From Sounds to Scenes: A Benchmark for Evaluating Context-Aware Auditory Scene Understanding in Large Audio Language Models

Anonymous ACL submission

Abstract

Recent Large Audio Language Models (LALMs) have achieved remarkable progress in audio perceptual tasks across individual acoustic layers, including speech, sound, and music. However, existing benchmarks predominantly evaluate these layers in isolation, overlooking the complex contextual relationships that arise when multiple acoustic sources co-occur in real-world auditory scenes. Real-world auditory interpretation requires **Context-Aware Auditory Scene Understanding (CASU)**: the ability to comprehend the holistic scene by integrating sound layers. To evaluate this capability, we introduce the CASU benchmark, which assesses whether Audio LLMs can interpret auditory scenes composed of speech, acoustic events (e.g., announcements), and background environments (e.g., traffic), and reason about the logical relationships between these layers. We propose a scalable pipeline for constructing time-accurate, semi-synthetic audio streams by composing real-world scene sounds with synthetic speech. Building on this data, we design four tasks that probe scene understanding: contextual question answering, entity extraction from the scene, speaker role inference, and counterfactual reasoning where scene is manipulated. Experiments across multiple LALMs demonstrate that effective auditory scene understanding requires integration over all auditory layers, rather than reliance on speech or sound alone, underscoring the necessity of CASU for advancing complex audio understanding in LALMs.

1 Introduction

Recent Large Audio-Language Models (LALMs) (Xu and Lin, 2025; Comanici et al., 2025) have demonstrated strong performance across a range of audio perceptual tasks, including Automatic Speech Recognition, sound event detection, and audio captioning. These models effectively bridge acoustic signals and reasoning, but existing benchmarks primarily evaluate such capabilities at the level of isolated acoustic layers, i.e, speech, sound, or music. As illustrated in Figure 1(a), audio understanding is often framed as a single-layer perception task(Sakshi et al., 2024; Clark et al., 2018). Even recent benchmarks that incorporate mixed-layer audio inputs (Figure 1(b)) typically treat co-occurring speech, events, and background sounds

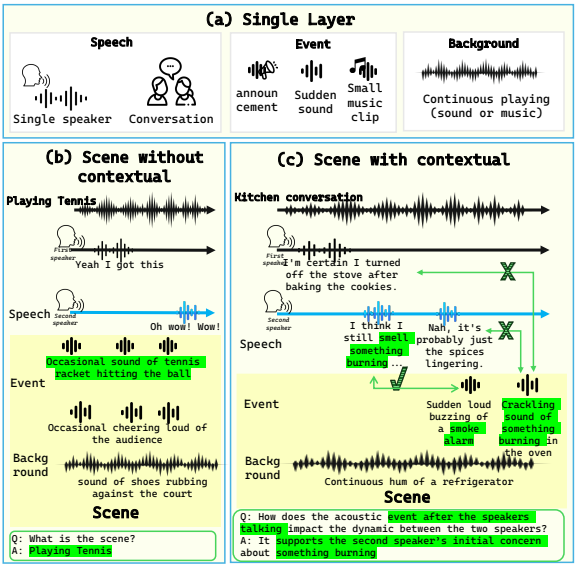


Figure 1: **Comparison of auditory understanding paradigms across benchmarks.** ***(a) Single-layer evaluation***, where prior work focuses on isolated acoustic layers, no cross-layer interactions. ***(b) Mixed-layer scenes without contextual reasoning***, introduced by recent benchmarks, where sounds of speech, events, and background co-occur, but contextual relationships are weakly expressed or not considered for answering questions. ***(c) Context-Aware Auditory Scene Understanding (CASU)***, where sounds of speech, events, and background jointly form a scene. In CASU, elements inside scene may confirm, contradict, or contextualize each other. Purple highlights indicate contextual relationships in the scene related to the question.

as loosely coupled(Ma et al., 2025; Kumar et al., 2025), without explicitly requiring models to reason about their contextual relationships.

In contrast, human auditory understanding is inherently scene-level and context-aware, integrating speech with surrounding acoustic events and background sounds to infer higher-order meaning(Lu et al., 2015). Consider the spoken phrase "It's time to go." If this speech coincides with a school bell (a transient event), it implies a routine dismissal; yet, if accompanied by an emergency siren, the same phrase becomes a critical warning to evacuate(Fuentes et al., 2022). Crucially, non-speech audio elements do not play fixed semantic roles: depending on context, they may function as irrelevant noise or impose decisive logical constraints on the interpretation of spoken content. Effective auditory understanding therefore requires models not only to recognize multiple sound sources, but also to reason about whether and how these layers interact(Du et al., 2025). This gap motivates **Context-Aware Auditory Scene Understanding (CASU)**, illustrated in Figure 1(c), which formalizes auditory scene understanding as a reasoning problem over structured interactions between sounds of speech, events, and background.

To evaluate this capability, we introduce the CASU benchmark, designed to assess whether LALMs can integrate distinct auditory layers to derive contextual meaning at the scene level. CASU employs a scalable semi-synthetic generation pipeline that composes real-world recordings of acoustic scenes (e.g., ambient traffic, room reverberation) and discrete events (e.g., sirens, glass breaking) with semantically controlled synthetic speech, enabling precise control over cross-layer relationships. Building on this pipeline, we define four tasks: contextual reasoning, entity extraction, role inference, and counterfactual reasoning. These tasks explicitly probe models' ability to reason about auditory scenes rather than rely on isolated or incidental cues.

We conduct CASU benchmark on a diverse set of state-of-the-art LALMs. Our results reveal a significant Perception-Understanding Gap: models that achieve near-human performance on single-layer tasks, such as high-fidelity speech transcription, frequently fail on scene understanding tasks that require resolving logical dependencies across other sounds. Furthermore, ablation studies demonstrate that the removal of any single acoustic layer - whether speech, event, or background - leads to a collapse on CASU tasks, confirming that true scene understanding relies on the holistic integration of all auditory dimensions.

In summary, our contributions are threefold:

(1) We introduce Context-Aware Auditory Scene Understanding (CASU), a new paradigm that frames auditory intelligence not as signal filtering, but as understanding and reasoning over the structured interactions among speech, events, and background sounds.

(2) We present a scalable semi-synthetic data generation pipeline that enables precise control over cross-layer contextual relationships, enabling systematical evaluation of scene understanding.

(3) We propose a benchmark with four carefully designed tasks and provide a comprehensive evaluation of state-of-the-art LALMs, revealing persistent limitations in multi-layer auditory understanding and highlighting key directions for future model development.

2 The CASU Benchmark

2.1 Dataset Construction

To rigorously evaluate context-aware auditory sense understanding, we require a dataset that offers precise control over the semantic relationships between acoustic layers. Natural 'in-the-wild' audio recordings rarely provide temporal annotations or semantic relationships among speech, events, and background sounds, making them unsuitable for evaluating fine-grained auditory context-aware reasoning. To address this, we introduce a scalable, semi-synthetic data generation pipeline (Fig. 2). This pipeline operates in a three-stage procedure: Script Generation, Layer Retrieval & Synthesis, and Time-Aligned Composition.

Stage 1: Script Generation and Semantic Blueprint. The script of each audio sample is a structured JSON script, generated by a GPT-4o(Achiam et al., 2023). We prompt the GPT to synthesize a complex auditory scenario defined by the superposition of three distinct semantic layers:

1. **Background Sound ($T_{\mathbf{B}}$):** The continuous environmental sound (e.g., *Indoor Private: Dimly Lit Basement. Constant low hum of a generator.*).

2. **Speech ($T_{\mathbf{P}}$):** Human Speech Content.

3. **Acoustic Events ($T_{\mathbf{E}}$):** Discrete, transient sounds that occur in the scene (e.g., *sudden*

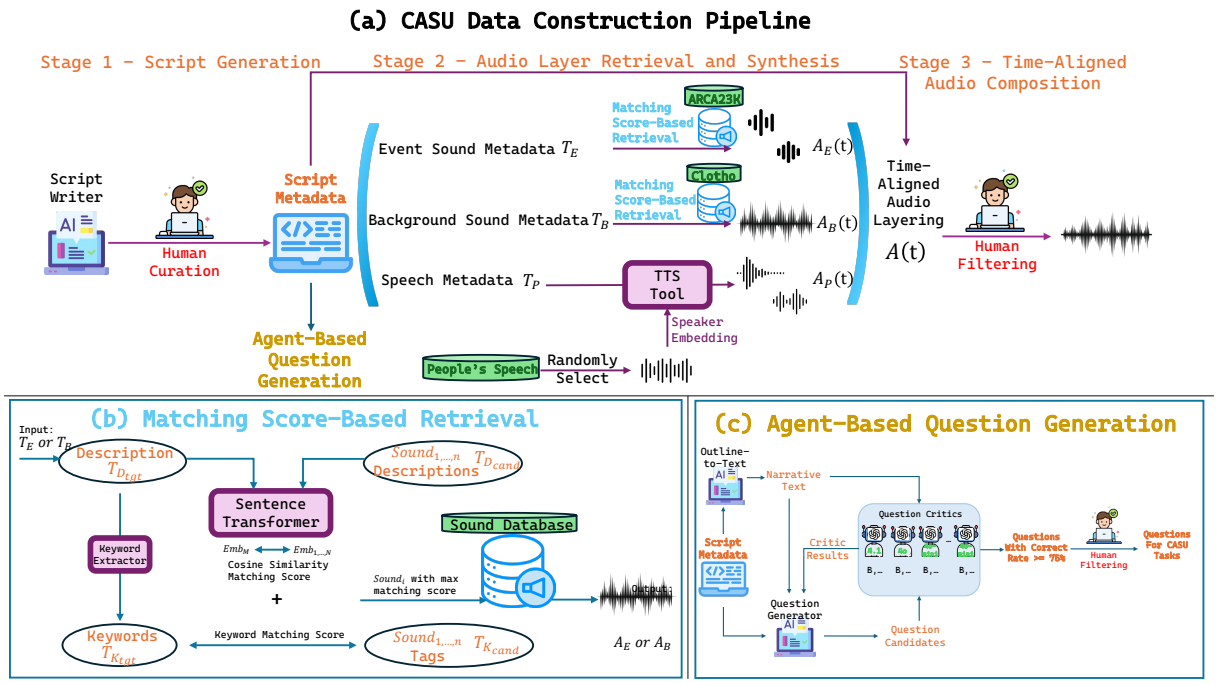


Figure 2: (a) The overview of the CASU data construction pipeline. In the pipeline, we propose (b) Matching-Score Based Retrieval for retriving the most suitable sound based on script, and (c) Agent-Based Question Generation, which can generate questions for all CASU tasks given script metadata.

loud bang).

As shown in the script example (Appendix A Fig 5), each component is assigned a text description T_D, a start timestamp t_{start}, and an end timestamp t_{end}. These scripts serve as the semantic blueprint that explicitly states the narrative contextual relationships across layers, which later enable systematic evaluation of scene understanding. A **Human Curation** step follows generation to filter for logical consistency and narrative reality.

Stage 2: Audio Layer Retrieval and Synthesis. Once the script is finalized, we instantiate the auditory layers using a hybrid approach of Speech Layer Synthesis and Non-Speech Layer Retrieval.

Speech Layer Synthesis ($\mathbf{A_P}$). To generate naturalistic speech, we utilize Zonos[1], a high-fidelity TTS tools. To ensure speaker diversity across the benchmark, we condition the TTS model on random speaker embeddings sampled from the **People's Speech** dataset(Galvez et al., 2021).

Non-Speech Layer Retrieval ($\mathbf{A_B}, \mathbf{A_E}$). For the background sound and event sound, we employ a **Matching Score-Based Retrieval** system (Fig. 2b) to select the optimal audio candidates from source datasets (**Clotho**(Drossos et al., 2020) for background scenes and **ARCA23K**(Iqbal et al., 2021) for short-duration events).

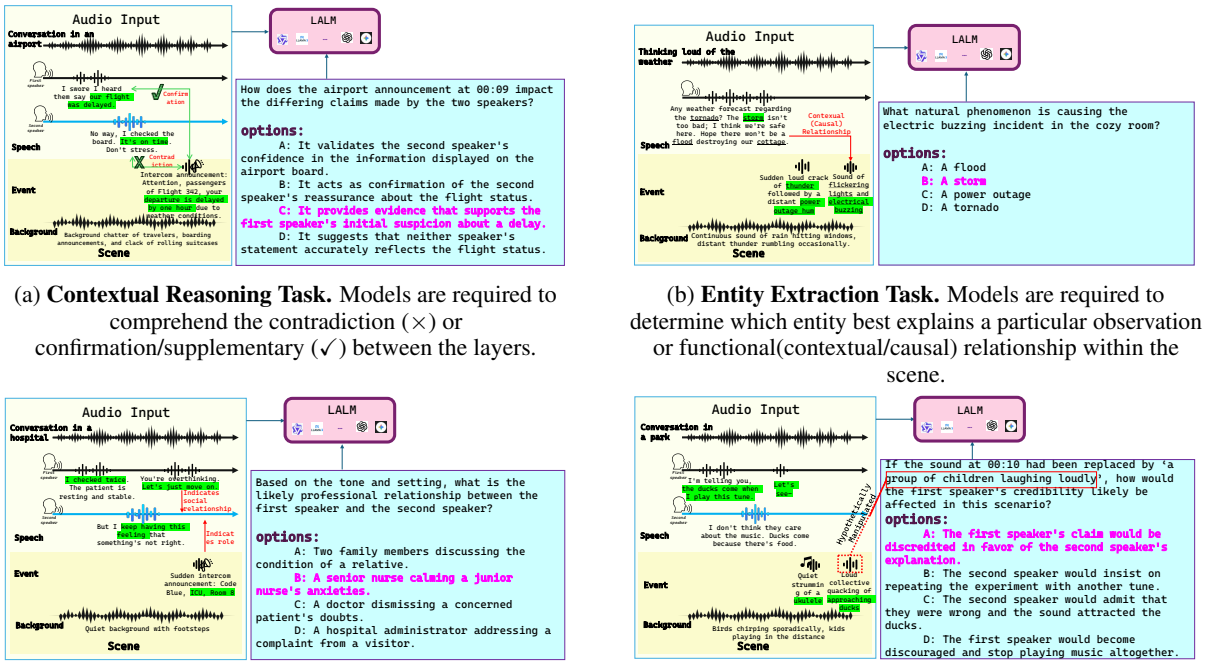
Matching Score-Based Retrieval. (Fig. 2b) Given a target description $T_{D_{\text{tgt}}}$ describing the sound T_B or T_E, we retrieve the most suitable candidate from a sound database where each audio sample is annotated with (i) a text description $T_{D_{\text{cand}}}$ and (ii) a set of categorical tags $T_{K_{\text{cand}}}$ that characterize its acoustic class. We first extract a set of keywords $T_{K_{\text{tgt}}}$ from $T_{D_{\text{tgt}}}$. Then, we compute semantic similarity at both description-level and keyword-level, as categorical tags are particularly informative for defining and disambiguating sound types (e.g., *glass breaking*, *engine*, *laughter*). The matching score S_{match} is thus formulated as a weighted combination of both similarity scores:

$$S_{\text{match}} = \alpha \cdot \text{Sim}_{\cos}\left(f_\theta(T_{D_{\text{tgt}}}), f_\theta(T_{D_{\text{cand}}})\right) + (1-\alpha) \cdot \text{Sim}_{\cos}\left(f_\theta(T_{K_{\text{tgt}}}), f_\theta(T_{K_{\text{cand}}})\right) \quad (1)$$

where $f_\theta(\cdot)$ denotes Sentence Transformer(Reimers and Gurevych, 2019) that maps text to semantic vectors, and $\text{Sim}_{\cos}$ represents the cosine similarity. The hyperparameter α balances semantic intent with keyword matching. The candidate with the maximal S_{match} is selected for synthesis, denoted as $\mathbf{A_B}$ or $\mathbf{A_E}$

Stage 3: Time-Aligned Audio Composition. The final audio stream $\mathbf{A}(t)$ is constructed simulating natural recording, i.e, treating the retrieved

[1]https://github.com/Zyphra/Zonos



(a) **Contextual Reasoning Task.** Models are required to comprehend the contradiction (×) or confirmation/supplementary (✓) between the layers.

(b) **Entity Extraction Task.** Models are required to determine which entity best explains a particular observation or functional(contextual/causal) relationship within the scene.

(c) **Role Inference Task.** Models are required to infer contextual roles or interpersonal relationships from the combination of speech and scene context.

(d) **Counterfactual Reasoning Task.** Models are required to reason about causal dependencies within the scene, and how a key interpretation or conclusion would change under an hypothetically altered acoustic evidence.

Figure 3: **CASU Task Suite.** We introduce four new tasks, Contextual Reasoning, Entity Extraction, Role Inference, and Counterfactual Reasoning, which test different capabilities of LALMs on scene understanding.

Background Sound as the base layer and aligning the Speech and Event layers at their designated temporal intervals defined in the script:

$$\mathbf{A}(t) = \mathbf{A}_{\mathbf{B}}(t) + \sum_{i=1}^{N_p} \mathbf{A}_{\mathbf{P}}^{(i)}(t - t_{P:\text{start}}^{(i)}) \cdot \mathbb{I}_{[t_{P:\text{start}}^{(i)}, t_{P:\text{end}}^{(i)}]} + \sum_{j=1}^{N_e} \mathbf{A}_{\mathbf{E}}^{(j)}(t - t_{E:\text{start}}^{(j)}) \cdot \mathbb{I}_{[t_{E:\text{start}}^{(j)}, t_{E:\text{end}}^{(j)}]} \quad (2)$$

where $\mathbb{I}_{[t_{P:\text{start}}, t_{P:\text{end}}]}$ and $\mathbb{I}_{[t_{E:\text{start}}, t_{E:\text{end}}]}$ are indicator functions that are 1 during the active interval of the specific speech segment or event and 0 otherwise. Following composition, a **Human Filtering** stage is applied to verify that the mixing levels are balanced and that the synthetic composition retains the logical causality intended by the script writer.

2.2 Tasks and Questions

A core objective of the CASU benchmark is to evaluate whether LALMs can interpret the contextual logic between speech, background, and events embedded within a complex scene. To this end, besides the **scene description** task, which requires LALM to recognize and describe all events and transcribe speeches in the audio, we also design four tasks that require models to integrate all contextual acoustic evidence and reason about logical, causal, and social aspects of the audio. These tasks move beyond traditional perceptual benchmarks, challenging models to perform deep semantic inference in auditory contexts where background sounds and events may either constrain or supplement speech topics or assertions. Fig. 3 provides representative examples for each task. In all cases, LALMs are given an auditory input $\mathbf{A}(t)$ of a scene and a text question Q (Appendix A Table 7).

Contextual Reasoning. The first task, *Contextual Reasoning*, assesses whether a model can reconcile spoken assertions with the broader acoustic context. In many real scenes, the meaning of a dialogue is contingent on background conditions or event that provide supplementary or contradictory evidence. In the example (Fig. 3a), a speaker's expressed confidence about a situation may be invalidated by an event cue that indicates otherwise, or conversely, may be demonstrated by a situational announcement.

Entity Extraction. Real-world audio understanding often involves identifying entities that are not directly stated but are implied by the sounds in the scene. The *Entity Extraction* task is designed to evaluate this capability. Given an audio (Fig. 3b), the query asks the model to determine which entity in the scene best explains a particular observation

or functional relationship within the scene.

Role Inference. Speeches, together with the sounds in the scene, encode social roles, speaker identities, and relational dynamics. To capture this dimension, the *Role Inference* task evaluates whether models can infer contextual roles or interpersonal relationships from the combination of speech and scene context. This task involves determining the most possible role for a solo speaker in the scene, or the social relationship between two speakers - for example (Fig. 3c), whether one speaker's choice of words indicate a professional authority over another, and what occupations of them are indicated by the background or event.

Counterfactual Reasoning. Finally, we introduce *Counterfactual Reasoning* as a means of evaluating a model's causal understanding of the auditory world. In this task (Fig. 3d), the associated question asks how a key interpretation or conclusion would change under an hypothetically altered acoustic evidence. Counterfactual reasoning lies at the heart of human causal inference: to understand not just what is, but what might have been if conditions were different. By introducing hypothetically counterfactual manipulations into the audio input, this task pushes models to reason about causal dependencies between environmental cues and semantic interpretations in a way that goes beyond descriptive understanding.

Question Generation. we employ an agent-based question generation framework for all 4 tasks illustrated in Fig. 2c(Wang et al., 2024). Starting from the structured script metadata of each audio scene, an LLM first produces a narrative textual description that state the scene containing temporal and causal relations. Based on the description and script, a question generation agent proposes multiple candidate questions aligned with the target task type. A panel of independent LLM-based critic agents will evaluate each candidate question by attempting to solve it under the same input conditions, retaining only those questions for which a sufficient consensus on the correct answer is achieved. Human filtering is followed to ensure logic validity, difficulty and category suitable. We defer statistics of dataset to Appendix A.1.

3 Experiments

Benchmarking Candidates. We compare a range of models for audio understanding, including (i) Qwen-series open-source models, with different versions and variants such as audio model Qwen-2-Audio-Instruct(Chu et al., 2024), and omni-models Qwen-2.5-Omni(Xu, 2025) and Qwen-3-Omni-Instruct(Xu and Lin, 2025); (ii) audio flamingo series models such as Audio-Flamingo-Chat(Kong et al., 2024), Audio Flamingo 2(Ghosh et al., 2025b), and Audio Flamingo 3(Ghosh et al., 2025a); (iii) other open-source LALMs such as Voxtral(Liu et al., 2025) and LLaMa-Omni(Fang et al., 2025), which are mainly trained on speech data; and SALMONN(Tang et al., 2024), LTU(Gong et al., 2023b), LTU-AS(Gong et al., 2023a), which are trained on diverse sound data; (iv) closed-source LALMs GPT-4o Audio(Achiam et al., 2023) and Gemini 2.0 Flash(Comanici et al., 2025); (iv) cascaded models with Qwen3-Omni-Captioner or GPT-4o-Transcribe as Captioner and GPT-4o or Qwen3-Instruct as Reansoner. Qwen3-Omni-Captioner has the same parameter as Qwen3-Omni-Instructor but can only be used for audio description. Reasoner will receive textual audio description as input and answer CASU questions.

Evaluation Metrics. For the **scene description task**, we have the narrative textual description of each audio (Fig. 2c) as reference text. Therefore, we use BLEU-4(Papineni et al., 2002) and BertScore(Reimers and Gurevych, 2019) compared to the reference text to evaluate the semantic alignment of audio description. To evaluate the perceptual precision on single layers (sound events, background, speeches), we use LLM-as-a-Judge(Zheng et al., 2023) to estimate the percentage of recognized event in the description compared to the audio script, and Word Error Rate (WER)(Ali and Renals, 2018) for speech transcription. For the four new **CASU scene understanding tasks** we propose, we report the accuracy on the questions. All questions are multiple-choice tests with 4 options.

4 Results and Discussions

4.1 Main Results

Table 1 reports the performance of a diverse set of LALMs on the CASU benchmark, covering both perceptual (scene description) and scene understanding tasks. Our analysis reveals that strong single-layer perception is necessary but insufficient for scene understanding.

Finding 1: The "Perception-Understanding Gap". Our results establish a nuanced relationship between perceptual fidelity and scene reasoning. First, the models achieving the highest accuracy

Table 1: CASU Benchmark results with different audio understanding models. **Bold** indicates best performance, underline indicates second-best performance. We report *1 - WER* instead of *WER* for the transcription performance, to make sure higher value represents better transcription performance.

Models	Size	Scene Description - Perception ↑				Scene Understanding - CASU (%) ↑			
		BLEU-4	BertScore	Event Match	1 - WER (Transcribe)	Contextual Reasoning	Entity Extraction	Role Inference	Counterfactual Reasoning
Audio Flamingo Chat(Kong et al., 2024)	3B	3.09	0.72	0.24	0.92	31.82	29.17	34.37	34.36
Audio Flamingo 2(Ghosh et al., 2025b)	3B	4.54	0.75	0.28	0.95	38.60	33.24	37.05	39.45
Audio Flamingo 3(Ghosh et al., 2025a)	7B	9.20	0.78	0.46	0.95	56.88	45.12	45.83	53.00
Qwen2-Audio-Instruct(Chu et al., 2024)	7B	1.66	0.76	0.41	0.95	44.02	40.81	40.00	37.81
Qwen2.5-Omni (Xu, 2025)	7B	4.48	0.73	0.44	0.96	62.06	61.10	51.37	65.38
Qwen3-Omni-30B-A3B-Instruct(Xu and Lin, 2025)	30B	**12.80**	0.81	0.65	0.96	71.18	68.85	63.51	74.50
LTU(Gong et al., 2023b)	7B	2.88	0.70	0.15	0.90	23.92	21.87	24.18	20.99
LTU-AS(Gong et al., 2023a)	7B	2.59	0.71	0.19	0.92	20.26	19.19	20.52	21.27
LLaMa-Omni(Fang et al., 2025)	8B	1.64	0.72	0.20	**0.98**	53.32	38.27	49.09	57.60
Mistral Voxtral(Liu et al., 2025)	24B	1.32	0.79	0.27	0.97	52.52	41.67	51.64	55.99
SALMONN(Tang et al., 2024)	13B	1.05	0.76	0.36	0.90	43.21	50.96	39.73	46.95
GPT-4o Audio(Achiam et al., 2023)	-	10.53	0.82	0.72	**0.98**	**74.02**	70.98	68.58	**74.96**
Gemini 2.0 Flash(Comanici et al., 2025)	-	9.19	**0.85**	**0.75**	**0.98**	73.30	**72.77**	**69.05**	73.83
Qwen3-Omni-30B-A3B-Captioner + Qwen3-30B-A3B-Instruct-2507	-	3.61	0.77	0.62	0.97	62.08	60.40	54.28	69.63
Qwen3-Omni-30B-A3B-Captioner + GPT-4o	-	3.61	0.77	0.62	0.97	63.03	59.70	54.97	68.46
GPT-4o-transcribe + Qwen3-30B-A3B-Instruct-2507	-	1.08	0.74	-	**0.98**	57.95	41.34	51.59	58.17
GPT-4o-transcribe + GPT-4o	-	1.08	0.74	-	**0.98**	59.77	41.75	52.46	58.10

on CASU tasks (*GPT-4o Audio, Gemini 2.0 Flash, Qwen3-Omni-Instruct*) also dominate perceptual metrics (BLEU > 9, Event Match > 0.6) However, a significant "Perception-Understanding Gap" emerges when looking at speech-centric models. Models like *Audio Flamingo 3* and *Mistral Voxtral* achieve near-ceiling speech transcription performance ($1 - WER \geq 0.95$) yet lag significantly in scene understanding, often trailing state-of-the-art models by 15%. This highlights that high-fidelity transcription is insufficient for CASU; the bottleneck for these models is likely their weaker event perception (Event Match < 0.5), preventing them from utilizing the acoustic scene as a logical constraint. Conversely, lower-fidelity transcription will lead to large performance drop (*SALMONN*). Finally, the task remains challenging: even the most capable models achieve $70 - 74\%$ accuracy, suggesting that current LALMs still struggle to fully resolve the complex inter-layer reasoning required by CASU. This demonstrates that accurate recognition of any single layer (speech or sound) is insufficient; effective scene understanding requires the holistic integration of speech, events, and background.

Finding 2: Joint Processing Beats Cascaded Reasoning. Omni-modal models (e.g., *Qwen3-Omni*, *GPT-4o Audio*) consistently outperform cascaded pipelines utilizing comparable backbones (*Qwen3-Omni-Captioner* + *Qwen3-Instruct*), particularly on the Counterfactual and Contextual Reasoning tasks. Cascaded systems, which convert audio to a textual description before reasoning, suffer from an "information loss" due to insufficient event match ($eventmatch = 0.62$) or transcription ($1 - WER = 0.97$). They inevitably discard subtle acoustic cues, such as the reverberation suggesting a large hall, which are tough to be written down for the audio description task. These lost signals often serve as the pivotal context needed to understand the scene or resolve logical ambiguities in CASU.

Finding 3: Different Layers in the Scene Play Different Roles, Not Just Noise. The performance variance on Entity Extraction vs. Role Inference reveals that CASU requires models to dynamically weigh contributions across auditory layers. Entity Extraction relies more on non-speech acoustic evidence. Models with strong sound recognition (*Qwen2.5-Omni*, *SALMONN*) excel here (*SALMONN* outperforms *Voxtral* by 9% on entity extraction, despite much worse performance on the other tasks), provided speech quality is not severely degraded. Role Inference is linguistically heavy but contextually grounded; it requires the model to infer social dynamics (e.g., "Customer" vs. "Server") which are often signaled by the timing of background events (e.g., a register beep after a sentence). Models with better transcription performance (*Voxtral*, *LLaMa-Omni*) perform good on this task. This validates our core hypothesis: effective LALMs must actively discern when an acoustic layer serves as a semantic anchor for the scene. Contextual Reasoning and Counterfactual Reasoning are the most demanding, requiring the model to disentangle layers. Models with low event perception performance or transcription performance (*LTU*, *LTU-AS*) collapse on both tasks.

4.2 Ablation Study on Audio Layers

To strictly verify the contribution of distinct auditory layers to scene understanding, we conducted

a component-wise ablation study using *Qwen2.5-Omni*. We masked the speech (A_P), events (A_E), and background (A_B) layers respectively with Gaussian noise, evaluating performance across all permutation subsets. Table 2 details the results.

Table 2: Results of Qwen2.5-Omni on audios without sounds of background(A_B)/event(A_E)/speech(A_P).

Audio Layers	Scene Understanding (%) ↑			
A_P, A_E, A_B	Contextual Reasoning	Entity Extraction	Role Inference	Counterfactual Reasoning
×, ×, ×	20.44	17.47	22.88	27.90
×, ×, ✓	24.89	26.17	31.93	21.16
×, ✓, ×	26.19	37.19	31.95	23.24
×, ✓, ✓	26.10	35.79	33.56	20.84
✓, ×, ×	57.53	53.41	49.81	55.18
✓, ×, ✓	58.97	54.11	49.06	56.75
✓, ✓, ×	60.58	60.46	51.47	64.33
✓, ✓, ✓	62.06	61.10	51.37	65.38

Finding 4: Speech serves as the semantic foundation, but the Scene provides the logical constraints. As shown in Table 2, speech (A_P) is the dominant carrier of information; removing it (rows 1-4) causes a performance collapse across all tasks, with accuracy down to near random guess levels (20-30%). However, relying on speech alone (Row 5: $A_P, \times, \times$) is insufficient for CASU. The complete removal of the acoustic scene (A_E and A_B) results in a significant performance degradation compared to the full model (Row 8), particularly in Counterfactual Reasoning (-10.2%) and *Entity Extraction* (-7.7%). This empirically proves that current LALMs do not solve these tasks by merely "guessing" from text; they actively utilize non-speech acoustic cues to ground their reasoning.

Events contribute to entity extraction. Adding discrete events (A_E) to the speech baseline (Row 5 vs. Row 7) yields the largest gain for Entity Extraction (53.41% → 60.46%). This confirms that models locate physical entities primarily through their acoustic signatures (e.g., a "dog" is confirmed by barking, not just the speaker mentioning a pet)

4.3 Scene Difficulty on Speech

Our dataset contains both two types of speech: One-Person Monologue and Inter-Person Conversation. Fig. 4 shows how different speech difficulty influence scene understanding.

Finding 5: Multi-Person Scene Setting Increases Understanding Complexity. We observe a consistent performance degradation when shifting from single-person monologues (Blue bars) to inter-person conversations (Orange bars) across all four tasks. This indicates that conversation scene

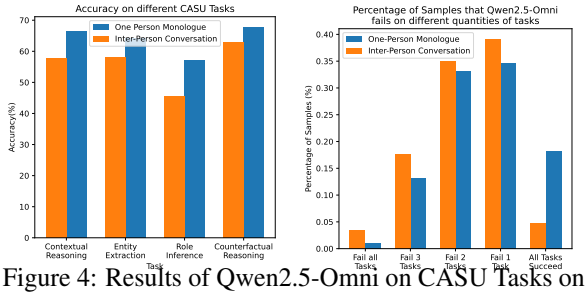


Figure 4: Results of Qwen2.5-Omni on CASU Tasks on different speech settings for the scene.

introduce significant difficulty on integrating other sound in the scene. As in Fig. 4b, models often fail on more percentage of audio samples. Role Inference task suffers the most, where accuracy drops by approximately 12%. This sharp decline validates the hypothesis that social reasoning is harder in conversations than monologue.

4.4 Influence of Scene Hints

To further investigate whether models struggle with percepting the scene, we evaluate performance when an explicit textual hint for the scene place is provided (e.g., "This audio is recorded in a [Location]"). Table 3 compares the performance with and without this auxiliary information.

Table 3: CASU Benchmark results with scene text hint.

Model	Prompt Condition	Scene Understanding (%) ↑			
		Contextual Reasoning	Entity Extraction	Role Inference	Counterfactual Reasoning
Qwen2.5-Omni	with hint	64.14	63.90	55.19	65.91
	without hint	62.06	61.10	51.37	65.38
GPT-4o Audio	with hint	74.15	72.36	70.01	74.42
	without hint	74.02	70.98	68.58	74.96

Finding 6: Textual Hint Reduces Acoustic Ambiguity. Providing an informative scene hint consistently improves performance, particularly for open-source models. *Qwen2.5-Omni* sees a performance boost across all tasks, with the most gain in Role Inference (+3.82%). This suggests that once the scene is explicitly defined (e.g., "hospital"), the model can more effectively retrieve the associated social schemas (e.g., Doctor/Patient dynamics) for reasoning. However, the benefit of textual hints is less for *GPT-4o Audio*. This indicates that stronger LALMs likely possess a robust internal representation of the auditory scene; they may have already implicitly inferred the context from the audio alone, rendering the textual hint redundant.

4.5 Error Analysis

We present two different types of errors in Table 4. The first one is due to LALMs' perception or rea-

Table 4: Illustration of different error types on the CASU scene understanding questions.

Error Type	Perception or Reasoning Error	Self-Reasoning Error
Question	What is the most plausible relationship between the first and the second speaker? A: Two strangers who happened to walk the same forest trail by accident. B: A parent calming their child's nighttime fears during a camping trip. C: Two friends on an adventurous night hike, one reassuring the other. D: A wildlife expert guiding a nervous tourist on a nature walk.	Which particular sound causes the interruption of the tranquil cooking environment in the kitchen? A: The sound of alarm, B: The blender malfunction noise, C: The beep from the smoke detector, D: The beep from the kettle with boiling water
Wrong Answer	Answer: B. Justification: The audio contains a child's voice expressing fear (*Perception Error*), which is a key indicator (*No Causal Relationship*) of a parent-child relationship.	Answer: C. Justification: Across all options, the beep from the smoke detector is more likely to interrupt ... This is evident from the sudden, loud, and disruptive nature (*Not cue in audio*) of the smoke detector's beep
Correct Answer	Answer: C. Justification: ...the interaction is 'alarm/dismissive' but the **conversation is comforting, suggesting familiarity and comfort (Friends)**. ...No adult-like dialogue so B is incorrect...	Answer: D. Justification: ... This is **evident from the sound of boiling water in the audio**, ..., causing a disruption in the otherwise peaceful scene.

soning error, and the second type is because the LALMs may use the information in the question or the internal knowledge for reasoning, not the cues or information in the audio.

5 Related Work

Large Audio-Language Models (LALMs). Recent advancements in LALMs have shifted from specialized single-task models to unified "omni-modal" architectures capable of processing speech, music, and sound events(Xu and Lin, 2025). Early works like SALMONN(Tang et al., 2024) and Qwen-Audio(Chu et al., 2024) demonstrated strong capabilities in separate perception tasks, such as automatic speech recognition (ASR) and audio captioning. More recently, models like GPT-4o Audio(Achiam et al., 2023), Gemini 2.0 Flash(Comanici et al., 2025), and Audio Flamingo(Ghosh et al., 2025a) have pushed the boundaries of general auditory reasoning. Despite these strides, most existing LALMs are trained and evaluated primarily on loosely coupled audio streams. They often treat non-speech sounds as secondary background signals rather than semantic anchors, limiting their ability to resolve scenarios where the acoustic scene (e.g., an emergency siren) fundamentally alters the logical interpretation of the spoken content.

Audio Understanding and Reasoning Benchmarks. Evaluating holistic audio intelligence remains a significant challenge. Initial benchmarks focused on specific verticals: OpenASQA(Gong et al., 2023a) for speech, CompA(Ghosh et al., 2023) for sound events, and MusicBench(Melechovsky et al., 2024) for musical understanding. Recognizing the need for broader evaluation, recent suites like AIR-Bench(Yang et al., 2024) and AudioBench(Wang et al., 2025) have unified diverse datasets to test fundamental tasks like ASR and scene classification. To probe deeper reasoning, MMAU(Sakshi et al., 2024) introduced massive multi-disciplinary QA pairs across speech, sound, and music, while MMAU-Pro(Kumar et al., 2025) and SAKURA(Yang et al., 2025) tested hierarchical reasoning. Similarly, Dynamic-SUPERB(Huang et al., 2024) expanded instruction-tuned evaluation to cover over 180 tasks. While these benchmarks assess the breadth of audio knowledge, they often treat mixed audio as a loose collection of signals. They rarely evaluate Context-Aware case, where the model must actively discern whether the background acts as noise or as a critical semantic anchor that alters the logic of the foreground speech. Our work fills this gap by formalizing scene-level interaction and reasoning.

6 Conclusion

We introduced Context-Aware Auditory Scene Understanding (CASU), a new paradigm that shifts audio evaluation from isolated perception to holistic reasoning over interacting acoustic layers. Our findings demonstrate that true auditory intelligence requires models to actively discern the semantic utility of non-speech sounds rather than treating them as mere noise. We hope CASU serves as a critical stepping stone for future LALMs to bridge the divide between recognizing signals and understanding real-world auditory scenes.

7 Limitations

Despite the rigorous semi-synthetic data generation pipeline together with human-in-the-lop filtering, our CASU benchmark might remain synthetic compared to fully human-labeled and curated datasets such as MMAR (Ma et al., 2025), MMAU-PRO (Kumar et al., 2025). However, our underlying pipeline is modular and configurable by design, facilitating future adaptation and generalization towards different domains and settings.

Our provided dataset contains audio scenes under 30 seconds, due to current LALM's limitation on processing audios. Some scene might be "expedited" (i.e, events happen soon after the speech, but in real world there might be a time gap), as we are trying to crowd real-world scene into 30 seconds clip. Without loss of generality, our proposed pipeline can simulate real world scene. And our pipeline can definitely generate longer audios if furture LALMs can process longer audios.

In addition, due to LALMs on audio processing limit ($< 30s$), introducing conversations with more than 2 persons make the audio hard-to-understand for human (lots of speech overlapping). Therefore, current audios are confined to 1-person monologue and 2-person conversational scenarios. However, our pipeline can serve as a robust baseline for future complex interactions. The constraint of 2-person conversation facilitates an incremental progression towards natural human interactions without introducing unnecessary confounding variables associated with intricate multi-party conversational settings. We leave the expansion to support N-person conversational dynamics for future works.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmed Ali and Steve Renals. 2018. Word error rate estimation for speech recognition: e-wer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 20–24. Association for Computational Linguistics (ACL).

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

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

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*.

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.

Henghui Du, Guangyao Li, Chang Zhou, Chunjie Zhang, Alan Zhao, and Di Hu. 2025. Crab: A unified audio-visual scene understanding model with explicit cooperation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18804–18814.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. LLaMA-Omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations*.

Magdalena Fuentes, Bea Steers, Pablo Zinemanas, Martin Rocamora, Luca Bondi, Julia Wilkins, Qianyi Shi, Yao Hou, Samarjit Das, Xavier Serra, and 1 others. 2022. Urban sound & sight: Dataset and benchmark for audio-visual urban scene understanding. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 141–145. IEEE.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *arXiv preprint arXiv:2111.09344*.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025a. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025b. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Forty-second International Conference on Machine Learning*.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Evuru, S Ramaneswaran, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2023. Compa: Addressing the gap in compositional reasoning in audio-language models. *arXiv preprint arXiv:2310.08753*.

Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023a. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

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

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, and 1 others. 2024. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12136–12140. IEEE.

T. Iqbal, Y. Cao, A. Bailey, M. D. Plumbley, and W. Wang. 2021. ARCA23K: An audio dataset for investigating open-set label noise. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, pages 201–205, Barcelona, Spain.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *International Conference on Machine Learning*, pages 25125–25148. PMLR.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeong-gon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, and 1 others. 2025. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*.

Alexander H Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, and 1 others. 2025. Voxtral. *arXiv preprint arXiv:2507.13264*.

Tong Lu, Gongyou Wang, and Feng Su. 2015. Context-based environmental audio event recognition for scene understanding. *Multimedia Systems*, 21(5):507–524.

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*.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8293–8316.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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*.

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

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. 2025. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Jin Xu. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu and Junyang Lin. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Chih-Kai Yang, Neo Ho, Yen-Ting Piao, and Hung-yi Lee. 2025. Sakura: On the multi-hop reasoning of large audio-language models based on speech and audio information. *arXiv preprint arXiv:2505.13237*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and 1 others. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Dataset

Our proposed CASU data construction pipeline can generate a large number of audios and questions. In this section, we report the statistics and examples of the dataset we use for the experiment.

A.1 Dataset Statistics

Table 5 summarizes the final statistics of the CASU dataset. It details the scale of the benchmark, which comprises over 1,500 unique audio scenes and 6,000 distinct questions. Notably, the average justification length of ~ 43 words highlights the depth of reasoning required, as models must explain the logical connections between the ~ 13.5 second audio clips and the ground truth answers.Table

Table 5: Statistics of CASU Benchmark Dataset

Statistics	**Number**
Total Audios	1557
Total Questions	6228
Task Categories	4
Avg. Ques. Length	20.70 words
Avg. Ans. Length	2 words
Avg. Justification Length	42.91 words
Avg. Audio Length	13.50 sec

Table 6 reports statistics from the human verification stage of our semi-synthetic generation pipeline. To ensure high fidelity, we filtered a total of 2,100 generated candidates, retaining only those that met strict semantic consistency criteria (a 74.14% acceptance rate). The table also reflects the complexity of the synthesized scenes, which average 2.42 distinct speech segments and nearly 2 retrieved background audio layers per sample.

Table 6: Statistics of CASU Data Construction Pipeline

Statistics	**Number**
Total Generated Audios	2100
Filtered Audios	543
Audio Kept Rate	74.14%
Avg. Speech Segment	2.42
Avg. Num of Retrieved Audio	1.93

A.2 Examples of Audio Script, Hint, and Questions.

Fig. 5 presents raw JSON scripts for two single CASU instances. This structured manifest serves as the blueprint for audio synthesis, explicitly defining the temporal and semantic relationships between the three acoustic layers:

1. Acoustic Scene (T_B): Sets the continuous environmental context (e.g., "Futuristic Suburban Street").

2. Speech (T_P): Defines the dialogue content and prosody (e.g., Speaker A's "Cheerful" tone vs. Speaker B's "Skeptical" tone).

3. Event (T_E): Crucially, this block encodes the logical constraints. In this specific example, the "Sudden rumble of thunder" at 00:10.50 is deliberately timed to occur immediately after Speaker A claims the weather is fine. The notes field explicitly captures this interaction ("Thunderstorm sound contradicts Speaker A"), ensuring that the generated ground-truth answer for the reasoning task is rigorously supported by the acoustic evidence. Some events contain "with speech" domain if the event contains some human talking (for example, an announcement).

We present concrete CASU question examples for two distinct auditory scenes.

Table 7 illustrates a "Security/Conference Room" scenario (Fig. 5 Left). Here, the logical conflict arises between the speaker's verbal assurance ("everything's under control") and the acoustic reality (a "high-pitched security alarm"). The model must prioritize the acoustic event over the speech to correctly answer the Contextual Reasoning and Role Inference questions.

Table 8 shows a conversation about weather (Fig. 5 Right). This example highlights a different logical structure: an environmental event ("sudden thunder") serves as empirical evidence that contradicts a speaker's reliance on a tool (the weather app), requiring the model to validate the skepticism of the second speaker.

Both tables detail the specific prompts for all four CASU tasks—Contextual Reasoning, Entity Extraction, Role Inference, and Counterfactual Reasoning. Crucially, the "Justification" column in both tables confirms that the ground truth is not a simple label, but a structured reasoning chain that explicitly binds the acoustic evidence to the semantic conclusion.

B Prompts

We present all prompts for the whole pipeline, data synthetic, and model test on this section.

Prompts for Audio Script Generation are presented at Table 9.

Prompts for Scene Description and Scene Understanding Tasks are presented at Table 10.

```
{
  "acoustic scene": {
    "content": "Indoor Office: Near-Future Conference Room",
    "start": "00:00.00",
    "end": "00:12.50",
    "notes": "Steady hum of high-tech air filtration, quiet ticking of a digital clock, and soft whirring of futuristic holographic projectors."
  },
  "speech": [
    {
      "content": "The security system hasn't triggered any alerts today.",
      "start": "00:02.50",
      "end": "00:05.00",
      "notes": "Spoken with assured calmness."
    },
    {
      "content": "Everything's under control.",
      "start": "00:06.00",
      "end": "00:07.50",
      "notes": "Said with quiet authority and a hint of pride."
    }
  ],
  "event": [
    {
      "content": "Sudden blare of a high-pitched security alarm",
      "start": "00:08.00",
      "end": "00:09.50",
      "notes": "Contradicts the speaker's statement by indicating a security breach.",
      "with speech": ""
    },
    {
      "content": "Urgent announcement via intercom",
      "with speech": "Intruder alert on the third floor.",
      "start": "00:10.00",
      "end": "00:12.00",
      "notes": "Solidifies the contradiction by highlighting an urgent situation."
    }
  ]
}
```



```
{
  "acoustic scene": {
    "content": "Futuristic Suburban Street: Afternoon",
    "start": "00:00.00",
    "end": "00:15.00",
    "notes": "Ambient sound of whispering trees, soft humming from robotic lawn mowers, and children playing with flying gadgets."
  },
  "speech": [
    {
      "content": "Speaker A: The forecast app predicted a clear day, perfect for a walk without getting wet.",
      "start": "00:01.00",
      "end": "00:04.00",
      "notes": "Cheerful and reassuring, echoing reliance on technology."
    },
    {
      "content": "Speaker B: I don't know, those clouds look ominous.",
      "start": "00:05.00",
      "end": "00:07.00",
      "notes": "Skeptical, expressing doubt based on observation."
    },
    {
      "content": "Speaker A: It's fine, just enjoy the day.",
      "start": "00:08.50",
      "end": "00:10.00",
      "notes": "Dismissing caution, urging relaxation."
    }
  ],
  "event": [
    {
      "content": "Sudden rumble of thunder followed by pattering rain",
      "start": "00:10.50",
      "end": "00:13.50",
      "notes": "Supporting Speaker B: Thunderstorm sound contradicts Speaker A's trust in the app, confirming Speaker B's instinct was right."
    },
    {
      "content": "Distant robotic voice from a nearby smart speaker: 'Severe weather warning activated.'",
      "start": "00:13.50",
      "end": "00:14.50",
      "notes": "Ultimate contradiction: The technology Speaker A trusted now advises caution, aligning with Speaker B's earlier observation."
    }
  ]
}
```


Figure 5: Script Examples

Table 7: Question Examples for Audio with Script Example 1

Task Type	Question	Answer with Justification
(Optional) Hint	-	The recording happens in a Near-Future Conference Room.
Contextual Reasoning	Which statement best describes how the acoustic events relate to the speaker's assurance that 'everything's under control'? A: The calm hum of air filtration and ticking clock ensures that the speaker's authority remains unchallenged. B: The background sounds harmonize with the speaker's assurance, reinforcing the message of control and security. C: The high-pitched alarm directly refutes the speaker's claim by introducing a contradictory element of crisis. D: The urgency of the intercom announcement follows the speaker's statement logically, confirming all was under control.	Answer: C. Justification: Option C correctly identifies the alarm as an immediate contradiction to the speaker's words. The other options falsely attribute agreement or logical continuation between the speaker's claim and subsequent events.
Entity Extraction	Based on the speaker's statements and the surrounding audio events, what is the most plausible role of the speaker in this scene? A: A senior facilities engineer responsible for building infrastructure monitoring. B. A corporate security supervisor giving a status update before an automated override. C. A project manager attempting to reassure colleagues despite lacking real-time authority. D. An AI-assisted meeting coordinator relaying outdated system diagnostics.	Answer: B. Justification: The speaker explicitly references security system status ("hasn't triggered any alerts") and speaks with quiet authority, suggesting formal responsibility over security rather than casual observation.
Role Inference	What specific sound event undermines the speaker's earlier assertion of everything being under control? A: The ticking of the digital clock. B: The high-pitched security alarm. C: The quiet whirring of holographic projectors. D: The hum of the air filtration systems.	Answer: B. Justification: The high-pitched security alarm is the sound that disrupts the calm and contradicts the assertion of control. The clock, projectors, and air filtration are background sounds providing ambiance
Counterfactual Reasoning	If the 'sudden blare of a high-pitched security alarm' had instead been replaced by the 'gentle chime of a visitor notification,' how would the speaker's emotional response most likely change? A: The speaker would maintain calmness and possibly smile, perceiving it as normal protocol. B: The speaker would likely become more anxious and question the adequacy of the security system. C: The speaker would immediately run out of the room in a panic. D: The speaker would express relief that the security breach was a false alarm.	Answer: A. Justification: The gentle chime likely indicates a routine event rather than an emergency. The speaker, already confident, would maintain composure. Option B and D represent unwarranted dramatic reactions given the mild nature of the chime.

Table 8: Question Examples for Audio with Script Example 2

Task Type	Question	Answer with Justification
(Optional) Hint	-	The recording happens in a Futuristic Suburban Street.
Contextual Reasoning	How does the acoustic event at 00:10.50 influence the first speaker's earlier confidence in the weather forecast?, A: The event reinforces the first speaker's trust in the forecast accuracy., B: The event validates the first speaker's emphasis on relying on technology over instinct. C: The event is an unrelated background noise that supports the serene setting described. D: The event provides immediate auditory evidence that contradicts the first speaker's assurance of a clear day.	Answer: D. Justification: The sudden thunder and rain (00:10.50) occur just after the first speaker assures it's a clear day, directly contradicting their reliance on the weather app.
Entity Extraction	What specific auditory element confirms the second speaker's skepticism about the weather despite the forecast's prediction? A: The whispering trees in the background. B. The sudden rumble of thunder and wind. C. The sudden rumble of thunder and rain. D. The new announcement about weather change.	Answer: C. Justification: The 'sudden rumble of thunder and rain' directly confirms the second speaker's skepticism and contradicts the first speaker's reliance on the forecast app for a clear day. No wind or announcement recognized in the scene.
Role Inference	Based on the interaction between the speakers and the environmental cues, what is the most likely relationship or dynamic between them? A: Two friends discussing their plans while ignoring weather alerts. B: A parent encouraging their cautious child on a family outing. C: A meteorologist providing insights to a curious learner. D: Two strangers debating the reliability of weather technology.	Answer: B. Justification: The background is casual ('Ambient sound of whispering trees') with both speakers engaged in a light and friendly exchange. The tone is 'cheerful/reassuring' (the first speake) vs. 'skeptical' (the second speaker), typical of friendly banter. The second speaker misinterprets the equal dynamic as hierarchical. Option C implies a formal exchange which is not evident. Option D suggests unfamiliarity contradicted by the casual tone.
Counterfactual Reasoning	If the sound at 00:10.50 had been replaced by 'a gentle breeze and birdsong', how would the power dynamic between the speakers likely change? A: The first speaker's credibility would be strengthened, as their reliance on the forecast app would seem justified. B:The second speaker would insist further on canceling the walk, fearing the weather might still change suddenly. C: The first speaker would start doubting their own advice and question the reliability of the forecast app. D: The second speaker would immediately feel triumphant and declare their observational skills superior to technology.	Answer: A. Justification: With a gentle breeze and birdsong instead of a thunderstorm, the weather would align with Speaker A's initial prediction, thus reinforcing their credibility and trust in the forecast app.

Table 9: Prompts for generating the audio script

Script Category	Prompt
Single Person Monologue	You are an expert scriptwriter and acoustic designer for a large-scale AI research project focused on Context-Aware Auditory Understanding (CAAU). Your sole task is to generate original, single-speaker audio synthesis scripts in a precise JSON array format. The scene can cover any type of situations, such as daily-life, suspense, sci-fi, in-house conversations, classroom, public space, and more. I. Purpose and Core Constraints (Rules) 1. Purpose: The generated scripts must feature a clear interplay between the speaker's words (Speech in the Audio) and the non-speech context (Context in the Audio, composed of Event and Scene). This context must be semantically active, either providing crucial supplementary evidence or logically contradicting the speech content. 2. Speaker Constraint: Each script must feature only one speaker. 3. Contradiction/Supplement (Most Important): The context must be very important for the whole audio understanding and provide essential information other than the speech. The context must be integral to the meaning. Do not include random background noise that merely interferes with, but does not influence, the understanding. 4. Counterfactual Logic: The scripts should be easily adapted for Counterfactual Causal QA (CFQA), meaning the removal or alteration of one key event or the scene would fundamentally change the audio's narrative outcome. 5. Timing: Ensure realistic timing. The full audio must be concise (under 15 seconds) and event timings must closely align with the corresponding speech segments. 6. Uniqueness: Generate new, original scenarios. Do not reuse the provided examples. 7. Sometimes, the event may contain some speech. For example, if an event is a Sudden intercom announcement with content 'Code Blue, ICU, Room 7', In this case, you should add "with speech": 'Code Blue, ICU, Room 7' for the event, and put Sudden intercom announcement at the content domain (see the last example). 8. Timeline: Please make sure, the events last long enough, and there are enough time between event and speech, in order to be recognized (unless this event must happen short after the speech). For example, if the audio lasts 12 seconds and the speech only takes up from 4s to 7s. For the remaining seconds (1s to 4s and 7s to 12s), if you want to insert events, it is better to assign the event far from the speech, unless this event must happen short after the speech. 9. Enforce the per-request "style brief" constraints: - Genres: [GENRES] - Povs: [POVS] - Time Periods: [TIME PERIODS] II. Output Format Structure Your output MUST be a valid, single JSON array containing new, unique scenarios. Do not include any text before or after the JSON block. Use the following strict JSON schema: ...[JSON SCRIPT FORMAT] III. FEW SHOW EXAMPLES ...[EXAMPLES]
Inter-Person Conversation	Role: You are an expert Narrative Designer and Acoustic Engineer specializing in high-fidelity AI training data. Your goal is to create complex auditory scenarios where the relationship between speech and environment is critical for reasoning. Task: Generate a JSON array containing 2 to 4 unique audio scripts. Each script must feature a two-person dialogue where the Non-Speech Context (Acoustic Scene S + Events E) provides information that is essential to interpreting the truth of the conversation. I. Narrative Requirements Contextual Logic: The background sounds must fulfill one of two logical roles: Resolving a Disagreement/Doubt: Speaker A doubts a claim made by Speaker B, and a background event occurs that proves Speaker A's suspicion was correct. No Counterfactual Audio: Do not describe "what might happen." Describe the actual sounds heard in the audio. 1. Purpose: The generated scripts must feature a clear interplay between the speakers' words (Speech in the Audio) and the non-speech context (Context in the Audio, composed of Event and Scene). This context must be semantically active, either providing crucial supplementary evidence or logically contradicting the speech content. 2. Speaker Constraint: Two-Person Interaction: Every script must involve two distinct speakers (Speaker A and Speaker B). For the output, they should also be called 'Speaker A' and 'Speaker B'. 3. Contextual Logic: The background sounds must fulfill one of two logical roles: - 3.1 Contradicting a Pact/Agreement: Both speakers agree on a fact (e.g., "We are safe"), but the context (background sounds or the sound) prove them wrong (e.g., sound of a predator/intruder). - 3.2 Resolving a Disagreement/Doubt: Speaker A doubts a claim made by Speaker B, and a background event occurs that proves Speaker A's suspicion was correct. The context must be integral to the meaning. Do not include random background noise that merely interferes with, but does not influence, the understanding. 4. Counterfactual Logic: The scripts should be easily adapted for Counterfactual Causal QA (CFQA), meaning the removal or alteration of one key event or the scene would fundamentally change the audio's narrative outcome. However, Do not describe "what might happen." Describe the actual sounds heard in the audio. 5. Timing: Ensure realistic timing. The full audio must be concise (under 15 seconds) and event

Table 10: Prompts for Scene Description and Scene Understanding

Script Category	Prompt
Scene Description	You are an audio understanding model. You will be given one audio clip that may contain speech, environmental sounds, events, background ambience, or any combination of these. Your task is to carefully listen to the audio and write a short, coherent narrative description of what is happening in the scene. Guidelines - Describe only what can be recognized from the audio (do not guess unseen visuals). - Identify and integrate: - - Environment / setting (if inferable from sound) - - Background sounds or ambience - - Events or sound actions - - Speech content, including tone or emotion if audible - Combine all elements into a natural, story-like paragraph, not a list. - Keep the description concise but vivid (typically 4-8 sentences). - If speech is present, quote it verbatim when possible. - If multiple sounds overlap, describe how they interact over time. - Do not mention technical details (e.g., microphones, audio quality). - Do not add information that cannot be inferred from sound. Output Format Requirement: Plain text only; One coherent paragraph; No bullet points, timestamps, or labels Objective: Produce a human-readable script that captures what is happening in the audio scene, as if narrating it to someone who cannot hear it.
Scene Understanding	You are an audio understanding and reasoning model. You will be given: * One audio clip, which may contain speech, environmental sounds, events, or background ambience, alone or in combination. * A scene description telling you in what place or in what case the audio happens (optional). * A question about the audio. * A set of multiple-choice options (A/B/C/D). Your task is to: 1. Listen carefully to the audio and understand what is happening. 2. Use the audio (and the scene description, if provided) to reason about the situation. 3. Select the single best answer from the options. 4. Provide a brief justification based strictly on what can be inferred from the audio. — Guidelines * Base your answer only on audible evidence and the given scene description. If you can't find evidence in the audio, make a random guess on (A/B/C/D). * Your reasoning should consider: * Speech content and who is speaking * Events or actions implied by sounds * Tone and intent * Background sounds / environment * Do not rely on your own knowledge or assumptions not supported by the audio, as there might be imaginations or settings that are not factual in real life. What you should do is to infer the logic inside the audio. * If the audio is ambiguous, choose the option that is most strongly supported by the evidence. Do not say that you can't understand the audio. * Do not describe the full scene unless needed for justification. * Do not mention that you are an AI or that you cannot hear—assume perfect access to the audio. — Output Format (Strict) Your output must be in plain text and follow this exact structure: Answer: X. Justification: <a concise explanation of why this option is correct based on the audio> Where: * X is one of A, B, C, or D. * The answer line must appear **first** and exactly in the format: 'Answer: X'. Objective Demonstrate accurate **audio comprehension and reasoning** to answer the question correctly, and clearly explain the decision using evidence from the audio. Below is the Question: Question: [[QUESTION]] Scene: [[SCENE]] Options: A: [[optionA]] B: [[optionB]] C: [[optionC]] D: [[optionD]]