

STAR-BENCH: PROBING DEEP SPATIO-TEMPORAL REASONING AS AUDIO 4D INTELLIGENCE

Zihan Liu^{1,2*}, Zhikang Niu^{3,5*}, Qiuyang Xiao³, Zhisheng Zheng³, Ruqi Yuan¹, Yuhang Zang^{2✉}, Yuhang Cao², Xiaoyi Dong^{2,4}, Jianze Liang², Xie Chen^{3,5}, Leilei Sun¹, Dahua Lin^{2,4}, Jiaqi Wang^{2,5✉}

¹ Beihang University, ² Shanghai AI Laboratory, ³ Shanghai Jiao Tong University,

⁴The Chinese University of Hong Kong, ⁵ Shanghai Innovation Institute

liuzihan@buaa.edu.cn, zangyuhang@pjlab.org.cn

Code: <https://github.com/InternLM/StarBench>

Benchmark: <https://huggingface.co/datasets/internlm/STAR-Bench>

Homepage: <https://internlm.github.io/StarBench>

ABSTRACT

Despite rapid progress in Multi-modal Large Language Models and Large Audio-Language Models, existing audio benchmarks largely test semantics that can be recovered from text captions, masking deficits in fine-grained perceptual reasoning. We formalize audio **4D intelligence** that is defined as reasoning over sound dynamics in time and 3D space, and introduce **STAR-Bench** to measure it. STAR-Bench combines a Foundational Acoustic Perception setting (six attributes under absolute and relative regimes) with a Holistic Spatio-Temporal Reasoning setting that includes segment reordering for continuous and discrete processes and spatial tasks spanning static localization, multi-source relations, and dynamic trajectories. Our data curation pipeline uses two methods to ensure high-quality samples. For foundational tasks, we use procedurally synthesized and physics-simulated audio. For holistic data, we follow a four-stage process that includes human annotation and final selection based on human performance. Unlike prior benchmarks where caption-only answering reduces accuracy slightly, STAR-Bench induces far larger drops (-31.5% temporal, -35.2% spatial), evidencing its focus on linguistically hard-to-describe cues. Evaluating 19 models reveals substantial gaps compared with humans and a capability hierarchy: closed-source models are bottlenecked by fine-grained perception, while open-source models lag across perception, knowledge, and reasoning. Our STAR-Bench provides critical insights and a clear path forward for developing future models with a more robust understanding of the physical world.

1 INTRODUCTION

As a fundamental modality of human perception, audio serves a pivotal role in communication, aesthetic appreciation, and situational awareness, complementing the limitations of visual perception. With the rise of Multimodal Large Language Models (MLLMs) (Comanici et al., 2025; Achiam et al., 2023) and especially Large Audio-Language Models (LALMs) (Chu et al., 2024; Goel et al., 2025), these models have shown impressive capabilities in understanding audio, representing a crucial step toward diverse applications such as embodied intelligence (Paul et al., 2022).

To drive progress, a series of audio benchmarks has been introduced (Yang et al., 2024; Sakshi et al., 2025), covering traditional tasks like Automatic Speech Recognition (ASR) and sound event classification. While some recent efforts are beginning to emphasize reasoning abilities (Ma et al., 2025; Kumar et al., 2025), we observe that existing benchmarks predominantly focus on coarse-grained semantic content, which is audio information that can be distilled into textual descriptions with minimal loss. As shown in the **left** part of Fig. 1, we first use Gemini 2.5 Pro (Comanici et al., 2025) to generate detailed audio captions for samples in recent representative audio benchmarks MMAU (test-mini) (Sakshi et al., 2025) and MMAR (Ma et al., 2025). We then prompt the model to answer questions based *only* on these audio captions, and its performance drops by only 5.9% and 9.0%,

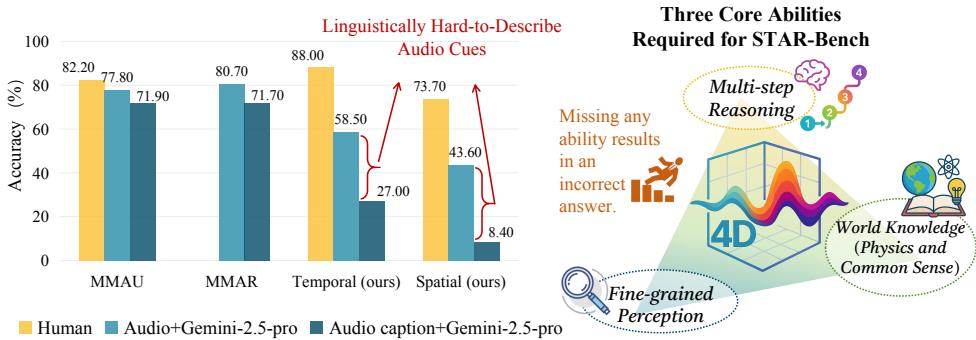


Figure 1: (**Left**): A comparison between humans and the Gemini 2.5 Pro with and without audio captions on various audio benchmarks. Our STAR-Bench evaluates linguistically hard-to-describe audio cues. See Sec. B.1 for audio caption details. (**Right**): The three core abilities required to solve tasks in the STAR-Bench benchmark.

respectively, compared to when it processes the raw audio. This result suggests that existing benchmarks primarily evaluate audio information that is **easily representable by text**. However, human auditory intelligence is not limited to this coarse-grained understanding. For example, humans can intuitively judge the water level in a container from the dynamic changes in the pouring sound, even without being able to precisely articulate the underlying acoustic features. Similarly, we can infer the trajectory and distance of a vehicle approaching from behind to ensure our safety. These abilities are rooted in deep reasoning of audio cues **that are difficult to represent linguistically**.

To capture this human-like audio competence, we propose a new paradigm, called **audio 4D intelligence**. This is defined as the ability to perform deep reasoning over the dynamics of **sound sources** in **time (1D)** and **three-dimensional space (3D)**, grounded in an understanding of the physical world. Mastering 4D audio intelligence is crucial for various applications. In embodied AI and robotics, for instance, agents must integrate fine-grained auditory cues to interact naturally with their surroundings, such as using sound to infer the trajectory of an object or to monitor the subtle operations of a machine. To systematically evaluate this paradigm and bridge the gap between current audio benchmarks and real-world auditory intelligence, we introduce the **Spatio-Temporal Audio Reasoning (STAR-BENCH)** benchmark.

STAR-BENCH is designed through a hierarchical task structure with two levels. At the **Foundational Acoustic Perception** level, we conduct a fine-grained, quantitative evaluation of six core audio attributes (pitch, loudness, duration, azimuth, elevation, distance) across both absolute perception ranges and relative discrimination sensitivity. We also introduce a **Holistic Spatio-Temporal Reasoning** level that evaluates an audio model’s ability to infer both event order and 3D scene structure. Temporal reasoning is tested via segment reordering that spans continuous processes and discrete event scripts, while spatial reasoning covers static localization, multi-source relations, and dynamic trajectory tracking. As shown in the **right** part of Fig. 1, every question in our holistic tasks is designed to probe a synthesis of three core pillars, such as multi-step reasoning. A failure in any one of these pillars will lead to an incorrect response. Our **data curation pipeline** couples procedurally synthesized, fully parameterized audio for foundational perception with large-scale real-world corpora for holistic reasoning. For the latter, we use a four-stage process including **human annotation** and **final selection by human performance** to ensure the high quality of our benchmark samples.

Our comprehensive evaluation of 19 models (16 open-source and 3 closed-source) reveals a clear capability hierarchy between the two groups. Leading closed-source models like Gemini 2.5 Pro excel in knowledge and reasoning, shifting their primary bottleneck to the more difficult challenge of fine-grained perception. In contrast, open-source models exhibit fundamental weaknesses across all three core capabilities. Through our detailed error analysis and ablation studies, we highlight several key insights for the future development of open-source audio models: 1) **Enhancing dense audio captioning**. Open-source models struggle to produce dense, fine-grained captions, which limits their perceptual sensitivity and ability to extract embedded knowledge. Bridging this gap is a crucial first step. 2) **Improving multi-audio reasoning**. Open-source models lag significantly in comparing, integrating, and grounding information across multiple audio clips. 3) **Moving beyond channel-averaged audio preprocessing**. The common practice of averaging multi-channel audio

Table 1: A comparative overview of our benchmark against other representative audio benchmarks. (✓: Fully supported, ○: Partially supported, ✗: Not supported)

Benchmark	Temporal Deep Reasoning	Spatial Deep Reasoning	Quantitative Attribute Evaluation	Robust Evaluation	Multi-Audio	Fully Human-Annotated	Fully Expert Verified
AIR-Bench [45]	✗	✗	✗	✗	✗	✗	✗
MMAU [30]	✗	✗	✗	✗	✗	✓	✓
Dynamic-SUPERB Phase-2 [16]	✗	✗	✗	✗	○	○	✗
MMAR [27]	✗	✗	✗	✗	✗	✓	✓
MMAU-Pro [20]	✗	✗	✗	✗	✓	✓	✓
STAR-BENCH (ours)	✓	✓	✓	✓	✓	✓	✓

into a mono signal is a major bottleneck for spatial reasoning. Developing architectures that natively process multi-channel cues is essential for unlocking genuine spatial awareness.

Our contributions are summarized as: **(1)** We formalize **audio 4D intelligence**, and empirically show that prior benchmarks largely probe text-representable semantics, motivating a shift toward fine-grained, non-linguistic auditory cues. **(2)** We introduce the STAR-BENCH with foundational acoustic perception and holistic spatio-temporal reasoning tasks, together with a rigorous curation pipeline with expert validation. **(3)** We provide a comprehensive evaluation of 19 LALMs/OLMs. Our analyses and standardized protocols establish strong baselines and testbeds for future research.

2 RELATED WORK

The recent progress of Large Audio-Language Models (LALMs)(Kong et al., 2024; Chu et al., 2024; Wu et al., 2025; Xiaomi, 2025) and Omni-Language Models (OLMs)(Xu et al., 2025; Yao et al., 2024; AI et al., 2025) has significantly advanced audio understanding. At the same time, it has spurred the development of numerous benchmarks to comprehensively evaluate their capabilities. Earlier benchmarks(Wang et al., 2024; Yang et al., 2024) mainly focused on semantic-level understanding tasks (transcription, captioning, and simple question answering), and recent benchmarks(Sakshi et al., 2025; Ma et al., 2025; Kumar et al., 2025) have begun to investigate logical audio reasoning tasks. However, existing benchmarks do not address 4D audio intelligence or deep spatio-temporal reasoning across multiple audio inputs, and instead remain limited to single-clip understanding and reasoning. To fill these gaps, we propose a benchmark designed for multi-audio and deep spatio-temporal reasoning, enabling more comprehensive evaluation of audio 4D intelligence. See Table 1 for a comparison with existing benchmarks, and Sec. A for further related works.

3 STAR-BENCH

Understanding dynamic sound sources in both time (1D) and three-dimensional space (3D) is a crucial skill for MLLMs to comprehend the physical world. To address this need, our benchmark, STAR-BENCH, is designed to comprehensively evaluate this 4D intelligence in the audio domain. As illustrated in Fig. 2, our evaluation has two complementary sub-tasks: (1) Foundational Acoustic Perception (Sec. 3.1), which uses procedurally synthesized audio to quantitatively profile a model’s basic perceptual abilities under controlled conditions, and (2) Holistic Spatio-Temporal Reasoning (Sec. 3.2), which uses real-world audio to evaluate more complex reasoning in dynamic and authentic scenarios. We also elaborate our data curation pipeline in the Sec. 3.3.

3.1 FOUNDATIONAL ACOUSTIC PERCEPTION

The Foundational Acoustic Perception task is motivated by the need for a robust, quantitative evaluation of the core perceptual abilities that underpin 4D audio intelligence. A model’s capacity for complex reasoning about dynamic audio scenes in the physical world is directly dependent on its ability to accurately perceive fundamental acoustic properties. Our foundational acoustic perception task systematically probes a model’s understanding of three critical auditory attributes: **Loudness**, **Pitch**, **Duration**, and the three spatial dimensions: **Azimuth**, **Elevation**, and **Distance**. Just as a solid understanding of grammar is required for writing a complex narrative, a model must be able to accurately perceive these core attributes before it can reason about the dynamic, spatial relationships of sound sources in the physical world. Without a firm grasp of these foundational elements, a model

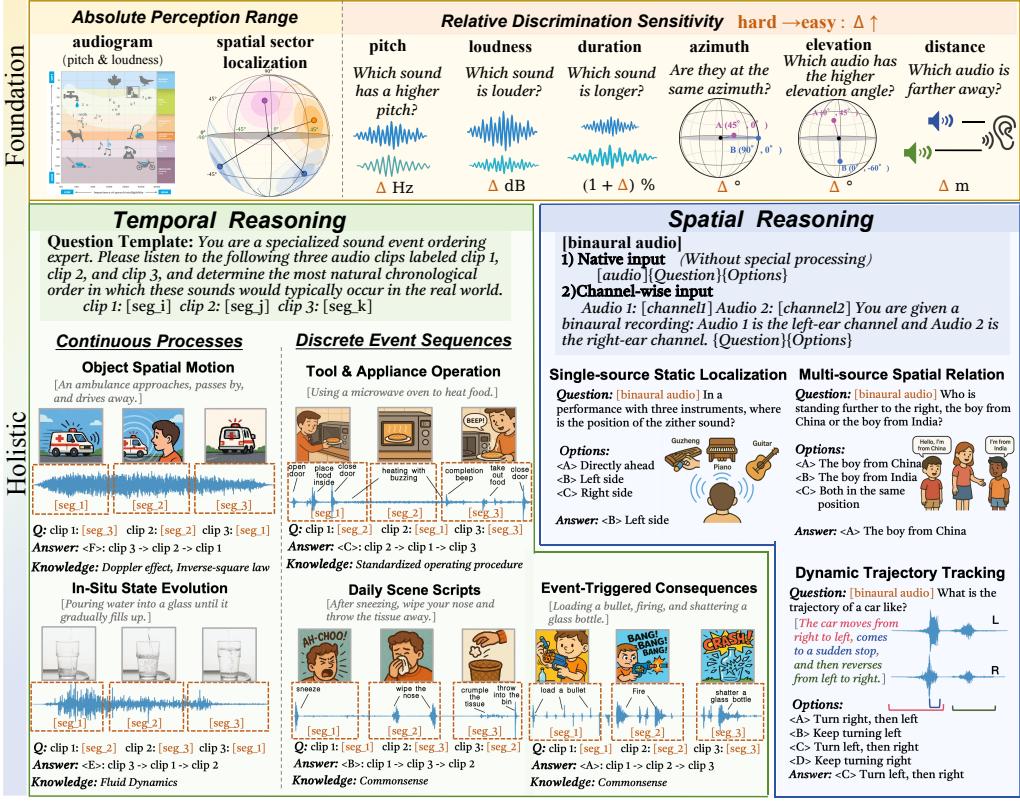


Figure 2: **Data examples from STAR-BENCH:** (1) the foundational perception task (upper) and (2) the holistic spatio-temporal reasoning task, which includes both temporal reasoning (bottom left) and spatial reasoning (bottom right). Zoom in for the best view.

cannot accurately interpret complex, real-world acoustic scenes, which require understanding how sounds change over time and move through space.

We employ a targeted synthesis strategy to generate precise evaluation samples in a controlled environment for the foundational perception task. For non-spatial attributes (Loudest, Pitch, Duration), we synthesize pure sine waves by directly specifying their parameters. For spatial attributes (Azimuth, Elevation, Distance), we use the Pyroomacoustics (Scheibler et al., 2018) physics-based simulation engine to render acoustic scenes. The targeted synthesis strategy allows us to investigate a model’s audio perceptual abilities under the following two sub-tasks:

1) Absolute Perception Range, which defines the sensory limits of MLLMs for acoustic attributes. For pitch and loudness, we adapt the design of human audiology tests to create an “audiogram” for the MLLMs. Specifically, we synthesize sine waves with frequencies ranging from 125 Hz to 8000 Hz and loudness levels from -10 to 110 dB HL and require the model to identify if a clear beep is in the first or second part of an audio clip, or if it’s not there at all. For spatial attributes, we design interval localization tasks that require the model to identify a sound’s azimuth within one of four 90° quadrants (from 0° to 360°), its elevation relative to ear-level (above, at, or below, from -90° to 90°), and its distance category (near, medium, or far, within a 0 - 10m range). Tab. 3 presents detailed examples of these absolute perception range tasks. Through these precise tasks, we establish the absolute limits of what the model can hear, which is crucial for developing AI systems that can safely and effectively interact with the physical world.

2) Relative Discrimination Sensitivity, which investigates how well a model can detect small changes in acoustic attributes. The ability to detect small changes allows a model to make nuanced judgments, like determining if a sound is getting louder or a pitch is rising. Analogous to measuring the human Just Noticeable Difference (JND), the relative discrimination task presents the model with an audio clip containing two sounds and requires it to compare them based on a specific attribute. We meticulously designed four to six distinct difficulty levels for each of the six attributes, as detailed in Tab. 3. Level 1 serves as a control group to test for random guessing, presenting identical sounds ($\Delta=0$) for non-spatial attributes and a sub-threshold difference for spatial ones. Subsequent

levels then introduce progressively larger differences, ranging from subtle variations perceptible to humans to more significant, real-world changes. By analyzing the model’s performance across these different levels of stimulus differences, we can quantitatively assess its discrimination sensitivity for each attribute.

3.2 HOLISTIC SPATIO-TEMPORAL REASONING

Building on the model’s fundamental audio perceptual abilities (Sec. 3.1), we further introduce holistic temporal reasoning (Sec. 3.2.1) and spatial reasoning (Sec. 3.2.2), which are designed to systematically evaluate a model’s reasoning ability that is required for audio 4D intelligence.

3.2.1 TEMPORAL REASONING TASKS

The core of temporal reasoning lies in understanding the intrinsic logic of event sequences, encompassing physical causality, functional procedures, or social conventions. To evaluate this capability, we design a novel **Audio Segment Reordering** setting. Specifically, we curate a collection of audio events characterized by strong sequential uniqueness, semantic clarity, and logical universality. Each event is segmented into three clips, which are then shuffled as inputs to the model. The models are required to restore the original temporal sequence based solely on the audio content. Our temporal reasoning tasks are organized into two meta-categories (continuous processes, discrete event sequences) and five subcategories based on their core logical principles.

The **continuous processes** assess a model’s ability to track the subtle, continuous evolution of acoustic features within a single, uninterrupted acoustic event. The **object spatial motion** subcategory reconstructs the spatio-temporal trajectory of moving sources (e.g., passing cars, airplanes) by interpreting key acoustic cues, such as the Doppler effect (frequency shifts indicating relative velocity) and the inverse-square law (loudness changes indicating distance). Besides, the **in-situ state evolution** subcategory assesses a model’s ability to track the intrinsic evolution of a stationary object’s state, a process governed by predictable trend patterns. These trend patterns arise from various underlying principles, including: *Fluid & Pneumatic Dynamics*, where the sound is governed by principles of turbulence, resonance, and pressure changes (e.g., a toilet flushing, water being poured); *Thermodynamic Processes*, involving irreversible state changes driven by heat (e.g., water boiling, food frying); *Energy Decay*, a process governed by resonant decay and frictional damping after a single excitation (e.g., a bell’s chime, an explosion’s echo); and complex *Biological Rhythms* that reflect an evolving physiological or emotional state.

The **discrete event sequences** category requires the model to understand the logical and temporal relationships between multiple, distinct acoustic events, which are governed by function, convention, or causality. The **tool & appliance operation** sub-category follows the standardized operating procedure for tools and appliances (e.g., a microwave, a power drill), where the sequence is correct when it follows the tool’s designed function. The **daily scene scripts** sub-category applies commonsense and contextual script knowledge to follow the conventional sequence of actions in a daily activity (e.g., brushing teeth, drinking water). The **event-triggered consequences** sub-category applies causal reasoning to infer that a trigger event (e.g., a firework explosion) will be followed by an automatic and irreversible outcome, whether physical (glass shattering) or social (a crowd cheering).

3.2.2 SPATIAL REASONING TASKS

Humans effortlessly perceive complex 3D auditory scenes (e.g., hearing a voice from behind, following an approaching car, or locating multiple speakers). Such an ability is fundamental for egocentric interaction and embodied AI systems, for instance, robots that navigate and interact with their surroundings. However, existing benchmarks focus primarily on the localization of static sound sources, whereas real-world scenarios demand reasoning that integrates both spatial and temporal cues. To address this gap, we organize the spatial reasoning task into three subcategories.

The **single-source static localization** evaluates the model’s ability to identify the direction of a target sound source among multiple static sources (e.g., judging whether a sound comes from the left or right). It assesses the basic spatial perception capability of the model and provides the foundation for more advanced reasoning. The **multi-source spatial relation** requires the model to determine the relative spatial relationships among multiple simultaneous sound sources (e.g., comparing the

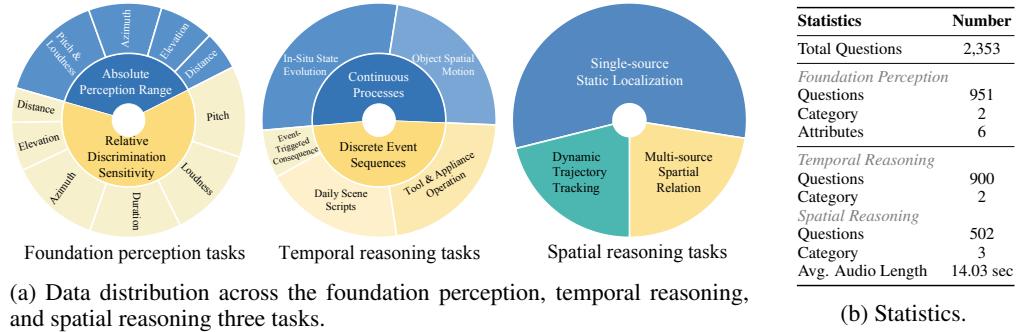


Figure 4: (a) The **data distribution** of STAR-BENCH across three main tasks. (b) **Data statistics** of our benchmark, including the total number of questions for each task and their sub-categories, and the average audio length for reasoning tasks.

placement of two speakers to decide which one is further to the right). Beyond localizing each source individually, the model must infer their spatial placement and choose the appropriate relational description from multiple candidates. The **dynamic trajectory tracking** introduces moving sound sources, which require the model to go beyond basic spatial perception to dynamically model spatio-temporal relations for reasoning about complex movement trajectories (e.g., tracking a passing car moving from left to right). This task extends spatial reasoning into the temporal domain and is more faithful to the complexity of real-world acoustic scenarios.

However, evaluating existing LALMs on multi-channel spatial tasks is challenging. The common practice of these models is to average multi-channel audio into a mono signal, resulting in the loss of substantial spatial information. We conduct a simple experiment as shown in Fig. 3. We construct 20 pseudo-stereo signals by assigning the original audio to the left channel and its additive inverse to the right. While human listeners could easily perform sound event classification on these signals, the models consistently failed due to signal cancellation during the mono conversion. The result confirms their lack of explicit support for genuine stereo audio processing. To provide a comprehensive assessment, we design two complementary strategies: **native input**, where the model directly processes stereo audio, follow their default processing pipeline to probe its intrinsic ability to exploit spatial cues; and the **channel-wise input**, where each channel is presented separately with explicit textual instructions, as shown in the bottom right of Fig. 2, and allow the model to approximate human-like use of interaural cues.

3.3 DATA CURATION PIPELINE

Our data curation pipeline integrates procedural synthesis with real-world data collection to ensure both comprehensive coverage and ecological validity. Fig. 4 shows the distribution and statistics of our STAR-BENCH. All audio for the *foundational perception* task is synthesized using precise parameterization or the Pyroomacoustics (Scheibler et al., 2018) physics-based simulator, providing complete control over acoustic parameters. Domain experts rigorously validate the task difficulty levels, which are then calibrated through human testing. For the *holistic spatio-temporal reasoning* task, the curation process comprises four key stages (see Fig. 5):

- 1) **Taxonomy Construction and Data Sourcing:** We build a hierarchical task taxonomy through a collaborative process involving domain experts and the Gemini 2.5 Pro (Comanici et al., 2025). This framework guides the sourcing of candidate data from large-scale, real-world audio libraries: Clotho (Drossos et al., 2019) and FSD50K (Fonseca et al., 2022) for temporal reasoning, and STARSS23 (Shimada et al., 2023), along with audio sourced from the internet for spatial reasoning.

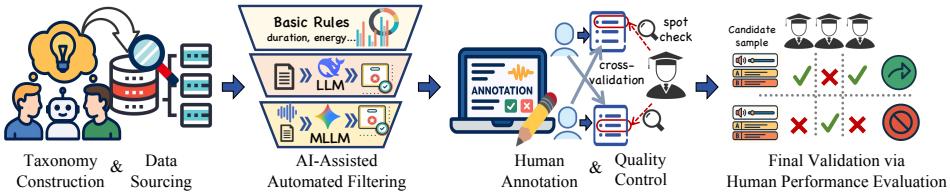


Figure 5: The four-stage **data annotation pipeline** for constructing our STAR-BENCH.

2) AI-Assisted Automated Filtering: This process employs an efficient three-stage funnel. First, we discard unsuitable samples based on basic properties like duration and energy. Next, an LLM (e.g., DeepSeek-V3 (Liu et al., 2024a)) performs an initial screening based on textual metadata, providing justifications for its decisions. Finally, a powerful multimodal model (e.g., Gemini 2.5 Pro (Comanici et al., 2025)) analyzes the audio, metadata, and the LLM’s outputs. The final step yields a judgment, a quality score, and a preliminary classification, further filtering irrelevant samples. The detailed prompts used to query the LLMs are provided in Sec. B.3.

3) Human Annotation and Quality Control: We recruit and train 10 undergraduate annotators to label the data using a professional platform. During this process, AI-generated information is provided as an auxiliary reference. To ensure high-quality labels, we implement a stringent two-round review process: the first round involves inter-annotator cross-validation until a consensus is reached, while the second consists of random spot-checks by three domain experts.

4) Final Validation via Human Performance Evaluation: To ensure all items in the benchmark are fair, unambiguous, and solvable by humans, we implement a final validation stage. In this phase, domain experts act as examinees and solve our tasks. Only items that are independently and correctly solved by at least two-thirds of the experts are retained. Our rigorous protocol ensures that all problems in our benchmark are well-posed and reliably solvable by human experts.

4 EVALUATION

Benchmarking Models. Our evaluation covers 19 models (16 open-source and 3 closed-source models). The open-source models span three categories: (1) Large Audio Language Models designed for universal audio-text understanding, including SALMONN (Tang et al., 2024), Qwen2-Audio Instruct (Chu et al., 2024), Audio Flamingo 3 (Goel et al., 2025) with its ‘think’ variant, DeSTA2.5-Audio (Lu et al., 2025), Kimi-Audio (KimiTeam et al., 2025), Step-Audio-2-mini (Wu et al., 2025), MidashengLM (Dinkel et al., 2025), and Xiaomi-MiMo-Audio (Xiaomi, 2025) with its ‘think’ variant; (2) a specialized model for spatial audio, BAT (Zheng et al., 2024); and (3) Omni Language Models with fully multimodal support, including Qwen-2.5-Omni (Xu et al., 2025), Phi4-MM (Abouelenin et al., 2025), Gemma-3n-E4B-it (Team et al., 2025), and Ming-Lite-Omni-1.5 (AI et al., 2025). We also include three leading closed-source models: Gemini 2.5 Pro (Comanici et al., 2025) (updated June 2025), Gemini 2.5 Flash (updated June 2025), and GPT-4o-audio-preview (Achiam et al., 2023) (version 2025-06-03).

Robust Evaluation. All questions in STAR-BENCH are presented as multiple-choice questions and evaluated using classification accuracy, with correctness determined via string matching of option labels or their full text. To ensure robustness, we evaluate each question multiple times under minor prompt perturbations, a strategy detailed in Sec. C. This approach yields two key metrics: **Average Accuracy (AA)**, the mean accuracy across all runs, and **All-Correct Rate (ACR)**, the proportion of questions answered correctly in every run, which serves as a stronger indicator of model reliability. Due to space limitations, we primarily report AA in the main text, while complete experimental results are available in Sec. D.

4.1 MAIN RESULT ANALYSIS

We present a comprehensive evaluation on STAR-BENCH, as shown in Tab. 2. Due to the space limit, detailed results on each task are provided in Sec. D. Our key findings are as follows:

STAR-BENCH is Challenging STAR-BENCH presents a considerable challenge for existing models. Human evaluators achieve high accuracy across all task categories (e.g., 75.6% on perception,

Table 2: Evaluation results of various models on STAR-BENCH. The best performance is highlighted in **bold**, and the second-best ones are underlined. MA (Macro Accuracy) denotes the unweighted mean of class-wise accuracies, while OA (Overall Accuracy) denotes the proportion of correctly answered instances. All reported values are AA (Average Accuracy across multiple runs) only; for ACR (All-Correct Rate), see Sec. D.

Models	Size	Foundational Perception			Temporal Reasoning			Spatial Reasoning			MA (%)
		Range	Sensitivity	MA	Continuous	Discrete	OA	Localization	Relation	Trajectory	
Random Guess	-	23.75	26.38	25.33	14.29	14.29	14.29	33.33	33.33	33.33	24.32
Human	-	79.42	74.55	75.60	90.12	85.51	88.00	70.00	80.00	77.00	73.72
SALMONN [33]	13B	27.32	25.48	26.22	14.88	13.30	14.15	26.15	28.61	39.94	29.62
Audio Flamingo 3 [14]	8.4B	31.79	35.72	34.15	9.23	8.01	8.67	37.22	38.35	44.03	38.91
Audio Flamingo 3 think [14]	8.4B	25.54	34.08	30.66	13.22	14.02	13.59	35.45	37.46	38.05	36.45
Qwen2-Audio-Instruct [7]	8.4B	29.88	26.47	27.84	13.29	12.10	12.74	21.32	24.78	15.09	20.78
DeStA2.5-Audio [26]	8.8B	29.87	19.79	23.82	16.53	17.39	16.93	23.67	34.81	37.74	29.15
BAT [51]	7B	22.81	6.25	12.87	0.00	0.00	0.00	0.00	0.00	0.00	4.29
Phi4-MM [1]	5.5B	19.14	29.85	25.56	16.74	16.99	16.85	33.10	27.14	34.28	32.01
Kimi-Audio [18]	7B	23.29	27.50	25.82	19.97	16.83	18.52	27.56	38.94	44.03	33.60
DiDashengLM [10]	7B	<u>36.94</u>	30.78	33.24	15.43	17.31	16.30	43.11	45.43	46.23	44.29
Step-Audio-2-mini [39]	7B	29.65	27.14	28.14	15.36	15.87	15.59	33.33	31.27	37.74	33.80
Gemma-3n-E4B-it [34]	7.5B	18.55	25.02	22.43	16.87	16.27	16.59	23.32	41.89	33.96	29.75
Ming-Lite-Omni-1.5 [3]	18.9B	26.76	26.76	26.76	17.08	15.54	16.37	20.14	35.10	38.36	27.35
Qwen-2.5-Omni [43]	7B	28.76	32.32	30.90	16.32	17.71	16.96	39.46	41.30	27.04	37.25
Xiaomi-MiMo-Audio [40]	7B	34.95	31.59	32.93	18.18	19.15	18.63	36.16	41.30	45.28	39.24
Xiaomi-MiMo-Audio-think [40]	7B	29.90	24.93	26.92	16.80	19.39	18.00	34.28	44.54	36.79	37.12
MinICPM-O-V2.6 [48]	8B	<u>31.02</u>	31.87	31.53	15.36	17.39	16.30	<u>29.92</u>	<u>43.36</u>	<u>38.36</u>	<u>34.73</u>
GPT-4o Audio [2]	-	27.58	34.55	31.76	15.91	23.56	19.44	41.81	43.97	39.94	41.70
Gemini 2.5 Flash [8]	-	33.46	<u>43.88</u>	<u>39.72</u>	27.55	34.38	30.70	24.62	43.07	22.64	28.35
Gemini 2.5 Pro [8]	-	39.90	51.13	46.64	54.88	62.74	58.52	40.87	48.97	45.28	43.62
<hr/>											

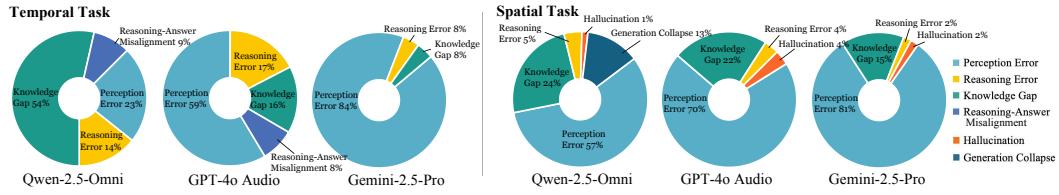


Figure 6: Error distribution across temporal and spatial Tasks.

88.0% on temporal, and 73.7% on spatial tasks), whereas all tested models fall well below this baseline. Most open-source models perform close to random guessing, and even the best closed-source model, Gemini 2.5 Pro, reaches only 49.59% average accuracy. In addition, model predictions on STAR-BENCH exhibit low reliability, as evidenced by the pronounced gap between their Average Accuracy (AA) and All-Correct-Rate (ACR) scores. A detailed discussion of this issue is provided in Sec. E.1. Although the underlying audio data for the temporal tasks (e.g., FSD50K, Clotho) is commonly used for model pre-training, our novel task formulation of temporal reasoning deliberately departs from conventional audio QA formats. This design allows for a more thorough evaluation of the integrated capabilities of current models.

A Clear Performance Gap between Closed-Source and Open-Source Models On the foundational perception and temporal tasks, Gemini 2.5 Pro establishes a commanding lead among all models. On spatial tasks, however, nearly all models, both closed- and open-source, perform poorly. As indicated by the prior experiment (Fig. 3), this is likely because most models (except BAT) discard multi-channel information during preprocessing, thereby losing key acoustic cues needed for spatial reasoning. Among closed-source models, Gemini 2.5 Pro surpasses Gemini 2.5 Flash, suggesting that stronger reasoning capabilities deliver substantial gains. In contrast, open-source models show the opposite pattern: the “think” modes of Audio Flamingo 3 and Xiaomi-MiMo-Audio perform worse than their no-thinking counterparts, implying that without sufficiently solid perceptual and knowledge foundations, reasoning can be ineffective or even detrimental.

4.2 DISCUSSION: WHY DO EXISTING MODELS STRUGGLE ON STAR-BENCH?

To better understand the underlying causes of the poor performance of existing models, we conduct a detailed error analysis along with a series of ablation studies. Due to space limitation, the ablation study on spatial reasoning is provided in Sec. E.2.

Error Analysis. We conduct a manual error analysis on 200 failed predictions sampled equally from temporal and spatial tasks of three representative models (Gemini 2.5 Pro, GPT-4o-audio, and Qwen-2.5-Omni). For temporal tasks, our analysis reveals a clear capability hierarchy across the models. The open-source Qwen-2.5-Omni shows major deficiencies in all three core abilities: its perception is coarse-grained and unable to capture subtle inter-segment distinctions, and a sub-

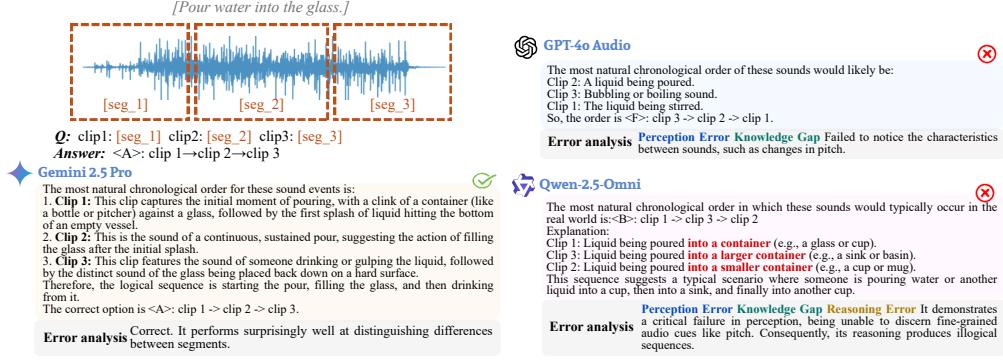


Figure 7: An error case in temporal reasoning task. More cases are provided in the Sec. F.

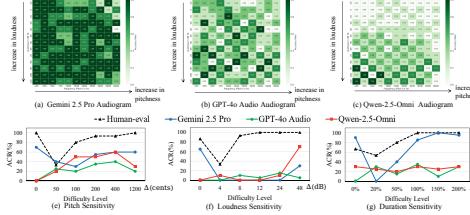


Figure 8: The sensitivity analysis in fine-grained perception.

stantial knowledge gap (54%) leads to reasoning that often appears specious due to the absence of physical-world grounding. GPT-4o-audio demonstrates stronger knowledge, but still suffers from perceptual and reasoning limitations, along with low-level issues such as misalignment between reasoning and final answers. In contrast, Gemini 2.5 Pro excels in knowledge and reasoning, shifting its primary bottleneck to the more advanced challenge of fine-grained perception (84%). As shown in Fig. 7, Gemini 2.5 Pro is the only model to succeed by providing a remarkably detailed description of acoustic nuances. Our finding suggests that the **advanced world knowledge is deeply embedded within detailed audio-text captioning**. While open-source models largely remain at a coarse semantic level (e.g., sound event classification), our analysis highlights that enabling them to generate fine-grained acoustic descriptions is critical toward more robust reasoning. On the other hand, most models demonstrate a lack of native spatial awareness in audio tasks, with weaknesses in perception, knowledge, and reasoning. Additionally, a prevalent type of error involves vision-centric hallucinations (e.g., “...based on the car’s trajectory in the video...”). This may be attributable to the models’ training on visual spatial tasks, leading them to misapply visual reasoning to auditory inputs.

Lack of Human-like Sensitivity in Fine-Grained Perception. To quantify the gap in perceptual sensitivity, we present model audiograms in Fig. 8 (a)(b)(c). Fig. 8 (e)(f)(g) further track the performance of both models and human subjects on the three core acoustic attributes (pitch, loudness, and duration) as task difficulty decreases. The results reveal a stark performance gap between all models and the human baseline, particularly in the perception of fine-grained loudness differences. A clear trend is visible even for the top-performing Gemini 2.5 Pro: its accuracy, while competent on easier tasks, plummets as perceptual granularity increases. This directly corroborates our error analysis, identifying fine-grained perception as its primary bottleneck. Notably, its performance on duration perception is an exception, showcasing **temporal grounding capabilities superior to those of other models** by accurately assessing audio segment lengths.

Ablation Study on Temporal Reasoning. To further pinpoint the specific limitations of temporal reasoning, we augment the baseline audio segment reordering task with two progressively easier settings: (1) + *Global Caption*, where a single sentence describing the overall scene is provided as a contextual guide; and (2) + *Uncut Audio*, where the complete, unsegmented audio track is offered as a reference, reducing the task to a straightforward process where the correct order can be determined simply by comparing and grounding each segment within the full audio. As shown in Fig. 9, Gemini 2.5 Pro’s performance scales effectively with task simplification, culminating in a near-perfect 99% accuracy in the + *Uncut Audio* setting. In contrast, the open-source models show minimal to no improvement across these settings. Their performance remains stagnant even when provided with

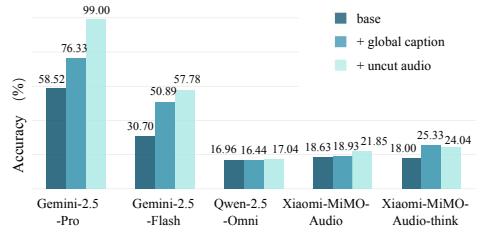


Figure 9: The ablation study on temporal reasoning.

the complete audio reference, despite the simplified nature of the task. This finding starkly exposes a core weakness in current open-source models: **a fundamental inability to effectively compare, ground, and integrate information from multiple audio inputs.**

5 CONCLUSION

We introduce STAR-BENCH, a comprehensive benchmark for evaluating 4D audio intelligence over time and 3D space. We use rigorous human annotation, consensus review, and expert validation to ensure the high quality of data samples. STAR-BENCH establishes standardized tasks and protocols for studying 4D audio intelligence, offering actionable diagnostics for model developers. We expect STAR-Bench to accelerate progress on advanced audio models and training with spatialized corpora, capabilities that are crucial for embodied agents.

REFERENCES

- Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahu Lin. Internlm2 technical report, 2024.
- Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*, 2025.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

- Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following, 2025. URL <https://arxiv.org/abs/2504.07957>.
- Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. Midashenglml: Efficient audio understanding with general audio captions. *arXiv preprint arXiv:2508.03983*, 2025.
- Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset, 2019. URL <https://arxiv.org/abs/1910.09387>.
- Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022.
- Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Forty-second International Conference on Machine Learning*, 2025.
- 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. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025.
- Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2023. doi: 10.1109/ASRU57964.2023.10389742.
- Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, et al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.
- KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report, 2025. URL <https://arxiv.org/abs/2504.18425>.
- Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *International Conference on Machine Learning*, pp. 25125–25148. PMLR, 2024.
- Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plicka, Miroslav Hlaváček, et al. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*, 2025.
- Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jiaqi Wang, and Dahua Lin. Beyond fixed-Training-free variable-length denoising for diffusion large language models, 2025. URL <https://arxiv.org/abs/2508.00819>.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024b. URL <https://arxiv.org/abs/2307.06281>.
- Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuan-jun Xiong, Dahua Lin, and Jiaqi Wang. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models, 2024c. URL <https://arxiv.org/abs/2410.17637>.
- Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual agentic reinforcement fine-tuning, 2025. URL <https://arxiv.org/abs/2505.14246>.
- Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, et al. Desta2. 5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *arXiv preprint arXiv:2507.02768*, 2025.
- Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*, 2025.
- Sudipta Paul, Amit Roy-Chowdhury, and Anoop Cherian. AVLEN: Audio-visual-language embodied navigation in 3d environments. In *NeurIPS*, 2022.
- Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. Vln-r1: Vision-language navigation via reinforcement fine-tuning, 2025. URL <https://arxiv.org/abs/2506.17221>.
- S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaseswaran Selvakumar, Oriol Nieto, Ramaṇi Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 351–355. IEEE, 2018.
- Kazuki Shimada, Archontis Politis, Parthasarathy Sudarsanam, Daniel A Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, et al. Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *Advances in neural information processing systems*, 36:72931–72957, 2023.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*, 2024.
- Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Sim-cot: Supervised implicit chain-of-thought, 2025a. URL <https://arxiv.org/abs/2509.20317>.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Videorope: What makes for good video rotary position embedding?, 2025b. URL <https://arxiv.org/abs/2502.05173>.

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, Yuxin Zhang, Zhao You, Brian Li, Changyi Wan, Hanpeng Hu, Jiangjie Zhen, Siyu Chen, Song Yuan, Xuelin Zhang, Yimin Jiang, Yu Zhou, Yuxiang Yang, Bingxin Li, Buyun Ma, Changhe Song, Dongqing Pang, Guoqiang Hu, Haiyang Sun, Kang An, Na Wang, Shuli Gao, Wei Ji, Wen Li, Wen Sun, Xuan Wen, Yong Ren, Yuankai Ma, Yufan Lu, Bin Wang, Bo Li, Changxin Miao, Che Liu, Chen Xu, Dapeng Shi, Dingyuan Hu, Donghang Wu, Enle Liu, Guanzhe Huang, Gulin Yan, Han Zhang, Hao Nie, Haonan Jia, Hongyu Zhou, Jianjian Sun, Jiaoren Wu, Jie Wu, Jie Yang, Jin Yang, Junzhe Lin, Kaixiang Li, Lei Yang, Liying Shi, Li Zhou, Longlong Gu, Ming Li, Mingliang Li, Mingxiao Li, Nan Wu, Qi Han, Qinyuan Tan, Shaoliang Pang, Shengjie Fan, Siqi Liu, Tiancheng Cao, Wanying Lu, Wenqing He, Wuxun Xie, Xu Zhao, Xueqi Li, Yanbo Yu, Yang Yang, Yi Liu, Yifan Lu, Yilei Wang, Yuanhao Ding, Yuanwei Liang, Yuanwei Lu, Yuchu Luo, Yuhe Yin, Yumeng Zhan, Yuxiang Zhang, Zidong Yang, Zixin Zhang, Binxing Jiao, Dixin Jiang, Heung-Yeung Shum, Jiansheng Chen, Jing Li, Xiangyu Zhang, and Yibo Zhu. Step-audio 2 technical report, 2025. URL <https://arxiv.org/abs/2507.16632>.

LLM-Core-Team Xiaomi. Mimo-audio: Audio language models are few-shot learners, 2025. URL <https://github.com/XiaomiMiMo/MiMo-Audio>.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction, 2025a. URL <https://arxiv.org/abs/2410.17247>.

Long Xing, Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jinsong Li, Shuangrui Ding, Weiming Zhang, Nenghai Yu, Jiaqi Wang, Feng Wu, and Dahua Lin. Scalecap: Inference-time scalable image captioning via dual-modality debiasing, 2025b. URL <https://arxiv.org/abs/2506.19848>.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

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

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025b.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025c.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Beichen Zhang, Yuhong Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Haodong Duan, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Booststep: Boosting mathematical capability of large language models via improved single-step reasoning, 2025a. URL <https://arxiv.org/abs/2501.03226>.

Zhixiong Zhang, Shuangrui Ding, Xiaoyi Dong, Songxin He, Jianfan Lin, Junsong Tang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Sec: Advancing complex video object segmentation via progressive concept construction, 2025b. URL <https://arxiv.org/abs/2507.15852>.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. Bat: Learning to reason about spatial sounds with large language models. In *International Conference on Machine Learning*, pp. 61454–61469. PMLR, 2024.

THE USE OF LARGE LANGUAGE MODELS

We used Gemini-2.5-Pro to assist in expanding and consolidating the taxonomy of tasks in our benchmark. Both DeepSeek-V3 and Gemini-2.5-Pro were utilized for the automated pre-screening of candidate data. The final task definitions and data samples are verified by humans. We also used GPT-4o to generate some of the illustrative figures presented in the paper, and used GPT-5 to polish the manuscript text. Only human-verified revisions are included in the final version.

A RELATED WORK

A.1 AUDIO LANGUAGE MODELS

With the advancements of large language models (LLMs) and multimodal language models (Yang et al., 2025a; Jiang et al., 2024; Achiam et al., 2023; Comanici et al., 2025; Cai et al., 2024; Touvron et al., 2023; Liu et al., 2024c; 2025; Zhang et al., 2025b; Qi et al., 2025; Xing et al., 2025b;a; Ding et al., 2025; Wei et al., 2025a;b; Li et al., 2025; Zhang et al., 2025a), recent research has increasingly focused on integrating audio perception with LLMs to enhance audio understanding and reasoning. Existing methods can be broadly grouped into two categories: Large Audio Language Models(LALMs) and Omni Language Models(OLMs).

Most LALMs combine a pre-trained audio encoder with an LLM backbone, where the two modalities are aligned via large-scale text-audio joint training. Notable models include LTU-AS (Gong et al., 2023), SALMONN (Tang et al., 2024), MidashengLM (Dinkel et al., 2025), Audio Flamingo series (Ghosh et al., 2025; Goel et al., 2025), Qwen-Audio series (Chu et al., 2023; 2024), Step-Audio (Wu et al., 2025) and Mimo-Audio (Xiaomi, 2025). These models have achieved remarkable performance across a wide range of audio understanding tasks, including automatic speech recognition(ASR), spoken question answering(SpokenQA), and automated audio captioning(AAC). In parallel, OLMs extend this paradigm to unify multimodal understanding with representative examples such as Qwen-2.5-Omni (Xu et al., 2025), Ming-Omni (AI et al., 2025), MiniCPM-O (Yao et al., 2024), Phi-4 (Abouelenin et al., 2025), GPT-4o (Achiam et al., 2023), and Gemini 2.5 (Comanici et al., 2025). Notably, they also achieve impressive performance on audio understanding and reasoning, highlighting their potential to bridge multimodal perception and advanced audio intelligence.

A.2 AUDIO BENCHMARKS

Existing audio benchmarks illustrate the rapid progress of multimodal evaluation but also expose limitations. AudioBench (Wang et al., 2024) and AIR-Bench (Yang et al., 2024) primarily focus on tasks such as automatic speech recognition (ASR), spoken question answering (SpokenQA), and audio captioning (AAC). These settings tend to reduce audio understanding to transcription or description, thereby neglecting the broader spectrum of acoustic reasoning. MMAU (Sakshi et al., 2025) and MMAR (Ma et al., 2025) further expand the scope, yet their results reveal an inherent weakness: LLMs with audio captions can achieve comparable performance to advanced LALMs, suggesting that these benchmarks probe little beyond language-level semantics. MMAU-Pro (Kumar et al., 2025) attempts to add temporal and spatial reasoning, but its scope is restricted to single-audio temporal reasoning and single static spatial reasoning.

Beyond the LALMs evaluation, multimodal benchmarks in video question answering (Cheng et al., 2025; Yang et al., 2025c) and embodied AI (Yang et al., 2025b) have emphasized temporal and spatial reasoning. However, these frameworks are predominantly grounded in the visual modality, leaving the audio modality underexplored. Real-world audio understanding frequently requires integrating information across multiple sound streams and reasoning about subtle changes in intensity, phase, or frequency—capabilities that existing benchmarks scarcely capture.

Our benchmark aims to address these gaps by introducing tasks that require **multi-audio input and cross-audio reasoning**, such as comparing or integrating information across multiple sound inputs, as well as **fine-grained spatio-temporal deep reasoning**, such as tracking how acoustic patterns evolve with underlying physical changes. Rather than being limited to surface-level semantics, the benchmark is designed to assess whether models can leverage raw audio cues to perform physically grounded reasoning across spatial and temporal dimensions.

Attribute	Range / Level	Example
Absolute Perception Range		
Pitch, Loudness	125 Hz - 8000 Hz -10dB - 110dB	[Audio] The audio you just heard is divided into two halves. Does a sound appear in the first half, the second half, or is it not present at all? (A) The first half (B) The second half (C) It is not present at all (D) Unable to determine
Azimuth	0° - 360°	[Audio] Given that 0° is directly in front and the angle increases clockwise, which azimuth range is the sound most likely coming from? (A) Front-Right (0°-90°) (B) Back-Right (90°-180°) (C) Back-Left (180°-270°) (D) Front-Left (270°-360°) (E) Unable to determine
Elevation	-90° - 90°	[Audio] Where does the sound seem to be coming from in terms of elevation, relative to ear level? (A) Above ear level (B) Below ear level (C) At ear level (D) Unable to determine
Distance	0 meter - 10 meters	[Audio] How far away does the sound seem to be? (A) Near (within about 0-3 meters) (B) Medium (around 3-8 meters) (C) Far (more than 8 meters) (D) Unable to determine
Relative Discrimination Sensitivity		
Pitch	0, 50, 100, 200, 400, 1200 (cents)	[Audio] Which sound has a higher pitch: the first sound, the second sound, or are they the same? (A) The first sound has a higher pitch (B) The second sound has a higher pitch (C) Both sounds are the same (D) Unable to determine
Loudness	0, 4, 8, 12, 24, 48 (dB)	[Audio] Which sound is louder: the first sound, the second sound, or are they the same? (A) The first sound is louder (B) The second sound is louder (C) Both sounds are the same (D) Unable to determine
Duration	0, 20, 50, 100, 150, 200 (%)	[Audio] Which sound is longer: the first sound, the second sound, or are they the same? (A) The first sound is longer (B) The second sound is longer (C) Both sounds are the same (D) Unable to determine
Azimuth	30, 60, 90, 120, 150, 180 (°)	Audio 1: [Audio_1] Audio 2: [Audio_2] Are Audio 1 and Audio 2 at the same azimuth? (Consider differences of less than 45° as the same.) (A) Same (B) Different (C) Unable to determine
Elevation	15, 90, 120, 150 (°)	Audio 1: [Audio_1] Audio 2: [Audio_2] Which audio has the higher elevation angle? (Consider differences of less than 45° as the same.) (A) Audio 1 is higher (B) Audio 2 is higher (C) Both are at the same elevation (D) Unable to determine
Distance	1-2, 4-5, 6-7, 8-9 (meters)	Audio 1: [Audio_1] Audio 2: [Audio_2] Which audio is farther away? (Consider differences of less than 3 meters as the same.) (A) Audio 1 is farther away (B) Audio 2 is farther away (C) Both audios are the same (D) Unable to determine

Table 3: Task examples of foundational acoustic perception.

B DETAILS OF DATA ANNOTATION

In this section, we present the details of data annotation.

B.1 PROMPTS FOR AUDIO CAPTIONING

The prompt for Gemini 2.5 Pro audio captioning: “Please provide a detailed description of the audio, including speech, music, environmental sounds, and any other noticeable elements. Be as specific as possible.”

B.2 DETAIL INFORMATION FOR FOUNDATIONAL ACOUSTIC PERCEPTION

Tab. 3 details the ranges and levels used for each acoustic attribute, alongside illustrative examples of our foundational acoustic perception tasks.

B.2.1 BINAURAL AUDIO SYNTHESIS

We generated binaural recordings for foundational perception tasks (azimuth, elevation, distance) in Pyroomacoustics (Scheibler et al., 2018) across three rectangular rooms—small ($4.0 \times 3.5 \times 2.8$ m), medium ($8.0 \times 6.0 \times 3.5$ m), and large ($20 \times 15 \times 8$ m)—each with a frequency-independent wall absorption coefficient of 0.25. Image-source reflections were modeled up to order 10 at 44.1 kHz (matched to the HRTF sampling rate). For each room, we evaluated two listener positions (distinct Cartesian coordinates) and oriented the head toward the +x axis. Binaural reception used a co-located two-microphone array at the listener position with ear-specific directivity derived from a measured SOFA HRTF¹ (MIT KEMAR, “normal pinna”; interpolation order 12, 1000 points), loaded via a local SOFA reader and applied to the left/right channels.

For each condition (room \times listener), sources were placed on a sphere centered at the listener (radii 1–10 m; configurable azimuth/elevation), and ear-specific BRIRs were computed. Mono source signals were drawn from three curated audio clips (“alarm,” “applause,” “telephones”), downmixed if necessary. Rendering was performed by convolving each dry signal with the left/right BRIRs after an early/late mix to emphasize distance cues: we preserved the first 80 ms and attenuated the late tail by 0.5. We then applied global peak normalization across the batch to avoid clipping while preserving inter-position level differences.

We discretized each attribute into fixed partitions to control dataset balance.

Absolute azimuth: Eight angles $\{30^\circ, 60^\circ, 120^\circ, 150^\circ, 210^\circ, 240^\circ, 300^\circ, 330^\circ\}$. For each angle we rendered all combinations of 3 rooms \times 2 listener positions \times 2 source clips, yielding $8 \times (3 \times 2 \times 2) = 96$ utterances. **Absolute elevation:** Six angles $\{-75^\circ, -45^\circ, -15^\circ, 15^\circ, 45^\circ, 75^\circ\}$. Per angle we rendered 3 rooms \times 2 listener positions \times 2 source clips, for $6 \times (3 \times 2 \times 2) = 72$ utterances.

¹https://sofacoustics.org/data/database/mit/mit_kemar_normal_pinna.sofa

Absolute distance: Radii from 1–10 m with a nonuniform allocation to emphasize near-field cues: for 1–7 m we generated 6 utterances per meter (42 total), and for 8–10 m we generated 3 per meter (9 total), giving $42 + 9 = 51$ utterances per (room \times listener) set.

Relative azimuth: Differences were multiples of 30° : $\{30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ, 180^\circ\}$ (6 levels), totaling $6 \times 20 = 120$ utterances. **Relative elevation:** Four difference angles $\{15^\circ, 90^\circ, 120^\circ, 150^\circ\}$ with 18, 17, 17, 12 utterances respectively (64 total). **Relative distance:** Four difference levels $\{1 - 2, 4 - 5, 6 - 7, 8 - 9\}$ m with counts per level $\{12, 12, 12, 9\}$, totaling 45 utterances.

B.3 PROMPT USED FOR AI-ASSISTED AUTOMATED FILTERING OF TEMPORAL TASK DATA

Fig. 10 and Fig. 11 present our carefully designed prompts, which leverage LLMs and MLLMs to filter candidate data that meet the requirements of audio segment reordering.

Role Setting:
You are a rigorous audio analysis expert, specializing in identifying dynamic audio with explicit temporal logic conforming to physical laws or strong causality. Your task is to screen suitable audio samples for high-standard “Audio Sequence Ordering Evaluation.” Analysis should rely on the audio itself, with text as auxiliary reference.

“Audio Clip Ordering” Evaluation Task:
Qualified audio is segmented into three clips, shuffled, and given to the model, which must reconstruct the sequence using only sound. Candidate audio must meet the following strict standards:

- 1.Strong sequence uniqueness:** The events in the audio must present a unique and clearly discernible temporal progression, with no possibility of alternative plausible orderings.
- 2.Semantic clarity:** Events in the audio must be easily identifiable by sound alone.
- 3.High logical universality:** The event sequence should conform to commonsense physical laws or strong causal relations, such that listeners from different backgrounds can reach a consistent understanding.

Note: Since the model sees no text, samples must be interpretable solely from sound.

Audio Classification Standards:

1. Sortable Single Event:

- Definition: Audio primarily represents a continuous event driven by a single process, exhibiting significant and predictable temporal dynamics.
- Possible categories include (but are not limited to):
 Spatial movement and distance variation. Physical processes in progression. Energy or state decay. Biological activity dynamics.
- Core judgment: The core judgment is whether the change is governed by a single continuous physical process, is significant, and is commonly recognized as unambiguous.
- Examples:
 The sound of pouring water. A ball bouncing to rest.

2. Sortable Multi-Event:

- Definition: Audio contains two or more independent events, where the events exhibit a strong causal relationship such that “A inevitably leads to B” or “A must precede B.”
- Core judgment: The core judgment is whether sub-events have distinct, separable acoustic features, follow a direct and widely recognized causal chain, and together form a concise, complete, and unambiguous process.
- Examples:
 Opening a bottle → pouring water → setting down the bottle.
 Cracking an egg → stirring → pouring into an oil pan.

Note: If event boundaries are unclear but the overall structure forms a continuous process, classify as “Single Event.” If there are clearly distinct stages with evident logical links, classify as “Multi-Event.”

#Special Exclusion Rules:

Filter out the following audio types, even if they exhibit some “dynamic change.” If they fail semantic clarity, sequence uniqueness, or logical universality, they must be labeled “Not Applicable.”

- Static or repetitive sounds.
- Sounds with intensity variation but without reflecting a physical law or typical action sequence.
- Weak or absent causality, or dependent on special background knowledge.
- Ambiguous, subjective, or interpretable in multiple ways.
- Ambient noise or atmosphere shifts without structured events.
- Dialogue or speech as main content.
- Poor quality or indistinct segmentation.

Features of High-Quality Samples:

- Clear event semantics: Each action or change is acoustically distinct and intuitively understandable.
- Unambiguous sequence: Temporal logic is indisputable, allowing only one reconstruction order.
- Narrativity and processuality: The audio presents a complete, coherent event chain consistent with human understanding and commonsense reasoning.

Figure 10: The prompt for our AI-assisted filtering process on temporal tasks.

Input
You will receive the following three components:

- 1.Text description: Labels, annotations, or metadata about the audio.
- 2.Preliminary judgment based on text: A text-only hypothesis about the audio’s attributes.
- 3.Audio file.

Note: Always base your classification primarily on the acoustic features of the audio itself, using text only as supplementary context.

#Output Format (Strictly follow the four-line structure)
Classification label: Sortable Single Event | Sortable Multi-Event | Uncertain | Not Applicable
Event description: For “Sortable Multi-Event,” use the format Event A → Event B → Event C to describe the strong causal chain. For “Single Event,” describe the core dynamic change. If “Not Applicable”, briefly summarize the content.
Reasoning: Explain classification basis, focusing on the three standards: semantic clarity, sequence uniqueness, and logical universality. Refer to specific acoustic features where possible.
Score: (Composite evaluation based on the three standards)
- 5 points: Events clear, sequence strong, logic robust, full structure or rich narrative (≥ 4 events).
- 4 points: Sequence clear, physics/logic reasonable, but semantics somewhat limited (≤ 3 events).
- 3 points: Some ambiguity, human listeners may also struggle to interpret
- 2 points: Content vague, minimal sequential logic, not recommended for ordering.
- 1 point: Events unclear, structure absent, unusable for ordering tasks.

Examples:

Input:
Text description: “A ping-pong ball landing.”
Preliminary judgment: –
Audio file: Sharp bouncing sounds, gradually weakening, intervals shortening, finally stopping
Output:
Classification label: Sortable Single Event
Event description: Ping-pong ball bouncing, sound intensity and interval decay until stop.
Reasoning: Clear physical energy decay, unique order, limited semantics.
Score: 4

Input:
Text description: “People counting down for New Year, then celebrating with fireworks.”
Preliminary judgment: –
Audio file: 10-second countdown → fireworks → cheering
Output:
Classification label: Sortable Multi-Event
Event description: Countdown (10s) → Fireworks → Cheering
Reasoning: Three distinct events, unique order, clear logic, strong semantics.
Score: 5

Input:
Text description: “A segment of continuous machine rumble.”
Preliminary judgment: –
Audio file: Indistinct machine operating noise
Output:
Classification label: Not Applicable
Event description: Continuous monotonous machine rumble without structural change.
Reasoning: Lacks identifiable dynamic events or temporal structure, unsuitable for ordering evaluation.
Score: 1

Figure 11: The prompt for our AI-assisted filtering process on temporal tasks.

C ROBUST EVALUATION

All questions in STAR-BENCH are presented as clear multiple-choice questions with well-formatted options. We adopt classification accuracy as the evaluation metric. To determine the correctness of a response, we employ string matching to extract either the chosen option label (e.g., <A>) or the full text content of the option from the model’s output.

Furthermore, we implement a robust evaluation strategy to ensure rigorous and reliable results. For perception and spatial tasks, we adopt the CircularEval method from MM-Bench (Liu et al., 2024b). Specifically, each question is presented to the model N times (N is the number of options), with the option order cyclically rotated in each run to mitigate potential positional biases. For temporal tasks, we conduct three runs per question with different temporal segment orders to evaluate the model’s robustness to sequence variations. Note that due to the significant API costs, GPT-4o Audio was evaluated only once per question. This strategy yields two key metrics: Average Accuracy (AA), the

mean accuracy across all evaluation runs, and All-Correct Rate (ACR), the proportion of questions answered correctly in every single run, which serves as a stronger indicator of model reliability.

For models that do not support multi-audio input (only Audio Flamingo 3 and its Think variant among the models we evaluated), we concatenate the audios with a 2-second silence and specify this in the prompt. In contrast, for models that support multiple audio inputs, we feed them sequentially with textual indices.

To establish a human performance baseline, we conduct a human evaluation on a randomly sampled subset of approximately 10% of the data from each task. This evaluation is performed by 10 university students, from whom we explicitly exclude anyone involved in data annotation or with domain-specific expertise, thereby ensuring a general, non-expert perspective.

D BREAKDOWN RESULTS

In this section, we present detailed results for perception, temporal reasoning, and spatial reasoning on STAR-BENCH, as shown in Tab. 4, Tab. 5, and Tab. 6.

E FURTHER ANALYSIS AND DISCUSSION

E.1 HIGH OUTPUT INSTABILITY AND CONCENTRATED PREDICTIONS

The reliability of model outputs on our benchmark is notably low, as evidenced by the stark contrast between their Average Accuracy (AA) and All-Correct-Rate (ACR) scores. Even the top-performing model, Gemini 2.5 Pro, exhibits an average drop of 25.01 percentage points from its AA to its ACR. This issue is even more pronounced for the majority of open-source models, which record an ACR near zero. This score indicates a complete failure to maintain consistent predictions under minor input perturbations. For these models, the instability often manifests as a tendency to concentrate predictions on a specific option, suggesting a reliance on superficial biases rather than genuine understanding.

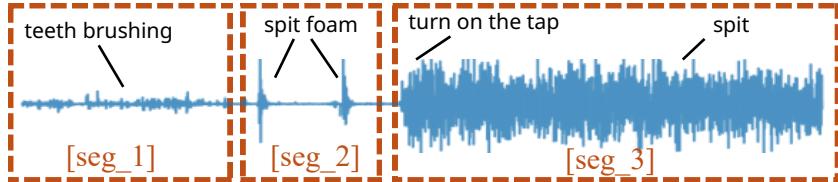
E.2 ABLATION STUDY ON SPATIAL REASONING.

As shown in Tab. 6, the results reveal a fundamental limitation of LALMs’ spatial understanding in perception. The **native input** inherently discards part of the multi-channel information during model preprocessing, which leads to a significant loss of spatial cues that are essential for fine-grained reasoning. On the other hand, the **channel-wise input** explicitly presents each channel with textual instructions, mitigating some of the information loss. However, as most models are not trained on multi-audio inputs, they struggle to align channel representations and to exploit interaural cues reliably. Overall, the gap between human and model performance highlights that spatial reasoning in audio remains an unsolved challenge. While channel-wise input can provide partial gains, neither strategy fully captures spatial dependencies, underscoring the need for an audio encoder that natively supports multi-channel audio input.

F CASE STUDY

In this section, we present several case studies of error analysis, including temporal reasoning (Figs. 12 to 17) and spatial reasoning (Fig. 18).

[Brushing of teeth vigorously and then turning on the sink water.]



Q: clip1: [seg_2] clip2: [seg_3] clip3: [seg_1]

Answer: <E>: clip 3 → clip 1 → clip 2

Gemini 2.5 Pro

The most natural chronological order for these sounds is <A>. Here's a breakdown of the events:

Clip 1: The sound of solid objects, most likely ice cubes, being dropped into a container like a blender pitcher.

Clip 2: The sound of a liquid being poured, which would typically follow adding the solid ingredients.

Clip 3: The sound of a blender running, which is the final step to mix the ice and the liquid together.

Therefore, the most logical sequence is dropping in the ice (clip 1), pouring in the liquid (clip 2), and then blending (clip 3).

Error analysis **Perception Error Reasoning Error** The core error is a fundamental scene misidentification: the audio depicts a teeth-brushing routine, not making a blended drink. All subsequent step-by-step analysis is therefore invalid.



GPT-4o Audio

Let's break down the sounds:

Clip 1 sounds like spitting or rinsing into a sink.

Clip 2 sounds like water from a faucet, which typically happens before or after brushing.

Clip 3 sounds like someone brushing their teeth.

In a natural choreography of events, you might first turn on the water to wet the toothbrush or rinse your mouth, then proceed to brushing your teeth, and finally, rinse or spit after brushing.

That would make the most natural order:

Clip 2 (water on) -> Clip 3 (brushing) -> Clip 1 (rinsing/spitting).

So the answer is D: clip 2 -> clip 3 -> clip 1.

Error analysis **Reasoning Error Perception Error** The failure to isolate the spitting sound from the water noise led to an incorrect event order.



Qwen-2.5-Omni

To determine the most natural chronological order of these sounds, let's analyze each clip:

1. **Clip 1: A door is being opened and closed. This typically happens at the beginning of a sequence of activities.**

2. **Clip 2: A toilet is being flushed. This usually follows the use of the bathroom.**

3. **Clip 3:** Someone is brushing their teeth. This is a common activity that often happens after using the bathroom.

Given this sequence, the most natural chronological order would be:

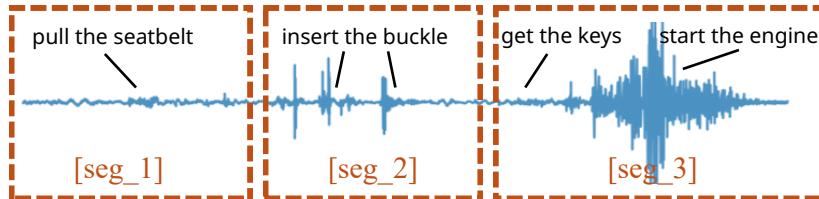
Open the door (Clip 1)- Use the bathroom (Clip 2)- Brush teeth (Clip 3)

Therefore, the correct answer is: <A>: clip 1 -> clip 2 -> clip 3

Error analysis **Perception Error Knowledge Gap** Qwen-2.5-Omni's perception is confined to individual clips, which prevents it from inferring sound events within the broader scene context. Consequently, seg_1 and seg_2 remain unassociated with seg_3, causing the model to engage in flawed or spurious reasoning.

Figure 12: An error case from the temporal reasoning task.

[A man fastens his seatbelt, then starts up the car.]



Q: clip1: [seg_3] clip2: [seg_1] clip3: [seg_2]

Answer: <D>: clip 2→clip 3→clip 1



Gemini 2.5 Pro



The most natural chronological order for these sounds is <F>. Here's a breakdown of the events:

1. Clip 3: This is the sound of a car door closing. This is one of the first actions taken after getting into a car.
 2. Clip 2: This is the sound of a person settling into the car seat. You can hear clothes rustling and the seat creaking as they get comfortable. This typically happens after the door is closed.
 3. Clip 1: This is the sound of keys being put into the ignition, followed by the car engine starting. This is the final step in the sequence before driving away.
- Therefore, the most logical real-world sequence is closing the door (Clip 3), settling into the seat (Clip 2), and then starting the car (Clip 1).

Error analysis Perception Error The first half of seg_2 is the sound of pulling the seatbelt, and the second half is the sound of inserting the buckle.



GPT-4o Audio



Let me break down the sounds.

Clip 2 sounds like a zipper, maybe someone zipping up a bag or coat.

Clip 3 is the sound of a car engine starting, along with some mechanical sounds.

Clip 1 sounds like a door closing.

In a natural sequence, you would expect someone to zip up their bag or jacket, then close a door after stepping out or in, and then finally start the car.

So the most natural chronological order would be: clip 2 (zipping) → clip 1 (door closing) → clip 3 (car starting).

That aligns with option E.

Error analysis Perception Error Reasoning-Answer Misalignment <E> should be clip3→ clip1 → clip2



Qwen-2.5-Omni



To determine the most natural chronological order of the sounds, let's analyze each clip:

1. Clip 1: This clip contains the sound of a car engine starting. This is typically the first action in a sequence of events when someone is about to drive.
2. Clip 2: This clip contains the sound of a seatbelt being fastened. This action usually happens after the car engine has started, as the driver or passenger needs to secure themselves before driving.
3. Clip 3: This clip contains the sound of a car door opening. This action usually happens before the car engine starts, as the driver or passenger needs to enter the car before starting the engine.

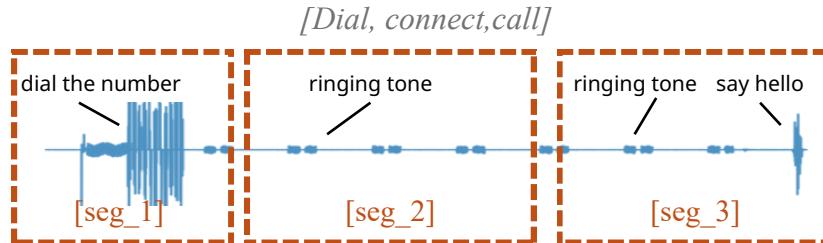
Given this analysis, the most natural chronological order is:

Clip 3: Car door opening - Clip 1: Car engine starting- Clip 2: Seatbelt being fastened

Therefore, the correct answer is:<E>: clip 3 -> clip 1 -> clip 2

Error analysis Perception Error Seg_3 is the sound of inserting the seatbelt buckle.

Figure 13: An error case from the temporal reasoning task.



Q: clip1: [seg_2] clip2: [seg_3] clip3: [seg_1]

Answer: <E>: clip 3→clip 1→clip 2



Gemini 2.5 Pro



The most natural chronological order for these sounds is:

1. **Clip 3:** The sound of a phone number being dialed on a touch-tone keypad. This is the first action taken to initiate a call.
2. **Clip 1:** The sound of a phone ringing (the ringback tone). This is what the caller hears after dialing and while waiting for the recipient to pick up.
3. **Clip 2:** The sound of someone answering the phone with "Hello?". This is the final step in the sequence, where the call is connected.

Therefore, the correct order is clip 3 -> clip 1 -> clip 2. The correct option is <E>.

Error analysis correct



GPT-4o Audio



The most natural chronological order for these sounds would typically be:

<A>: clip 1 -> clip 2 -> clip 3

This is because clip 1 sounds like a phone ringing, clip 2 has someone answering the phone with "hello," and **clip 3 is a busy signal**, which can occur after the call is ended or fails to connect.

Error analysis **Perception Error** Seg_3 is the sound of pressing numbers, not a busy signal.



Qwen-2.5-Omni



The most natural chronological order in which these sounds would typically occur in the real world is:

<A>: clip 1 -> clip 2 -> clip 3

Explanation:

1. The first sound (clip 1) is a telephone ringing, indicating that someone is calling.
2. The second sound (clip 2) is a person answering the phone, saying "hello."
3. The third sound (clip 3) is a **dial tone, which typically occurs after the call is connected and the call is being made.**

Error analysis **Reasoning Error** Seg_3's explanation is problematic: the dial tone occurs before the call is connected, not after.

Figure 14: An error case from the temporal reasoning task.

[An ambulance approached and then moved away.]

Q: clip1: [seg_3] clip2: [seg_1] clip3: [seg_2]

Answer: <D>: clip 2→clip 3→clip 1

Gemini 2.5 Pro ✨ ✓

The most natural chronological order for these sound events is <D>: clip 2 -> clip 3 -> clip 1.

Here is a breakdown of the events as they would likely occur:

1. **Clip 2:** This clip captures the sound of an emergency siren that is clearly approaching, followed by the abrupt, jarring sound of a collision or a car horn blaring, signifying a crash. This is the inciting incident.
2. **Clip 3:** This clip features a loud, steady siren. This represents the immediate aftermath of the crash, with the emergency vehicle now at the scene, its siren still active.
3. **Clip 1:** This clip features the sound of a siren fading into the distance (the Doppler effect is noticeable as the pitch drops). The sound of the damaged car's horn continues to blare but also seems to be part of a more distant, ambient soundscape. This suggests time has passed, and the emergency vehicle is now leaving the scene.

Error analysis correct

GPT-4o Audio ❌

All three clips sound like sirens approaching, but there's a noticeable difference in pitch and intensity that suggests a changing distance. It's common for sirens to start from far away and then approach closer, raising the volume and changing the pitch. From what I heard:

One clip likely represents a more distant or quieter siren.
Another one sounds like it's getting closer.
The final one should be the closest and the loudest.

Putting that together in a natural order, the most logical sequence of events in real life would be the siren starting far away and gradually approaching closer. Hence, the best chronological order would be where the siren is initially distant, moves closer, and finally is the closest. Looking at the pitch and dynamics of the sirens: Clip 2 sounds furthest away, **clip 1 seems closer than clip 2 but not the closest**, and clip 3 seems the closest.
So, the most natural chronological order would be clip 2 -> clip 1 -> clip 3. That corresponds to option C.

Error analysis Perception Error Seg_3 decreases in volume.

Qwen-2.5-Omni ❌

To determine the most natural chronological order of the sounds in the audio clips, let's analyze each clip:

Clip 1: An emergency vehicle siren is heard, indicating a situation requiring immediate attention.

Clip 2: Another emergency vehicle siren is heard, suggesting a second emergency vehicle is approaching or responding.

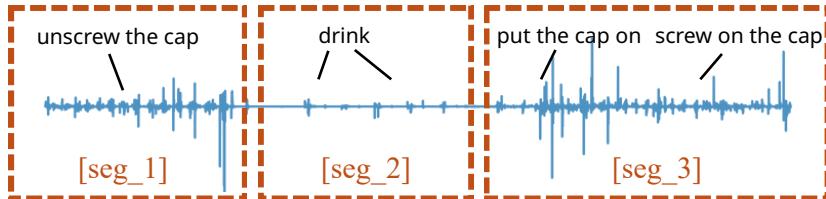
Clip 3: A third emergency vehicle siren is heard, indicating a third emergency vehicle is also responding.

Given this sequence, the most natural chronological order would be: First, the first emergency vehicle arrives (Clip 1). Then, the second emergency vehicle arrives (Clip 2). Finally, the third emergency vehicle arrives (Clip 3). Therefore, the correct answer is: <A>: clip 1 -> clip 2 -> clip 3

Error analysis Knowledge Gap The descriptions disregard fundamental audio characteristics, making no attempt to leverage pitch or volume variations to infer movement, thereby rendering the sequence meaningless.

Figure 15: An error case from the temporal reasoning task.

[A person opens a jar, quickly gulps the water and then closes the jar.]



Q: clip1: [seg_1] clip2: [seg_3] clip3: [seg_2]

Answer: : clip 1→clip 3→clip 2



Gemini 2.5 Pro

The most natural chronological order for these sounds is <F>: clip 1 -> clip 2 -> clip 3. X

Here is a breakdown of the events in that order:

1. **Clip 1:** This clip features the sound of a container, possibly with a powder or solid inside, being shaken. This is followed by the sound of the container being opened (like a cap being twisted off). This is a common preparatory step, like shaking a bottle of drink mix before opening it.
2. **Clip 2:** This clip begins with **the distinct sound of a liquid being poured from a bottle**. This would logically follow opening the container in clip 3, as one would pour a liquid (like water) into the mix. The clip also contains a spraying sound, but the pouring is the key transitional sound from clip 3.
3. **Clip 3:** This clip is the sound of someone drinking and swallowing the liquid. This is the final action of consumption after the drink has been prepared.

Error analysis Perception Error Seg_3 is the sound of closing the lid, not pouring.



GPT-4o Audio



Let's break down the sounds logically.

Clip 1 sounds like ice cubes being placed into a glass.

Clip 2 sounds like liquid being poured into that glass.

Clip 3 sounds like someone sipping or drinking from the glass.

In a natural sequence of events, you would first put ice in the glass, then pour the liquid, and finally drink from it.

Therefore, the most natural chronological order is: clip 1 (ice first), clip 2 (liquid poured next), and clip 3 (drinking the drink last).

That corresponds to option A: clip 1 -> clip 2 -> clip 3.

Error analysis Perception Error Descriptions of both seg_1 and seg_3 are incorrect.



Qwen-2.5-Omni



To determine the most natural chronological order of the sounds, let's analyze each clip:

1. **Clip 1:** Sounds like someone is stirring or mixing something in a container, possibly a glass or a cup.
2. **Clip 2:** Similar to Clip 1, it also sounds like stirring or mixing, but the sound is slightly different, possibly indicating a different type of container or a different substance.
3. **Clip 3:** Sounds like someone is drinking from a glass or a cup.

Given these observations, the most natural sequence would be:

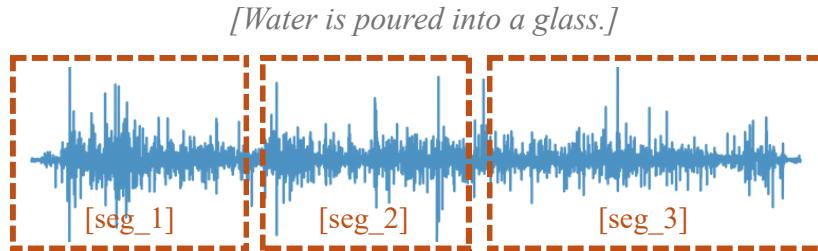
1. **Stirring or mixing (Clip 1)**
2. **Stirring or mixing (Clip 2)**

3. Drinking (Clip 3)

Therefore, the most natural chronological order is: <A>: clip 1 -> clip 2 -> clip 3

Error analysis Reasoning Error Simply placing the two mixing segments together lacks a rational explanation.

Figure 16: An error case from the temporal reasoning task.



Q: clip1: [seg_3] clip2: [seg_2] clip3: [seg_1]

Answer: <E>: clip 3→clip 2→clip 1



Gemini 2.5 Pro

Based on the sounds in the audio clips, the most natural chronological order is:



1. **Clip 2:** This is the sound of a carbonated beverage being poured vigorously into a glass, characterized by the "glugging" sound as air replaces the liquid in the bottle. This represents the main part of the pour.

2. **Clip 1:** This sounds like the continuation or finishing of the pour, perhaps "topping off" the glass after the initial foam has settled. The sound is less vigorous than in clip 2.

3. **Clip 3: This is the distinct sound of someone taking a large gulp or swallowing the drink that has just been poured.**

Therefore, the logical sequence of events is pouring the drink, finishing the pour, and then drinking it. The correct option is <C>.

Error analysis **Perception Error** Seg_1 is not the sound of gulping, but the crisp sound of water being poured into a glass.



GPT-4o Audio



"The most natural chronological order in which these sounds would typically occur is:
: clip 1 (**blender sound**) -> clip 3 (pouring liquid) -> clip 2 (**sipping**)

Error analysis **Perception Error** All segments are sounds of pouring water; the perception of both seg_3 and seg_2 is entirely incorrect.



Qwen-2.5-Omni



To determine the most natural chronological order of these sound events, let's analyze each clip:

Clip 1: Liquid being poured into a container.

Clip 2: Liquid being poured into a container.

Clip 3: Liquid being poured into a container.

Given that all three clips involve the same action (pouring liquid into a container), the most natural chronological order would be based on the sequence of events typically observed in a real-world scenario.

If we assume that the containers are being filled in a sequential manner, the most logical order would be:

1. First container is filled.

2. Second container is filled.

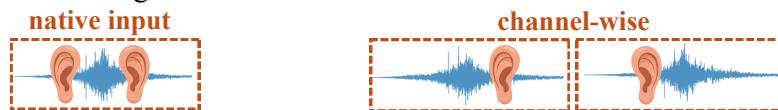
3. Third container is filled.

Therefore, the most natural chronological order is: <A>: clip 1 -> clip 2 -> clip 3

Error analysis **Perception Error Reasoning Error** It cannot perceive differences between segments, and can only output coarse-grained captions, leading to an inability to reason correctly.

Figure 17: An error case from the temporal reasoning task.

Q: What is the movement trajectory of the racing car?
Options: <A> From left to right From right to left <C> Remains unchanged
Answer: : From right to left



◆ **Gemini 2.5 Pro** X
Based on the audio, the sound of the race car **starts in the left channel, moves across the stereo field, and then fades out in the right channel**. This indicates the car is moving from left to right past the recording point.
Therefore, the correct answer is <A>: From left to right.

Error analysis

native: **Perception Error** the channels interfere with each other, causing the model's basic perception to fail

channel-wise: **Perception Error Knowledge Gap** It indicates that the model lacks knowledge for processing spatial audio tasks, relying only on volume to perceive spatial audio, which leads to incorrect answers.

◆ **GPT-4o Audio** X
The racing trajectory of a race car typically involves taking the optimal racing line through each corner to maintain speed and minimize lap time. It often involves starting wide on the approach, cutting close to the apex of the turn, and then moving wide again on the exit, to achieve the smoothest and fastest path. **So, the trajectory changes dynamically depending on the turns and straights of the track.**

◆ **Qwen-2.5-Omni** X
The audio does not provide information about the direction of the race car's movement, so it is impossible to determine the racing trajectory.
Therefore, the correct answer is C: Remains unchanged.

Error analysis

Perception Error Generation Collapse Hallucination

It indicates that during the binaural audio processing, the channels interfere with each other, causing the model's basic perception to fail and resulting in hallucinated outputs.

The correct answer is <A>: **From left to right**. Here's a breakdown of why:

1. **Audio 1 (Left Ear):** The sound of the race car starts, **becomes very loud, and then fades away**.
2. **Audio 2 (Right Ear):** The sound is initially quieter than in the left ear. **It becomes loud as the car passes the center point and then remains louder than the left ear as the car moves away.**

By comparing the two channels, we can hear that the sound is first prominent in the left ear (Audio 1) and then becomes prominent in the right ear (Audio 2). **This change in volume and timing between the ears creates the auditory illusion of the car traveling from the listener's left to their right.**

Based on the binaural cues in the audio recording, the correct answer is **B. From right to left.**

Based on the audio, the sound of the race car moves from the left channel to the right channel, which means the car is moving from the left to the right side. The correct option is <A>: From left to right.

Error analysis

Hallucination Knowledge Gap

Perception Error

It is shown that the model lacks knowledge about spatial audio tasks, making it unable to carry out the correct reasoning process, which in turn leads to hallucinated outputs.

The racing trajectory of a race car would be <A>: From left to right.

Figure 18: An error case from the spatial reasoning task.