EvA: An Evidence-First Audio Understanding Paradigm for LALMs

Anonymous ACL submission

Abstract

Large Audio Language Models (LALMs) still struggle in complex acoustic scenes, primarily due to weak perceptual grounding. We identify an evidence bottleneck: state-of-the-art systems often fail more on evidence extraction than on downstream reasoning, indicating that the core limitation lies in upstream acoustic representation rather than reasoning policy. To address this, we propose EvA (Evidence-First Audio), a dual-path architecture that combines Whisper and CED-Base with non-compressive, time-aligned fusion. EvA first performs hierarchical cross-layer aggregation within CED to retain multi-scale acoustic cues, then aligns aggregated features to the Whisper timeline and integrates them via inject-and-add fusion without changing sequence length. We further build EvA-Perception, a large-scale open-source training set with about 54K event-ordered captions (150 h) and about 500K QA pairs. Under a unified 0-shot protocol, EvA achieves the best open-source *Perception* scores on MMAU/MMAR/MMSU and improves over Kimi-Audio-7B across all reported metrics, with the largest gains on perception-heavy splits. These results support the evidence-first hypothesis that improving upstream evidence fidelity is key to stronger audio understanding.

1 Introduction

Large Audio Language Models (LALMs) aim to empower machines with the ability to listen, understand, and reason from sound. While recent systems like Qwen2.5-Omni (Xu et al., 2025) and Kimi-Audio (Ding et al., 2025) have demonstrated impressive performance on various benchmarks, their capabilities degrade sharply when confronted with complex acoustic scenes involving overlapping events, faint signals, or fine-grained temporal cues.

We argue this fragility stems not from a deficit in reasoning but from a more fundamental breakdown

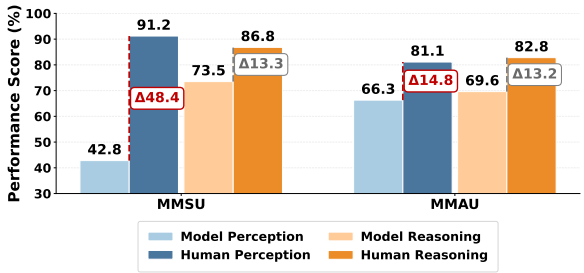

Figure 1: Perception–reasoning performance gap comparison between model and human. Model performance are averaged over Qwen2.5–Omni and Kimi–Audio–7B.

in acoustic perception. As illustrated in Figure 1, the performance gap between leading LALMs and humans is larger for perception-centric tasks than for reasoning-centric ones. On the MMSU benchmark, the model-human gap in perception is a stark 48.4 points (42.8% vs. 91.2%), whereas the reasoning gap is a comparatively narrow 13.3 points. This disparity reveals what we term the "evidence bottleneck": the primary performance ceiling is not a model's cognitive capacity, but its foundational inability to accurately perceive and represent acoustic evidence for subsequent reasoning.

We posit this "evidence bottleneck" is a systemic limitation rooted in information-theoretic principles, arising from three flaws in the prevailing LALM design. First, **reasoning-centric optimization**: Post-training methods like SFT and GRPO refine textual reasoning but are fundamentally incapable of correcting upstream perception errors. Once acoustic details are lost in the initial encoding stages, no downstream process can recover them. This places a hard ceiling on performance, irrespective of the reasoning model's power. Second, **frequency information attrition**: Dominant encoders process spectrograms as 1D sequences, collapsing the frequency dimension and discarding crucial spectral cues. Third, **flawed alignment interfaces** that manifest in two modes: lossy temporal com-

pression via modules like Q-Formers (Tang et al., 2023), or misaligned feature concatenation that provides no unified temporal coordinate for the LLM to exploit (Ghosh et al., 2024).

To systematically dismantle this bottleneck, we introduce **EvA (Evidence-First Audio)**, an architectural paradigm designed for **maximal evidence preservation**. EvA processes audio through two parallel, complementary streams: a Whisper encoder for phonetic content and a generalist audio encoder (CED-Base (Dinkel et al., 2024)) for the rich tapestry of non-speech sounds. Its core innovation is a two-stage, non-compressive fusion process that directly counteracts the information loss identified above. First, it performs **hierarchical evidence aggregation** within the generalist encoder, fusing features from intermediate layers to create a multi-scale representation that retains both local and global acoustic details. Second, this enriched acoustic sequence is integrated with the speech features via a **time-aware additive fusion**, which preserves the full temporal resolution of both streams and provides the LLM with a unified, high-fidelity timeline of all acoustic events.

To realize this architecture's potential, we develop EvA-Perception, a large-scale audio QA dataset of 54K high-fidelity captions and 500K QA pairs. Generated using temporal annotations from AudioSet-Strong (Hershey et al., 2021), it is explicitly designed to train models to ground reasoning in fine-grained acoustic evidence. Our model, fine-tuned on only 380 hours of audio, achieves strong and consistent improvements on complex audio understanding benchmarks and acoustic-scene evaluation, including MMAU (Sakshi et al., 2024), MMAR (Ma et al., 2025), MMSU (Wang et al., 2025), and CochlScene (Jeong and Park, 2022). Crucially, as hypothesized, the largest performance gains occur on the most perception-heavy tasks, directly validating our evidence-first approach.

Our main contributions are summarized as follows:

- **Problem Diagnosis** We identify and empirically validate the "evidence bottleneck" in SOTA LALMs, arguing from information-theoretic principles that poor acoustic grounding, not reasoning, is the primary performance limiter.
- **The EvA Architecture** We propose EvA, a dual-stream paradigm for evidence preservation, featuring a novel, non-compressive fusion process that hierarchically aggregates multi-level acoustic features and integrates them while maintaining full temporal fidelity.
- **Open-source Dataset and Strong Model** We release EvA-Perception, a high-temporal-precision dataset for perceptual training, and the EvA model, which achieves the best open-source *Perception* results on MMAU/MMAR/MMSU.

2 Related Works

2.1 Large Audio Language Models

Large Audio Language Models (LALMs) have progressed rapidly, with systems such as Qwen2-Audio (Chu et al., 2024), Qwen2.5-Omni (Xu et al., 2025), and Kimi-Audio (Ding et al., 2025). Most of these models rely on a single ASR-centric Whisper encoder (Radford et al., 2023), which is effective for speech but less robust for general non-speech audio. Prior dual-path designs partly address this issue, but SALMONN (Tang et al., 2023) introduces lossy Q-Former compression and GAMA (Ghosh et al., 2024) uses a fusion interface with weak temporal coherence. A detailed structural comparison with SALMONN-style Q-Former fusion is provided in Appendix A.11. These limits motivate our dual-path, non-compressive, time-aligned fusion design.

2.2 Audio Understanding Benchmarks

Audio-language evaluation has shifted from captioning datasets such as AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020) to more demanding benchmarks including MMAU (Sakshi et al., 2024), MMAR (Ma et al., 2025), and MMSU (Wang et al., 2025). These benchmarks require both fine-grained perception and higher-level reasoning over complex acoustic scenes. In this setting, performance is often constrained by evidence quality before reasoning, which directly motivates our evidence-first perspective and the construction of EvA-Perception.

3 Motivation: the Evidence Bottleneck

The perception–reasoning gap in Sec. 1 suggests an architectural bottleneck rather than a pure training issue. We analyze the LALM pipeline through a conceptual information-flow lens to identify where acoustic evidence is attenuated. We first examine the constraints of single-path models (Sec. 3.1),

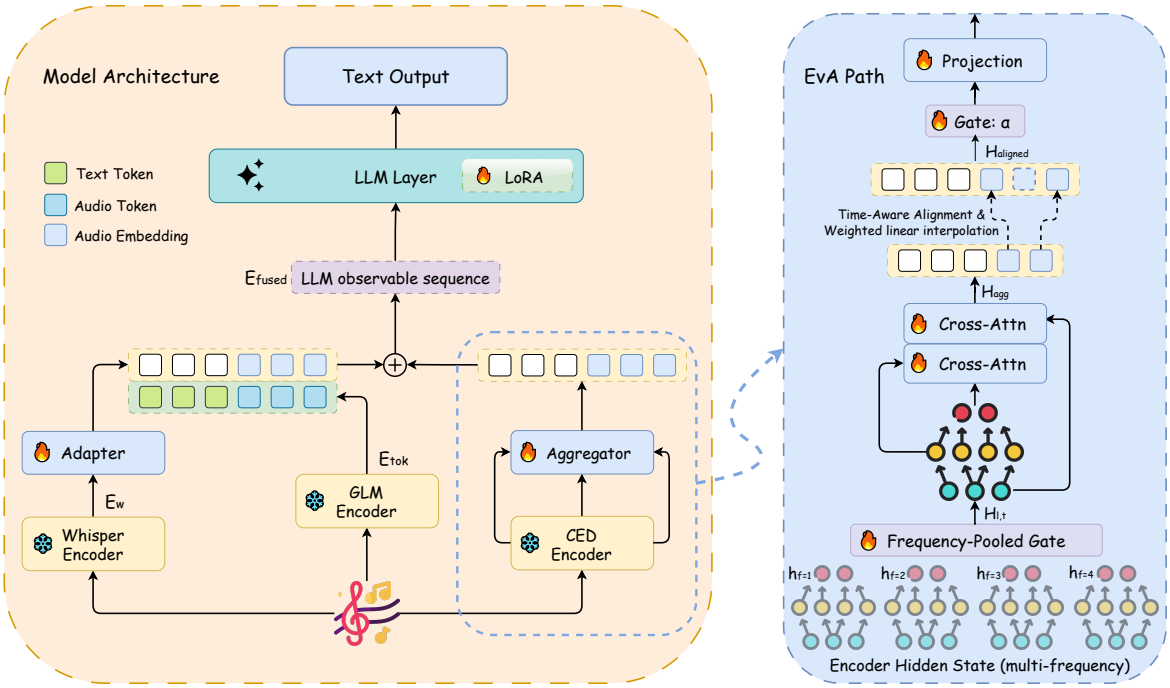

Figure 2: **Model architecture.** The left half of the left panel shows the Kimi-Audio backbone, while the right half illustrates the additional EvA Path modules. Audio is encoded into four frequency-band features by the CED Encoder, pooled across frequencies, fused via cross-attention, temporally aligned with Whisper, and integrated through gated additive fusion.

then show how dual-path architectures can relax these constraints in practice (Sec. 3.2). Formal derivations are provided in Appendix A.4.

Notation. We define the following variables: Z — **the latent acoustic evidence** required to answer the downstream task(e.g., event identities, temporal boundaries and ordering); X — the raw audio waveform; H — the encoder's hidden representation, O — the final representation passed to the language model, and Y — the model's textual output. Mutual information between any two variables is denoted as $I(\cdot;\cdot)$. We only assume that a joint distribution over (Z, X) exists; this minimal assumption suffices for *qualitative* reasoning about information flow. Mutual information is denoted $I(\cdot;\cdot)$.

3.1 Single-Path LALMs: An Information Constraint

Conditioned on fixed parameters after training, the inference-time forward pass is a composition of *deterministic* mappings of X:

$$H = E_\theta(X), \qquad O = P_\theta(H), \qquad Y = \pi_\theta(O).$$

Under this setting, the data-processing inequality (DPI) for deterministic functions implies

$$I(Z;Y) \le I(Z;O) \le I(Z;H) \le I(Z;X). \tag{1}$$

We use Eq. (1) *conceptually* to compare *relative* information retention across stages/paths: post-training that optimizes only the policy π_θ (e.g., SFT/GRPO(Group Relative Policy Optimization)) can improve how well the model *exploits* evidence already present in O, but it does not *restore* acoustic details attenuated upstream in $X \to H \to O$. This observation motivates architectural choices that prioritize evidence retention before the LLM (see Sec. 3.3). A derivation of (1) under fixed parameters appear in Appendix A.4.

3.2 Dual Paths Solution: Possible more information

Consider two complementary perception paths over the same input X: a speech-centric path producing $O_1 = P_1(E_1(X))$ (e.g., Whisper) and a general-audio path producing $O_2 = P_2(E_2(X))$ (e.g., CED-Base). The LLM receives the joint observation (O_1, O_2).

By the chain rule, $I(Z; O_1, O_2) = I(Z; O_1) + I(Z; O_2 \,|\, O_1) \ge I(Z; O_1)$. We cite this identity

to articulate the *complementarity intuition*: when the second path contributes cues not already captured by the first, the joint observation is—in a qualitative sense—no less informative than either stream alone. We do not estimate $I(\cdot\,;\cdot)$ nor claim quantitative increases or bounds in the main text; empirical ablations in Table 3 align with this intuition.

3.3 Implications for LALM Design

This perspective leads to two practical guidelines:
(i) Prioritize the perceptual front-end. The primary bottleneck often lies in evidence retention through the encoder/fusion stack rather than in the LLM policy; improving upstream access to fine-grained cues is therefore crucial.
(ii) Favor time-aligned, non-compressive fusion. Fusion interfaces that preserve temporal resolution and avoid heavy compression are aimed at minimizing avoidable information loss. In our system, the general-audio stream is aligned to the speech timeline and injected via an add-based mechanism with a learnable gate, which is structure-preserving and sequence-length neutral.

4 Method: Evidence-First Audio Understanding Paradigm

4.1 Architecture

EvA adopts a dual-path architecture on top of the **Kimi-Audio-7B** backbone that delivers time-aligned, structure-preserving acoustic evidence. As shown in Figure 2, the raw waveform is encoded by two complementary encoders, a speech-centric Whisper path and a generalist CED path. Their evidence is aligned to the token timeline and injected into the backbone LLM input space without changing sequence length. Details on initialization and the freezing/training policy are deferred to Sec. 4.3.

Complementary Dual Encoders. We employ two encoders that capture distinct but complementary information channels. (i) The **Whisper encoder** (E_W) is pre-trained on large-scale ASR corpora and excels at extracting robust, high-level linguistic features and provides a strong foundation for speech-related tasks. (ii) The **CED-Base encoder** (E_C), a Vision Transformer (ViT) based model trained for general audio event recognition, exposes richer non-linguistic cues (background events, music, transients). We extract hidden states from its shallow, middle, and final layers to harvest multi-scale semantics. These two encoders provide complementary views of the acoustic scene, corresponding to the distinct information channels O_1 and O_2 in our information-theoretic analysis, forming a comprehensive perceptual basis for the downstream LLM.

Hierarchical Evidence Aggregation. Standard encoders, which only expose their final-layer features, create an internal information bottleneck long before the LLM. The Data Processing Inequality dictates that these final features cannot be more informative than the collection of intermediate representations. To mitigate this loss, we introduce a hierarchical aggregation process that fuses and harvests features across the frequency domain and from multiple network depths.

First, in the **frequency domain**, we leverage the fact that the ViT-based CED encoder's feature maps implicitly retain a frequency axis. For the raw feature maps $\tilde{\mathbf{h}}_l \in \mathbb{R}^{B\times T\times F\times D_c}$ extracted from layer $l \in \{4, 8, L\}$ (where F is the number of frequency bands), we apply a lightweight gated attention mechanism. This performs a learnable, weighted pooling across the frequency bands for each time step:

$$\mathbf{h}_{l,t} = \sum_{f=1}^{F} \mathrm{softmax}(\mathrm{gate}(\tilde{\mathbf{h}}_{l,t,f})) \cdot \tilde{\mathbf{h}}_{l,t,f} \qquad (2)$$

This operation dynamically focuses on the most salient frequency bands at each moment, compressing the 2D feature map into a more informative 1D temporal sequence, which we denote as $\mathbf{H}_l \in \mathbb{R}^{B\times T\times D_c}$.

Second, in the **cross-layer domain**, we fuse these frequency-aggregated features, $\mathbf{H}_4, \mathbf{H}_8$, and $\mathbf{H}_L$, using a two-stage cascaded cross-attention mechanism implemented in our *Aggregator*. It first enriches the high-level semantic features $\mathbf{H}_L$ with mid-level temporal details from $\mathbf{H}_8$, and then grounds the resulting representation with low-level acoustic patterns from $\mathbf{H}_4$:

$$\mathbf{H}' = \mathrm{Norm}(\mathrm{CrossAttn}(\mathbf{H}_L, \mathbf{H}_8, \mathbf{H}_8) + \mathbf{H}_L) \qquad (3)$$

$$\mathbf{H}_{\mathrm{agg}} = \mathrm{Norm}(\mathrm{CrossAttn}(\mathbf{H}', \mathbf{H}_4, \mathbf{H}_4) + \mathbf{H}') \qquad (4)$$

Here, $\mathrm{CrossAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ denotes cross-attention with query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$; $\mathrm{Norm}(\cdot)$ denotes layer normalization. This cascaded, two-stage aggregation process produces a informative feature sequence $\mathbf{H}_{\mathrm{agg}}$ that embodies

acoustic evidence integrated across both multiple frequency bands and multiple levels of abstraction.

Time-Aware Alignment and Inject-and-Add Fusion. The final and most critical step is to integrate the evidence from both encoder paths without creating a fusion bottleneck. A key challenge is the temporal mismatch: the LLM-aligned Whisper features have a stride of **80 ms**, whereas the effective stride of our aggregated CED features $\mathbf{H}_{\text{agg}}$ is coarser at **160 ms**. To reconcile this, we upsample the CED evidence onto the Whisper timeline using a **time-aware linear interpolation**. This method respects the true mel-frame timestamps of each feature window. For each Whisper token's timestamp, we identify its two nearest neighbors in the CED sequence and compute a weighted average, carefully accounting for the temporal coverage of each feature window to avoid phase drift and preserve transients. The full algorithm is detailed in Appendix A.5. This process yields a temporally-aligned CED feature sequence, $\mathbf{H}_{\text{aligned}} \in \mathbb{R}^{B \times T_w \times D_c}$, that now shares the same timeline as the Whisper features.

Both the original Whisper features $\mathbf{E}_W$ and the aligned CED features $\mathbf{H}_{\text{aligned}}$ are then passed through separate lightweight projection heads (Proj_W and Proj_C respectively) to map them to the LLM's hidden dimension. Finally, they are integrated using our **inject-and-add** strategy. We chose this approach for three key principles: **(1) Efficiency:** simple vector addition incurs minimal computational overhead. **(2) Structural Compatibility:** it preserves the original sequence length and causality, requiring no modification to the LLM backbone. **(3) Controllability:** it allows for stable training via a learnable gate.

The final fused embedding $\mathbf{E}_{\text{fused}}$ is computed based on a mask $\mathbf{M}$ that identifies audio token positions:

$$\mathbf{E}_{\text{audio}}[i] = \big(\mathbf{E}_{\text{tok}}[i] + \text{Proj}_W(\mathbf{E}_W[i])\big) \cdot \sqrt{2} + \alpha \cdot \text{Proj}_C(\mathbf{H}_{\text{aligned}}[i]) \tag{5}$$

$$\mathbf{E}_{\text{fused}}[i] = \begin{cases} \mathbf{E}_{\text{audio}}[i], & \mathbf{M}[i] = 1 \\ \mathbf{E}_{\text{tok}}[i], & \mathbf{M}[i] = 0 \end{cases} \tag{6}$$

where $\mathbf{E}_{\text{tok}}$ are the initial token embeddings, and α is a learnable scalar gate initialized to a small value. This allows the model to gradually incorporate the generalist evidence without perturbing the LLM's pre-trained knowledge during early training stages. This strategy enriches each audio token locally, thereby circumventing the information bottlenecks typical of heavy, compressive fusion modules.

4.2 EvA-Perception: A Dataset for Evidence-Grounded Training

Perception-heavy failures in LALMs often stem from weakly grounded supervision: generic audio-caption lack temporal information or fine-grained high-fidelity cues (as shown in Table 1). **EvA-Captions**, a core component of EvA-Perception, couples temporal-order evidence with instruction-style supervision, guiding models to extract diverse evidence with temporal information from high-precision manual labels and then reason to mitigate bias.

Construction We build diverse complementary resources via a multi-expert pipeline (details and prompts in Appendix A.7). *AudioSet-Strong* (Hershey et al., 2021) provides audio resource and time-localized manual soft labels as acoustic priors, then *Gemini-2.5-Pro* (Comanici et al., 2025) transforms them into event-ordered, natural-language captions; *Whisper* (Radford et al., 2023) contributes ASR for speech presence; *OpenMu* (Zhao et al., 2024) adds music-oriented details; *Qwen-2.5-VL-72B* (Bai et al., 2025b) provides visual cues used exclusively to disambiguate underspecified audio events; and *QwQ-32B* (Team, 2025) consolidates all descriptions into a single fine-grained coherent caption while preserving temporal order.

Results This pipeline produces ∼**54K** fine-grained captions (**150 h**) and ∼**500K** QA pairs (closed/open 2:3; see Appendix A.7). We build:

1. **EvA-Captions & EvA-QA**: event-ordered captions (∼54K / 150 h) and corresponding QA pairs (∼500K).

2. **EvA-Alignment & EvA-Perception**: aggregated datasets for encoder alignment and SFT, respectively (detailed composition in Appendix A.7).

4.3 Training Strategy

We perform a two-stage training strategy to first stably integrate our fusion module with the pre-trained backbone, and then to fine-tune the entire system for complex, evidence-grounded reasoning.

Backbone Initialization. We build EvA on the public **Kimi-Audio-7B** backbone, keeping the pre-trained Whisper encoder and CED-Base frozen.

Name	# of Audio/QA	Avg. Caps Len	Visual	Music	Speech	Integration	Temporal Info
AudioCaps (Kim et al., 2019)	46k/46k	9.03	✗	✗	✗	✗	✗
Clotho (Drossos et al., 2020)	5k/5k	11.00	✗	✗	✗	✗	✗
LAION-Audio-630K (Wu et al., 2023)	630k/630k	7.30	✗	✗	✗	✗	✗
WavCaps (Mei et al., 2024)	403k/403k	7.80	✗	✗	✗	✗	✗
AudioSetCaps (Bai et al., 2025a)	1.9M/1.9M	28.00	✗	✗	✗	✗	✗
Auto-ACD (Sun et al., 2024)	1.5M/1.5M	18.10	✓	✗	✗	✓	✗
CompA-R (Ghosh et al., 2024)	62k/200k	18.00	✓	✗	✗	✓	✗
FusionAudio-1.2M (Chen et al., 2025)	1.2M/6M	47.18	✓	✓	✓	✓	✗
EvA-Caps/QA	54K/500K	**67.99**	✓	✓	✓	✓	✓

Table 1: Comparison of open-source audio caption datasets.

This avoids costly re-pretraining, preserves encoder distributions and tokenization, and ensures fair comparison by isolating the effect of our evidence-first fusion.

Stage 1: Alignment Tuning. In this stage, only the newly introduced modules (the CED Aggregator and projection heads) are trained, using next-token cross-entropy on text tokens. The objective is to learn the mapping from the generalist audio encoder's feature space to the LLM's input embedding space, without disrupting the model's pre-trained weights. The small initialization of the gate α in Eq. 6 is critical for stability.

Stage 2: Instruction Fine-Tuning. We fully train CED Aggregator and Whisper adapter; the LLM backbone is updated via **LoRA**, while both encoders stay frozen. We continue to use the same text-only cross-entropy objective on training set.

5 Experiments

In this section, we evaluate our method on the different benchmarks. Beyond the official leaderboards, we also report results under unified *Perception* and *Reasoning* splits to better reflect the effect of our method. EvA shows the strongest improvements on *Perception* subsets across the three main benchmarks, consistent with our evidence-first design.

5.1 Experimental Setup

Benchmarks. We evaluate mainly on three audio understanding benchmarks—**MMAU**, **MMAR**, and **MMSU**—and one acoustic-scene benchmark, **CochlScene**, which together cover perception-oriented understanding and scene-level recognition. To directly test our central hypothesis regarding the evidence bottleneck, we categorize the sub-tasks of each benchmark into two primary axes: **Perception** and **Reasoning**. This categorization allows us to quantify performance gains specifically on tasks that depend on acoustic perception versus those that test abstract reasoning, the detailed categorization can be found in Appendix A.9. For completeness, we also report results under each benchmark's original categories and describe our special handling of answer ordering in Table 6.

Baselines. We compare to strong general understanding systems and reasoning systems: *GPT-4o-Audio*, *Qwen2-Audio* (Chu et al., 2024), *Qwen2.5-Omni* (Xu et al., 2025), Audio-Flamingo-3 (Goel et al., 2025), Step-Audio-2-mini (Wu et al., 2025), *Kimi-Audio* (Ding et al., 2025), *Audio-Reasoner* (Xie et al., 2025b) and *R1-AQA* (Li et al., 2025). All baselines are run under a unified inference protocol, and numbers reported are from our reproduction to ensure fairness.

Implementation Details. Training follows the procedure outlined in Sec. 4.3. In **Stage 1 (Alignment)**, we use EvA-Alignment for alignment, training only the CED Aggregator with all other encoders frozen. We use a learning rate of 1×10^{-3}, train for 5 epochs, with a per-device batch size of 2 and gradient accumulation of 8, resulting in a global batch size of 128. In **Stage 2 (SFT)**, EvA-Perception is used for instruction tuning. Both the CED Aggregator and the Whisper adapter are trained, while the LLM backbone is fine-tuned via **LoRA** with a rank of 64, $\alpha = 64$, and a dropout rate of 0.05. The learning rate is reduced to 5×10^{-5} with 2 epochs, a per-device batch size of 2, and gradient accumulation of 16, yielding a global batch size of 256. The maximum sequence length for training and inference is set to 1024. Both training and decoding use greedy sampling (temperature 0) with a max length of 1024. Each stage takes approximately 12 hours to complete on 8×A100 GPUs. For the detail of training setting, please refer to Appendix A.6.

Model	Arch.	MMAU		MMAR		MMSU		CochlScene
		Perc.	Reas.	Perc.	Reas.	Perc.	Reas.	
GPT-4o-Audio	Proprietary	60.68	59.94	51.12	60.69	43.25	75.04	–
Qwen2.5-Omni-7B	Single-Path	73.99	69.57	48.68	59.20	40.37	66.31	80.38
Audio-Flamingo-3	Single-Path	76.78	**74.00**	59.23	**61.64**	45.63	77.86	75.57
Step-Audio-2-mini	Single-Path	69.97	71.62	54.58	61.01	44.36	**78.32**	83.24
Audio-Reasoner	Single-Path	66.56	65.73	44.34	38.23	40.73	68.38	–
R1-AQA	Single-Path	73.07	65.29	48.75	50.19	41.68	71.94	76.30
Kimi-Audio-7B-Instruct	Base Model	67.80	62.19	56.79	58.72	45.47	71.85	86.17
EvA	**EvA Dual-Path**	**78.64**	67.35	**59.79**	59.45	**47.52**	75.41	**87.04**

Table 2: Main results on benchmarks. EvA achieves strong open-source performance with particularly large gains on *Perception*. All numbers are from our unified reproduction under 0-shot setting. The exact definition of perception and reasoning in benchmarks and reproduction setting can be seen in Sec. 5.1.

5.2 Main Results

As shown in Table 2, EvA obtains **78.64/67.35** on MMAU (Perception/Reasoning), **59.79/59.45** on MMAR, **47.52/75.41** on MMSU, and **87.04** on CochlScene. Notably, EvA achieves the best open-source *Perception* scores on all three main benchmarks (MMAU/MMAR/MMSU). Relative to Kimi-Audio-7B, EvA improves all reported metrics (+10.84/+5.16 on MMAU, +3.00/+0.73 on MMAR, +2.05/+3.56 on MMSU, and +0.87 on CochlScene), with the largest gains concentrated on perception splits. EvA also attains the competitive CochlScene result in our comparison (**87.04**), showing that the evidence-first design transfers to specialized acoustic-scene classification. These results suggest that EvA's primary advantage comes from stronger evidence extraction, while reasoning performance remains competitive.

Case Study We illustrate differences using AudioCaps case studies (Fig. 3). Compared to Qwen2.5-Omni and Kimi-Audio, **EvA** produces captions that are both faithful and more expressive. In the case, EvA captures sequential tone shifts (calm speech → child excitement → laughter), while baselines misinterpret events. The examples show how EvA's evidence-first fusion enhances both accuracy and richness in captioning.

5.3 Ablation Studies

Setup. We study the proposed CED-based fusion path from two perspectives. First, we assess its **overall contribution** by comparing variants that (do not) use the CED branch during *alignment* (Stage 1) and *perception SFT* (Stage 2). Second, we analyze the necessity of **internal design** of the CED Aggregator, including the frequency-gated pooling over bands and the top–down cross-layer fusion across CED layers. Stage 1 variants are trained on **EvA-Alignment** and evaluated on *AudioCaps* using CLAP metrics, while Stage 2 variants are trained on **EvA-Perception** from the same aligned checkpoint and evaluated on MMAU, MMAR, and MMSU.

Overall effect of the CED branch. Compared to the S0 base model without any CED path, enabling the CED Aggregator during Stage 1 alignment substantially improves CLAP retrieval on *AudioCaps*: all Stage 1 variants that use the CED branch outperform the S0 backbone, while masking the CED path at inference time yields consistently worse scores than using it throughout (Table 3, top). Starting from the same aligned checkpoint, keeping the CED stream active in Stage 2 perception SFT further boosts MMAU and MMAR over masking the CED path (Table 3, bottom), indicating that the CED encoder contributes complementary non-speech evidence that remains useful after alignment. The speech-heavy MMSU benchmark shows a mild reversal, which is consistent with our focus on general audio rather than speech-specialized training.

Ablation of modules in the CED branch. Within the CED branch, we next ablate the Aggregator design (Table 3, top). Removing the top–down cross-layer fusion and using only the last CED layer feature (S1(4)) leads to a marked drop in all CLAP metrics with a comparable parameter budget. This directly supports our claim that intermediate CED layers supply complementary acoustic information that is lost when relying solely on

Case: AudioCaps/103394.wav
Human Reference: "A man is speaking followed by a child speaking and then laughter",
"A man speaking as a man talks over an intercom in the background followed by a girl talking then a group of people laughing and a man and woman talk"

Qwen2.5-Omni: man and child speaking
Kimi-Audio: a man speaks and a baby cries
EvA (Ours): **A man speaks** calmly and clearly, followed by a **child speaking** excitedly. **A group of people laugh**, and the **man speaks again** in a friendly tone

Figure 3: Qualitative comparison of captions generated by different models on AudioCaps examples.

Stage	Setting	Trainables	Start Ckpt	AudioCaps (CLAP)		
				Cos	R@1	R@5
S0	Base Model	N/A	N/A	14.61	9.27	24.97
S1(1)	w/o CED path	Adapter	S0	35.40	18.50	43.60
S1(2)	mask CED path in inf	Adapter & CED Agg.	S0	34.37	17.76	41.44
S1(3)	w/o frequency pooling	Adapter & CED Agg.	S0	35.54	21.24	49.61
S1(4)	w/o crossing fusion	Adapter & CED Agg.	S0	28.63	11.82	29.74
S1(5)	Q-former	Adapter & CED Q-former	S0	36.24	20.08	47.36
S1(0)	**EvA Dual-Path**	Adapter & CED Agg.	S0	**36.77**	**22.77**	**49.81**
Stage	**Setting**	**Trainables**	**Start Ckpt**	**MMAU**	**MMAR**	**MMSU**
S0	Base Model	N/A	N/A	65.33	49.21	43.36
S2(1)	mask CED path in inf	Adapter & LLM	S1(0)	75.85	54.59	**48.18**
S2(0)	**EvA Dual-Path**	Adapter & CED Agg. & LLM	S1(0)	**80.19**	**55.26**	47.44

Table 3: Ablations of the EvA fusion path. On *AudioCaps*, we use the CLAP encoder (Wu et al., 2023) to embed text/audio; **Cos** is cosine similarity, and higher is better for all CLAP metrics. On MMAU/MMAR/MMSU, we report *Perception* accuracy (Sec. 5.1). *Adapter* denotes the Whisper adapter; *LLM* denotes LoRA on the backbone; "*mask CED path*" disables the CED branch at fusion.

the final layer. Replacing our hierarchical Aggregator with a window-level Q-Former (S1(5)), similar in spirit to SALMONN, improves over weaker baselines but still underperforms the full EvA configuration (S1(0)), indicating that non-compressive, multi-level fusion is more effective than aggressive window compression in our setting. We also analyze the structural shortcoming of Q-Former module simply compared with Aggregator in Appendix A.11.

We further examine the frequency-gated pooling over four sub-bands. Simply averaging over bands (S1(3)) instead of using a learnable gate consistently degrades CLAP scores, showing that the gated pooling helps the model reweight spectral regions. To probe which bands matter, we perform a coarse band-masking diagnostic by zeroing or keeping individual bands within the same gate (Appendix A.10). Across the benchmarks, masking any single band or keeping only one band is always worse than using the full range, and no individual band dominates across all tasks.

Taken together, these ablations suggest that (i) EvA benefits from multi-layer CED features rather than only the last layer, and (ii) the four sub-bands provide complementary cues, making EvA a reasonable and robust design choice rather than a rough module assembly.

6 Conclusion

In this work, we identified the evidence bottleneck as a critical limitation in Large Audio Language Models (LALMs), where perceptual grounding fails in complex acoustic scenes. To address this, we introduced EvA, an evidence-first paradigm that combines a dual-encoder architecture with a novel, non-compressive and temporally faithful fusion mechanism to preserve acoustic evidence. Our approach, supported by the EvA-Perception dataset, achieves strong performance on benchmarks like MMAU, MMAR, and MMSU, with the largest gains concentrated in perception-heavy tasks, validating the importance of prioritizing evidence fidelity for advanced audio understanding.

7 Limitations

While EvA advances evidence-grounded audio understanding, several limitations remain: our curated audio–text training corpus currently uses English-only captions, even though the paired audio and evaluation benchmarks contain multilingual inputs, so a more systematic multilingual textual supervision and evaluation is still missing; temporal reasoning is constrained by soft event boundaries from AudioSet-Strong, and music analysis lacks expert-level concepts like pitch or harmony. Addressing these challenges represents key directions for future work.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and 1 others. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D Plumbley, Woon-Seng Gan, and Jianfeng Chen. 2025a. Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models. *IEEE Transactions on Audio, Speech and Language Processing*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Shunian Chen, Xinyuan Xie, Zheshu Chen, Liyan Zhao, Owen Lee, Zhan Su, Qilin Sun, and Benyou Wang. 2025. Fusionaudio-1.2 m: Towards fine-grained audio captioning with multimodal contextual fusion. *arXiv preprint arXiv:2506.01111*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang. 2024. Ced: Consistent ensemble distillation for audio tagging. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 291–295. IEEE.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Preprint*, arXiv:2507.08128.

Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. 2021. The benefit of temporally-strong labels in audio event classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 366–370. IEEE.

Il-Young Jeong and Jeongsoo Park. 2022. Cochlscene: Acquisition of acoustic scene data using crowdsourcing. *Preprint*, arXiv:2211.02289.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. 2025. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, and 1 others. 2025. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*.

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.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2023. Mustango: Toward controllable text-to-music generation. *arXiv preprint arXiv:2311.08355*.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

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.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie. 2024. Auto-acd: A large-scale dataset for audio-language representation learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5025–5034.

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*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 90 others. 2025. Step-audio 2 technical report. *Preprint*, arXiv:2507.16632.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. 2025a. 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, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025b. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*.

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*.

Chao-Han Huck Yang, Sreyan Ghosh, Qing Wang, Jaeyeon Kim, Hengyi Hong, Sonal Kumar, Guirui Zhong, Zhifeng Kong, S Sakshi, Vaibhavi Lokegaonkar, and 1 others. 2025. Multi-domain audio question answering toward acoustic content reasoning in the dcase 2025 challenge. *arXiv preprint arXiv:2505.07365*.

Mengjie Zhao, Zhi Zhong, Zhuoyuan Mao, Shiqi Yang, Wei-Hsiang Liao, Shusuke Takahashi, Hiromi Wakaki, and Yuki Mitsufuji. 2024. Openmu: Your swiss army knife for music understanding. *arXiv preprint arXiv:2410.15573*.

A Appendix

A.1 Potential Risk

Because the model and training resources are built from publicly available online content, EvA may inherit latent moral, ethical, or racial biases. We therefore recommend strengthened human review and supervision when using the system, especially in high-stakes or sensitive scenarios.

A.2 Access Statement

All training and evaluation data used in this work are collected from publicly available open-source datasets on the Internet. We do not use private user recordings or non-public personal data.

A.3 Use of LLMs

Large language models were utilized for specific tasks, such as assisting with coding and providing grammar checks and language refinement during the writing of this paper. All scientific content, including research design, experimentation, data analysis, and conclusions, was independently conducted by the authors without LLM involvement in the core research process.

A.4 Notes for the Information-Flow View

DPI under deterministic inference. After training, condition on fixed parameters θ. Let $H = E_\theta(X), O = P_\theta(H), Y = \pi_\theta(O)$ be deterministic mappings of X. For any measurable f, DPI gives $I(Z; f(X)) \le I(Z; X)$. Applied stage-wise to the composition,

$$I(Z;Y) \le I(Z;O) \le I(Z;H) \le I(Z;X).$$

If inference includes independent randomness U (e.g., stochastic decoding), write $Y = g(O, U)$ with $U \perp Z \,|\, O$, so $I(Z;Y) \le I(Z;O,U) = I(Z;O)$.

Chain-rule identity (for intuition). For random variables Z, O_1, O_2, the chain rule gives $I(Z;O_1,O_2) = I(Z;O_1) + I(Z;O_2 \,|\, O_1) \ge I(Z;O_1)$. We use this only to express complementarity succinctly; no quantitative claim is made.

A.5 Algorithm for Time-Aware Coverage-Weighted Linear Interpolation

Assumptions and Notation. We denote the aggregated CED sequence as $\mathbf{H}_{\text{agg}} \in \mathbb{R}^{T_c \times D}$ with per-feature centers $t_c[0], \ldots, t_c[T_c - 1]$ (monotonically increasing), and the target Whisper centers $t_w[0], \ldots, t_w[T_w - 1]$ (also monotonic). All timestamps share the same unit (e.g., mel frames or milliseconds). A small constant $\varepsilon > 0$ (we use 10^{-8}) is added for numerical stability. When $T_c \le 1$, we simply repeat the sole vector to length T_w.

Coverage weights. For each CED window ℓ, its coverage weight $c[\ell] \in [0, 1]$ measures the fraction of the window overlapping valid (non-padded) audio. Concretely, if a window starts at start_ℓ and ends at end_ℓ with window size t_{sz}, and the valid audio spans $[0, T_{\text{mel}} - 1]$, then

$$c[\ell] \;=\; \frac{\max\big(0,\, \min(\text{end}_\ell,\, T_{\text{mel}} - 1) - \text{start}_\ell + 1\big)}{t_{\text{sz}}}\,.$$

Thus $c = 1$ for fully valid windows and decreases as padding overlap grows.

Target centers for Whisper. Let $step_{mel}$ be the mel frames per Whisper token and $center_{mel}$ its center offset. For $k = 0, \ldots, T_w - 1$, we use

$$t_w[k] = k \cdot step_{mel} + center_{mel}. \qquad (7)$$

In our implementation, $step_{mel} = 8$ and $center_{mel} = 4$.

Discussion. By reweighting both neighbors with $c[\ell]$ in the numerator *and* renormalizing by the weighted sum in the denominator, features largely sourced from padded/silent regions contribute less to the aligned representation, especially near boundaries. In practice we use a vectorized implementation (single search-sorted, broad-casted arithmetic) that avoids Python loops while preserving the above semantics.

A.6 Training setting

As shown in Table 4, we summarize the training hyperparameters for both stages. Notably, the CED Aggregator is trained from scratch in Stage 1, while the Whisper adapter and LLM are only updated in Stage 2. Both stages use a cosine learning rate schedule with a small warmup, and we keep the encoders frozen throughout to preserve their pretrained representations and ensure stable training of the new modules. The LoRA configuration for the LLM is chosen to balance expressiveness with parameter efficiency, targeting the query, key, value, and output projection matrices without modifying the MLP layers. We also prepare for DeepSpeed ZeRO-3 optimization but keep it off by default to allow flexibility in resource-constrained settings.

Algorithm 1 Time-Aware, Coverage-Weighted Linear Interpolation (0-based indexing)

Require: Aggregated CED features $\mathbf{H}_{\text{agg}} \in \mathbb{R}^{T_c \times D}$
Require: CED centers $t_c[0..T_c-1]$ (monotonic), Whisper centers $t_w[0..T_w-1]$ (monotonic)
Require: Coverage weights $c[0..T_c-1]$, with $c[\ell] \in [0,1]$
Require: Stability constant $\varepsilon > 0$ (e.g., 10^{-8})
Ensure: Aligned features $\mathbf{H}_{\text{aligned}} \in \mathbb{R}^{T_w \times D}$

Initialize $\mathbf{H}_{\text{aligned}}$ as a tensor of shape (T_w, D)
if $T_c \leq 1$ **then**
  return $\mathbf{H}_{\text{aligned}} \leftarrow$ repeat the sole vector to length T_w
end if
for $k \leftarrow 0$ to $T_w - 1$ **do**
  ▷ Locate neighbors of $t_w[k]$ in t_c via binary search
  $r \leftarrow \text{searchsorted}(t_c,\, t_w[k])$
  $r \leftarrow \text{clamp}(r,\, 1,\, T_c - 1), \quad l \leftarrow r - 1$
  ▷ Linear interpolation factor with clamping
  $\alpha \leftarrow \dfrac{t_w[k] - t_c[l]}{t_c[r] - t_c[l] + \varepsilon};$
  $\alpha \leftarrow \text{clamp}(\alpha,\, 0,\, 1)$
  ▷ Coverage-weighted, normalized interpolation
  $\mathbf{x}_L, \mathbf{x}_R \leftarrow \mathbf{H}_{\text{agg}}[l],\ \mathbf{H}_{\text{agg}}[r]$
  $c_L, c_R \leftarrow c[l],\ c[r]$
  $\text{num} \leftarrow (1-\alpha)\,(c_L\,\mathbf{x}_L) \ + \ \alpha\,(c_R\,\mathbf{x}_R)$
  $\text{den} \leftarrow (1-\alpha)\,c_L \ + \ \alpha\,c_R \ + \ \varepsilon$
  $\mathbf{H}_{\text{aligned}}[k] \leftarrow \text{num}/\text{den}$
end for
return $\mathbf{H}_{\text{aligned}}$

	Stage 1 (Alignment)	**Stage 2 (SFT, LoRA)**
Trainable modules	CED Aggregator	CED Aggregator, Whisper adapter; LLM via LoRA
Dataset	EvA-Alignment	EvA-Perception
Epochs	5	2
Per-device batch	2	2
Grad. accumulation	8	16
Global batch (8×A100)	128	256
Optimizer	AdamW (β_2=0.95, wd 0.1)	AdamW (β_2=0.95, wd 0.1)
LR / schedule / warmup	1×10^{-3} / cosine / 1%	5×10^{-5} / cosine / 1%
Max seq length	512	1024
LoRA (LLM)	–	r=64, α=64, dropout = 0.05; targets=`q,k,v,o` (`include_mlp`=False)
Modules to save	–	`model.vq_adaptor`
Checkpoint export	split every epoch (keep last 5)	split every epoch (keep last 5)
Distributed setup	torchrun, 8×A100-80GB	torchrun, 8×A100-80GB
DeepSpeed ZeRO-3	prepared (config available), *off* by default	prepared, *off* by default
Runtime (wall-clock)	∼12h	∼12h

Table 4: Training hyperparameters. Encoders (Whisper, CED-Base) are frozen throughout.

A.7 Data Construction

As shown in Figure 4 and Figure 5, we provide the detailed instruction for the construction of our EvA-Perception dataset, which is designed to train the model for rigorous multimodal information integration and pure audio description generation. The instruction emphasizes the importance of controlling cross-modal interference, performing cross-reasoning and correction, and generating a purely audio-focused description while marking potential ambiguities based solely on auditory perception. The input sources include audio tags, audio descriptions, speech content (ASR), music descriptions, and video descriptions, each with specific guidelines on how to use them for accurate audio captioning. The processing steps outline a systematic approach to multimodal parsing, auditory fact determination, ambiguity inference, reliability as-

sessment, and final caption generation.

A.8 Cases in EvA-Captions and EvA-QA

We provide two cases from our EvA-Captions and EvA-QA datasets in Figure 6 and Figure 7, respectively. These cases illustrate the complexity and richness of the audio descriptions and the corresponding QA pairs, showcasing the model's ability to handle intricate acoustic scenes and extract meaningful information for question answering tasks. The captions contain detailed descriptions of the audio content, while the QA pairs test various aspects of understanding, such as emotional effects, vocal presence, sound dominance, and the contribution of specific elements to the overall atmosphere.

A.9 Introduction on Benchmarks

MMAU covers broad audio modalities with a mixture of perceptual, information-extraction, and reasoning questions; **MMAR** emphasizes multi-step inference across hierarchical layers; **MMSU** targets spoken language understanding, including fine-grained linguistic and paralinguistic phenomena. **Reasoning**. Here, *MMSU*'s native tags are kept, *MMAU*'s *information extraction/reasoning* map to Perception/Reasoning, and *MMAR*'s *Signal+Perception/Semantic+Culture* likewise.

We discover that, in the original MMAU test set, the reference answers are distributed unevenly across choice positions: **A: 39.5%, B: 27.1%, C: 20.8%, and D: 12.6%.** Such imbalance may bias evaluation for models with positional preferences (e.g., favoring earlier options). To mitigate this artifact, we randomized the order of choices, ensuring that the final distribution of correct answers is balanced across positions. All reported MMAU results in the main paper are based on this balanced setting.

A.10 Additional Analysis: Frequency-Band Ablation of the CED Path

To better understand how EvA exploits spectral cues, we perform an exploratory ablation on the CED branch by masking coarse frequency bands at inference time. The CED encoder partition the 64 Mel bins into four contiguous groups and treat them as approximate 2 kHz bands: $[0, 2), [2, 4), [4, 6), [6, 8)$ kHz. On top of the frozen CED encoder, we insert a band mask before the Aggregator: for each configuration, a binary vector of length four determines which bands are zeroed out and which are kept. All experiments share the same single-encoder EvA backbone; we only change the band mask at inference without re-training.

Table 8 reports the results on multi-task benchmarks (MMAU, MMAR, MMSU) and a specialized scene benchmark (CochlScene). We consider both "*mask X*" (drop one or two bands and keep the others) and "*left X*" (keep only one band and mask the rest), and compare them with the full 0–8 kHz setting used in EvA.

Broadband cues are consistently better than any subset. Across all benchmarks, any masking or single-band configuration yields lower scores than using the full 0–8 kHz range. The full-band EvA remains the best-performing setting in all cases, which supports the interpretation that EvA relies on broadband, complementary spectral information rather than a single dominant frequency region.

CochlScene shows mild sensitivity to low frequencies. On CochlScene, masking the lowest band (0–2 kHz) tends to be among the weaker configurations, and keeping only a single narrow band is also not optimal. Together, these observations suggest that low-frequency components provide useful contextual cues for ambient scenes, but are most effective when combined with mid- and high-band information. We view this as a mild trend rather than a strong claim, since the absolute differences between band configurations remain relatively small.

Similarly, the aggregated MMAU, MMAR and MMSU scores are designed to combine heterogeneous tasks, making them less sensitive to any single frequency band; here the main takeaway is simply that broadband CED fusion is more reliable than any restricted band subset. These results corroborate our main conclusion that EvA benefits from broadband, multi-band CED features, and we do not observe evidence that the model's performance is dominated by a single narrow frequency region.

A.11 Structural comparison with Q-Former–based fusion

In this appendix, we clarify the structural differences between EvA's CED Aggregator and a Q-Former–based fusion scheme (as instantiated by our SALMONN-style variant), and relate them to the temporal behavior illustrated in Figure 8.

Token length and temporal granularity. Figure 8 compares the audio token sequence length

Dataset	Constituent Sources	Modality	Quantity
EvA-Alignment	EvA-Captions	Sound, Speech, Music	53,934
	AudioTime (Xie et al., 2025a)	Sound	5,000
	CommonVoice (Ardila et al., 2019)	Speech	20,000
	MusicBench (Melechovsky et al., 2023)	Music	30,000
	MusicCaps (Agostinelli et al., 2023)	Music	4,852
Total			113,786
EvA-Perception	EvA-Captions	Sound, Speech, Music	53,934
	EvA-QA	Sound, Speech, Music	525,673
	AudioSkills: Counting-QA (Goel et al., 2025)	Sound	46,266
	ESC50 (Piczak, 2015)	Sound	2,000
	AudioTime (Xie et al., 2025a)	Sound	5,000
	DCASE2025_T5 (Yang et al., 2025)	Sound	10,687
	CommonVoice (Ardila et al., 2019)	Speech	20,000
	MusicBench (Melechovsky et al., 2023)	Music	30,000
	MusicCaps (Agostinelli et al., 2023)	Music	4,852
	AudioSkills: MagnaTagATune (Goel et al., 2025)	Music	364,760
Total			1,063,172

Table 5: Composition of EvA-Perception and EvA-Alignment datasets.

Model	Sound	Speech	Music	Avg.
Qwen2-Audio	58.86	47.75	44.31	50.30
Qwen2.5-Omni	73.87	65.47	67.96	69.10
Kimi-Audio	74.77	62.35	64.24	67.19
Audio-Reasoner	65.77	66.07	66.77	65.00
R1-AQA	74.47	65.17	66.77	68.80
EvA(Ours)	**80.78**	**68.47**	**74.65**	**74.63**

Table 6: Main results on MMAU-mini-test.

Model	**Single Modality**			**Mixed Modality**				**Avg.**
	Sound	Speech	Music	S-M	S-S	M-S	S-M-S	
Qwen2-Audio	52.73	42.86	34.95	36.36	50.46	45.12	50.00	44.80
Qwen2.5-Omni	59.39	61.22	48.06	54.55	61.01	**64.63**	58.33	58.30
Kimi-Audio	55.76	59.86	45.15	36.36	61.01	54.88	45.83	55.40
Audio-Reasoner	50.30	49.66	38.35	36.36	56.42	48.78	50.00	48.70
R1-AQA	**60.00**	51.36	42.23	54.55	57.80	52.44	45.83	52.30
EvA(Ours)	55.76	**63.01**	**50.25**	**63.64**	**63.76**	57.32	**79.17**	**59.30**

Table 7: Main results on MMAR.

Instruction for Caption Generation

Rigorous Multimodal Information Integration and Pure Audio Description Expert
Core Task
You are an expert specializing in audio information processing. Your goal is: to integrate and analyze textual descriptions from multiple modalities, strictly controlling cross-modal interference, and to perform cross-reasoning and correction. Ultimately, you should generate a **purely audio-focused**, temporally ordered, accurate, and detailed description of the audio content in fluent English, while marking potential ambiguities that are **only based on auditory perception**. **Absolutely forbidden:** including any visual information, specific speech transcript content, or ambiguity annotations derived from audio-video inconsistencies in the final output.
Input Sources (may contain errors, hallucinations, or incompleteness):

- **Audio Tags:** A set of frame-level sound category labels annotated by humans. These represent the most prominent acoustical features perceptible to humans, with **high reliability**. Format: `[start time, end time, event]`, e.g., `[start time: 9.0, end time: 10.0, event: Generic impact sounds]`.
- **Audio Description:** A textual description of the audio content (may include sound events, environmental sounds, music, human voice characteristics, etc.). This is an **important basis** for describing audio facts and must be cross-validated with tags and music description.
- **Speech Content (ASR):** Automatic speech recognition results. This is used **only** to confirm the existence of human voices, judge non-content features (e.g., speech vs. nonverbal sounds, emotional tone), and assist in inferring possible scenes or events. **The specific text content must never appear in the final output** (not quoted, summarized, or paraphrased). If empty, it means no obvious human voice or only non-speech behaviors (e.g., breathing, crying, background chatter).
- **Music Description:** Contains information about music elements (features, instruments, rhythm, etc.) and other scene-related sounds. **Musical features are highly reliable. If empty, it means no clear music.** Non-musical descriptions here are lower priority and secondary to audio tags, audio descriptions, and ASR.
- **Video Description:** Visual scene description. Used **only under strict conditions** (see Step 2 "Positive Correction") to disambiguate uncertain auditory sources and identify inconsistencies between hearing and vision. **Never** speculate or describe sound sources, positions, or visual actions based solely on video. If empty, no visual assistance is available.

Processing Steps:

1. **Multimodal Parsing:** Extract key sound events, source characteristics, environment, and music elements from each source. Give priority to audio tags. From ASR, only detect voice presence and non-content features, possibly aiding environment/emotion inference, but never include speech text itself.
2. **Auditory Fact Determination and Cross-modal Correction:**
   - Base facts primarily on: **Audio Tags** $>$ **Music Description (music part)** $\approx$ **Audio Description** $>$ **Speech presence (ASR)** $>$ **Non-music part of Music Description**.
   - Apply **Positive Correction with video** only if audio information is ambiguous and video provides clear, reasonable confirmation of the specific sound source (e.g., generic rumble corrected to airplane noise if airplane is explicitly shown). If tags already specify the type, no correction applies.
   - If video does not support or contradicts, never override auditory facts; only internally mark inconsistency for conservative phrasing later.
   - Adopt extreme conservatism when conflicts remain unresolved: omit or cautiously phrase uncertain elements.
3. **Pure Auditory Ambiguity Inference:** List ambiguities only from hearing, such as similarity of sounds (e.g., car vs. plane), multiple possible sources, or common perceptual misinterpretations. **Never include visual-based ambiguities.**
4. **Reliability Assessment:** If audio facts are extremely scarce, or sources are severely conflicting and cannot yield reliable auditory facts, directly output: `UNCERTAIN_AUDIO_INFORMATION_DETECTED`.
5. **Final Audio Caption Generation:** If reliable:
   - Generate a fluent, precise, audio-focused English description that preserves event order, number of occurrences, auditory features (sound type, timbre, rhythm, loudness, space, etc.).
   - Integrate confirmed auditory facts, using cautious wording for uncertain elements ("sounds like", "a sound resembling ... is heard", "potentially ...").
   - Strictly exclude: visual details, speech text, or speculation not supported by input sources.
   - Express emotional tone if strongly supported by audio (e.g., voice emotion or music mood).

Output Format: Normally, output must be JSON:

```
{
  "Potential ambiguities": [
    "Ambiguity description 1 based solely on auditory perception.",
    "Ambiguity description 2 based solely on auditory perception."
  ],
  "Audio caption": "Final audio description focusing solely on audible elements and their
  auditory characteristics, detailed and fluent English. Use conservative language when
  uncertain."
}
```

In the special case of Step 4 failure, output only: `UNCERTAIN_AUDIO_INFORMATION_DETECTED`.

Figure 4: Instruction for Caption Generation.

Instruction for QA Generation

You are a data generator that converts a single audio caption into QA pairs. Write as if you actually listened to the audio.

Hard rules:

1. **Grounding:** Use **only** the given description as grounding. Do not invent facts beyond what the description supports. Do not produce ASR-style verbatim transcripts.
2. **Output format:** Return JSON Lines (JSONL). Each line is one JSON object. No markdown, no backticks, no extra prose, no blank lines.
3. **JSON validity:** Use straight ASCII quotes. No trailing commas. All keys required. Types must match the schema.
4. **Brevity:** Closed-ended answers must be short (yes/no, true/false, a label, a small set, or a number) or MCQ (A–D, answer is a single letter).
5. **Evidence:** `support_span` must not be empty. Provide exact words/phrases taken from the description that justify the answer. Do not add commentary.
6. **Difficulty tagging:** Use the boolean field `ishard`. Set it to true **only** for the most difficult items in the batch; otherwise false.
7. **No meta-reference to source:** In both `question` and `answer`, never mention the existence of any "caption/description/text". Write from an *audio-listener* perspective, not a reader perspective.

Schema (each JSON object must match):

```
{
  "question": string,
  "answer": string,
  "answer_style": "close" | "open",
  "type": one of [
    "presence", "counting", "temporal_order", "concurrency", "trend",
    "comparative", "scene_spatial", "causality", "mood_style",
    "submodality_decomposition", "music_feature", "speech_structure",
    "cross_modal_integration", "higher_order_semantics",
    "aesthetic_function", "scene_inference",
    "rhythm_structure", "tension_dynamics"
  ],
  "submodality": array of ["speech","music","sound"],
  "support_span": array of strings (never empty),
  "confidence": number in [0,1],
  "ishard": boolean
}
```

Closed-ended conventions:

- Yes/No answers must be exactly `"Yes"` or `"No"`.
- True/False answers must be exactly `"True"` or `"False"`.
- MCQ: include the options concisely in the `question`; set `answer` to exactly one letter among `"A","B","C","D"`.
- Counts are integers as strings (e.g., `"2"`).
- Options or sets should be concise strings (e.g., `"drums"`, `"drums, guitar"`).

Design goals:

- Favor integrative, cross-modality questions that combine speech/music/sound cues.
- Include deep semantic understanding: structural roles, function vs. ornament, implied emotion arcs, rhythm-harmony interplay, tension build/release.
- Keep questions diverse (no near-duplicates) and tightly grounded by the description's content—yet never *mention* the description explicitly.

Figure 5: Instruction for QA Generation.

after feature fusion in SALMONN versus EvA. The Q-Former in SALMONN maps a long sequence of encoder features to a much shorter set of latent queries, thereby introducing temporal compression. In contrast, EvA preserves full temporal resolution: Time-Aware Alignment produces H_{aligned} on the Whisper timeline, and the inject-and-add fusion in Eq. (6) yields E_{fused} without reducing sequence length. This non-compressive design ensures that short transient events and fine-grained temporal structure remain available to the downstream LLM, which is particularly important for perception-heavy benchmarks.

Access to acoustic evidence. Q-Former architectures typically only consume the encoder's final-layer features. For audio encoders, these top-layer representations are often semantically collapsed

Case1

Captions:
The audio opens with a lively live performance featuring rhythmic electric guitars, prominent bass tones, punchy percussion, **and** a mellow synth undertone, all layered with sustained crowd energy. A male vocalist delivers a passionate vocal passage that blends with the driving instrumentation **before** transitioning into instrumental segments. A high-pitched female vocalist emerges briefly, **adding** melodic contrast to the arrangement. The track culminates in an abrupt crescendo of an intense, unified crowd shout that escalates sharply **before** terminating abruptly, leaving a sudden silence.

QA pairs:
Q: What is the primary emotional effect created by the abrupt final shout and silence?
A: A jarring sense of closure with heightened tension release

Q: How many distinct vocalists are present in the audio?
A: 2

Q: What type of vocal presence is heard after the initial male vocalist's passage?
A: A high-pitched female vocalist

Q: What is the primary emotional effect created by the abrupt final shout and silence?
A: A jarring sense of closure with heightened tension release

Figure 6: Instruction for QA Generation.

Case1

Captions:
Persistent wind noise dominates throughout the recording with intermittent radio transmissions containing male speech. The radio communication **begins** abruptly with fragmented phrases mentioning 'Blackbird,' **followed** by three distinct mechanical ticks at approximately the same temporal proximity. A sharp burst of static noise momentarily **interrupts** the transmission **before** resuming with the male voice reiterating 'Blackbird' amidst continuing wind interference. The radio remains active **throughout the recording** with overlapping wind noise maintaining consistent background presence.

QA pairs:
Q: Which sound consistently overlays the entire recording?
A: Wind noise

Q: Is there a sharp burst of static noise present in the audio?
A: Yes

Q: How does the recurrence of 'Blackbird' contribute to the overall atmosphere of tension?
A: It creates a sense of urgency and repetition amid unstable communication, suggesting unresolved or critical information being transmitted through persistent interference.

Figure 7: Instruction for QA Generation.

and may lose low- and mid-level cues that are crucial for environmental sound and event recognition. EvA introduces an explicit cross-layer bypass that aggregates multiple CED layers, H_4, H_8, H_L, into H_{agg} via the two-stage cascaded cross-attention. This allows the fusion module to reuse shallow, mid-level, and high-level acoustic information instead of relying solely on the last encoder layer.

Interaction mechanism. Q-Former modules rely on a bank of static learnable queries that are shared across all inputs: the same latent queries attend to encoder features regardless of the specific audio content. In contrast, EvA performs content-adaptive cross-layer retrieval inside the CED path: H_L first attends to H_8, and the resulting representation then attends to H_4, forming a top–down hierarchy $H_L \rightarrow H_8 \rightarrow H_4$. This hierarchical attention enables the model to selectively recover fine-grained evidence from lower layers conditioned on the current high-level context, which is not possible when only the final encoder layer is exposed to a fixed query set.

Setting	MMAU	MMAR	MMSU	CochlScene
mask 0-2kHz	72.20	55.53	63.72	60.67
mask 2-4kHz	72.30	55.63	63.72	60.91
mask 4-6kHz	72.30	56.24	63.82	60.61
mask 6-8kHz	72.40	55.33	63.88	61.00
mask 0-4kHz	72.30	54.93	63.11	60.67
mask 4-8kHz	71.90	55.43	63.98	60.78
left 0-2kHz	72.10	55.73	63.34	60.34
left 2-4kHz	72.50	54.23	63.52	60.41
left 4-6kHz	71.90	54.12	63.38	60.02
left 6-8kHz	72.80	55.33	62.81	60.27
use 0-8kHz	**73.90**	**59.76**	**62.24**	**74.94**

Table 8: Main results on various audio benchmarks. All numbers are average scores.

Summary. Taken together, these structural differences—multi-layer access, content-adaptive cross-layer retrieval, and non-compressive temporal fusion—provide a different inductive bias from Q-Former–based designs. This is consistent with our ablations in Table 3, where the SALMONN-style Q-Former variant improves over weaker baselines but still underperforms EvA's hierarchical Aggregator on perception benchmarks.

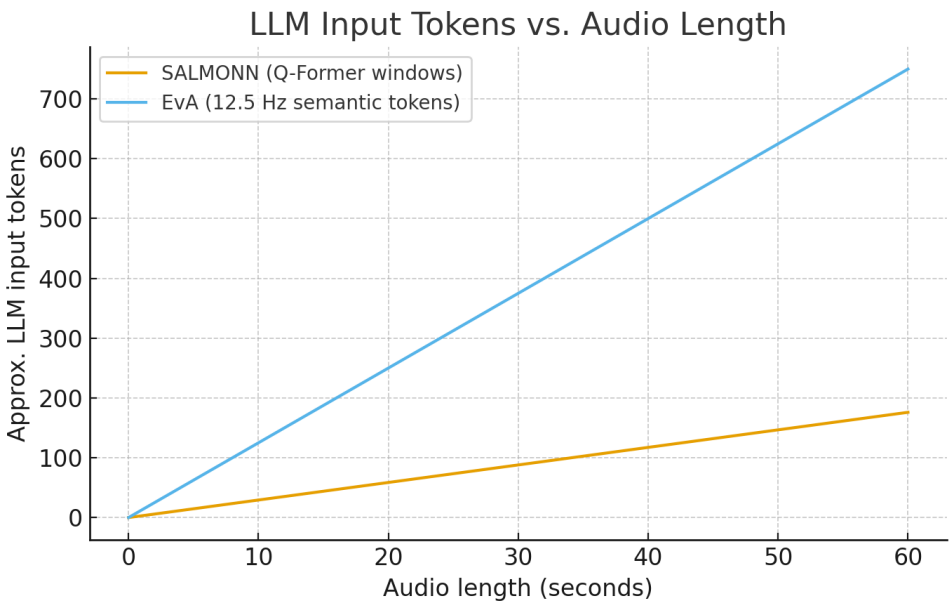

Figure 8: Comparison of audio token sequence length after feature fusion. SALMONN's Q-Former compresses audio tokens into a shorter latent sequence, while EvA preserves sequence length via inject-and-add fusion.