transcription or high-level acoustic captioning. A model must accurately determine what events occur, when they occur (temporal order), how long they last (duration), and how often they appear (count). Crucially, any generated response $\mathcal{R}$ should remain strictly grounded in the acoustic evidence provided by $\mathcal{A}$. Our diagnostic analysis of state-of-the-art omni-modal LALMs (e.g., Qwen2.5-Omni) reveals that current systems frequently violate this grounding requirement. Instead of relying on the actual audio signal, the model often falls back on language priors, generating responses that are semantically plausible yet acoustically incorrect. These deviations manifest through a compact set of systematic failure modes that directly affect temporal grounding and degrade performance on downstream applications such as acoustic monitoring, multimedia forensics, and event analytics. To characterize these behaviors, we introduce a four-category taxonomy of hallucinations that captures the dominant and practically actionable errors exhibited by contemporary LALMs. This taxonomy aligns closely with the controlled perturbations introduced by our data generation pipeline, enabling precise attribution and measurement. Notably, even the SOTA LALM we evaluated demonstrates all categories of these failure modes.

-Event Omission: The model fails to mention perceptible events present in the audio (e.g., a faint hiss before a loud impact is ignored). Omission harms recall of the acoustic scene and often reflects attention bias toward high-energy events.

-False Event Identity (Fabrication & Misclassification): The model either (a) fabricates events that are not present in the audio or (b) assigns the wrong

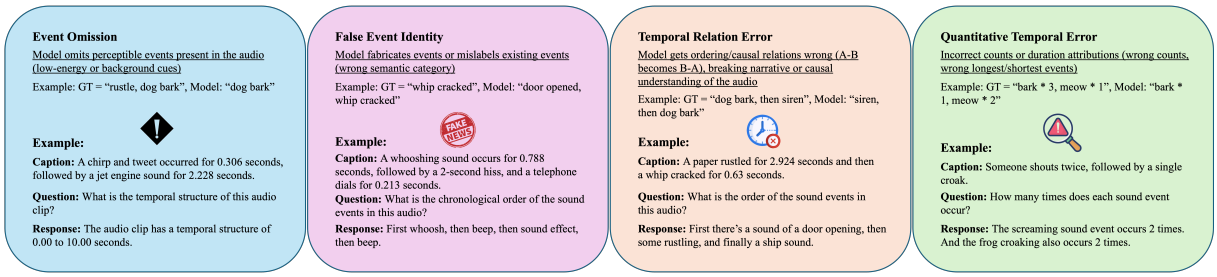

Figure 2: Four principal hallucination types observed in temporal audio reasoning: Event Omission, False Event Identity, Temporal Relation Error, and Quantitative Temporal Error. Each panel illustrates the definition of the failure mode along with a representative hallucination example produced by Qwen2.5-Omni on diagnostic testing.

semantic label to an existing event (e.g., calling a cap gun a "whip" or inventing a "door opening" that never occurs). Both behaviors reflect errors in mapping acoustic evidence to semantic labels.

-Temporal Relation Error: The model gets relations between events wrong: it swaps the chronological order, or misstates causal/sequence relations (e.g., saying B→A when A→B). These errors break narrative or causal understanding of the audio.

-Quantitative Temporal Error: The model errs on quantitative temporal attributes: incorrect counts of repeated events (count bias) or wrong comparative duration attributions (e.g., saying the shorter event is longest). These mistakes indicate failure in simple temporal arithmetic and duration comparison

We quantify these deficiencies using a specialized diagnostic evaluation. Figure 2 reveals that Qwen2.5-Omni struggles significantly across all dimensions, showing an Omission Rate over 40% and a Boundary IoU under 0.4. These results confirm that standard SFT is inadequate for complex acoustic scenes. The model frequently defaults to language priors rather than maintaining strict auditory grounding. This structural weakness necessitates a post-training strategy that penalizes ungrounded generation. Motivated by this, we propose AHA, a framework designed to enforce alignment and mitigate these complex fine-grained reasoning failures.

4 AHA: A Unified Post-Training Framework for Audio Hallucination

4.1 Dataset Formulation and Unified Design

We introduce **AHA**, a unified dataset designed to both *align* and *evaluate* LALMs on hallucination behaviors arising from complex acoustic scenes. While AHA is grounded in audio data, a key design consideration is that many existing LLMs do not directly accept audio as input. Instead, these models typically operate on textual descriptions or instructions derived from audio content. Accordingly, AHA is constructed around caption-level representations that serve as a proxy for acoustic input, enabling the study and mitigation of audio-related hallucinations in a model-agnostic manner. All audio samples in AHA are sourced from existing captioned audio corpora, specifically ***AudioTime*** (Xie et al., 2025), which contains 20,000 training and 2,000 testing data samples in four categories: duration, frequency, ordering and timestamp. Each data sample consists of a temporally structured audio clip paired with a descriptive caption. We used a subset of AudioTime as our data source, which contains all samples with multiple audio events in the audio clip, resulting in 11,507 training and 1,251 testing samples. Formally, AHA is defined over a shared set of audio-question pairs:

$$\mathcal{D} = \{(a_i, q_i)\}_{i=1}^{N}, \tag{1}$$

where a_i denotes an audio clip and q_i denotes a hallucination-sensitive question generated from its caption-level description. The audio signal a_i is used to establish ground truth and human verification, whereas model-facing supervision is provided through caption-derived instructions. From this shared pool, we derive two complementary views:

$$\mathcal{D}^{\text{align}} = \{(a_i, q_i, r_i^+, r_i^-)\}, \tag{2}$$

$$\mathcal{D}^{\text{eval}} = \{(a_i, q_i, y_i^*, \tau(q_i))\}, \tag{3}$$

where r_i^+ and r_i^- denote preferred and dispreferred responses, y_i^* is a ground-truth answer, and $\tau(q_i)$ denotes hallucination type annotations. Both views share the same underlying audio samples, questions, and hallucination taxonomy, ensuring that alignment and evaluation target consistent hallucination phenomena.

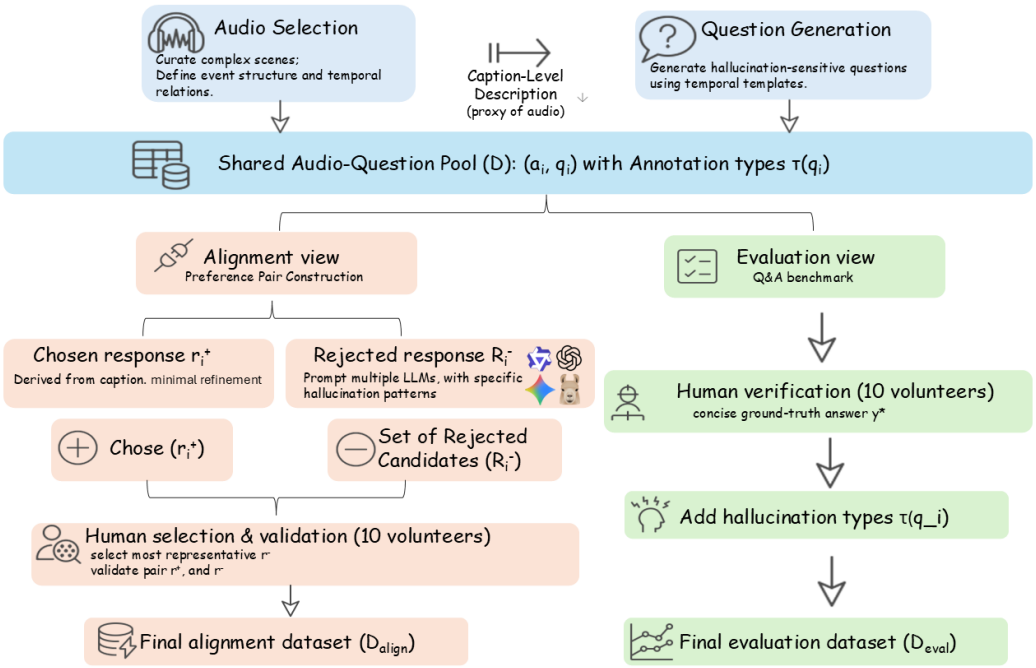

Figure 3: **The unified data construction pipeline for AHA.** The process begins by establishing a shared **Audio-Question Pool** ($\mathcal{D}$) derived from complex acoustic scenes and hallucination-sensitive fine-grained reasoning templates. From this foundation, the pipeline bifurcates into two complementary views: (Left) The **Alignment View** constructs preference pairs for DPO by contrasting caption-derived chosen responses (r_i^+) against LLM-generated rejected responses (r_i^-) that contain specific hallucination patterns. (Right) The **Evaluation View** establishes a rigorous QA benchmark by collecting human-verified ground-truth (y_i^*) and annotating fine-grained hallucination types ($\tau(q_i)$).

4.2 Hallucination-Oriented Construction

Audio selection. Audio samples are curated from existing captioned audio corpora with an explicit focus on hallucination-prone scenarios. We prioritize clips that contain multiple sequential or overlapping sound events, frequent scene transitions, and non-trivial temporal dependencies. Although the audio signal is not directly consumed by most LLMs, it serves as the authoritative source for defining event structure, temporal relations, and annotation validity.

Question generation. Given a caption-level description of an audio clip, we programmatically generate hallucination-sensitive questions by sampling from a predefined set of *fine-grained reasoning templates*. These templates are designed to elicit hallucination behaviors commonly observed when models reason over incomplete or abstracted representations of audio content. The templates are organized into the following categories: **1) Explicit temporal ordering**, e.g., identifying first/last events or exact event sequences; **2) Temporal logic and counterfactual reasoning**, e.g., order verification or hypothetical trimming; **3) Temporal counting and duration comparison**, e.g., event frequency or longest/shortest event. Each template may include symbolic placeholders (e.g., *Event A*, *Event B*), which are instantiated using event mentions inferred from the caption. Formally, each question q_i is associated with one or more hallucination types:

$$\tau(q_i) \subseteq \Big\{\textsc{Omission},\ \textsc{FalseIdentity},\ \textsc{TemporalOrder},\ \textsc{Quantitative}\Big\}. \tag{4}$$

These questions serve as a shared intermediate representation reused by both the alignment and evaluation views of AHA, enabling consistent supervision despite differences in model input modalities. (see Appendix B.3 for the list of question templates)

4.3 Alignment: Counterfactual Hard Negative Synthesis

We construct preference pairs via counterfactual hard negative synthesis. For each audio-question

pair (a_i, q_i), the **chosen response** r_i^+ derives directly from the ground truth caption to ensure factual accuracy. To generate the **rejected responses**, we employ a structured prompting strategy using external text-only LLMs. We explicitly instruct the model to synthesize counterfactual errors corresponding to the hallucination taxonomy defined in Section 4.2. For instance, the prompt directs the model to swap the chronological order of events, modify specific counts, or fabricate non-existent sounds based on the ground truth (see Appendix B.3 for prompt templates). This process yields a set of negative candidates:

$$\mathcal{R}_i^- = \{r_i^{-(1)}, \ldots, r_i^{-(K)}\}. \tag{5}$$

These candidates serve as **hard negatives** because they mimic the linguistic fluency of the chosen response while strictly contradicting the temporal logic of the audio.

Human selection and validation. From the candidate set $\mathcal{R}_i^-$, **10 human volunteers** with experience in audio understanding select the single most representative hallucinated response, denoted as r_i^-. Selection criteria prioritize responses that are linguistically plausible under caption-only reasoning yet clearly incorrect given the acoustic evidence. This ensures the final pair provides a sharp contrast for optimization. The resulting alignment dataset is formalized as:

$$\mathcal{D}^{\text{align}} = \{(a_i, q_i, r_i^+, r_i^-)\}_{i=1}^{N_{\text{align}}}. \tag{6}$$

This preference-based supervision provides a reliable signal for Direct Preference Optimization (DPO), explicitly teaching models to prefer fine-grained reasoning over hallucinations.

4.4 AHA Eval: Hallucination QA Benchmark

The evaluation view of AHA reuses the same audio-question pairs as the alignment view but differs in supervision format. For a subset of (a_i, q_i), a group of **10 human volunteers** manually verify a concise ground-truth answer:

$$y_i^* = \text{Ans}(a_i, q_i), \tag{7}$$

which is strictly supported by perceptible acoustic evidence and avoids unnecessary temporal metadata unless explicitly required by the question. Each evaluation instance is additionally annotated with hallucination types $\tau(q_i)$, enabling fine-grained, category-level diagnostic evaluation:

$$\mathcal{D}^{\text{eval}} = \{(a_i, q_i, y_i^*, \tau(q_i))\}_{i=1}^{N_{\text{eval}}}. \tag{8}$$

5 Experiments & Results

We evaluate our alignment approach across a mixture of in-domain and out-of-domain benchmarks to assess both (A) reduction in hallucinations and (B) general improvements in fine-grained reasoning and broader audio understanding. We compare 6 models: Qwen2.5-Omni (base), Qwen2.5-Omni-DPO (β=0.3), Kimi-Audio, Audio Flamingo 3, GPT-4o-mini and Gemini-2.5. Experiments use both AHA-Eval and a suite of public audio benchmarks: MMAU-test, MMAU-test-mini, MMAU-Pro, and MMAR.

5.1 Alignment Post-Training

We perform preference alignment on **Qwen2.5-Omni** using our dataset $\mathcal{D}^{\text{align}}$. We adopt DPO (Rafailov et al., 2023) to steer the model toward audio-grounded reasoning. Unlike RLHF, DPO optimizes the policy directly on preference data without requiring a separate reward model. Let $x = (a, q)$ denote the audio-question input. For each instance, the dataset provides a preferred response r^+ (ground-truth) and a rejected response r^- (hallucination). We optimize the policy π_θ against the frozen reference model π_{ref} by minimizing:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,r^+,r^-)\sim\mathcal{D}^{\text{align}}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(r^+|x)}{\pi_{\text{ref}}(r^+|x)} - \beta\log\frac{\pi_\theta(r^-|x)}{\pi_{\text{ref}}(r^-|x)}\Big)\Big] \tag{9}$$

where σ is the sigmoid function and β controls the divergence from π_{ref}. This objective increases the likelihood of grounded responses while explicitly suppressing the specific hallucination in r^-.

Implementation Details. We use Qwen2.5-Omni-7B as our foundation model. To reduce computational cost, we apply Low-rank adaptation (LoRA) (Hu et al., 2022) in our DPO post-training. Following the effective hyperparameter choices for LoRA (Schulman and Lab, 2025), We set the rank $r = 16$ and scaling factor $\alpha = 32$, and attached LoRA layers to all attention and MLP layer in the model. The reference model is frozen in 16-bit precision. We train for 8 epochs using the AdamW optimizer with a learning rate of 2×10^{-6} and a global batch size of 16. The DPO parameter β is set to 0.3. Experiments are conducted on 4 NVIDIA A100 GPUs. For brevity, we refer to this aligned model as **Qwen-Audio-AHA** in the subsequent

Table 1: **Unified Evaluation Results.** **Panel A** reports hallucination error rates (%) on the AHA benchmark (**lower is better**). **Panel B** reports accuracy (%) on public benchmarks (**higher is better**). Blue rows highlight our aligned model's absolute performance. Green rows quantify the substantial improvement (Δ) over the baseline.

PANEL A: AHA EVAL BENCHMARK (ERROR RATE %, LOWER IS BETTER)				
Model	**Event Omission**	**False Identity**	**Temp. Relation**	**Quant. Temporal**
GPT-4o	77.8	72.2	34.1	63.6
Gemini 2.5	74.7	80.8	34.8	58.3
Kimi-Audio	84.7	**57.6**	35.6	72.5
Audio Flamingo 3	61.4	81.6	40.0	75.3
Qwen2.5-Omni (Base)	70.6	70.6	30.5	69.6
Qwen-Audio-AHA (Ours)	**53.8**	64.1	**15.9**	**52.6**
Reduction (Δ)	**▼ -16.8**	**▼ -6.5**	**▼ -14.6**	**▼ -17.0**

PANEL B: PUBLIC BENCHMARKS (ACCURACY %, HIGHER IS BETTER)				
Model	**MMAU-test-mini**	**MMAU-test**	**MMAU-Pro**	**MMAR**
GPT-4o	57.8	54.5	41.7	49.2
Gemini 2.5	70.5	67.1	**54.9**	57.3
Kimi-Audio	54.9	56.1	50.1	49.1
Audio Flamingo 3	74.5	72.4	25.0	54.1
Qwen2.5-Omni (Base)	74.6	71.2	54.2	58.1
Qwen-Audio-AHA (Ours)	**76.4**	**72.5**	54.8	**59.7**
Improvement (Δ)	**▲ +1.8**	**▲ +1.3**	**▲ +0.6**	**▲ +1.6**

sections. (More details refer to appendix A)

5.2 AHA Benchmark Evaluation

AHA provides a controlled setting to quantify hallucination behaviors arising from complex fine-grained reasoning (Section 4). Each test instance is annotated with one or more of the taxonomy-defined hallucination types. We leverage LLM-as-a-judge by providing ground-truth audio captions to a GPT-4o client, which then label model responses across the four dimensions. This allows us to calculate precise metrics: **event omission rate**, **false-identity rate**, **ordering error rate** and **counting error rate**. We report hallucination rates of the generated questions, which follow the unified design of AHA (Sections 4.1–4.4). As shown in Table 1 (Panel A), current SOTA models exhibit severe grounding failures. Even strong baselines like Gemini and GPT-4o exceed 74% in Event Omission rates. Kimi-Audio, while robust in identity, suffers an 84.7% omission rate. These results underscore that standard training paradigms often fail to ensure strict adherence to acoustic evidence.

Hallucination Reduction. Across all four hallucination dimensions, Qwen-Audio-AHA substantially reduces error rates compared to the Qwen2.5-Omni base model. The most significant absolute gains were observed in Quantitative Temporal Error and Event Omission, where we achieved reductions of 17.0% and 16.8% respectively. In quantitative fine-grained reasoning, where baselines often return incorrect event frequencies or durations, our aligned model lowered the error rate from 69.6% to 52.6%. Similarly, in Temporal Relation reasoning, the most challenging category for many models, Qwen-Audio-AHA nearly halved the base model's error rate, dropping it from 30.5% to 15.9%. These results indicate that DPO effectively teaches the model to prioritize acoustic grounding over the linguistically plausible but factually incorrect priors found in caption-only models (Table 1).

Quality of Fine-Grained Reasoning. To illustrate the impact of alignment on model output, we conducted qualitative case studies using the AHA dataset. We found that Qwen-Audio-AHA consistently produces more concise and acoustically faithful responses. While the base model frequently exhibits "hallucinations of granularity", such as fabricating overlapping events or misstating causal sequences, the aligned model generates results that adherence to the temporal metadata derived from the audio. This finding motivates using AHA as a diagnostic hallucination benchmark, where precise event tuples extracted from captions allow scoring.

5.3 Generalization to Public Benchmarks

To evaluate the impact of our alignment on broad multimodal audio understanding, we evaluated

Qwen-Audio-AHA on four standard benchmarks: MMAU-test, MMAU-test-mini, MMAU-Pro, and MMAR. We strictly adhered to official protocols for all datasets.

Performance Consistency and Mutual Benefits. Table 1 (Panel B) presents the evaluation results. It is important to note a *granularity mismatch*: while AHA targets *fine-grained* temporal hallucinations, public benchmarks primarily assess *coarse-grained* perception, which may dilute the visibility of our improvements. Typically, such targeted optimization incurs an "alignment tax" (Ouyang et al., 2022), where general capabilities degrade under specific constraints. However, **Qwen-Audio-AHA** defies this trend, exhibiting *positive transfer* across the board. We observe consistent accuracy gains: **+1.8%** on MMAU-test-mini, **+1.3%** on MMAU-test, and **+1.0%** on MMAR. These metrics confirm that our framework enhances model precision on general tasks without overfitting to the preference dataset.

Generalization of Fine-grained Reasoning Capabilities. Notably, the reasoning capabilities learned from our alignment transfer effectively to public benchmarks. We observe significant improvements in tasks that require complex reasoning. This indicates that the **AHA** framework does not simply memorize caption templates. Instead, it equips the model with robust **complex fine-grained reasoning**. As a result, the model achieves better performance on diverse, out-of-distribution public benchmarks.

6 Analysis

Does hallucination alignment improve general reasoning? Contrary to concerns that alignment might degrade general capabilities, our results indicate the opposite. As detailed in Table 2, Qwen-Audio-AHA achieves substantial gains on the **Temporal Event Reasoning** subsets of public benchmarks, peaking at **+8.3%** on MMAU-Test. Notably, these specific gains far outpace the model's average improvement on general benchmarks (+0.6% to +1.8%). This discrepancy confirms that our method does not merely suppress text, but it also refines the internal representation of the audio timeline. Consequently, the model becomes far more effective at resolving complex event sequences that previously triggered "hallucinations of granularity".

Table 2: **Complex Temporal/Event Reasoning Accuracies.** Blue : Ours. Green : Improvement (Δ).

COMPLEX FINE-GRAINED REASONING ACCURACIES (HIGHER IS BETTER)			
Model	**MMAU-Mini**	**MMAU-Test**	**MMAU-Pro**
Qwen2.5-Omni (Base)	70.8	59.5	64.4
Qwen-Audio-AHA (Ours)	**77.1**	**67.8**	**65.0**
Improv. (Δ)	▲ +6.3	▲ +8.3	▲ +0.6

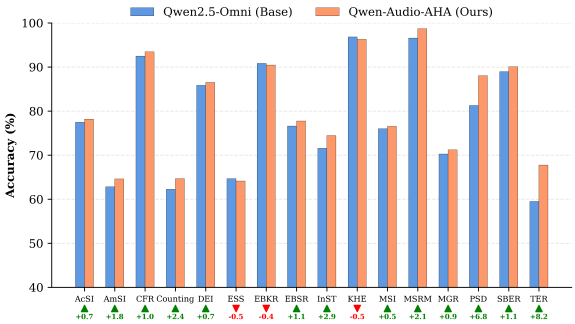

Figure 4: Categorical accuracies before and after alignment in MMAU-test benchmark. Most dimensions have improved accuracy after alignment, especially **Temporal Event Reasoning (TER)** and **Phonological Sequence Decoding (PSD)**, while stagnation or minor degradation are observed in a few subcategories.

How effective is alignment across different taxonomies? Beyond overall gains, the MMAU-test breakdown (Fig. 4) shows improvements across most dimensions, led by **TER** (+8.2%) and **PSD** (+6.8%). Subjective categories like **Emotion State Summarisation (ESS)** (-0.5%) saw slight drops (note that our alignment data excludes emotion-specific examples). Crucially, this broad success indicates that the precise temporal grounding enforced by **AHA** yields a positive transfer effect, enhancing fundamental representation and boosting performance across most taxonomies beyond just specific hallucination mitigation.(For full names of those categories, refer to appendix C)

7 Conclusion

In this paper, we introduced **AHA** to mitigate hallucinations in LALMs. By synthesizing **counterfactual hard negatives** based on a fine-grained taxonomy, we effectively force models to prioritize acoustic evidence over linguistic priors. For evaluation, we established **AHA-Eval**, a diagnostic benchmark tailored for detecting these hallucinations. Finally, Our aligned model demonstrates significant improvements, even outperforming SOTA methods on both diagnostic and public benchmarks.

8 Limitations

While **Qwen-Audio-AHA** demonstrates significant advancements in aligning audio-language models, our study highlights a critical bottleneck in the broader research landscape: the inadequacy of current automated evaluation protocols. We observe that even the most advanced models (e.g., GPT-5, Gemini-3) used as judges exhibit rigid sensitivity that often penalizes valid, high-level reasoning.

Caption: A helicopter is heard once, followed by a single chuckle or chortle.

Question: How many distinct sound events occur in this sequence?

Model Prediction: Well, there are two distinct sound events. First, there's the sound of an engine idling, and then there's the sound of people laughing. So, two in total.

LLM's Judgement:
- Event Omission: ❌
- False Event Identity: ❌
- Temporal Relation Error: ✅
- Quantitative Temporal Error: ✅

Figure 5: A critical failure case of automated LLM evaluation. The LALM correctly identifies the acoustic event as "people laughing", which is semantically equivalent to the caption's "chuckle or chortle". However, the LLM judge fails to bridge this semantic gap, incorrectly penalizing the model with **Event Omission** and **False Event Identity** errors. This highlights the risk of false positives in current hallucination benchmarks.

The "Semantic Equivalence" Trap in LLM Judging. A primary failure mode of current evaluation pipelines is the inability to recognize semantic equivalence across different lexical choices. As illustrated in Figure 5, the ground truth describes a sound as a "chuckle or chortle", while our model predicts "people laughing". Despite these terms referring to the same fundamental acoustic event, the LLM judge lacks the acoustic intuition to bridge this semantic gap, incorrectly flagging the response as a **False Event Identity**. Similar spurious errors arise with synonyms like "shout" versus "scream", or "engine noise" versus "idling". This suggests that current reported hallucination rates may be inflated by false positives, where models are penalized not for factual errors, but for describing the truth with different vocabulary than the caption.

Bias Against Reasoning Depth. Furthermore, current benchmarks typically favor a specific "depth" of description, which usually adheres strictly to the caption's surface forms and restricts the evaluation of multi-dimensional reasoning. Audio events can be validly described via onomatopoeia (e.g., "clicking"), the immediate action ("typing"), or the underlying scene ("someone working at a desk"). We find that when LALMs provide more granular acoustic details or higher-level scene inferences than the ground truth, they are often unfairly penalized by metrics that rely on rigid n-gram overlap. This limitation reflects a gap in the community's infrastructure rather than a model deficit. We argue that the field must move beyond single-dimension accuracy and develop dynamic benchmarks capable of appreciating the diverse but correct interpretations of complex acoustic scenes.

References

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and 1 others. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

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 preprint arXiv:2311.07919*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Rita Frieske and Bertram E Shi. 2024. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *arXiv preprint arXiv:2401.01572*.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, and 1 others. 2024. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow,

Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

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, and 1 others. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

John Schulman and Thinking Machines Lab. 2025. Lora without regret. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/lora/.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. 2025. Audiotime: A temporally-aligned audio-text benchmark dataset. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

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. *arXiv preprint arXiv:2305.11000*.

A Implementation Details

We provide in Table 3 the hyperparameters we used in training Qwen-Audio-AHA:

B AHA Data Generation Details

We provide additional details regarding the construction of the AHA training and evaluation datasets. The generation process followed a structured pipeline designed to elicit and penalize specific audio grounding failures.

B.1 Question Paradigm Selection

We initially curated a set of candidate question paradigms focusing on event identification, quantitative counting, and complex temporal relationships. These paradigms were synthesized from common query structures found in public open-ended QA datasets, such as **MECAT-QA** and multiple-choice benchmarks like **MMAU**. By adapting these existing structures, we ensured that the AHA questions reflect realistic challenges in general audio understanding while maintaining a specialized focus on fine-grained reasoning.

B.2 Automated Preference Construction

Using audio-caption pair from the **AudioTime** corpus as the foundation, we employed **GPT-4o** to instantiate the shared question pool ($\mathcal{D}$). For each instance, the model was prompted to select an appropriate question template and generate a corresponding preference pair based on the following rules:

- **Chosen Response:** A response that accurately answers the query while remaining strictly grounded in the caption-derived ground truth.
- **Rejected Response:** A response that appears linguistically plausible but deliberately incorporates one or more taxonomy-defined halluci-

Table 3: **Hyperparameters for training Qwen-Audio-AHA**

Hyperparameter	value
LoRA rank	16
LoRA α	32
LoRA dropout	0.1
LoRA bias	None
LoRA target layers	r".*model.*(q_projlk_projlv_projlo_projlup_projldown_projlgate_proj)"
DPO β	0.3
DPO warmup ratio	0.03
n_epochs	8
learning rate	2e-6
weight decay	0
Adam $\beta 1$	0.9
Adam $\beta 2$	0.95
dtype	BF16
per device train batch size	1
gradient accumulation steps	16

nations: **Event Omission**, **False Event Identity**, **Temporal Relation Error**, or **Quantitative Temporal Error**.

B.3 Prompting and Quality Control

The generation prompt included exhaustive definitions and detailed examples for each hallucination category to guide the judge model toward generating high-fidelity negative samples. The final output was formatted as a structured JSON object to facilitate downstream processing. Following the automated generation phase, human volunteers performed a manual audit of the samples to verify their acoustic accuracy, linguistic plausibility, and categorical diversity. Any instances found to be ambiguous or misaligned with the acoustic evidence were discarded to ensure the integrity of the **AHA-Eval** benchmark. We present in Figure 6 the prompt and Figure 7 candidate questions we used for AHA data generation.

C MMAU-test benchmark Category Abbreviations

For aesthetics reason, we put abbreviations in our categorical demonstration of MMAU-test result. We list the names of those abbreviations proposed by the official MMAU-test benchmark in Table 4.

Context: You are an expert audio analyst. I will provide you with a ground truth description of an audio clip that contains a temporal sequence (e.g., "Gunshots then screaming").

Task:
- Select a question from the list below:
{questions}
- Optionally, slightly modify the question but keep exactly the same sementic meaning.
- If "Event A" or "Event B" appear in the question template, replace it with the events infered from the caption.

- Generate the "Chosen" Response: Answer the question accurately based on the ground truth.
  - Do not include timestamp or duration information unless the question explicitly asked.
  - Do not give very long answers.
- Generate the "Rejected" Response: Answer the question plausibly, but makes one or more mistakes listed below:
  - Event Omission: The answer fails to mention perceptible events present in the audio (e.g., a faint hiss before a loud impact is ignored).
  - False Event Identity: The answer either (a) fabricates events that are not present in the audio or (b) assigns the wrong semantic label to an existing event (e.g., calling a cap gun a “whip” or inventing a “door opening” that never occurs).
  - Temporal Relation Error: The answer gets relations between events wrong. It swaps the chronological order, or misstates causal/sequence relations (e.g., saying B->A when A->B).
  - Quantitative Temporal Error: The answer errs on quantitative temporal attributes: incorrect counts of repeated events (count bias) or wrong comparative duration attributions (e.g., saying the shorter event is longest).

- Both "Chosen" and "Rejected" responses should directly answer the question. Do not include extra explanation or summary unless the question asked.
- Keep the responses as short as possible.
- Return a JSON object with schema: {{"prompt": "...", "chosen": "...", "rejected": "..."}}, where "prompt" should contain the selected question, "chosen" should contain the "Chosen" response, and "rejected" should contain the "Rejected" response.

Ground Truth:{caption}

Figure 6: The prompt used for AHA data generation.

Table 4: **MMAU-test benchmark Category Abbreviations.**

Abbreviation	Name
AcSI	Acoustic Source Inference
AmSI	Ambient Sound Interpretation
CFR	Conversational Fact Retrieval
Counting	Counting
DEI	Dissonant Emotion Interpretation
ESS	Emotion State Summarisation
EBKR	Event-Based Knowledge Retrieval
EBSR	Event-Based Sound Reasoning
InST	Instrumentation
KHE	Key highlight Extraction
MSI	Melodic Structure Interpretation
MSRM	Multi Speaker Role Mapping
MGR	Musical Genre Reasoning
PSD	Phonological Sequence Decoding
SBER	Sound-Based Event Recognition
TER	Temporal Event Reasoning

Category 1: Explicit Ordering (The Core Task)
"What is the chronological order of the sound events in this audio?"
"What is the first sound event in this audio?"
"What is the last sound event in this audio?"
"Describe the sequence of sounds from start to finish."
"Does the {Event A} occur before or after the {Event B}?"
"List the sound events in the exact order they appear."
"Does the audio start with {Event A} or {Event B}?"

Category 2: Temporal Logic & Reasoning (Harder)
"True or False: The {Event B} is heard before the {Event A}."
"If I trimmed the first half of this audio, which sound event would be removed?"
"Which sound event dominates the beginning of the recording?"
"What is the temporal structure of this audio clip?"
"Can you identify the sound of {Event A} in the sequence?"

Category 3: Counting & Frequency (Temporal Distribution)
"How many distinct sound events occur in this sequence?"
"Does {Event A} occur more than once?"
"Which sound event has the longest duration?"
"Given the audio sample, which sound can be heard the longest?"
"Given the audio sample, which sound can be heard the shortest?"
"For the given audio, identify the sound with the longest duration."
"For the given audio, identify the sound with the shortest duration."
"Which sound event has the shortest duration?"
"How many times does each sound event occor?"

Figure 7: Candidate questions for AHA data generation.