FROM ATTENTION WEIGHTS TO AUDIO WRITES: MITIGATING HALLUCINATIONS IN LARGE AUDIO LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large audio language models (LALMs) may describe sound events or attributes that are absent from the input recording, resulting in hallucinations. Existing attention interventions mitigate these errors by sustaining attention to audio, but attention weights alone do not determine the vector contributions of audio positions to the attention output. To address this limitation, we introduce the Audio Relative Write Ratio (ARWR), which measures the norm of the vector update from audio positions relative to the sum of update norms from audio and other positions. By incorporating value aggregation and output projection, ARWR provides a more accurate characterization of the contribution of audio positions to the attention output. Building on this diagnostic, we propose ARWR-Cal, a framework for hallucination mitigation through residual calibration of the final hidden state. At each generation step, a lightweight low-rank calibrator uses the source vectors and final hidden state to predict a residual adjustment. ARWR controls the permitted adjustment magnitude. Experiments show ARWR outperforms audio attention in hallucination diagnosis, while ARWR-Cal improves average accuracy by 8.2 percentage points over the strongest baseline across three datasets and three LALMs. Code is available at https://anonymous.4open.science/r/ARWR-Cal.

1 INTRODUCTION

Large audio language models (LALMs) enable natural language interaction with speech and environmental sounds (Chu et al., 2024; Deshmukh et al., 2023; Gong et al., 2024), as well as music (Tang et al., 2024), allowing users to query auditory scenes. Recent multimodal assistants, such as Qwen2.5-Omni (Xu et al., 2025), integrate these capabilities with other inputs to generate context-aware responses. However, LALMs can still hallucinate sound events or attributes unsupported by the recording (Hsu et al., 2025; Chen et al., 2026; Kuan et al., 2024), making hallucination mitigation a central challenge for reliable audio understanding.

To mitigate hallucinations in LALMs, recent studies use decoding or supervised learning to ground LALM responses in acoustic evidence. AAD (Hsu et al., 2025) contrasts audio-conditioned and reference predictions to favor audio-sensitive candidates. AHA (Chen et al., 2026) uses preference alignment with counterfactual audio negatives to discourage linguistically plausible but acoustically unsupported answers. For reasoning over long audiovisual recordings, MPRC (Dong et al., 2026) estimates claim support by analyzing responses to modality perturbations, without updating the backbone. Recent work identifies declining audio attention as a potential contributor to hallucinations, motivating interventions such as AttnAdapter (Du & Zhong, 2026) that rescale attention across input segments and decoding steps to maintain access to audio information and mitigate hallucinations.

However, audio attention mass measures the weights assigned to audio positions, not their vector contributions to the decoder. As shown in Figure 1, we found that even when the model attends to the audio, it may still generate answers unsupported by the recording. Analyses of Transformer vector norms (Kobayashi et al., 2020; 2021) show that the attention output depends not only on these weights but also on the value representations and output projection. Different value vectors can therefore yield different decoder updates under identical attention weights. Thus, attention allocation alone does not establish whether a generated answer is genuinely grounded in the recording.

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

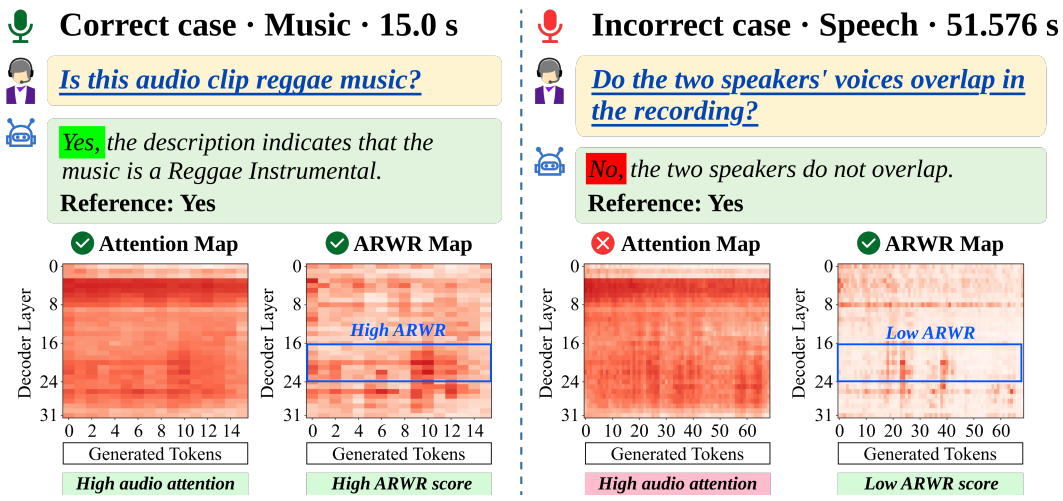

Figure 1: **Audio attention and ARWR for correct and hallucinated answers.** Both cases show high audio attention, while ARWR is high for the correct answer and low for the hallucinated answer. This contrast illustrates how ARWR can reveal hallucination risk that is not apparent from audio attention alone.

To address this limitation, we define the Audio Relative Write Ratio (ARWR) as the norm of the vector aggregated from audio positions divided by the summed norms of this vector and the vector from all remaining positions. By incorporating value aggregation and output projection, ARWR measures the strength of audio writes relative to writes from other sources. Building on this measure, we propose ARWR-Cal, a framework that uses ARWR to assess hallucination risk and guide correction of the model's predictions. At each generation step, ARWR-Cal uses a lightweight low-rank calibrator to predict a residual adjustment to the final hidden state based on the source vectors and that hidden state. The adjusted state passes through the frozen readout to update next-token vocabulary scores. ARWR bounds the adjustment magnitude, with lower values permitting larger corrections. We train the calibrator on supervised audio question answering data while keeping the LALM backbone and vocabulary projection frozen.

Our contributions are summarized as follows.

- We introduce ARWR, a diagnostic that quantifies the relative strength of writes from audio positions in LALMs. By incorporating value aggregation and output projection, ARWR measures aspects of the attention output that attention weights alone miss.
- We propose ARWR-Cal, a hallucination correction framework that uses ARWR to bound residual adjustments predicted by a lightweight calibrator while keeping the LALM frozen.
- We evaluate ARWR for hallucination diagnosis on Clotho-AQA and ARWR-Cal for hallucination mitigation across three datasets and three LALMs. The results show that ARWR outperforms audio attention in hallucination diagnosis, while ARWR-Cal improves average accuracy by 8.2 percentage points over the strongest baseline, validating the effectiveness of ARWR-guided residual calibration in improving answer accuracy.

2 Related Work

2.1 Hallucination Diagnosis

Hallucination diagnosis estimates answer reliability from model outputs or internal signals (Azaria & Mitchell, 2023). SelfCheckGPT (Manakul et al., 2023) identifies potential factual errors by examining inconsistencies across sampled responses. Semantic entropy (Farquhar et al., 2024) measures uncertainty over clusters of semantically equivalent answers, separating variation in meaning from variation in wording. Moving beyond output text, INSIDE (Chen et al., 2024) uses the eigenvalues

of a covariance matrix of response representations to assess consistency in the hidden space. These methods assess reliability by comparing multiple responses to the same input.

Other studies instead examine attention during generation and its relation to hallucinations. Lookback Lens (Chuang et al., 2024) addresses contextual hallucinations in LLMs using a classifier trained on relative attention to input context and previously generated text. For LALMs, Du & Zhong (2026) link hallucinations to declining audio attention, highlighting the importance of maintaining access to acoustic evidence. In VLMs, LVLMs-Saliency (Zhang et al., 2026) combines attention with gradients, linking hallucinations to the reduced influence of preceding output tokens on subsequent predictions. Attention allocation alone, however, does not fully determine vector contributions, as attention outputs also depend on the value representations and their output projection (Kobayashi et al., 2020; 2021). Consequently, substantial attention to audio can coexist with weak vector writes, limiting attention mass as a diagnostic of the strength of audio contributions. ARWR instead compares the norm of the write aggregated from audio positions with the sum of the norms of writes from audio and other positions, after value aggregation and output projection.

2.2 Hallucination Mitigation

Hallucination mitigation follows two approaches: supervised adaptation and inference time intervention. Among supervised adaptation methods, AHA (Chen et al., 2026) uses counterfactual negative responses to favor acoustically supported answers, while AHA-Guard (Seth et al., 2026) provides preference pairs targeting misleading queries and manipulated audio. PCLM (Pang et al., 2026) combines prompt-conditioned mixing of audio encoder layers with preference optimization to improve paralinguistic understanding. Other methods use supervision to learn intervention directions or calibrate evidential support. ITI (Li et al., 2023) learns truth-related directions and shifts activations in selected attention heads. MPRC (Dong et al., 2026) learns to assess support for audiovisual claims from native predictions and responses to audio perturbations.

Inference time interventions adjust candidate scores or attention during generation. AAD (Hsu et al., 2025) contrasts predictions with and without audio, while AVCD (Jung et al., 2025) uses attention to select modality perturbations for audiovisual contrastive decoding. DeCo (Wang et al., 2025) refines candidate scores using dynamically selected earlier layers rather than different input conditions. Attention and saliency guide a complementary set of interventions. AttnAdapter (Du & Zhong, 2026) rescales attention across input segments and decoding steps to maintain access to audio information. In VLMs, OPERA (Huang et al., 2024) counters excessive reliance on summary tokens through decoding penalties and reconsideration of earlier token choices. SGRS and LocoRE (Zhang et al., 2026) further link diagnosis to intervention by respectively filtering candidates using saliency and reinforcing attention to recent outputs. Compared with these approaches, ARWR-Cal uses the vector contributions from audio and other positions, together with the final hidden state, to predict a residual adjustment to that state, while ARWR controls the permitted adjustment magnitude.

3 Audio Relative Write Ratio

We characterize the contribution of audio during answer generation through the vector updates supplied by audio token positions. We first decompose the attention outputs by source position, then define the Audio Relative Write Ratio (ARWR) to quantify the relative strength of vector writes from audio positions across decoding steps.

3.1 From Attention Weights to Source Writes

Given an audio recording a, a question q, and an answer prefix $y_{<t}$, a LALM predicts the next answer token y_t. At generation step t, let $\mathcal{I}_t$ denote the positions accessible under the attention mask to the query at the last position of the current input sequence. We partition these positions into $\mathcal{A}$, the positions at which audio embeddings are inserted into the decoder input, and $\mathcal{N}_t = \mathcal{I}_t \setminus \mathcal{A}$, containing all remaining positions, including the question, instructions, special tokens, and previously generated text.

Source decomposition. For each of the H attention heads at layer l, let $\alpha^h_{l,t,j}$ be the attention weight on position j, normalized over $\mathcal{I}_t$, and $\mathbf{v}^h_{l,j} \in \mathbb{R}^{d_h}$ its value vector, where d_h is the per-head

value dimension. We aggregate values separately over the audio and remaining positions:

$$\mathbf{s}_{l,t}^{A,h} = \sum_{j \in \mathcal{A}} \alpha_{l,t,j}^{h} \mathbf{v}_{l,j}^{h},$$
$$\mathbf{s}_{l,t}^{N,h} = \sum_{j \in \mathcal{N}_t} \alpha_{l,t,j}^{h} \mathbf{v}_{l,j}^{h}. \tag{1}$$

Projecting the concatenated per-head aggregates gives the audio and nonaudio *source writes*:

$$\mathbf{u}_{l,t}^{A} = \mathbf{W}_{l}^{O} \operatorname{Concat}_{h=1}^{H}\big(\mathbf{s}_{l,t}^{A,h}\big),$$
$$\mathbf{u}_{l,t}^{N} = \mathbf{W}_{l}^{O} \operatorname{Concat}_{h=1}^{H}\big(\mathbf{s}_{l,t}^{N,h}\big) + \mathbf{b}_{l}^{O}, \tag{2}$$

where $\mathbf{W}_{l}^{O} \in \mathbb{R}^{d \times H d_h}$ projects into the d-dimensional hidden space and $\mathbf{b}_{l}^{O} \in \mathbb{R}^{d}$ is the output bias. The vectors $\mathbf{u}_{l,t}^{A}$ and $\mathbf{u}_{l,t}^{N}$ denote the audio and nonaudio *source writes*, respectively. Including the output bias in the latter yields $\mathbf{o}_{l,t} = \mathbf{u}_{l,t}^{A} + \mathbf{u}_{l,t}^{N}$, the attention update added to the residual stream.

Attention mass. Audio attention mass is the total weight assigned to audio positions, averaged across heads:

$$m_{l,t}^{A} = \frac{1}{H} \sum_{h=1}^{H} \sum_{j \in \mathcal{A}} \alpha_{l,t,j}^{h}. \tag{3}$$

Vector analyses by Kobayashi et al. (2020) characterize attention outputs through both attention weights and transformed value vectors. For fixed attention weights, changes in the magnitudes or directions of audio value vectors can alter $\mathbf{u}_{l,t}^{A}$ while leaving $m_{l,t}^{A}$ unchanged. Thus, the source writes in equation 2 capture vector contributions to the residual stream that attention mass alone does not describe.

3.2 Definition of Audio Relative Write Ratio

We define the ARWR at layer l and generation step t as

$$r_{l,t} = \frac{\|\mathbf{u}_{l,t}^{A}\|_2}{\|\mathbf{u}_{l,t}^{A}\|_2 + \|\mathbf{u}_{l,t}^{N}\|_2 + \epsilon}, \tag{4}$$

where $\|\cdot\|_2$ denotes the Euclidean norm and $\epsilon > 0$ is a numerical stabilization constant. The score $r_{l,t} \in [0, 1]$ quantifies the relative strength of the audio write. Lower values correspond to weaker audio writes relative to nonaudio writes.

Layer aggregation. For a D-layer decoder, we use $\mathcal{L} = [D/2, 3D/4)$ for ARWR aggregation and source-vector extraction. We average the ratios over these layers to obtain the generation-step score:

$$r_t = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} r_{l,t}. \tag{5}$$

ARWR-Cal uses r_t to set the permitted correction magnitude, with lower values allowing larger residual adjustments.

4 ARWR-Cal: Residual Calibration Guided by Audio Writes

4.1 Overview

Building on ARWR, we propose ARWR-Cal to mitigate hallucinations in LALMs through residual calibration of the final decoder hidden state. As shown in Figure 2, it combines an ARWR-guided magnitude controller with a lightweight low-rank residual calibrator. At each generation step, the calibrator predicts a residual adjustment from audio and nonaudio source writes extracted from selected decoder layers, together with the final hidden state. ARWR computed from these writes determines a continuous correction budget that constrains the adjustment magnitude. The predicted adjustment is bounded by this budget and added to the final hidden state, which is then passed through the frozen final normalization and vocabulary projection to obtain the calibrated prediction. We train the calibrator on supervised audio question-answering data while keeping the LALM backbone frozen.

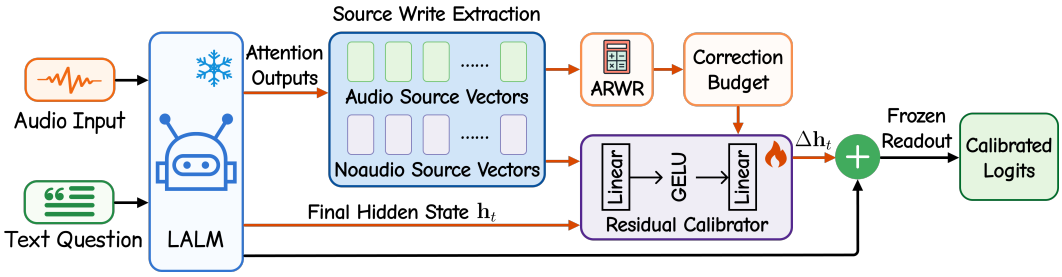

Figure 2: **Overview of ARWR-Cal.** ARWR-Cal uses audio and non-audio source representations to guide output calibration. ARWR controls the correction budget, while a lightweight calibrator adjusts the final hidden state to produce calibrated logits through the frozen readout. Only the calibrator is trained.

4.2 ARWR-Guided Correction Budget

Before calibrating model predictions, we establish a correction budget to regulate the intervention strength at each generation step. We use the generation-step ARWR score r_t in equation 5 to control the correction budget, allowing larger adjustments when audio writes are relatively weak.

Continuous correction budget. We map r_t to a scalar correction budget

$$b_t = 0.5 + \mathrm{clip}\left(\frac{\tau - r_t}{\tau}, 0, 1\right), \tag{6}$$

where $\mathrm{clip}(\cdot, 0, 1)$ restricts its argument to $[0, 1]$ and $\tau > 0$ is a reference threshold. We estimate τ from the 75th percentile of ARWR scores in the calibration training split and keep it fixed during inference. For $r_t < \tau$, the budget increases linearly as ARWR decreases.

Magnitude constraint. Let $\mathbf{h}_t \in \mathbb{R}^d$ denote the final decoder hidden state before the final RMS normalization, and let $\Delta\mathbf{h}_t \in \mathbb{R}^d$ denote the residual adjustment. The correction budget sets an upper bound on the adjustment norm:

$$\|\Delta\mathbf{h}_t\|_2 \le \rho\, b_t\, \|\mathbf{h}_t\|_2, \tag{7}$$

where $\rho > 0$ is the base correction scale relative to the hidden state norm.

4.3 Residual Calibrator

Given the correction budget, we use a lightweight residual calibrator to predict an adjustment to the final hidden state. The calibrator conditions this adjustment on the source vectors and the hidden state to be corrected.

Calibration inputs. The audio and nonaudio source vectors describe the updates contributed by the two groups of token positions, while $\mathbf{h}_t$ represents the decoder state used for the next prediction. We combine these inputs into a calibration feature:

$$\mathbf{x}_t = \phi\big((\mathbf{u}^A_{l,t})_{l\in\mathcal{L}}, (\mathbf{u}^N_{l,t})_{l\in\mathcal{L}}, \mathbf{h}_t\big) \in \mathbb{R}^m, \tag{8}$$

where l indexes the decoder layers in the selected set $\mathcal{L}$, and m is the feature dimension. The fixed feature map ϕ applies RMS normalization and a shared Gaussian projection to each input vector, then concatenates the projected features and log RMS magnitudes. The concatenated features are standardized using training statistics and clipped to a fixed range. This construction provides the calibrator with features of the vector components as well as their magnitudes.

Low-rank residual prediction. We implement the calibrator as a two-layer multilayer perceptron (MLP). It maps $\mathbf{x}_t$ to a predicted residual vector

$$\mathbf{g}_t = \mathbf{W}_{\mathrm{up}}\ \mathrm{GELU}(\mathbf{W}_{\mathrm{down}}\mathbf{x}_t + \mathbf{b}_{\mathrm{down}}), \tag{9}$$

where $\mathbf{W}_{\mathrm{down}} \in \mathbb{R}^{k\times m}$ projects the input features into a k-dimensional intermediate representation, and $\mathbf{b}_{\mathrm{down}} \in \mathbb{R}^k$ is its bias. $\mathbf{W}_{\mathrm{up}} \in \mathbb{R}^{d\times k}$ maps this representation into the decoder hidden space to produce the residual adjustment vector.

Budgeted calibration. Using the budget b_t from equation 6, we obtain the applied adjustment:

$$\Delta \mathbf{h}_t = \rho \, b_t \, \|\mathbf{h}_t\|_2 \frac{\mathbf{g}_t}{\sqrt{1 + \|\mathbf{g}_t\|_2^2}}. \tag{10}$$

The smooth normalization bounds the predicted vector's norm, ensuring that the adjustment satisfies equation 7. We add the adjustment to the hidden state and compute calibrated logits through the frozen final normalization and vocabulary projection:

$$\begin{aligned} \widetilde{\mathbf{h}}_t &= \mathbf{h}_t + \Delta \mathbf{h}_t, \\ \widetilde{\mathbf{z}}_t &= \mathbf{W}_{\text{voc}} \, \mathrm{RMSNorm}\big(\widetilde{\mathbf{h}}_t\big), \end{aligned} \tag{11}$$

where $\mathbf{W}_{\text{voc}} \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the vocabulary projection and $\mathcal{V}$ is the token vocabulary.

4.4 Training Objective

We train the residual calibrator using labeled audio question answering examples. The LALM backbone, final normalization, and vocabulary projection remain frozen. The training objective combines answer supervision, prediction preservation, and adjustment regularization:

$$\mathcal{J}(\theta) = \mathcal{L}_{\text{CE}} + \lambda_{\text{p}} \mathcal{L}_{\text{preserve}} + \lambda_{\text{n}} \mathcal{L}_{\text{norm}}, \tag{12}$$

where $\lambda_{\text{p}}, \lambda_n \geq 0$ control the contributions of the two regularization terms. Let $\mathcal{D} = \{(a_i, q_i, y_i^\star)\}_{i=1}^n$ denote the calibration training set. For each example i, let t_i be the supervised prediction position and $y_{i,t_i}^\star$ the corresponding reference token. We denote the original and calibrated vocabulary distributions at this position by p_i and $\widetilde{p}_i$, evaluated under the same input prefix and decoding transformations.

Answer supervision. The cross entropy loss supervises the calibrated prediction using the reference token:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{n} \sum_{i=1}^{n} \log \widetilde{p}_i(y_{i,t_i}^\star). \tag{13}$$

Prediction preservation. We penalize changes to the original prediction distribution on correctly answered training examples:

$$\mathcal{L}_{\text{preserve}} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} D_{\text{KL}}(p_i \,\|\, \widetilde{p}_i)\,, \tag{14}$$

where $\mathcal{C}$ indexes these examples and D_{KL} is the Kullback–Leibler divergence; the loss is zero when $\mathcal{C}$ is empty.

Adjustment regularization. We encourage small corrections by penalizing the squared adjustment norm relative to the hidden state norm:

$$\mathcal{L}_{\text{norm}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\|\Delta \mathbf{h}_i\|_2^2}{\|\mathbf{h}_i\|_2^2}, \tag{15}$$

where $\mathbf{h}_i$ and $\Delta \mathbf{h}_i$ denote the hidden state and its adjustment at position t_i.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on three datasets: AudioHallucination (Kuan et al., 2024), HalluAudio (Zhao et al., 2026), and Clotho-AQA (Lipping et al., 2022). AudioHallucination evaluates sound object presence through binary questions with random, popular, and adversarial negative sampling. HalluAudio evaluates hallucinations across speech, environmental sounds, and music using binary, multiple-choice, and open-ended tasks. Clotho-AQA contains crowdsourced questions about environmental recordings, covering both yes/no and other open-ended questions.

Table 1: Performance on three audio benchmarks (%). Higher is better for all metrics. **Bold** indicates the best result within each model and metric.

	AudioHallucination						HalluAudio			Clotho-AQA
	Adversarial		Popular		Random		Speech	Sound	Music	
Method	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	Acc.	Acc.	Acc.
Qwen2-Audio-7B-Instruct										
Base	48.6	24.7	50.7	30.5	55.9	28.1	34.6	59.7	48.6	49.0
Prompt-only	46.8	27.8	50.3	36.1	59.2	39.0	32.4	60.7	58.1	51.4
AAD	45.8	36.2	49.7	43.7	70.1	62.8	32.5	62.0	50.1	53.0
AttnAdapter	48.5	29.6	50.6	34.7	62.4	44.5	34.1	61.7	55.9	52.3
AHA	68.7	71.7	75.3	79.4	**79.6**	**81.3**	37.3	62.3	57.7	51.3
ARWR-Cal (Ours)	**74.2**	**73.7**	**81.5**	**82.5**	70.0	71.8	**62.8**	**85.6**	**72.0**	**55.1**
SALMONN-7B										
Base	51.7	12.3	52.6	16.5	54.1	16.3	38.0	48.2	60.1	49.5
Prompt-only	51.8	22.0	56.2	35.0	60.9	36.6	34.2	51.2	54.2	50.5
AAD	52.3	29.4	58.3	43.7	63.3	43.1	35.8	51.3	61.4	51.2
AttnAdapter	51.1	20.3	54.2	30.0	61.6	38.8	39.0	52.1	49.4	48.5
AHA	64.8	61.8	74.8	75.5	**74.2**	**71.0**	36.3	54.1	54.1	51.3
ARWR-Cal (Ours)	**67.5**	**69.1**	**76.1**	**79.0**	65.0	68.9	**65.6**	**81.4**	**73.1**	**52.8**
SALMONN-13B										
Base	53.6	26.5	57.8	38.1	61.2	37.5	42.8	62.3	58.0	52.3
Prompt-only	52.4	26.4	56.3	37.1	67.1	51.5	42.9	60.3	59.8	50.9
AAD	51.6	32.0	56.8	44.0	71.1	60.1	42.9	62.0	60.8	51.6
AttnAdapter	51.9	29.9	54.7	37.2	67.3	53.2	40.3	60.0	59.5	48.9
AHA	62.4	63.4	71.6	74.7	**78.6**	**78.5**	42.7	60.3	59.6	52.8
ARWR-Cal (Ours)	**69.5**	**69.2**	**79.3**	**81.1**	69.3	70.7	**64.4**	**85.7**	**72.7**	**54.0**

Models and baselines. We evaluate three LALMs: Qwen2-Audio-7B-Instruct (Chu et al., 2024), SALMONN-7B, and SALMONN-13B (Tang et al., 2024). We compare ARWR-Cal with five baselines: Base, Prompt-only, Audio-Aware Decoding (AAD) (Hsu et al., 2025), AttnAdapter (Du & Zhong, 2026), and AHA (Chen et al., 2026). Detailed baseline settings are provided in the appendix.

Implementation details. We fit a separate ARWR-Cal calibrator for each model and dataset. The calibrator uses a bottleneck dimension of $k = 8$, a fixed 32-dimensional projection for each input vector, and a base relative correction scale of $\rho = 0.2$. In the loss function, we set $\lambda_{\mathrm{p}} = 2$ and $\lambda_n = 1$. We use the 75th training percentile of ARWR as the default budget threshold. The calibrator is trained for up to 20 epochs using AdamW with a learning rate of 3×10^{-3}, weight decay of 10^{-3}, and batch size 24. Experiments are conducted on two NVIDIA A100 (80GB) GPUs.

5.2 Main Results

Table 1 compares ARWR-Cal with the baselines across three datasets and three models.

AudioHallucination. ARWR-Cal achieves the highest accuracy and F1 on the adversarial and popular subsets for all three backbones. Relative to AHA, the strongest baseline in these settings, accuracy increases by 1.3–7.7 percentage points and F1 by 2.0–7.3 points. The improvements in both metrics indicate better recognition of absent sound objects alongside higher overall answer accuracy in these two settings. On the random subset, ARWR-Cal improves over Base for every backbone, but AHA retains the highest accuracy and F1, showing that the relative advantage depends on the negative sampling strategy.

HalluAudio. ARWR-Cal achieves the highest accuracy in all nine combinations of backbone and audio domain. Compared with the strongest baseline for each combination, accuracy improves by 21.5–26.6 percentage points on speech, 23.3–27.3 points on sound, and 11.7–13.9 points on music.

Table 2: Ablation results for Qwen2-Audio-7B-Instruct on three audio benchmarks (%). *w/o Adaptive Budget* fixes $b_t = 0.5$. *w/o Source Features* predicts the residual from hidden-state features alone while retaining the ARWR-guided budget. Higher is better for all metrics.

	AudioHallucination						HalluAudio			Clotho-AQA
	Adversarial		Popular		Random		Speech	Sound	Music	
Method	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	Acc.	Acc.	Acc.
w/o Adaptive Budget	70.7	68.7	75.3	74.9	**77.4**	**77.5**	53.6	83.2	68.7	51.3
w/o Source Features	67.0	64.5	73.8	73.8	72.5	70.6	54.3	76.4	65.4	53.5
ARWR-Cal (Ours)	**74.2**	**73.7**	**81.5**	**82.5**	70.0	71.8	**62.8**	**85.6**	**72.0**	**55.1**

The consistent gains across both model families and all three domains support the effectiveness of ARWR-guided residual calibration for hallucination mitigation across diverse audio content.

Clotho-AQA. ARWR-Cal achieves the highest accuracy for all three backbones: 55.1% for Qwen2-Audio-7B-Instruct, 52.8% for SALMONN-7B, and 54.0% for SALMONN-13B. These results exceed the strongest respective baselines by 2.1, 1.5, and 1.2 percentage points, with gains over Base ranging from 1.7 to 6.1 points. The gains on Clotho-AQA support the applicability of ARWR-Cal to broader audio question answering.

5.3 Ablation Study

We conduct two ablation studies on Qwen2-Audio-7B-Instruct and SALMONN-7B across three datasets to evaluate the contributions of the ARWR-guided correction budget and the audio and nonaudio source vectors used as calibrator inputs. **w/o Adaptive Budget** replaces the ARWR-guided budget with a constant $b_t = 0.5$, while retaining all calibrator inputs. **w/o Source Features** removes features derived from the audio and nonaudio source vectors from the calibrator input, retaining only features of the final hidden state. Table 2 reports the results for Qwen2-Audio-7B-Instruct. Full results are provided in Appendix C.1.

Effect of the adaptive budget. On HalluAudio, the ARWR-guided budget improves accuracy over the fixed budget by 9.2, 2.4, and 3.3 percentage points on speech, sound, and music, respectively. It also improves adversarial and popular accuracy on AudioHallucination by 3.5 and 6.2 points. These gains support allowing larger corrections at lower ARWR in these settings.

Effect of source features. Including audio and nonaudio source features improves HalluAudio accuracy by 6.6–9.2 percentage points across the three domains. On the adversarial and popular subsets of AudioHallucination, accuracy increases by 7.2 and 7.7 points and F1 by 9.2 and 8.7 points, respectively. With the budget rule unchanged, these gains support using source features alongside the hidden state for residual prediction.

Meanwhile, the complete model does not lead in every setting. On the random subset of AudioHallucination, the fixed budget achieves 77.4% accuracy and 77.5% F1, compared with 70.0% and 71.8% for the complete model. On Clotho-AQA, the complete model yields the best ablation result of 55.1%, compared with 53.5% when removing source features. These differences show that the benefits of adaptive budgets and source features are task dependent, highlighting the need to balance correction capacity with preservation of useful predictions.

5.4 Evaluating ARWR for Hallucination Risk Assessment

To evaluate whether ARWR reflects hallucination risk more effectively than audio attention mass, we randomly sample 1,000 answers generated by Qwen2-Audio-7B-Instruct on Clotho-AQA. For each metric, we independently rank the same answers by decreasing score and divide them into five groups of 200, ordered from low to high predicted risk. We compute the mean diagnostic score and the proportion of answers labeled as hallucinations in each group, keeping the hallucination labels

identical across the two comparisons. Figure 3 plots these groupwise scores and hallucination rates.

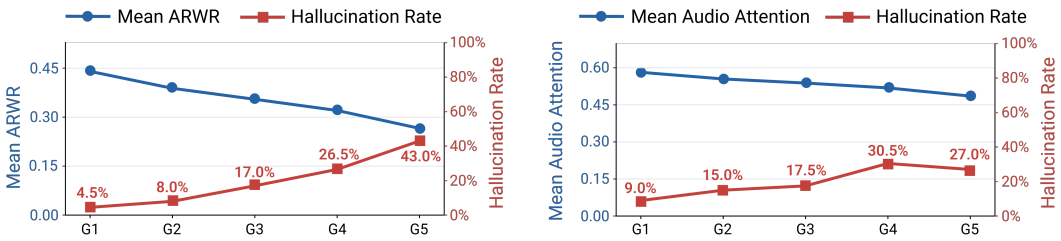

Figure 3: **Hallucination rates across risk groups.** We randomly sample 1,000 answers generated by Qwen2-Audio-7B-Instruct on Clotho-AQA. **Left:** Answers are ranked by decreasing ARWR and divided into five groups of 200; the curves show the mean ARWR and observed hallucination rate in each group. **Right:** The same answers are independently ranked by decreasing audio attention mass and divided into five groups of 200; the curves show the mean audio attention mass and observed hallucination rate in each group.

In addition to the groupwise analysis, we compute AUROC and AP on the same individual answers to quantify risk discrimination. AUROC reflects the overall ability to distinguish hallucinated from non-hallucinated answers across risk thresholds, while AP reflects how effectively hallucinated answers are concentrated toward the top of the risk ranking. Higher values of both metrics indicate better hallucination risk discrimination.

Results and analysis. As ARWR decreases, the hallucination rate rises monotonically from 4.5% in the lowest predicted risk group to 43.0% in the highest. Attention-based grouping also reveals higher hallucination rates toward the high-risk end, but the rate peaks at 30.5% in the fourth group and falls to 27.0% in the fifth. At the same selection size of 200 answers, the highest-risk ARWR group has a hallucination rate 16.0 percentage points higher than its attention counterpart. These results show that ARWR provides a more consistent ordering of risk and concentrates more hallucinations in the highest-risk group on this sample. This finding supports the diagnostic value of ARWR, which incorporates value aggregation and output projection, rather than relying solely on attention allocation.

The answer-level metrics in Table 3 further support this observation. ARWR achieves an AUROC of 0.7431 and AP of 0.3670, compared with 0.6295 and 0.2600 for audio attention mass. The paired AUROC difference is 0.1135, with a 95% confidence interval of [0.0824, 0.1448]. These results indicate that ARWR provides stronger discrimination of false-affirmative risk.

Table 3: Answer-level hallucination risk discrimination on the same answers used in Figure 3.

Diagnostic	AUROC ↑	AP ↑
Audio attention mass	0.6295	0.2600
ARWR	**0.7431**	**0.3670**

6 Conclusion

In this paper, we present ARWR-Cal, a framework for mitigating hallucinations in LALMs through residual calibration. We introduce ARWR to address the limitations of attention weights in characterizing audio contributions by measuring the relative strength of audio writes after value aggregation and output projection. Building on this measure, ARWR-Cal sets a correction budget that bounds the residual adjustments predicted by a lightweight calibrator while keeping the LALM frozen. Across three datasets and three LALMs, ARWR-Cal improves average accuracy by 8.2 percentage points over AHA, the strongest baseline overall, demonstrating its effectiveness in improving audio question answering. Future work will refine correction budgets to better preserve correct predictions and evaluate robustness under distribution shifts.

AI USE STATEMENT

In this work, we used generative AI tools to polish the language of the manuscript, including improving grammar, clarity, and readability. We have not used generative AI tools for any other tasks in this work. We have reviewed all AI-assisted work to ensure accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide a link to an anonymous code repository in the abstract. The appendix provides further implementation details, including baseline implementations, hyperparameter configurations, supplementary training settings, and random seeds, to support reproduction of our experiments.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, December 2023. URL https://aclanthology.org/2023.findings-emnlp.68/.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms'internal states retain the power of hallucination detection. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 3056–3076, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0d1986a61e30e5fa408c81216a616e20-Paper-Conference.pdf.

Yanxi Chen, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Xin Li, Peijie Qiu, Hao Wang, Xuanzhao Dong, Yujian Xiong, Anderson Schneider, Yuriy Nevmyvaka, and Yalin Wang. AHA: Aligning large audio-language models for reasoning hallucinations via counterfactual hard negatives. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29294–29306, July 2026. doi: 10.18653/v1/2026.findings-acl.1464. URL https://aclanthology.org/2026.findings-acl.1464/.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1419–1436, November 2024. URL https://aclanthology.org/2024.emnlp-main.84/.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=gJLAfO4KUq.

Zixuan Dong, Jiafu Tang, Zhide Lei, Zhe Cao, Zijie Zhang, Yanghai Wang, Shihao Li, Xiaodong Wang, Baoyun Peng, and Jiaheng Liu. Omnihalluc-l: Counterfactual benchmarking and modality-perturbation reliability calibration for long-form omni hallucination, 2026. URL https://arxiv.org/abs/2606.03614.

Hangyu Du and Jingxing Zhong. Keeping models listening: Segment- and time-aware attention rescaling at decoding time. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16267–16271, 2026. doi: 10.1109/ICASSP55912.2026.11460793.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. URL https://www.nature.com/articles/s41586-024-07421-0.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nBZBPXdJlC.

Tzu-Wen Hsu, Ke-Han Lu, Cheng-Han Chiang, and Hung-Yi Lee. Reducing object hallucination in large audio-language models via audio-aware decoding. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, 2025. doi: 10.1109/ASRU65441.2025.11434595.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. AVCD: Mitigating hallucinations in audio-visual large language models through contrastive decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=3zx087XPtz.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075, November 2020. doi: 10.18653/v1/2020.emnlp-main.574. URL https://aclanthology.org/2020.emnlp-main.574/.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4547–4568, November 2021. URL https://aclanthology.org/2021.emnlp-main.373/.

Chun-Yi Kuan, Wei-Ping Huang, and Hung yi Lee. Understanding Sounds, Missing the Questions: The Challenge of Object Hallucination in Large Audio-Language Models. In *Interspeech 2024*, pp. 4144–4148, 2024. doi: 10.21437/Interspeech.2024-1076.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-aqa: A crowdsourced dataset for audio question answering, 2022. URL https://arxiv.org/abs/2204.09634.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, December 2023. URL https://aclanthology.org/2023.emnlp-main.557/.

Jiacheng Pang, Ashutosh Chaubey, and Mohammad Soleymani. Do audio LLMs listen or read? analyzing and mitigating paralinguistic failures with voxparadox. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=v7rYbRR9Zw.

Ashish Seth, Sonal Kumar, Ramaneswaran Selvakumar, Nishit Anand, Utkarsh Tyagi, Prem Seetharaman, Ramani Duraiswami, and Dinesh Manocha. Audio hallucination attacks: Probing the reliability of large audio language models, 2026. URL https://arxiv.org/abs/2603.29263.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 16607–16629, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/476ab8f369e489c04187ba84f68cfa68-Paper-Conference.pdf.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM can see? dynamic correction decoding for hallucination mitigation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4z3IguA4Zg.

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

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, and Hao Tang. Hallucination begins where saliency drops. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sjnErRHXf3.

Feiyu Zhao, Yiming Chen, Wenhuan Lu, Daipeng Zhang, Xianghu Yue, and Jianguo Wei. HalluAudio: A comprehensive benchmark for hallucination detection in large audio-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38797–38816, July 2026. doi: 10.18653/v1/2026.acl-long.1797. URL https://aclanthology.org/2026.acl-long.1797/.

A Dataset Details

AudioHallucination. AudioHallucination (Kuan et al., 2024) evaluates object hallucination through binary questions about whether a particular sound is present in a recording. The benchmark is constructed from AudioCaps recordings with AudioSet sound-event labels: positive questions refer to annotated sound objects, while negative questions refer to objects absent from the annotations. It uses three negative sampling strategies. *Random* sampling selects absent objects at random; *popular* sampling selects the most frequent absent objects in the dataset; and *adversarial* sampling selects absent objects that frequently co-occur with the annotated objects. The original construction balances positive and negative questions within each setting. We report accuracy and F1 separately for all three settings to examine performance under different types of negative queries.

HalluAudio. HalluAudio (Zhao et al., 2026) contains over 5,000 human-verified question–answer pairs spanning speech, environmental sounds, and music. Its recordings are drawn from Common Voice, FSD50K, GTZAN, and the Mridangam Strokes and Tonics datasets. The benchmark combines binary judgments, multiple-choice questions, and open-ended questions to probe sound presence, temporal order, counting, relative loudness, and musical attributes. Questions are constructed from source annotations and augmented with mismatched attributes, adversarial prompts, and mixed-audio conditions to test whether answers remain grounded in the recording. We report accuracy separately for speech, sound, and music to assess how calibration behaves across audio domains.

Clotho-AQA. Clotho-AQA (Lipping et al., 2022) is a crowdsourced audio question answering dataset built from 1,991 environmental recordings selected from Clotho, each lasting 15–30 seconds. Six questions are collected for each recording: two intended to elicit "yes," two "no," and two other single-word answers. Each question receives answers from three annotators who are distinct from the question authors, providing multiple references that can capture differences in interpretation and answer specificity. The combination of binary and non-binary questions complements the hallucination benchmarks by assessing calibration on broader questions about everyday sounds.

Data splits. We partition the data by audio and reserve 10% as a calibration pool to evaluate ARWR-Cal with limited task-specific supervision. This pool is further divided into training and validation sets at an 80:20 ratio, while the remaining data are used for testing. All questions associated with the same audio are kept in the same partition. We fit a separate calibrator for each backbone–dataset pair while keeping the backbone frozen throughout.

B Implementation and Training Details

B.1 Baseline Implementations

Base provides the original recording and question using each model's native input template and generates answers through standard decoding without additional grounding instructions or correction. Prompt-only prepends the instruction *"Focus on the given audio and answer the following question."* to the original question, while retaining the remaining Base settings.

For AAD (Hsu et al., 2025), we follow the authors' official decoding implementation. For AttnAdapter (Du & Zhong, 2026), we implement the attention rescaling algorithm described by the authors, which adjusts attention across input segments and decoding steps. Following their validation-based tuning approach, we conduct a predefined two-stage search for each backbone. We select the configuration with the highest mean accuracy across the three datasets' validation sets and keep it fixed during test evaluation.

AHA adaptation. We adapt AHA (Chen et al., 2026) to our three datasets by constructing preference pairs from their existing training partitions instead of using the original AHA training corpus. Each pair retains the original recording and question, with a reference answer as the preferred response. For questions with explicit labels, we construct contradictory responses directly, such as reversing an unambiguous Yes/No label. For other questions, GPT-4o generates rejected responses using the official AHA generation prompt adapted to keep the question and reference answer fixed. Generation is grounded in available audio captions or, when captions are unavailable, facts from the training questions and answers. We discard candidates that alter the question or reference answer, or match a reference under our equivalence checks, and retain one rejected response per distinct reference with its annotation frequency as the weight. Clotho-AQA questions with both Yes and No as valid references are excluded from preference training; validation and test partitions remain unchanged.

AHA training. We implement AHA's direct preference optimization objective and train rank-16 LoRA adapters in the decoder attention and feedforward projections. The pretrained parameters, including the original SALMONN adapters, remain frozen and define the reference policy. The objective uses summed answer token log probabilities, including the end-of-sequence token, with preference coefficient $\beta = 0.3$. Training uses AdamW with a learning rate of 2×10^{-6} and an effective batch size of 16 for eight epochs. Checkpoint selection uses answer accuracy on the original validation partition.

B.2 Supplementary ARWR-Cal Training Details

We cache source features and final hidden states from the frozen backbone and optimize the calibrator through the frozen readout. The calibrator's output projection is initialized to zero, so training starts from the original model predictions. Feature normalization statistics and the 75th-percentile ARWR threshold are estimated only from the training partition.

For a D-layer decoder, we use $\mathcal{L} = [D/2, 3D/4)$ as a fixed heuristic to observe audio contributions between initial input processing and final answer readout. By examining layers beyond the early stages, we aim to measure audio writes after the prediction query has been contextualized by the question and generated prefix. We also avoid relying solely on the deepest layers, where acoustic evidence may already be carried by nonaudio representations.

Table 4 summarizes the default configuration.

B.3 Evaluation Metrics

Accuracy (Acc.). Accuracy measures the proportion of predictions that agree with the reference answers:

$$\text{Acc.} = \frac{1}{M} \sum_{j=1}^{M} \mathbb{I}[\hat{y}_j = y_j], \tag{16}$$

Table 4: Default ARWR-Cal hyperparameters.

Hyperparameter	Value
Random seed	42
Intermediate dimension k	8
Correction scale ρ	0.2
ARWR threshold quantile	0.75
Decoder-layer set $\mathcal{L}$	$[D/2, 3D/4)$
Optimizer	AdamW
Learning rate	3×10^{-3}
Weight decay	10^{-3}
Maximum training epochs	20
Question batch size	24
Adjustment regularization weight λ_n	1.0

where M is the number of prediction–reference comparisons, $\hat{y}_j$ and y_j are the extracted prediction and reference label, and $\mathbb{I}[\cdot]$ is the indicator function.

F1-score. For AudioHallucination, we report F1 with *no* as the positive class to assess recognition of absent sound objects. F1 is the harmonic mean of precision P_{no} and recall R_{no}:

$$\text{F1} = \frac{2P_{\text{no}}R_{\text{no}}}{P_{\text{no}} + R_{\text{no}}}. \tag{17}$$

Here, P_{no} is the fraction of no predictions that are correct, and R_{no} is the fraction of no-reference questions correctly predicted as no. Ratios with zero denominators are set to zero; higher F1 indicates a better balance between precision and recall.

AUROC. AUROC measures how well a diagnostic ranks answers labeled as hallucinations above other answers in terms of risk. Let s_i denote the ARWR or audio attention mass score and $h_i \in \{0, 1\}$ the hallucination label. For binary questions with unanimous yes/no references and parsable predictions, $h_i = 1$ when the prediction is yes and all references are no, and $h_i = 0$ otherwise. We use $z_i = -s_i$ as the risk score and compute

$$\text{AUROC} = \frac{1}{n_1 n_0} \sum_{i:h_i=1} \sum_{j:h_j=0} \left(\mathbb{I}[z_i > z_j] + \frac{1}{2}\mathbb{I}[z_i = z_j] \right), \tag{18}$$

where n_1 and n_0 are the numbers of answers with labels 1 and 0, respectively, and ties receive half credit. Higher AUROC indicates better risk discrimination, with 0.5 corresponding to chance-level ordering.

Average precision (AP). AP summarizes precision across recall levels as answers are selected in decreasing order of risk:

$$\text{AP} = \sum_{k=1}^{K} (R_k - R_{k-1}) P_k, \qquad R_0 = 0, \tag{19}$$

where P_k and R_k are hallucination-detection precision and recall after including all answers at or above the k-th distinct risk threshold, with thresholds traversed from high to low. Higher AP indicates better retrieval of hallucinations from the risk ranking.

C Additional Experimental Results

C.1 Full Results of Ablation Studies

Table 5 presents the ablation results for Qwen2-Audio-7B-Instruct and SALMONN-7B across the three datasets, using the variants defined in Section 5.3. The SALMONN-7B results complement the Qwen2-Audio analysis in the main text and show how the effects of the two components vary between these backbones.

Table 5: Ablation results for Qwen2-Audio-7B-Instruct and SALMONN-7B on three audio benchmarks (%). *w/o Adaptive Budget* fixes $b_t = 0.5$. *w/o Source Features* predicts the residual from hidden-state features alone while retaining the ARWR-guided budget. Higher is better for all metrics; bold denotes the best result within each model and metric.

	AudioHallucination						HalluAudio			Clotho-AQA
	Adversarial		Popular		Random		Speech	Sound	Music	
Method	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	Acc.	Acc.	Acc.
Qwen2-Audio-7B-Instruct										
w/o Adaptive Budget	70.7	68.7	75.3	74.9	**77.4**	**77.5**	53.6	83.2	68.7	51.3
w/o Source Features	67.0	64.5	73.8	73.8	72.5	70.6	54.3	76.4	65.4	53.5
ARWR-Cal (Ours)	**74.2**	**73.7**	**81.5**	**82.5**	70.0	71.8	**62.8**	**85.6**	**72.0**	**55.1**
SALMONN-7B										
w/o Adaptive Budget	67.3	67.5	**78.6**	**81.0**	64.1	66.2	49.0	71.5	51.1	51.8
w/o Source Features	62.4	55.9	71.4	69.8	64.3	57.4	51.0	76.5	55.4	51.1
ARWR-Cal (Ours)	**67.5**	**69.1**	76.1	79.0	**65.0**	**68.9**	**65.6**	**81.4**	**73.1**	**52.8**

Effect of the ARWR-guided budget. With SALMONN-7B, replacing the fixed budget with the ARWR-guided budget increases HalluAudio accuracy by 16.6, 9.9, and 22.0 percentage points on speech, sound, and music, respectively. These gains exceed the corresponding Qwen2-Audio improvements of 9.2, 2.4, and 3.3 points. On AudioHallucination, however, SALMONN-7B gains only 0.2 points in adversarial accuracy and 1.6 points in F1, while popular accuracy and F1 decrease by 2.5 and 2.0 points, respectively. The contrast indicates that the benefit of the ARWR-guided correction allowance depends on the task as well as the backbone.

Effect of audio and nonaudio source features. For SALMONN-7B, including source features raises adversarial and popular accuracy by 5.1 and 4.7 percentage points, and F1 by 13.2 and 9.2 points. On HalluAudio, the corresponding gains are 14.6 points on speech, 4.9 on sound, and 17.7 on music. Together with the Qwen2-Audio results, these comparisons support providing source vectors alongside the hidden state to predict the residual adjustment. The contribution is most pronounced on speech and music for SALMONN-7B, while its smaller sound gain shows that source features do not contribute equally across domains.

C.2 Supplementary Hallucination Risk Assessment Experiments

We supplement the main risk-group experiment with two additional configurations on Clotho-AQA using Qwen2-Audio-7B-Instruct: random sampling and audio-balanced sampling, both with seed 20260920. Each configuration selects 1,000 answers without replacement from the same eligible pool, using neither diagnostic scores nor hallucination labels to determine selection. Audio-balanced sampling distributes questions as evenly as possible across recordings; with 1,743 eligible recordings, this sample contains exactly one answer from each of 1,000 distinct recordings. We retain the same labels, score computation, and five groups of 200 answers as in the main experiment, and compute AUROC and AP as described in Appendix B.3.

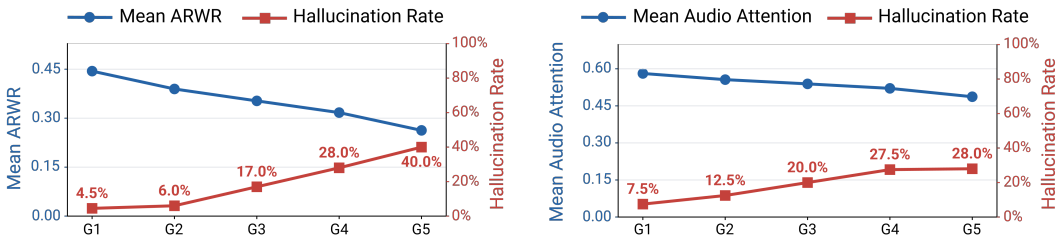

Figure 4: **Hallucination risk assessment under random sampling.** We sample 1,000 answers generated by Qwen2-Audio-7B-Instruct on Clotho-AQA. The same answers are independently ranked by decreasing ARWR and audio attention mass and divided into five groups of 200. The curves show the mean diagnostic score and hallucination rate within each group.

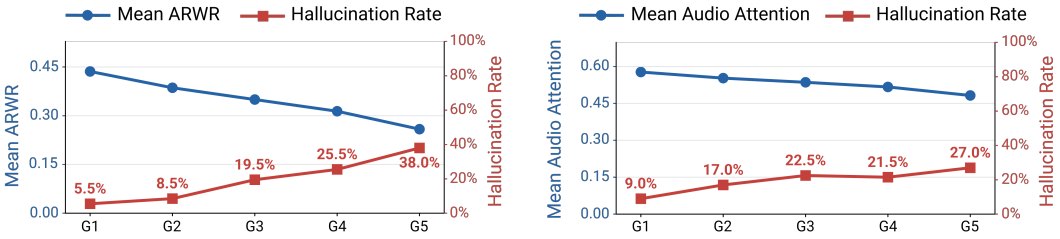

Figure 5: **Hallucination risk assessment under audio-balanced sampling.** We sample one answer from each of 1,000 distinct Clotho-AQA recordings, with answers generated by Qwen2-Audio-7B-Instruct. The same answers are independently ranked by decreasing ARWR and audio attention mass and divided into five groups of 200. The curves show the mean diagnostic score and hallucination rate within each group.

Table 6: Supplementary risk assessment on Clotho-AQA with Qwen2-Audio-7B-Instruct. AUROC and AP use the 0–1 scale. Higher is better.

Diagnostic	N	AUROC	AP
Random sampling			
Attention mass	1,000	0.6474	0.2699
ARWR	1,000	**0.7436**	**0.3618**
Audio-balanced sampling			
Attention mass	1,000	0.6129	0.2675
ARWR	1,000	**0.7131**	**0.3486**

Results and analysis. Figures 4 and 5 show that ARWR yields monotonically increasing hallucination rates under both sampling configurations. For the additional random sample, the rate rises from 4.5% to 40.0%, and the highest-risk group exceeds its attention-based counterpart by 12.0 percentage points. Under audio-balanced sampling, the hallucination rate increases from 5.5% to 38.0% as ARWR decreases, with the highest-risk rate exceeding that of attention mass by 11.0 percentage points. Table 6 further shows that ARWR achieves higher AUROC and AP under both configurations, with gains of 0.0962–0.1002 and 0.0810–0.0919, respectively.

D Correction Cases and Error Analysis

To examine when ARWR-Cal corrects an erroneous answer and when it alters a correct prediction, we compare Base, Prompt-only, and ARWR-Cal on Clotho-AQA using Qwen2-Audio-7B-Instruct. All three methods use the same recordings and questions with greedy decoding; Prompt-only additionally controls for the audio-grounding instruction used by ARWR-Cal. Table 7 presents four test examples with unanimous binary reference answers.

Selective correction within the same recording. Cases A and B ask about different sound events in the same recording. In Case A, both Base and Prompt-only incorrectly report paper being ripped, whereas ARWR-Cal rejects this unsupported event. In Case B, ARWR-Cal preserves the correct answer that an alarm rings. The contrast illustrates the desired balance between error correction and prediction preservation: an erroneous claim can be removed while a supported claim about the same recording is retained. This behavior is consistent with conditioning the residual on the current source vectors and decoder state, which allows the adjustment to depend on the question and prediction context.

Corrections in both answer directions. In Case C, both baselines deny that the engine slowly idles down, while ARWR-Cal changes the answer to yes, matching the references. Together with the yes-to-no correction in Case A, this example shows that successful calibration can require opposite changes in the answer decision. This motivates the separation between correction magnitude and direction in ARWR-Cal: the ARWR score determines the permitted adjustment magnitude, while

Table 7: Correction and error cases on Clotho-AQA with Qwen2-Audio-7B-Instruct. Ref. denotes the unanimous reference answer. Base and Prompt-only columns show answer labels; ARWR-Cal responses are shown in full. Cases A–B share one recording, and C–D share another.

Case	Question	Ref.	Base	Prompt-only	ARWR-Cal response
A	Is paper being ripped?	No	Yes	Yes	No, paper is not being ripped.
B	Does an alarm ring?	Yes	Yes	Yes	Yes, an alarm rings.
C	Does the engine slowly idle down?	Yes	No	No	Yes, the engine slowly idles down.
D	Does the engine stop making sound?	No	No	No	Yes, the engine stops making sound at the end of the recording.

the learned residual determines how the prediction changes. Suppressing affirmative answers alone would fail to recover the temporal change queried in Case C.

Misjudging closely related temporal properties. Case D uses the same recording as Case C but asks whether the engine stops making sound. Both baselines correctly answer no, whereas ARWR-Cal changes the answer to yes. An engine gradually idling down does not imply that it becomes silent, so correcting one temporal judgment does not ensure correctness on a closely related question. The paired cases expose a limitation of magnitude control: bounding a residual does not ensure that its direction preserves distinctions between related claims. They also illustrate why prediction preservation matters alongside error correction, particularly when similar acoustic evidence supports different answers to different questions.

E Discussion

E.1 Limitations

- **High ARWR does not guarantee correctness.** ARWR summarizes the relative magnitude of source writes after value aggregation and output projection. This magnitude alone cannot determine whether the acoustic information supports a particular claim or whether its contribution favors a correct prediction. Groups with lower ARWR scores have higher hallucination rates. However, these results do not show that increasing audio write strength will make an answer correct.
- **Supervision and model access.** ARWR-Cal requires supervised calibration for each backbone and dataset pair and access to internal model states. Although the backbone remains frozen, the calibrator and its feature normalization statistics are fitted to the corresponding training distribution. Their effectiveness may therefore depend on the audio domain and question distribution, and transfer to unseen domains remains untested. Access to source vectors and the final hidden state also limits direct use with models available only through APIs that return generated text.
- **Prediction preservation.** The norm constraint controls correction magnitude but does not guarantee preservation of an initially correct answer. Even a small hidden-state adjustment can change the selected token when competing candidates have similar logits. The prediction-preservation loss discourages such changes on correctly answered training examples, but offers no guarantee for unseen inputs. The mixed results across model and dataset combinations in Table 1 highlight the need to balance correction with prediction preservation.

E.2 Future Work

- **Cross-domain transfer.** Training on multiple audio datasets could help ARWR-Cal learn calibration patterns shared across domains. Currently, we fit a separate calibrator for each backbone–dataset pair, so its correction behavior is learned from a specific task distribution. We plan to train a shared calibrator across datasets for each backbone and adapt it to new tasks with limited additional supervision. This could reduce dependence on task-specific labels and make calibration easier to transfer across audio domains.

- **Causal validation of audio contributions.** ARWR captures relative audio write strength, but it does not establish which acoustic cues support the model's answer. A causal analysis could selectively modify relevant sounds and intervene on their source writes to examine how these changes affect the answer. Comparing these interventions with perturbations of irrelevant sounds would help identify the contributions that drive the model's decision. This could guide refinements to ARWR that better reflect the evidence needed to answer the question.
- **Adaptive correction budgets.** ARWR-Cal currently sets the correction budget through a fixed mapping from ARWR, without considering how sensitive the current prediction is to an adjustment. We plan to learn a more flexible mapping that also considers the stability of candidate predictions. In particular, the minimum allowance and its growth as ARWR decreases could be adapted to the input. The aim is to retain enough correction strength for unsupported answers while reducing unnecessary changes to answers that are already correct.