

# 000 001 002 003 004 005 006 007 008 009 010 STAR-BENCH: PROBING DEEP SPATIO-TEMPORAL 002 REASONING AS AUDIO 4D INTELLIGENCE

005 **Anonymous authors**

006 Paper under double-blind review

## 009 ABSTRACT

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

## 032 1 INTRODUCTION

033 As a fundamental modality of human perception, audio serves a pivotal role in communication, aes-  
034 thetic appreciation, and situational awareness, complementing the limitations of visual perception.  
035 With the rise of Multimodal Large Language Models (MLLMs) (Comanici et al., 2025; Achiam  
036 et al., 2023) and especially Large Audio-Language Models (LALMs) (Chu et al., 2024; Goel et al.,  
037 2025), these models have shown impressive capabilities in understanding audio, representing a cru-  
038 cial step toward diverse applications such as embodied intelligence (Paul et al., 2022).

039 To drive progress, a series of audio benchmarks has been introduced (Yang et al., 2024; Sakshi et al.,  
040 2025), covering traditional tasks like Automatic Speech Recognition (ASR) and sound event classi-  
041 fication. While some recent efforts are beginning to emphasize reasoning abilities (Ma et al., 2025;  
042 Kumar et al., 2025), we observe that existing benchmarks predominantly focus on coarse-grained  
043 semantic content, which is audio information that can be distilled into textual descriptions with min-  
044 imal loss. As shown in the **left** part of Fig. 1, we first use Gemini 2.5 Pro (Comanici et al., 2025)  
045 to generate detailed audio captions for samples in recent representative audio benchmarks MMAU  
046 (test-mini) (Sakshi et al., 2025) and MMAR (Ma et al., 2025). We then prompt the model to answer  
047 questions based *only* on these audio captions, and its performance drops by only 5.9% and 9.0%,  
048 respectively, compared to when it processes the raw audio. This result suggests that existing bench-  
049 marks primarily evaluate audio information that is **easily representable by text**. However, human  
050 auditory intelligence is not limited to this coarse-grained understanding. For example, humans can  
051 intuitively judge the water level in a container from the dynamic changes in the pouring sound, even  
052 without being able to precisely articulate the underlying acoustic features. Similarly, we can infer  
053 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**.

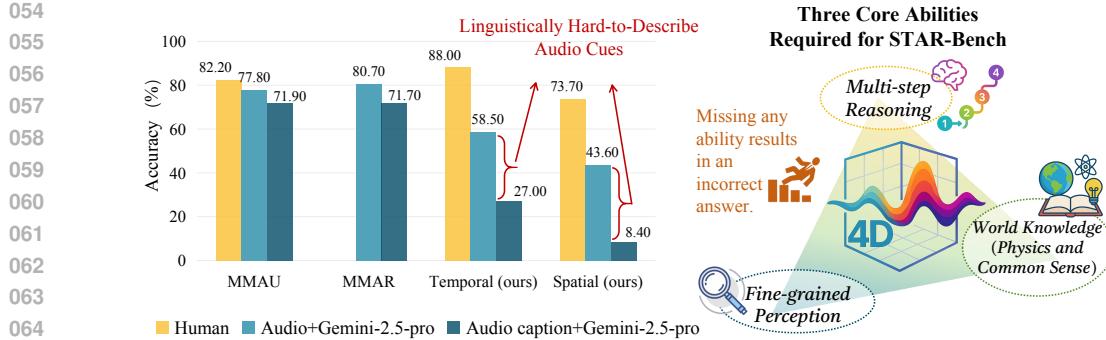


Figure 1: (Left): A comparison between humans and the Gemini 2.5 Pro with and without audio captions on various audio benchmarks. Our STAR-Bench evaluates linguistically hard-to-describe audio cues. See Appendix B.1 for audio caption details. (Right): The three core abilities required to solve tasks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

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 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 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.

108 Table 1: A comparative overview of our benchmark against other representative audio benchmarks.  
 109 (✓: Fully supported, ○: Partially supported or limited amount, ✗: 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]	✗	○	✗	✗	✓	✓	✓
<b>STAR-BENCH (ours)</b>	✓	✓	✓	✓	✓	✓	✓

## 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 largely overlook audio 4D intelligence. Although some advanced benchmarks do touch upon spatio-temporal aspects, their coverage remains limited in both scale and depth. While MMAU Sakshi et al. (2025), MMAU-Pro Kumar et al. (2025) and MMAR Ma et al. (2025) contain temporal questions, they mainly involve identifying the timing or ordering of events (e.g., when a sound occurs, which event comes first). These are primarily perceptual-layer tasks. By contrast, our “temporal deep reasoning” tasks require understanding physical principles or causal dynamics across segments (e.g., inferring how a process evolves over time or how one event implies another), which cannot be solved by local timing cues alone. In addition, the spatial tasks in MMAR and MMAU-Pro are often restricted to single-source localization, and many items do not necessitate meaningful use of stereo cues (e.g., simple arriving vs. departing judgments). In contrast, STAR-BENCH introduces a hierarchical design covering three sub-tasks in complex scenes and explicitly emphasizes stereo-cue-based reasoning.

A comparative overview of STAR-BENCH and prior benchmarks is presented in Tab. 1. STAR-BENCH evaluates deep spatio-temporal reasoning through tasks that go beyond surface-level perception and instead require applying physical or causal knowledge, performing multi-step reasoning in complex real-world scenarios, and integrating information across multiple clips or events. STAR-BENCH rests on a hierarchical and comprehensive task design. In addition, a rigorous data curation pipeline ensures high-quality samples, and robust evaluation strengthens the reliability of the results.

## 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

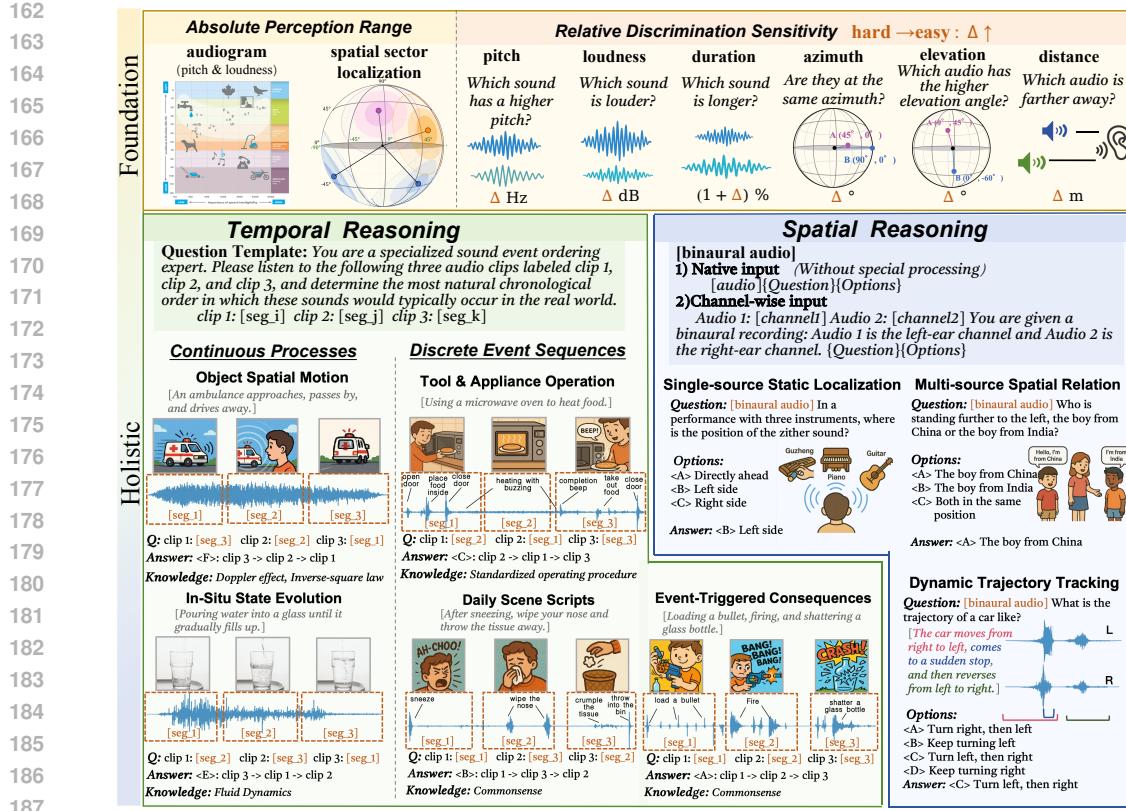


Figure 2: **STAR-BENCH**: (1) the foundational perception task (upper) and (2) the holistic spatio-temporal reasoning task, which includes both temporal reasoning (bottom left) and spatial reasoning (bottom right). Zoom in for the best view.

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 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 (Loudness, 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  $90^\circ$  quadrants (from  $0^\circ$  to  $360^\circ$ ), its elevation relative to ear-level (above, at, or below, from  $-90^\circ$  to  $90^\circ$ ), 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.

216 **2) Relative Discrimination Sensitivity**, which investigates how well a model can detect small  
 217 changes in acoustic attributes. The ability to detect small changes allows a model to make nuanced  
 218 judgments, like determining if a sound is getting louder or a pitch is rising. Analogous to measuring  
 219 the human Just Noticeable Difference (JND), the relative discrimination task presents the model with  
 220 an audio clip containing two sounds and requires it to compare them based on a specific attribute.  
 221 We meticulously designed four to six distinct difficulty levels for each of the six attributes, as de-  
 222 tailed in Tab. 3. Level 1 serves as a control group to test for random guessing, presenting identical  
 223 sounds ( $\Delta=0$ ) for non-spatial attributes and a sub-threshold difference for spatial ones. Subsequent  
 224 levels then introduce progressively larger differences, ranging from subtle variations perceptible to  
 225 humans to more significant, real-world changes. By analyzing the model’s performance across these  
 226 different levels of stimulus differences, we can quantitatively assess its discrimination sensitivity for  
 227 each attribute.  
 228

### 229 3.2 HOLISTIC SPATIO-TEMPORAL REASONING

230 Building on the model’s fundamental audio perceptual abilities (Sec. 3.1), we further introduce  
 231 holistic temporal reasoning (Sec. 3.2.1) and spatial reasoning (Sec. 3.2.2), which are designed to  
 232 systematically evaluate a model’s reasoning ability that is required for audio 4D intelligence.  
 233

#### 234 3.2.1 TEMPORAL REASONING TASKS

235 The core of temporal reasoning lies in understanding the intrinsic logic of event sequences, encom-  
 236 passing physical causality, functional procedures, or social conventions. To evaluate this capability,  
 237 we design a novel **Audio Segment Reordering** setting. Specifically, we curate a collection of au-  
 238 dio events characterized by strong sequential uniqueness, semantic clarity, and logical universality.  
 239 Each event is segmented into three clips, which are then shuffled as inputs to the model. The models  
 240 are required to restore the original temporal sequence based solely on the audio content. Our tem-  
 241 poral reasoning tasks are organized into two meta-categories (continuous processes, discrete event  
 242 sequences) and five subcategories based on their core logical principles.  
 243

244 The **continuous processes** assess a model’s ability to track the subtle, continuous evolution of acous-  
 245 tic features within a single, uninterrupted acoustic event. The **object spatial motion** subcategory  
 246 reconstructs the spatio-temporal trajectory of moving sources (e.g., passing cars, airplanes) by inter-  
 247 preting key acoustic cues, such as the Doppler effect (frequency shifts indicating relative velocity)  
 248 and the inverse-square law (loudness changes indicating distance). Besides, the **in-situ state evo-**  
 249 **lution** subcategory assesses a model’s ability to track the intrinsic evolution of a stationary object’s  
 250 state, a process governed by predictable trend patterns. These trend patterns arise from various un-  
 251 derlying principles, including: *Fluid & Pneumatic Dynamics*, where the sound is governed by prin-  
 252 ciples of turbulence, resonance, and pressure changes (e.g., a toilet flushing, water being poured);  
 253 *Thermodynamic Processes*, involving irreversible state changes driven by heat (e.g., water boiling,  
 254 food frying); *Energy Decay*, a process governed by resonant decay and frictional damping after a  
 255 single excitation (e.g., a bell’s chime, an explosion’s echo); and complex *Biological Rhythms* that  
 256 reflect an evolving physiological or emotional state.  
 257

258 The **discrete event sequences** category requires the model to understand the logical and temporal  
 259 relationships between multiple, distinct acoustic events, which are governed by function, conven-  
 260 tion, or causality. The **tool & appliance operation** sub-category follows the standardized operating  
 261 procedure for tools and appliances (e.g., a microwave, a power drill), where the sequence is correct  
 262 when it follows the tool’s designed function. The **daily scene scripts** sub-category applies com-  
 263 monsense and contextual script knowledge to follow the conventional sequence of actions in a daily  
 264 activity (e.g., brushing teeth, drinking water). The **event-triggered consequences** sub-category ap-  
 265 plies causal reasoning to infer that a trigger event (e.g., a firework explosion) will be followed by an  
 266 automatic and irreversible outcome, whether physical (glass shattering) or social (a crowd cheering).  
 267

#### 268 3.2.2 SPATIAL REASONING TASKS

269 Humans effortlessly perceive complex 3D auditory scenes (e.g., hearing a voice from behind, follow-  
 270 ing an approaching car, or locating multiple speakers). Such an ability is fundamental for egocen-  
 271 tric interaction and embodied AI systems, for instance, robots that navigate and interact with their  
 272 surroundings. However, existing benchmarks focus primarily on the localization of static sound  
 273

270 sources, whereas real-world scenarios demand reasoning that integrates both spatial and temporal  
 271 cues. To address this gap, we organize the spatial reasoning task into three subcategories.  
 272

273 The **single-source static localization** evaluates the model’s ability to identify the direction of a target  
 274 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  
 275 for more advanced reasoning. The **multi-source spatial relation** requires the model to determine  
 276 the relative spatial relationships among multiple simultaneous sound sources (e.g., comparing the  
 277 placement of two speakers to decide which one is further to the right). Beyond localizing each  
 278 source individually, the model must infer their spatial placement and choose the appropriate relational  
 279 description from multiple candidates. The **dynamic trajectory tracking** introduces moving sound  
 280 sources, which require the model to go beyond basic spatial perception to dynamically model  
 281 spatio-temporal relations for reasoning about complex movement trajectories (e.g., tracking a passing  
 282 car moving from left to right). This task extends spatial reasoning into the temporal domain and  
 283 is more faithful to the complexity of real-world acoustic scenarios.

284 However, evaluating existing LALMs on multi-channel spatial tasks is chal-  
 285 lenging. The common practice of these  
 286 models is to average multi-channel au-  
 287 dio into a mono signal, resulting in  
 288 the loss of substantial spatial infor-  
 289 mation. We conduct a simple experiment  
 290 as shown in Fig. 3. We construct 20  
 291 pseudo-stereo signals by assigning the  
 292 original audio to the left channel and  
 293 its additive inverse to the right. While  
 294 human listeners could easily perform  
 295 sound event classification on these sig-  
 296 nals, the models consistently failed due to signal cancellation during the mono conversion. The  
 297 result confirms their lack of explicit support for genuine stereo audio processing. To provide a com-  
 298 prehensive assessment, we adopt two complementary strategies. The first is the **native input** setting,  
 299 where the model directly processes stereo audio using its default pipeline. This allows us to probe  
 300 its intrinsic ability to exploit spatial cues. The second is the **channel-wise input** setting, where the  
 301 left and right channels are presented separately with explicit textual instructions, as shown in the  
 302 bottom right of Fig. 2. This configuration serves as an ablation study to examine whether current  
 303 models have any spatial capability when the binaural information is preserved at the input.

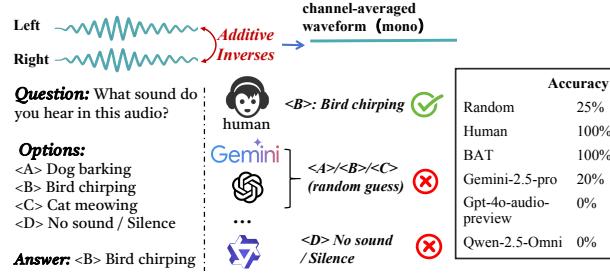
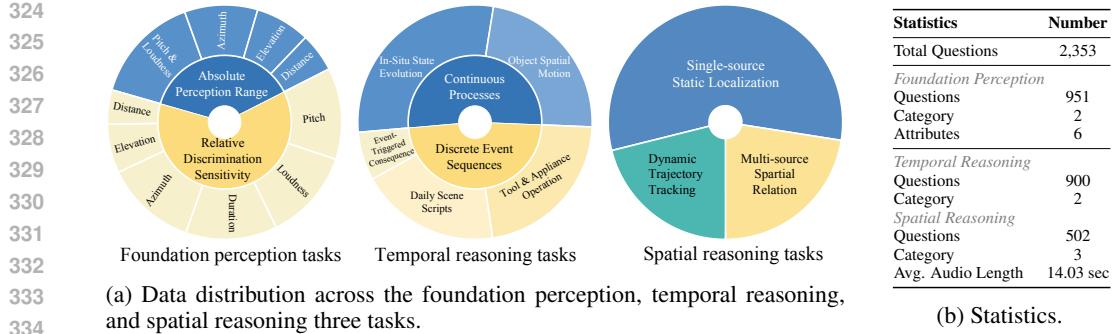


Figure 3: Audio preprocessing in existing models results in the loss of dual-channel information.

### 3.3 DATA CURATION PIPELINE

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

- 315 1) **Taxonomy Construction and Data Sourcing:** We build a hierarchical task taxonomy through a  
 316 collaborative process involving domain experts and the Gemini 2.5 Pro (Comanici et al., 2025). This  
 317 framework guides the sourcing of candidate data from large-scale, real-world audio libraries: Clotho  
 318 (Drossos et al., 2019) and FSD50K (Fonseca et al., 2022) for temporal reasoning, and STARSS23  
 319 (Shimada et al., 2023), along with audio sourced from the internet for spatial reasoning.
- 320 2) **AI-Assisted Automated Filtering:** This process employs an efficient three-stage funnel. First,  
 321 we discard unsuitable samples based on basic properties like duration and energy. Next, an LLM  
 322 (e.g., DeepSeek-V3 (Liu et al., 2024a)) performs an initial screening based on textual metadata,  
 323 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) Data distribution across the foundation perception, temporal reasoning, and spatial reasoning three tasks.

(b) Statistics.

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.

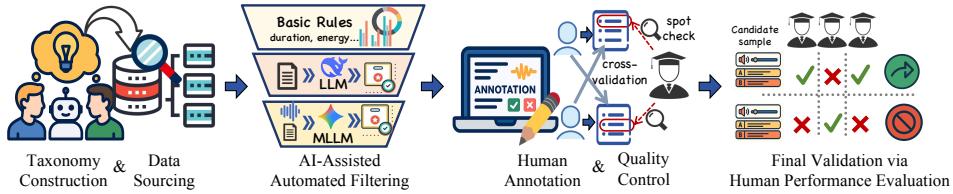


Figure 5: The four-stage **data annotation pipeline** for constructing our STAR-BENCH.

a judgment, a quality score, and a preliminary classification, further filtering irrelevant samples. The detailed prompts used to query the LLMs are provided in Appendix B.3.1.

**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. [More details are provided in Appendix B.3.2.](#)

**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 Appendix C. This approach yields two key metrics: **Average Accuracy (AA)**, the mean accuracy across all runs, and **All-Correct Rate (ACR)**, the

378  
379  
380  
381  
382  
Table 2: Evaluation results of various models on STAR-BENCH. The best performance is high-  
lighted in **bold**, and the second-best ones are underlined. MA (Macro Accuracy) denotes the un-  
weighted 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 Appendix 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
MiDashengLM [10]	7B	<u>36.94</u>	30.78	33.24	15.43	17.31	16.30	<b>43.11</b>	<u>45.43</u>	<b>46.23</b>	<b>44.29</b>
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-1.5 [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	31.02	31.87	31.53	15.36	17.39	16.30	<u>29.92</u>	43.36	38.36	34.73
GPT-4o Audio [2]	-	27.58	34.55	31.76	15.91	23.56	19.44	<u>41.81</u>	43.97	39.94	41.70
Gemini 2.5 Flash [8]	-	33.46	43.88	<u>39.72</u>	27.55	<u>34.38</u>	30.70	24.62	43.07	22.64	28.35
Gemini 2.5 Pro [8]	-	<b>39.90</b>	<b>51.13</b>	<b>46.64</b>	<b>54.88</b>	<b>62.74</b>	<b>58.52</b>	40.87	<b>48.97</b>	<b>45.28</b>	<b>43.62</b>

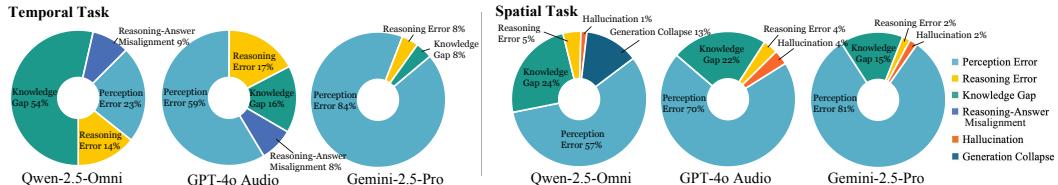


Figure 6: Error distribution across temporal and spatial Tasks.

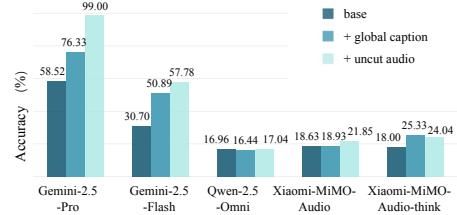
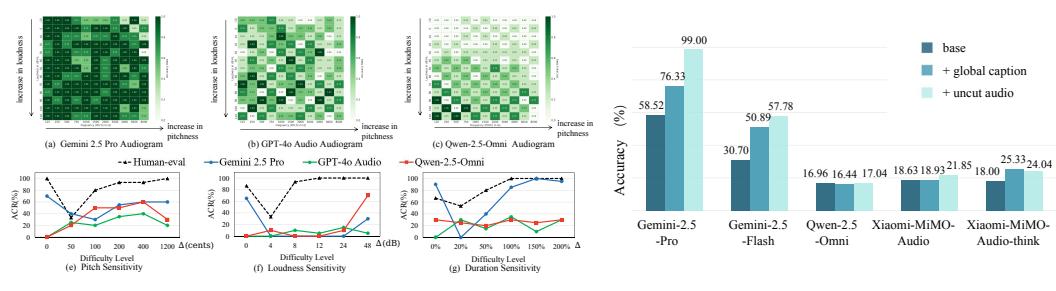
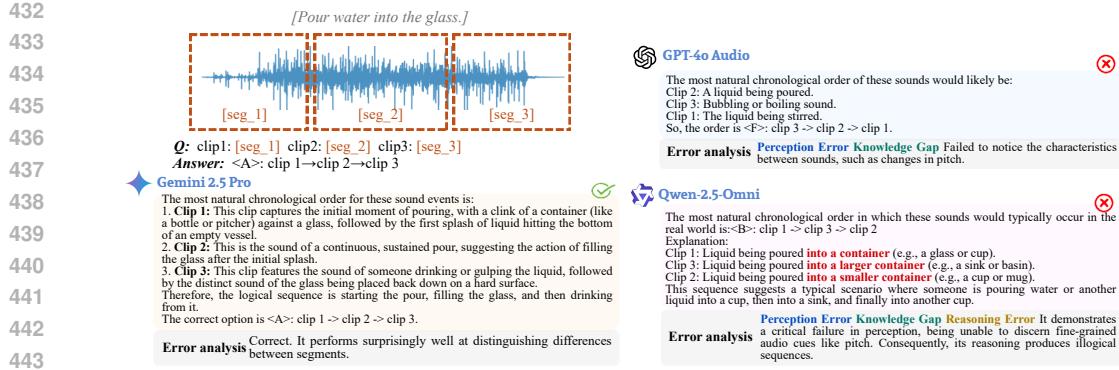
405 proportion of questions answered correctly in every run, which serves as a stronger indicator of  
406 model reliability. Due to space limitations, we primarily report AA in the main text, while complete  
407 experimental results are available in Appendix D.

#### 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 Appendix 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, 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 Appendix 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. **Meanwhile, this design also serves as a diagnostic lens on the limitations of current training pipelines.** The poor performance across models suggests that existing training paradigms often centered on clip-level tagging, QA, or captioning over linguistically salient cues (e.g., using FSD50K for sound event recognition) and do not equip models with the abilities needed for audio 4D intelligence.

**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, sug-



gesting 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 Appendix 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 substantial 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 Range and Sensitivity in Foundational Perception.** To quantify the gap in perceptual range and sensitivity, we provide detailed visualizations of model performance on our

486 foundational perception tasks in Fig. 8. The first row of Fig. 8 presents audiograms that compare  
 487 model coverage across the pitch-loudness space. Gemini 2.5 Pro achieves a much broader coverage  
 488 than the other two models, where greener regions indicate higher accuracy and the covered area  
 489 reflects the perceptual range. In contrast, human listeners with normal hearing are expected to  
 490 achieve near-full coverage, underscoring the gap between current models and human perceptual  
 491 abilities in terms of range. The second row of Fig. 8 further track the performance of both models  
 492 and human subjects on the three core acoustic attributes (pitch, loudness, and duration) as task  
 493 difficulty decreases. The results reveal a stark performance gap between all models and the human  
 494 baseline, particularly in the perception of fine-grained loudness differences. A clear trend is visible  
 495 even for the top-performing Gemini 2.5 Pro: its accuracy, while competent on easier tasks, plummets  
 496 as perceptual granularity increases. This directly corroborates our error analysis, identifying fine-  
 497 grained perception as its primary bottleneck. Notably, its performance on duration perception is  
 498 an exception, showcasing **temporal grounding capabilities superior to those of other models** by  
 499 accurately assessing audio segment lengths.

500 **Ablation Study on Temporal Reasoning.** To further pinpoint the specific limitations of temporal  
 501 reasoning, we augment the baseline audio segment reordering task with two progressively easier  
 502 settings: (1) + *Global Caption*, where a single sentence describing the overall scene is provided as a  
 503 contextual guide; and (2) + *Uncut Audio*, where the complete, unsegmented audio track is offered as  
 504 a reference, reducing the task to a straightforward process where the correct order can be determined  
 505 simply by comparing and grounding each segment within the full audio. As shown in Fig. 9, Gemini  
 506 2.5 Pro’s performance scales effectively with task simplification, culminating in a near-perfect 99%  
 507 accuracy in the + *Uncut Audio* setting. In contrast, the open-source models show minimal to no  
 508 improvement across these settings. Their performance remains stagnant even when provided with  
 509 the complete audio reference, despite the simplified nature of the task. This finding starkly exposes  
 510 a core weakness in current open-source models: **a fundamental inability to effectively compare, ground, and integrate information from multiple audio inputs.**

## 5 CONCLUSION

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

520  
 521  
 522  
 523  
 524  
 525  
 526  
 527  
 528  
 529  
 530  
 531  
 532  
 533  
 534  
 535  
 536  
 537  
 538  
 539

540 ETHICS STATEMENT  
541

542 Our study primarily relies on datasets from open-source research communities and publicly avail-  
543 able online resources, as described in detail in the main text. These datasets do not involve private  
544 information, sensitive content, or material that could raise concerns related to safety, discrimina-  
545 tion, or harmful societal impact. All annotation and evaluation tasks were carried out by university  
546 volunteers who participated on a voluntary basis. No human subjects were placed at risk, and no  
547 personally identifiable information was collected during the course of this research.

548  
549 REPRODUCIBILITY STATEMENT  
550

551 We provide a detailed description of the construction process of our benchmark dataset and evalua-  
552 tion pipeline in the main text. To facilitate reproducibility, we will release the benchmark dataset  
553 as well as the evaluation code to the community. Clear documentation and step-by-step instructions  
554 are included to ensure that other researchers can replicate our experiments and verify the reported  
555 results.

556  
557 REFERENCES  
558

559 Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin  
560 Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical  
561 report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint*  
562 *arXiv:2503.01743*, 2025.

563 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-  
564 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical  
565 report. *arXiv preprint arXiv:2303.08774*, 2023.

566 Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunx-  
567 iang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal  
568 model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.

569 Zheng Cai, Maosong Cao, Haojong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui  
570 Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye  
571 Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting  
572 Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li,  
573 Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun  
574 Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang  
575 Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song,  
576 Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang,  
577 Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong,  
578 Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia  
579 Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo  
580 Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui  
581 Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou,  
582 Xipeng Qiu, Yu Qiao, and Dahu Lin. Internlm2 technical report, 2024.

583 Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking  
584 video-llms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*, 2025.

585 Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and  
586 Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale  
587 audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

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

591 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit  
592 Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the

594 frontier with advanced reasoning, multimodality, long context, and next generation agentic capa-  
 595 bilities. *arXiv preprint arXiv:2507.06261*, 2025.

596

597 Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan  
 598 Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction  
 599 following, 2025. URL <https://arxiv.org/abs/2504.07957>.

600

601 Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang  
 602 Xiao, Junbo Zhang, and Jiahao Zhou. Midashenglm: Efficient audio understanding with general  
 603 audio captions. *arXiv preprint arXiv:2508.03983*, 2025.

604

605 Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset,  
 606 2019. URL <https://arxiv.org/abs/1910.09387>.

607

608 Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: an open  
 609 dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language  
 Processing*, 30:829–852, 2022.

610

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

614

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

619

620 Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. Joint audio and  
 621 speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Work-  
 shop (ASRU)*, pp. 1–8, 2023. doi: 10.1109/ASRU57964.2023.10389742.

622

623 Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-  
 624 Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, et al. Dynamic-superb phase-2: A collab-  
 625 oratively expanding benchmark for measuring the capabilities of spoken language models with  
 626 180 tasks. In *The Thirteenth International Conference on Learning Representations*, 2025.

627

628 Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bam-  
 629 ford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al.  
 630 Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

631

632 KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen,  
 633 Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu,  
 634 Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing  
 635 Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang,  
 636 Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin,  
 637 Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report, 2025. URL <https://arxiv.org/abs/2504.18425>.

638

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

642

643 Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand,  
 644 Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. Mmau-pro: A chal-  
 645 lenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv  
 preprint arXiv:2508.13992*, 2025.

646

647 Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jiaqi Wang, and Dahua Lin. Beyond fixed:  
 648 Training-free variable-length denoising for diffusion large language models, 2025. URL <https://arxiv.org/abs/2508.00819>.

648 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,  
 649 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint*  
 650 *arXiv:2412.19437*, 2024a.

651

652 Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan,  
 653 Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal  
 654 model an all-around player?, 2024b. URL <https://arxiv.org/abs/2307.06281>.

655 Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuan-  
 656 jun Xiong, Dahua Lin, and Jiaqi Wang. Mia-dpo: Multi-image augmented direct preference opti-  
 657 mization for large vision-language models, 2024c. URL <https://arxiv.org/abs/2410.17637>.

658

659 Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan,  
 660 Dahua Lin, and Jiaqi Wang. Visual agentic reinforcement fine-tuning, 2025. URL <https://arxiv.org/abs/2505.14246>.

661

662 Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang,  
 663 Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, et al. Desta2. 5-audio: Toward  
 664 general-purpose large audio language model with self-generated cross-modal alignment. *arXiv*  
 665 *preprint arXiv:2507.02768*, 2025.

666

667 Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen,  
 668 Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. Mmar: A challenging benchmark for deep rea-  
 669 soning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*, 2025.

670

671 Sudipta Paul, Amit Roy-Chowdhury, and Anoop Cherian. AVLEN: Audio-visual-language embed-  
 672 ed navigation in 3d environments. In *NeurIPS*, 2022.

673

674 Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. Vln-r1: Vision-  
 675 language navigation via reinforcement fine-tuning, 2025. URL <https://arxiv.org/abs/2506.17221>.

676

677 S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ra-  
 678 mani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio  
 679 understanding and reasoning benchmark. In *The Thirteenth International Conference on Learn-  
 680 ing Representations*, 2025.

681

682 Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio  
 683 room simulation and array processing algorithms. In *2018 IEEE international conference on  
 684 acoustics, speech and signal processing (ICASSP)*, pp. 351–355. IEEE, 2018.

685

686 Kazuki Shimada, Archontis Politis, Parthasarathy Sudarsanam, Daniel A Krause, Kengo Uchida,  
 687 Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, et al.  
 688 Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal anno-  
 689 tations of sound events. *Advances in neural information processing systems*, 36:72931–72957,  
 690 2023.

691

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

695

696 Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej,  
 697 Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical  
 698 report. *arXiv preprint arXiv:2503.19786*, 2025.

699

700 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée  
 701 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and  
 702 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

703

704 Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi  
 705 Aw, and Nancy F Chen. Audiobench: A universal benchmark for audio large language models.  
 706 *arXiv preprint arXiv:2406.16020*, 2024.

702 Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and  
 703 Dahua Lin. Sim-cot: Supervised implicit chain-of-thought, 2025a. URL <https://arxiv.org/abs/2509.20317>.

704

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

708

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

724

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

727

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

731

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

735

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

738

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

742

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

746

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

751

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

755

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

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

759

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

763

764

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

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810 THE USE OF LARGE LANGUAGE MODELS  
811812 We used Gemini-2.5-Pro to assist in expanding and consolidating the taxonomy of tasks in our  
813 benchmark. Both DeepSeek-V3 and Gemini-2.5-Pro were utilized for the automated pre-screening  
814 of candidate data. The final task definitions and data samples are verified by humans. We also used  
815 GPT-4o to generate some of the illustrative figures presented in the paper, and used GPT-5 to polish  
816 the manuscript text. Only human-verified revisions are included in the final version.  
817818  
819 A RELATED WORK  
820821 A.1 AUDIO LANGUAGE MODELS  
822823 With the advancements of large language models (LLMs) and multimodal language models (Yang  
824 et al., 2025a; Jiang et al., 2024; Achiam et al., 2023; Comanici et al., 2025; Cai et al., 2024; Touvron  
825 et al., 2023; Liu et al., 2024c; 2025; Zhang et al., 2025b; Qi et al., 2025; Xing et al., 2025b;a;  
826 Ding et al., 2025; Wei et al., 2025a;b; Li et al., 2025; Zhang et al., 2025a), recent research has  
827 increasingly focused on integrating audio perception with LLMs to enhance audio understanding and  
828 reasoning. Existing methods can be broadly grouped into two categories: Large Audio Language  
829 Models(LALMs) and Omni Language Models(OLMs).  
830831 Most LALMs combine a pre-trained audio encoder with an LLM backbone, where the two modalities  
832 are aligned via large-scale text-audio joint training. Notable models include LTU-AS (Gong  
833 et al., 2023), SALMONN (Tang et al., 2024), MidashengLM (Dinkel et al., 2025), Audio Flamingo  
834 series (Ghosh et al., 2025; Goel et al., 2025), Qwen-Audio series (Chu et al., 2023; 2024), Step-  
835 Audio (Wu et al., 2025) and Mimo-Audio (Xiaomi, 2025). These models have achieved remarkable  
836 performance across a wide range of audio understanding tasks, including automatic speech recogni-  
837 tion(ASR), spoken question answering(SpokenQA), and automated audio captioning(AAC). In  
838 parallel, OLMs extend this paradigm to unify multimodal understanding with representative exam-  
839 ples such as Qwen-2.5-Omni (Xu et al., 2025), Ming-Omni (AI et al., 2025),MiniCPM-O (Yao et al.,  
840 2024), Phi-4 (Abouelenin et al., 2025), GPT-4o (Achiam et al., 2023), and Gemini 2.5 (Comanici  
841 et al., 2025). Notably, they also achieve impressive performance on audio understanding and reason-  
842 ing, highlighting their potential to bridge multimodal perception and advanced audio intelligence.  
843844 A.2 AUDIO BENCHMARKS  
845846 Existing audio benchmarks illustrate the rapid progress of multimodal evaluation but also expose  
847 limitations. AudioBench (Wang et al., 2024) and AIR-Bench (Yang et al., 2024) primarily focus  
848 on tasks such as automatic speech recognition (ASR), spoken question answering (SpokenQA),  
849 and audio captioning (AAC). These settings tend to reduce audio understanding to transcription or  
850 description, thereby neglecting the broader spectrum of acoustic reasoning. MMAU (Sakshi et al.,  
851 2025) and MMAR (Ma et al., 2025) further extend the evaluation scope. However, their results reveal  
852 an inherent weakness—LLMs equipped with audio captions can perform on par with advanced  
853 LALMs, suggesting that such benchmarks still probe little beyond language-level semantics.  
854855 Although some advanced benchmarks, such as MMAR (Ma et al., 2025) and MMAU-Pro (Kumar  
856 et al., 2025), do touch upon spatio-temporal aspects, their coverage remains limited in both scale and  
857 depth. For instance, their temporal analysis is typically reduced to identifying the timing or order  
858 of events occurring in the audio, while spatial analysis is often limited to localizing a single sound  
859 source. In contrast, our benchmark systematically evaluates models’ temporal and spatial deep  
860 reasoning capabilities within complex, real-world physical contexts, requiring them to infer causal  
861 and dynamic relationships. Beyond audio benchmarks, multimodal benchmarks in video question  
862 answering (Cheng et al., 2025; Yang et al., 2025c) and embodied AI (Yang et al., 2025b) have em-  
863 phasized temporal and spatial reasoning. However, these frameworks are predominantly grounded  
864 in the visual modality, where exploration of the audio modality remains comparatively limited. In  
865 real-world scenarios, audio understanding often depends on integrating information across multiple  
866 sound streams and reasoning about subtle changes in intensity, phase, or frequency—capabilities  
867 that existing benchmarks scarcely capture.  
868

Attribute	Range / Level	Example
<b>Absolute Perception Range</b>		
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
<b>Relative Discrimination Sensitivity</b>		
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.

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.

## 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 (4.0×3.5×2.8 m), medium (8.0×6.0×3.5 m), and large (20×15×8 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 HRTF<sup>1</sup> (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 × 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

<sup>1</sup>[https://sofacoustics.org/data/database/mit/mit\\_kemar\\_normal\\_pinna.sofa](https://sofacoustics.org/data/database/mit/mit_kemar_normal_pinna.sofa)

918 late tail by 0.5. We then applied global peak normalization across the batch to avoid clipping while  
 919 preserving inter-position level differences.  
 920

921 We discretized each attribute into fixed partitions to control dataset balance.  
 922

923 **Absolute azimuth:** Eight angles  $\{30^\circ, 60^\circ, 120^\circ, 150^\circ, 210^\circ, 240^\circ, 300^\circ, 330^\circ\}$ . For each angle  
 924 we rendered all combinations of 3 rooms  $\times$  2 listener positions  $\times$  2 source clips, yielding  $8 \times (3 \times$   
 925  $2 \times 2) = 96$  utterances. **Absolute elevation:** Six angles  $\{-75^\circ, -45^\circ, -15^\circ, 15^\circ, 45^\circ, 75^\circ\}$ . Per  
 926 angle we rendered 3 rooms  $\times$  2 listener positions  $\times$  2 source clips, for  $6 \times (3 \times 2 \times 2) = 72$  utterances.  
 927 **Absolute distance:** Radii from 1–10 m with a nonuniform allocation to emphasize near-field cues:  
 928 for 1–7 m we generated 6 utterances per meter (42 total), and for 8–10 m we generated 3 per meter  
 929 (9 total), giving  $42 + 9 = 51$  utterances per (room  $\times$  listener) set.  
 930

931 **Relative azimuth:** Differences were multiples of  $30^\circ$ :  $\{30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ, 180^\circ\}$  (6  
 932 levels), totaling  $6 \times 20 = 120$  utterances. **Relative elevation:** Four difference angles  
 933  $\{15^\circ, 90^\circ, 120^\circ, 150^\circ\}$  with 18, 17, 17, 12 utterances respectively (64 total). **Relative distance:**  
 934 Four difference levels  $\{1-2, 4-5, 6-7, 8-9\}$  m with counts per level  $\{12, 12, 12, 9\}$ , totaling  
 935 45 utterances.  
 936

### 937 B.3 DETAILS OF THE CURATION PROCESS FOR REASONING TASKS

#### 938 B.3.1 PROMPT USED FOR AI-ASSISTED FILTERING OF TEMPORAL TASK DATA

939 Fig. 10 and Fig. 11 present our carefully designed prompts, which leverage Gemini 2.5 Pro to filter  
 940 candidate data that meet the requirements of audio segment reordering. **Briefly, we feed the audio,**  
 941 **its metadata, and our task description, and ask Gemini 2.5 Pro to decide, under our strict criteria**  
 942 **of strong sequence uniqueness, semantic clarity, and high logical universality, (i) whether the audio**  
 943 **is suitable for a reordering task, (ii) whether it reflects a continuous or discrete process, (iii) the**  
 944 **reasoning behind its judgment, and (iv) a quality score.** We adopt a conservative filtering strategy,  
 945 discarding only samples explicitly marked as “not applicable”. All remaining clips, along with the  
 946 model’s analysis, are then passed to professional annotators for verification and annotation. A prior  
 947 LLM-based filtering step follows a similar procedure, but without audio input.  
 948

#### 949 B.3.2 DETAILS OF HUMAN ANNOTATION AND QUALITY CONTROL

950 Following automated filtering, each candidate sample undergoes a rigorous, multi-stage human an-  
 951 notation and quality control process to ensure high data quality and annotation consistency. This  
 952 process is as follows:

953 (1) **Systematic Training:** All annotators received detailed written guidelines and completed a trial  
 954 annotation of 10 samples. These trials were meticulously reviewed by experts to ensure a unified  
 955 understanding of the criteria.  
 956

957 (2) **Inter-annotator Cross-validation:**

958 (i) *Initial Annotation:* A sample is first annotated by Annotator A. The annotation content  
 959 includes:

- 960 • For Temporal Reasoning: Task compliance checks, segment boundary delineation,  
 961 textual descriptions for sub-clips and the global audio, scene classification, and audio  
 962 quality scoring.
- 963 • For Spatial Reasoning: Selecting appropriate segments, task classification, and the  
 964 generation of a question, the correct answer, and distractor options for the multiple-  
 965 choice format.

966 (ii) *Review and Flagging:* The annotated sample is then fully reviewed by Annotator B, who  
 967 flags any inconsistencies with detailed comments and marks the sample as “failed”.  
 968

969 (iii) *Consensus through Negotiation:* Annotators A and B then discuss all flagged issues to  
 970 reach a consensus and apply corrections. During the discussions, primary sources of  
 971 ambiguity are as follows:

- 972 • For Temporal Reasoning: (a) The reasonableness of the segmented clip boundaries.  
 973 (b) The existence of multiple logically plausible orderings for the segmented clips.  
 974 (c) Significant discrepancies in audio quality scores.(d) Adherence to formatting and  
 975 content guidelines for the captions.

972                   • For Spatial Reasoning: (a) Whether the spatial perception presented in the audio un-  
 973                   ambiguously aligns with the annotated answer. (b) Whether the constructed question-  
 974                   answer pair clearly necessitates the use of audio spatial cues for resolution. (c) Po-  
 975                   tential ambiguity in mapping the event name mentioned in the question to a specific  
 976                   sound in the audio. (d) The appropriate difficulty and plausibility of the distractor  
 977                   options.

978                   (iv) *Expert Arbitration*: In the cases where a consensus cannot be reached, the sample is  
 979                   escalated to an expert panel for a final decision. If the experts still cannot agree, the  
 980                   sample is discarded.

981                   (3) **Expert Spot-check:** After passing cross-validation, a random 10% of samples undergo a final  
 982                   quality check by experts to ensure consistency and accuracy. Any discovered issues are then  
 983                   sent back for revision.

984

985

986

987                   **# Role Setting:**

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

991                   **# “Audio Clip Ordering” Evaluation Task:**

992                   Qualified audio is segmented into three clips, shuffled, and given to the model, which must reconstruct the  
 993                   sequence using only sound. Candidate audio must meet the following strict standards:

994                   1. **Strong sequence uniqueness:** The events in the audio must present a unique and clearly discernible temporal  
 995                   progression, with no possibility of alternative plausible orderings.

996                   2. **Semantic clarity:** Events in the audio must be easily identifiable by sound alone.

997                   3. **High logical universality:** The event sequence should conform to commonsense physical laws or strong causal  
 998                   relations, such that listeners from different backgrounds can reach a consistent understanding.

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

1000                   **# Audio Classification Standards:**

1001                   1. **Sortable Single Event:**

1002                   - Definition: Audio primarily represents a continuous event driven by a single process, exhibiting significant and  
 1003                   predictable temporal dynamics.

1004                   - Possible categories include (but are not limited to):

1005                   - Spatial movement and distance variation. Physical processes in progression. Energy or state decay. Biological  
 1006                   activity dynamics.

1007                   - Core judgment: The core judgment is whether the change is governed by a single continuous physical process,  
 1008                   is significant, and is commonly recognized as unambiguous.

1009                   - Examples:

1010                   - The sound of pouring water. A ball bouncing to rest.

1011                   2. **Sortable Multi-Event:**

1012                   - Definition: Audio contains two or more independent events, where the events exhibit a strong causal  
 1013                   relationship such that “A inevitably leads to B” or “A must precede B.”

1014                   - Core judgment: The core judgment is whether sub-events have distinct, separable acoustic features, follow a  
 1015                   direct and widely recognized causal chain, and together form a concise, complete, and unambiguous process.

1016                   - Examples:

1017                   - Opening a bottle → pouring water → setting down the bottle.

1018                   - Cracking an egg → stirring → pouring into an oil pan.

1019                   **Note:** If event boundaries are unclear but the overall structure forms a continuous process, classify as “Single  
 1020                   Event.” If there are clearly distinct stages with evident logical links, classify as “Multi-Event.”

1021                   **#Special Exclusion Rules:**

1022                   Filter out the following audio types, even if they exhibit some “dynamic change.” If they fail semantic clarity,  
 1023                   sequence uniqueness, or logical universality, they must be labeled “Not Applicable.”

1024                   - Static or repetitive sounds.

1025                   - Sounds with intensity variation but without reflecting a physical law or typical action sequence.

1026                   - Weak or absent causality, or dependent on special background knowledge.

1027                   - Ambiguous, subjective, or interpretable in multiple ways.

1028                   - Ambient noise or atmosphere shifts without structured events.

1029                   - Dialogue or speech as main content.

1030                   - Poor quality or indistinct segmentation.

1031                   **# Features of High-Quality Samples:**

1032                   - Clear event semantics: Each action or change is acoustically distinct and intuitively understandable.

1033                   - Unambiguous sequence: Temporal logic is indisputable, allowing only one reconstruction order.

1034                   - Narrativity and processuality: The audio presents a complete, coherent event chain consistent with human  
 1035                   understanding and commonsense reasoning.

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

1026  
1027 **# Input**  
1028 You will receive the following three components:  
1029 1.Text description: Labels, annotations, or metadata about the audio.  
1030 2.Preliminary judgment based on text: A text-only hypothesis about the audio's attributes.  
1031 3.Audio file.  
1032 **Note:** Always base your classification primarily on the acoustic features of the audio itself, using text only as  
1033 supplementary context.

1034 **#Output Format (Strictly follow the four-line structure)**  
1035 **Classification label:** Sortable Single Event | Sortable Multi-Event | Uncertain | Not Applicable  
1036 **Event description:** For “Sortable Multi-Event,” use the format Event A → Event B → Event C to describe the  
1037 strong causal chain. For “Single Event,” describe the core dynamic change. If “Not Applicable”, briefly  
1038 summarize the content.  
1039 **Reasoning:** Explain classification basis, focusing on the three standards: semantic clarity, sequence uniqueness,  
1040 and logical universality. Refer to specific acoustic features where possible.  
1041 **Score:** (Composite evaluation based on the three standards)  
1042 - 5 points: Events clear, sequence strong, logic robust, full structure or rich narrative (≥4 events).  
1043 - 4 points: Sequence clear, physics/logic reasonable, but semantics somewhat limited (≤3 events).  
1044 - 3 points: Some ambiguity, human listeners may also struggle to interpret  
1045 - 2 points: Content vague, minimal sequential logic, not recommended for ordering.  
1046 - 1 point: Events unclear, structure absent, unusable for ordering tasks.

1047 **# Examples:**  
1048 **Input:**  
1049 Text description: “A ping-pong ball landing.”  
1050 Preliminary judgment: –  
1051 Audio file: Sharp bouncing sounds, gradually weakening, intervals shortening, finally stopping  
1052 **Output:**  
1053 Classification label: Sortable Single Event  
1054 Event description: Ping-pong ball bouncing, sound intensity and interval decay until stop.  
1055 Reasoning: Clear physical energy decay, unique order, limited semantics.  
1056 Score: 4

1057 **Input:**  
1058 Text description: “People counting down for New Year, then celebrating with fireworks.”  
1059 Preliminary judgment: –  
1060 Audio file: 10-second countdown → fireworks → cheering  
1061 **Output:**  
1062 Classification label: Sortable Multi-Event  
1063 Event description: Countdown (10s) → Fireworks → Cheering  
1064 Reasoning: Three distinct events, unique order, clear logic, strong semantics.  
Score: 5

1065 **Input:**  
1066 Text description: “A segment of continuous machine rumble.”  
1067 Preliminary judgment: –  
1068 Audio file: Indistinct machine operating noise  
1069 **Output:**  
1070 Classification label: Not Applicable  
1071 Event description: Continuous monotonous machine rumble without structural change.  
1072 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

1070 All questions in STAR-BENCH are presented as clear multiple-choice questions with well-formatted  
1071 options. We adopt classification accuracy as the evaluation metric. To determine the correctness of  
1072 a response, we employ string matching to extract either the chosen option label (e.g., <A>) or the  
1073 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

1080 mean accuracy across all evaluation runs, and All-Correct Rate (ACR), the proportion of questions  
 1081 answered correctly in every single run, which serves as a stronger indicator of model reliability.  
 1082

1083 For models that do not support multi-audio input (only Audio Flamingo 3 and its Think variant  
 1084 among the models we evaluated), we concatenate the audios with a 2-second silence and specify this  
 1085 in the prompt. In contrast, for models that support multiple audio inputs, we feed them sequentially  
 1086 with textual indices.

1087 To establish a human performance baseline, we conduct a human evaluation on a randomly sam-  
 1088 pled subset of approximately 10% of the data from each task. This evaluation is performed by 10  
 1089 university students, from whom we explicitly exclude anyone involved in data annotation or with  
 1090 domain-specific expertise, thereby ensuring a general, non-expert perspective.

## 1092 D BREAKDOWN RESULTS

1093 In this section, we present detailed results for perception, temporal reasoning, and spatial reasoning  
 1094 on STAR-BENCH, as shown in Tab. 4, Tab. 5, and Tab. 6.

## 1097 E FURTHER ANALYSIS AND DISCUSSION

### 1099 E.1 HIGH OUTPUT INSTABILITY AND CONCENTRATED PREDICTIONS

1100 The reliability of model outputs on our benchmark is notably low, as evidenced by the stark contrast  
 1101 between their Average Accuracy (AA) and All-Correct-Rate (ACR) scores. Even the top-performing  
 1102 model, Gemini 2.5 Pro, exhibits an average drop of 25.01 percentage points from its AA to its ACR.  
 1103 This issue is even more pronounced for the majority of open-source models, which record an ACR  
 1104 near zero. This score indicates a complete failure to maintain consistent predictions under minor  
 1105 input perturbations. For these models, the instability often manifests as a tendency to concentrate  
 1106 predictions on a specific option, suggesting a reliance on superficial biases rather than genuine un-  
 1107 derstanding.

### 1109 E.2 ABLATION STUDY ON SPATIAL REASONING.

1110 As shown in Tab. 6, the results reveal a fundamental limitation of LALMs’ in spatial perception. The  
 1111 **native input** inherently discards part of the multi-channel information during model preprocessing,  
 1112 which leads to a significant loss of spatial cues that are essential for fine-grained reasoning. On  
 1113 the other hand, the **channel-wise input** explicitly presents each channel with textual instructions,  
 1114 mitigating some of the information loss. Despite this, most existing models are not trained to handle  
 1115 multi-audio inputs. As a result, they consistently struggle to align channel representations and fail to  
 1116 make reliable use of interaural differences. Overall, the pronounced gap between human and model  
 1117 performance highlights that spatial reasoning in audio remains an unsolved challenge, underscoring  
 1118 the need for audio encoders that natively support multi-channel audio input.

## 1120 F CASE STUDY

1121 In this section, we present several case studies of error analysis, including temporal reasoning  
 1122 (Figs. 12 to 17) and spatial reasoning (Fig. 18).

1134

1135 Table 4: Results for the foundational perception task. Each cell reports AA / ACR: Average Accu-  
1136 racy (AA; overall accuracy across all runs) / All-Correct Rate (ACR; proportion of samples that are  
1137 correct on every run). The best model in each category is shown in **bold**, and the second best is  
1138 underlined.

1139

Model	Size	Absolute Perception Range				Relative Discrimination Sensitivity					MA (%)
		Pitch&Loudbness	Azimuth	Elevation	Distance	Pitch	Loudbness	Duration	Azimuth	Elevation	
Random Guess	—	25.00/0.39	20.00/0.03	25.00/0.39	25.00/0.39	25.00/0.39	25.00/0.39	25.00/0.39	33.33/3.7	25.00/0.39	25.33/0.68
Human	—	98.67/—	73.33/—	66.67/—	70.00/—	83.33/—	85.56/—	83.33/—	83.33/—	38.09/—	73.68/—
SALMONN	13B	14.34/0.00	25.83/0.63	35.76/0.00	33.33/0.00	31.04/0.00	25.00/0.00	28.54/0.00	31.39/3.89	24.15/0.00	12.77/0.00
Audio Flamingo 3	8.4B	37.59/0.00	<b>27.92</b> /3.13	28.82/0.00	32.84/0.00	42.50/1.67	28.96/0.00	34.79/0.00	38.61/6.67	<b>33.90</b> /0.00	35.56/0.00
Audio Flamingo 3 think	8.4B	51.75/6.99	8.75/0.00	33.33/1.04	8.33/0.00	36.04/8.33	<b>45.63</b> /2.50	59.38/38.33	<u>41.11</u> /4.17	12.29/0.00	10.00/0.00
Qwen2-Audio-Instruct	8.4B	35.66/1.40	22.50/0.00	<b>48.61</b> /10.76	12.75/0.98	35.63/0.00	16.25/0.00	26.46/0.00	35.00/8.06	21.61/1.69	23.88/0.00
DeSTa2.5-Audio	8.8B	16.96/0.00	21.25/0.42	45.49/1.39	35.78/1.74	11.67/0.00	11.25/0.00	22.71/0.00	33.06/7.78	10.59/0.00	29.44/0.00
BAT	7B	0.00/0.00	26.04/26.06	41.67/41.67	23.53/23.53	0.00/0.00	0.00/0.00	37.50/37.50	0.00/0.00	0.00/0.00	12.87/12.87
Ph4-MM	5.5B	9.44/0.00	24.17/0.00	15.97/0.00	26.96/0.00	24.38/0.00	30.00/0.00	27.92/0.00	36.94/0.00	32.62/0.00	27.22/0.00
Kimi-Audio	7B	18.71/0.00	18.12/0.00	38.19/0.00	18.13/0.00	24.38/0.00	32.29/0.00	34.17/0.83	39.72/3.89	25.00/0.85	9.44/0.00
MidDashengLM	7B	<b>48.95</b> /33.57	20.63/0.00	<b>48.26</b> /11.81	29.90/0.98	40.00/34.17	17.08/0.83	23.54/7.50	34.72/8.61	27.12/1.69	42.22/6.11
Step-Audio-2-mini	7B	37.32/0.00	20.00/0.00	31.60/0.69	29.41/0.00	25.00/0.00	29.17/0.00	20.00/0.00	31.36/0.00	25.00/0.00	28.14/0.07
Gemma-3n-E4B-it	7.5B	7.18/0.00	24.38/4.17	21.00/0.00	17.00/0.00	30.95/0.00	30.75/0.00	15.00/5.82	40.56/0.00	24.73/0.00	23.31/0.00
Ming-Lite-Omni-1.5	18.9B	28.67/0.00	20.21/0.00	27.78/0.00	30.39/3.92	16.57/16.67	16.67/16.67	17.00/16.67	<b>41.07</b> /0.28	32.00/0.00	36.11/1.00
Qwen-2.5-Omni	7B	27.45/3.59	18.33/0.21	17.57/1.47	<b>41.67</b> /1.47	48.13/35.00	39.79/15.00	38.33/26.67	16.11/0.28	11.02/0.00	40.56/2.78
Xiaomi-MiMo-Audio	7B	36.71/5.59	18.54/19.17	48.26/3.82	36.27/2.94	46.04/24.17	36.46/0.83	17.70/16.67	40.56/2.22	20.98/0.00	27.78/1.67
Xiaomi-MiMo-Audio-think	7B	43.01/14.69	11.67/0.00	25.69/0.00	39.21/4.90	28.13/3.33	15.21/1.67	22.17/1.67	29.44/2.50	21.88/0.45	32.22/1.67
MiniCPM-O-v2.6	8B	46.33/8.39	24.58/0.21	23.26/0.35	29.90/0.00	38.13/3.33	38.96/4.17	32.08/3.33	37.22/2.78	22.10/0.22	22.78/0.00
GPT-4o Audio	—	<b>45.28</b> /—	16.67/—	44.44/—	3.92/—	43.33/—	36.04/—	46.46/—	29.58/—	11.86/—	40.00/—
Gemini 2.5 Flash	—	<b>62.59</b> /18.19	12.50/0.00	18.06/0.35	40.69/1.47	<b>48.54</b> /21.67	<b>40.83</b> /6.67	<b>63.13</b> /27.50	37.08/9.17	25.42/0.85	<b>48.33</b> /4.44
Gemini 2.5 Pro	—	<b>86.71</b> /62.94	25.83/1.25	5.88/0.00	41.18/5.88	<b>63.33</b> /52.50	33.75/15.83	<b>78.96</b> /68.33	37.08/13.75	29.24/6.36	<b>64.44</b> /12.22

1150

1151

1152

1153 Table 5: Results for the temporal reasoning task. Each cell reports AA / ACR: Average Accuracy  
1154 (AA; overall accuracy across all runs) / All-Correct Rate (ACR; proportion of samples that are  
1155 correct on every run). The best model in each category is shown in **bold**, and the second best is  
1156 underlined.

1157

Model	Size	Continuous Processes		Discrete Event Sequences			OA (%)
		Object Spatial Motion	In-Situ State Evolution	Tool & Appliance Operation	Daily Scene Scripts	Event-Triggered Consequences	
Random Guess	—	14.29/0.00	14.29/0.00	14.29/0.00	14.29/0.00	14.29/0.00	14.29/0.00
Human	—	91.11/—	88.89/—	87.88/—	83.33/—	83.33/—	88.00/—
SALMONN	13B	13.88/0.74	16.12/0.00	13.56/1.96	13.15/1.11	12.50/0.00	14.15/0.89
Audio Flamingo 3	8.4B	8.55/0.00	10.08/0.47	8.66/0.98	7.22/1.11	8.33/3.13	8.67/0.67
Audio Flamingo 3 think	8.4B	14.37/0.00	11.78/0.93	15.36/1.47	12.96/2.22	11.46/0.00	13.59/1.00
Qwen2-Audio-Instruct	8.4B	12.89/0.00	13.80/0.93	12.09/0.00	12.22/1.11	11.46/0.00	12.74/0.44
DeSTa2.5-Audio	8.8B	16.98/0.37	15.97/1.40	19.93/1.47	15.56/0.56	11.46/0.00	16.93/0.89
BAT	7B	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00
Ph4-MM	5.5B	17.72/0.00	15.50/0.47	16.34/0.98	17.04/3.89	20.83/3.13	16.85/1.22
Kimi-Audio	7B	18.71/1.49	21.55/2.33	18.63/0.49	15.19/2.22	14.58/0.00	18.52/1.56
MidDashengLM	7B	17.10/0.37	13.33/0.00	17.16/1.96	16.67/2.22	21.88/0.00	16.30/1.00
Step-Audio-2-mini	7B	16.11/0.37	14.42/0.00	15.52/0.00	16.30/0.00	15.63/0.00	15.59/0.11
Gemma-3n-E4B-it	7.5B	17.10/0.00	16.59/0.00	17.81/0.00	13.70/0.00	20.83/0.00	16.59/0.00
Ming-Lite-Omni-1.5	18.9B	17.47/1.12	16.59/0.47	13.89/0.00	17.59/1.11	14.58/0.00	16.37/0.67
Qwen-2.5-Omni	7B	17.10/0.37	15.35/0.93	19.77/1.47	16.48/0.56	11.46/0.00	16.96/0.78
Xiaomi-MiMo-Audio	7B	18.22/0.00	18.14/0.47	17.16/0.98	20.19/2.22	26.04/3.13	18.63/0.89
Xiaomi-MiMo-Audio-think	7B	16.36/0.37	17.36/0.47	19.93/1.96	18.70/2.22	19.79/0.00	18.00/1.11
MiniCPM-O-v2.6	8B	16.23/0.00	14.26/0.93	17.48/0.49	17.78/0.56	14.58/0.00	16.30/0.44
GPT-4o Audio	—	15.61/—	16.28/—	24.02/—	22.78/—	25.00/—	19.44/—
Gemini 2.5 Flash	—	30.86/3.35	<b>23.41</b> /3.72	<b>38.07</b> /12.75	30.19/7.22	<b>34.38</b> /9.38	30.70/6.56
Gemini 2.5 Pro	—	<b>63.82</b> /38.66	<b>43.72</b> /17.67	<b>69.77</b> /46.08	<b>57.22</b> /38.33	<b>48.96</b> /28.13	<b>58.52</b> /34.89

1169

1170

1171

1172 Table 6: Results for the spatial reasoning task using native and channel-wise audio input. Each  
1173 cell reports AA / ACR: Average Accuracy (AA; overall accuracy across all runs) / All-Correct Rate  
1174 (ACR; proportion of samples that are correct on every run). The best model in each category is  
1175 shown in **bold**, and the second best is underlined.

1176

Model	Size	Single-Source Static Localization		Multi-Source Spatial Relation		Dynamic Trajectory Tracking		OA (%)
		Native Input	Channel-wise Input	Native Input	Channel-wise Input	Native Input	Channel-wise Input	
Random Guess	—	33.33/3.70	—	33.33/3.70	—	33.33/3.70	—	33.33/3.70
Human	—	70.00/—	—	80.00/—	—	77.00/—	—	73.72/—
SALMONN	13B	26.15/3.18	26.62/3.18	28.61/4.42	29.50/5.31	39.94/0.94	38.36/0.94	29.62/2.99
Audio Flamingo 3	8.4B	37.22/1.77	<b>42.87</b> /2.12	38.35/4.42	46.31/10.62	44.03/4.72	<b>46.23</b> /0.94	38.91/2.99
Audio Flamingo 3 think	8.4B	35.45/7.42	<b>42.87</b> /13.78	37.46/23.01	46.02/23.01	38.05/18.87	<b>37.11</b> /19.81	36.45/13.35
Qwen2-Audio-Instruct	8.4B	21.32/8.48	6.36/1.77	24.78/3.54	12.09/4.42	15.09/0.94	11.64/2.83	20.78/5.78
DeSTa2.5-Audio	8.8B	23.67/2.83	20.38/4.59	34.81/9.73	41.30/19.47	37.74/10.38	32.08/21.70	29.15/5.98
BAT	7B	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00
Ph4-MM	5.5B	33.10/0.35	32.63/0.35	27.14/0.88	29.79/0.88	34.28/0.94	33.02/0.00	32.01/0.59
Kimi-Audio	7B	27.56/3.53	16.49/3.53	38.94/15.04	22.42/8.85	44.03/7.55	40.25/8.49	33.60/6.97
MidDashengLM	7B	<b>43.11</b> /15.19	37.22/17.67	<b>45.43</b> /23.89	42.77/16.81	<b>46.23</b> /30.19	45.60/21.70	<b>44.29</b> /20.32
Step-Audio-2-mini	7B	33.33/0.00	33.33/0.00	31.27/0.00	37.46/0.00	37.74/6.38	35.22/2.83	33.80/1.34
Gemma-3n-E4B-it	7.5B	23.32/1.41	28.27/6.01	41.89/15.04	36.58/7.96	33.96/5.66	40.57/8.49	29.75/5.37
Ming-Lite-Omni-1.5	18.9B	20.14/6.36	34.63/6.01	35.10/9.73	33.04/9.73	38.36/18.87	39.94/20.75	27.35/9.76
Qwen-2.5-Omni	7B	39.46/7.07	36.98/15.19	41.30/18.58	35.10/15.93	27.04/17.92	34.59/8.49	37.25/11.95
Xiaomi-MiMo-Audio	7B	36.16/0.71	41.58/5.65	41.30/5.31	38.05/4.42	<b>45.28</b> /9.43	44.34/9.43	39.24/3.58
Xiaomi-MiMo-Audio-think	7B	34.28/7.42	25.44/2.83	44.54/14.16	37.76/7.96	36.79/7.55	27.99/3.77	37.12/8.96
MiniCPM-O-v2.6	8B	29.92/3.18	27.92/2.83	43.36/11.50	39.53/12.39	38.36/26.42	35.53/17.92	34.73/9.96
GPT-4o Audio	—	<b>41.81</b> /—	<b>42.76</b> /—	43.07/—	<b>54.87</b> /—	39.94/—	42.45/—	41.70/—
Gemini 2.5 Flash	—	24.62/4.95	40.75/7.42	43.07/15.93	43.07/17.70	22.64/2.83	40.57/11.32	28.35/6.97
Gemini 2.5 Pro	—	40.87/10.95	34.98/11.66	<b>48.97</b> /25.66	49.26/14.15	<b>45.28</b> /14.15	<b>47.17</b> /7.55	40.77/12.75

1188

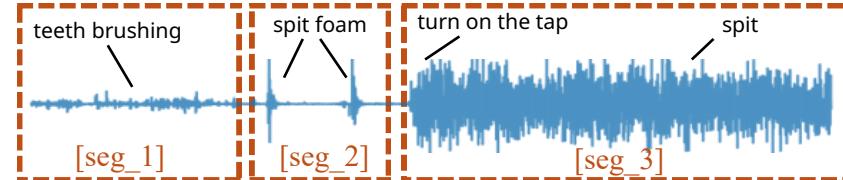
1189

1190

1191

1192 *[Brushing of teeth vigorously and then turning on the sink water.]*

1193

1194 *Q: clip1: [seg\_2] clip2: [seg\_3] clip3: [seg\_1]*1195 *Answer: <E>: clip 3 → clip 1 → clip 2*

1196

1197

1198

1199

1200

1201

### 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).

**Perception Error Reasoning Error** The core error is a fundamental

**Error analysis** 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

**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.

1238

1239

1240

1241

1242

1243

1244

1245

1246

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

1247

1248

1249

1250

1251

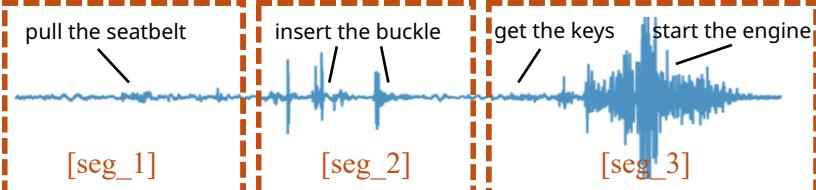
1252

1253

1254

1255

1256

*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.

1290

1291

1292

1293

1294

1295

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

1296

1297

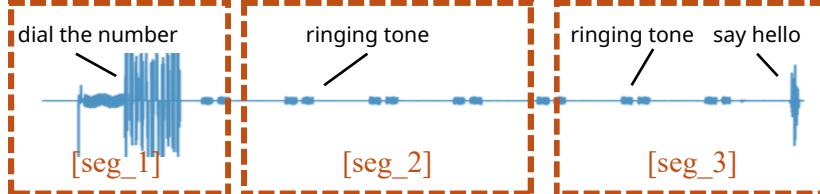
1298

1299

1300

*[Dial, connect, call]*

1301

1302 *Q:* clip1: [seg\_2] clip2: [seg\_3] clip3: [seg\_1]1303 *Answer:* <E>: clip 3→clip 1→clip 2

1304

1305

1306

1307

1308

1309

1310

1311

**Gemini 2.5 Pro**

1312 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.

1313 Therefore, the correct order is clip 3 → clip 1 → clip 2. The correct option is &lt;E&gt;.

1314

1315

1316

1317

1318

1319

1320

1321

1322

**GPT-4o Audio**

1323 The most natural chronological order for these sounds would typically be:

&lt;A&gt;: clip 1 → clip 2 → clip 3

1324

1325

1326

1327

1328

1329

1330

1331

**Qwen-2.5-Omni**

1332 The most natural chronological order in which these sounds would typically occur in the real world is:

&lt;A&gt;: clip 1 → clip 2 → clip 3

1333 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.**

1334

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

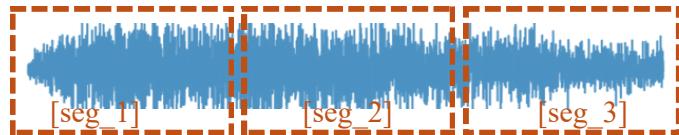
1348

1349

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

1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359

[An ambulance approached and then moved away.]



1360 **Q:** clip1: [seg\_3] clip2: [seg\_1] clip3: [seg\_2]  
1361 **Answer:** <D>: clip 2→clip 3→clip 1



### Gemini 2.5 Pro



1363 The most natural chronological order for these sound events is <D>: clip 2 -> clip 3 -> clip 1.  
1364

1365 Here is a breakdown of the events as they would likely occur:

1366 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.

1367 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.

1368 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.

1372 **Error analysis** correct



### GPT-4o Audio



1374 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:

1378 One clip likely represents a more distant or quieter siren.

1379 Another one sounds like it's getting closer.

1380 The final one should be the closest and the loudest.

1381 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.

1382 So, the most natural chronological order would be clip 2 -> clip 1 -> clip 3. That corresponds to option C.

1386 **Error analysis Perception Error** Seg\_3 decreases in volume.



### Qwen-2.5-Omni



1387 To determine the most natural chronological order of the sounds in the audio clips, let's analyze each clip:

1388 **Clip 1:** An emergency vehicle siren is heard, indicating a situation requiring immediate attention.

1389 **Clip 2:** Another emergency vehicle siren is heard, suggesting a second emergency vehicle is approaching or responding.

1390 **Clip 3:** A third emergency vehicle siren is heard, indicating a third emergency vehicle is also responding.

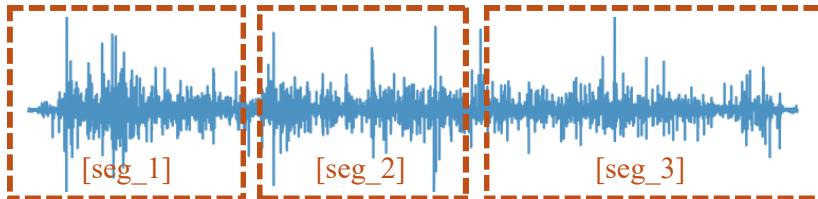
1391 Given this sequence, the most natural chronological order would be: First, the first emergency vehicle arrives (Clip 1). Then, the second emergency vehicle arrives (Clip 2). Finally, the third emergency vehicle arrives (Clip 3). Therefore, the correct answer is: <A>: clip 1 -> clip 2 -> clip 3

1392 **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.

1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403

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



1458  
1459  
1460  
1461  
1462*[Water is poured into a glass.]*1463  
1464  
1465  
1466  
1467  
1468  
1469**Q:** clip1: [seg\_3] clip2: [seg\_2] clip3: [seg\_1]**Answer:** <E>: clip 3→clip 2→clip 1

1470

1471

1472

**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 &lt;C&gt;.

1480  
1481  
1482  
1483**Error analysis** **Perception Error** Seg\_1 is not the sound of gulping, but the crisp sound of water being poured into a glass.

1484

**GPT-4o Audio**"The most natural chronological order in which these sounds would typically occur is:  
<B>: clip 1 (**blender sound**) -> clip 3 (pouring liquid) -> clip 2 (**sipping**)1485  
1486  
1487**Error analysis** **Perception Error** All segments are sounds of pouring water; the perception of both seg\_3 and seg\_2 is entirely incorrect.

1490

**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.1491  
1492  
1493

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.

1494  
1495  
1496

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.**

1500

1501

1502

Therefore, the most natural chronological order is: &lt;A&gt;: clip 1 -&gt; clip 2 -&gt; clip 3

1503  
1504  
1505**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.

1506

1507

1508

1509

1510

1511

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

1512

1513

1514

1515

1516 **Q:** What is the movement trajectory of the racing car?1517 **Options:** <A> From left to right <B> From right to left <C> Remains unchanged1518 **Answer:** <B>: From right to left

1519

1520

1521

1522

1523

1524

1525

1526

1527

1528

1529

1530

1531

1532

1533

1534

1535

1536

1537

1538

1539

1540

1541

1542

1543

1544

1545

1546

1547

1548

1549

1550

1551

1552

1553

1554

1555

1556

1557

1558

1559

1560

1561

1562

1563

1564

1565



### Gemini 2.5 Pro

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.



### channel-wise



#### 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.

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**.



### GPT-4o Audio

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**.

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 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.



### Qwen-2.5-Omni

**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**.

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

#### 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.

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