Evaluating Time Awareness and Cross-modal Active Perception of Large Models via 4D Escape Room Task

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have made rapid progress toward unified Omni models that integrate language, vision, and audio. However, existing environments largely focus on 2D or 3D visual context and vision-language tasks, offering limited support for temporally dependent auditory signals and selective cross-modal integration, where different modalities may provide complementary or interfering information, which are essential capabilities for realistic multimodal reasoning. As a result, whether models can actively coordinate modalities and reason under time-varying, irreversible conditions remains underexplored. To this end, we introduce **EscapeCraft-4D**, a customizable 4D environment for assessing selective cross-modal perception and time awareness in Omni models. It incorporates trigger-based auditory sources, temporally transient evidence, and location-dependent cues, requiring agents to perform spatio-temporal reasoning and proactive multimodal integration under time constraints. Building on this environment, we curate a benchmark to evaluate corresponding abilities across powerful models. Evaluation results suggest that models struggle with modality bias, and reveal significant gaps in current model's ability to integrate multiple modalities under time constraints. Further in-depth analysis uncovers how multiple modalities interact and jointly influence model decisions in complex multimodal reasoning environments.

1 Introduction

In recent years, Multimodal Large Language Models (MLLMs) (Alayrac et al., 2022; Liu et al., 2023; OpenAI, 2025; Team, 2025a,c) have made rapid progress in effectively bridging heterogeneous perceptual modalities and enabling coherent reasoning across them, facilitating the application of MLLMs to a wide range of tasks, especially in Vision-Language (VL) including visual commonsense reasoning (Zhou et al., 2024a; Kil et al., 2024), spatial reasoning (Yang et al., 2025; Zhan et al., 2025) and embodied manipulation and planning (Driess et al., 2023; Li et al., 2024b). Toward more realistic scenarios involving more than two modalities and time-aware interactions, prior work has commonly adopted agentic approaches that integrate isolated modules for different modalities (Wu et al., 2023; Shen et al., 2024; Huang et al., 2023). More recent advances have shifted towards Omni models that process multiple modality streams within a unified architecture (Xu et al., 2025; Wang et al., 2025a; Team, 2024a; Zhang et al., 2025b).

Despite this progress, the ability of current models to deeply integrate multiple modalities while preserving sensitivity to temporal signals remains underexplored. Existing evaluation paradigms focus primarily on static settings of 2D images or 3D volumetric scenes, often omitting the dimension of ***time***. This deficiency motivates the development of a 4-dimensional (4D) setup that incorporates dynamic and time-dependent factors to fully evaluate the capabilities to handle realistic tasks of Omni models. Furthermore, current multimodal benchmarks primarily emphasizes VL contexts, frequently relying on silent video clips and thereby overlooking the participation of auditory modality (Li et al., 2023, 2024a; Fu et al., 2024). These benchmarks present two major shortcomings: First, they fall short on adequately evaluating ***active perception across modalities*** (Bajcsy, 1988; Chaplot et al., 2020; Wilson et al., 2025), an essential capability that proactively determines *when*, *where*, and *how* to selectively acquire multimodal information and suppress misleading information. Second, they overlook a critical aspect of real-life environments, ***time awareness***, the ability to reason and act upon time-varying cues and irreversible information availability (Wang and Zhao, 2024; Chen et al., 2025), beyond conventional temporal grounding of events to timestamps or intervals (Gao et al., 2017; Cai et al., 2024). Yet, dedicated evaluation

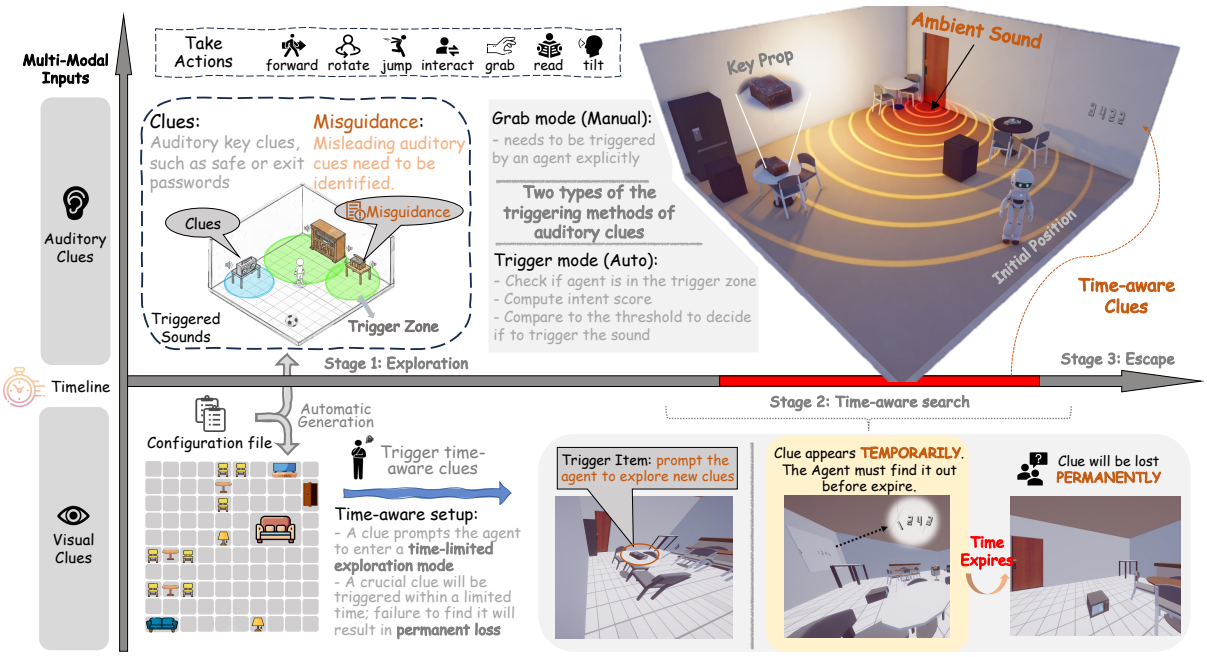

Figure 1: Our EscapeCraft-4D environment setup. It incorporates auditory, visual, and time-aware clues to evaluate multimodal reasoning. The system is designed to test the abilities of agents to coordinate cross-modal perception, such as leveraging spatially grounded auditory cues and visually presented clues or making time-aware decisions under strict temporal constraints. This figure illustrates the sequencially arranged tasks according to three stages: exploration, time-aware search, and finally escape, with dynamic auditory cues playing central roles in guiding the decision-making process. The trigger system dynamically introduces *time-aware clues* that challenge the agent to act within limited timeframes, showcasing the integration of auditory and visual modalities.

and development of such time-aware reasoning, acting and planning capabilities remain limited in the current scope.

In this paper, we address the above limitations by introducing **EscapeCraft-4D**, a simulated environment that augments a 3D world with an explicit temporal dimension. It serves as a testbed for more realistic 4D tasks and advanced multimodal reasoning and acting. Specifically, we incorporate time awareness through time-varying visual cues and auditory signals, both of which naturally encode temporal information and require agents to reason under dynamic and irreversible conditions.

Moreover, EscapeCraft-4D is a highly customizable environment with more than two modalities, *i.e.*, language, vision, and particularly audio. Notably, the inclusion of audio enables more complex design of cross-modal active perception where not all modalities are equally informative, requiring selective multimodal integration. Our environment design establishes the 4D paradigm through two key mechanisms as visualised in Figure 1, specifically: 1) It is *temporally dynamic*, requiring MLLMs to model temporal variations when aligning information across modalities; 2) It contains continuously active audio sources that provide location-dependent cues. Together, these properties require models to ground visual and auditory inputs within a unified spatio-temporal coordinate system and to selectively integrate multimodal information during decision making. Our contributions can be summarized as follows:

- We introduce **EscapeCraft-4D**, a 4D environment that enables explicit evaluation of active cross-modal perception and time-aware reasoning in Omni models and agents.
- Our results reveal significant gaps in current Omni models when incorporating multiple modalities under time constraints, such as susceptibility to misleading auditory cues, and limited time-aware interaction.
- Analysis shows auditory cues strongly guide exploration and multi-hop reasoning, while failures stem from missed spatio-temporal triggers, highlighting the need for cross-modal, time-aware coordination in complex tasks.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) build on earlier vision-language models (VLMs) and

audio-language models (ALMs). A representative early milestone is CLIP (Radford et al., 2021), which learns aligned image-text representations via contrastive training in a shared embedding space. Subsequent work Flamingo (Alayrac et al., 2022) augments a pretrained language model with visual conditioning modules, enabling in-context few-shot learning from interleaved image-text prompts. More recent open-source systems, BLIP (Li et al., 2022), LLaVA style visual instruction tuning (Liu et al., 2023), MiniCPM-V (Yao et al., 2024), and Qwen2.5-VL (Bai et al., 2025), further advance toward general purpose understanding and generation. Proprietary models such as GPT-4V (Team, 2024c) and Gemini (Team, 2024b) also demonstrate robust multimodal capabilities at scale. In parallel, ALMs condition language understanding and generation on audio signals, enabling models to *hear* speech and general sounds (Zhang et al., 2023b; Ghosh et al., 2024; Chu et al., 2024; Tang et al., 2024; Kong et al., 2024; Zhang et al., 2023a). These efforts move toward unified language-centric frameworks that integrate heterogeneous perceptual modalities.

Recent research has further progressed toward an Omni paradigm that unifies vision, audio, and language with agentic pipeline (Tao et al., 2026) or within a single model (e.g., GPT-4o (OpenAI, 2024), Qwen3-Omni (Xu et al., 2025), and VITA-1.5 (Fu et al., 2025)). By jointly modeling these modalities, Omni models capture richer contextual cues, such as StreamOmni (Zhang et al., 2025b), MGMOmni (Wang et al., 2025a), LongCat-Flash-Omni (Team, 2025b), and Gemini 2.5 (Team, 2025a). This trend represents a key step toward general purpose multimodal intelligence.

2.2 Time-Aware Benchmarks and Perishable Evidence Collection

In interactive and embodied reasoning tasks, information is frequently transient rather than persistent. In many scenarios, critical evidences and cues are triggered only after specific events and remains accessible for a limited time (Zhou et al., 2024b), e.g., a notification briefly appearing on a phone screen before disappearing. (Zhang et al., 2025a) introduces causal dependency that ties decision quality to decision timing. Such constraint on sequential actions requires the agent to reason on both action selection and prioritization, where misaligned decisions on the sequence can lead to irreversible information loss and mistakes. Meanwhile, existing environments and benchmarks incorporate time mainly for action ordering (Shridhar et al., 2020; Srivastava et al., 2021; Zhang et al., 2024), where task-relevant information (e.g. actions required) remains stable rather than shaped by decision timing. The agent thus can defer exploration without lasting consequences, leaving time-aware reasoning and perishable evidence collection underexplored. In contrast, time-aware benchmarks evaluate whether agents recognize and respond to temporal changes in evidence availability, where delayed decisions can lead to irreversible information loss (e.g. a password that will disappear forever after a period of time).

3 Environment and Task

3.1 Overview of EscapeCraft-4D

The room-escape task in this study is inspired by real-world escape room games and built on prior multimodal settings introduced in (Wang et al., 2025b; Lim et al., 2025). It is designed to evaluate the multimodal reasoning and planning capabilities of MLLMs. Following the 3D EscapeCraft environment (Wang et al., 2025b) that requires models to autonomously explore, perceive, and interact with the environment to achieve long-horizon goals, we introduce a fourth dimension, *time*, by incorporating audio components as shown in Figure 1. Vision, language, and audio are three most commonly perceived and used modalities in human activities (Stein and Meredith, 1993; Baltrušaitis et al., 2018; Liang et al., 2024). However, most current frameworks fail to address the time-awareness and lack mechanisms to integrate these modalities over time. Our integration of the audio signals and time-variant cues forces models to reason, act and planning within a specific duration, bridging the gap between static multimodal perception and dynamic, time-sensitive agency.

We especially include audio modality through two distinct categories, *ambient audio* and *trigger audio*. Specifically, following the task design protocol in (Wang et al., 2025b), auditory cues are integrated directly into the prop chain so that each reasoning step may depend on both visual evidence and spatially grounded audio signals. To account for the temporal nature of these signals, we further expand the escape tasks to include time-variant cues, enabling the evaluation of time-awareness of models during acting and planning. This section details the integration of the audio modality and

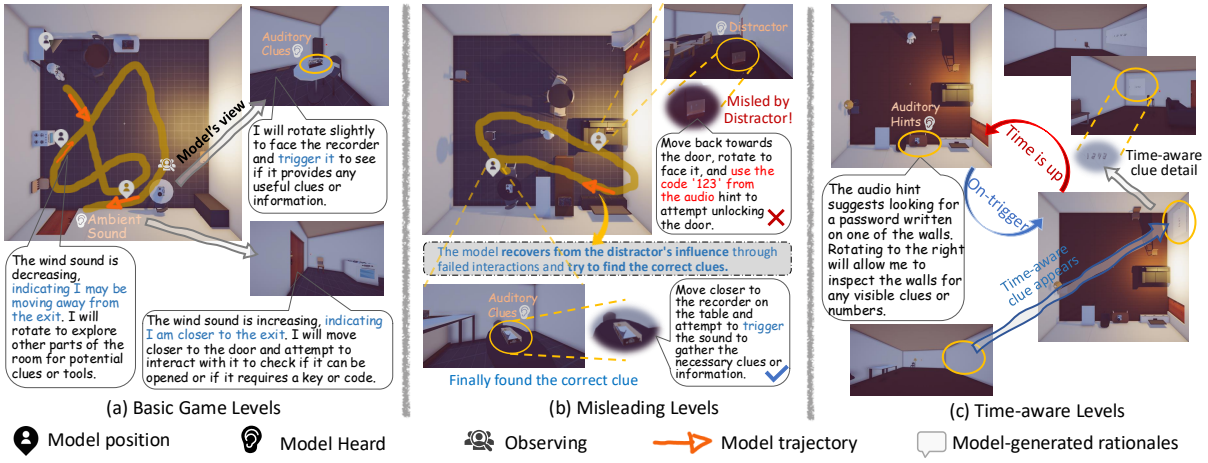

Figure 2: Examples of our designed levels with successful runs from GPT-4o. (a) Basic game design included both auditory and visual clues. (b) Misleading levels included distractors which would play misleading auditory clues. Models should distinguish which auditory clue was genuinely useful and which was the distractor. (C) Time-aware level design included clues only visible in limited time. Models should find these clues in the time limit or the clue would disappear.

the design of corresponding tasks.

3.2 Auditory Modality Integration

3.2.1 Ambient Audio

This audio signal is continuously emitted from a designated object (e.g., the sound of wind blowing outside the exit door) and spatially grounded to that object, as illustrated by the "Ambient Sound" example in the upper right part of Figure 1. Its perceived loudness is a deterministic function of the relative position from the agent to the source object, monotonically increasing as the agent approaches the source and attenuating with distance until inaudible beyond a predefined range. Although these sounds do not directly reveal symbolic content (e.g., passwords), they provide a gradient signal that can be exploited to (i) guide the direction of search, (ii) prioritize exploration targets, and (iii) coordinate visual attention with spatial movement. Thus, exploration shifts from purely visual scanning to selective cross-modal perception guided by audio.

3.2.2 Trigger Audio

This represents the time-variant signal in our environment that enables the evaluation of time-awareness. It is bound to specific interactable objects (e.g., a recorder, as visualized in the "Triggered Sounds" example in the upper left part of Figure 1), and is activated only when the agent issues an explicit trigger request while remaining inactive for the rest of time. To make it a non-trivial prop-chain element rather than a information channel that can be triggered randomly and unexpectedly, activation is gated by physical conditions, the agent must satisfy proximity and orientation thresholds before interaction. Once triggered, the object plays a short speech message that may contain (a) password to or instruction for following steps or (b) auditory deceptive content (e.g., plausible but irrelevant numbers).

Compared to visually dominant setups (Wang et al., 2025b), our audio- and time-augmented setting shifts the challenge from passive visual search to selective cross-modal perception and also has the ability to evaluate dynamic audio-visual fusion under deception. This design stresses (i) proximity-conditioned information acquisition, (ii) embodied verification of auditory claims, and (iii) strategic exploration policies that integrate continuous guide (ambient) and discrete prompt (triggered) auditory cues in the multi-hop reasoning.

3.3 Time-aware Tasks and Levels

We introduce time-aware levels where task-critical evidence (e.g., the password to unlock the exit) is explicitly time-varying or non-persistent rather than statically available throughout exploration. In these scenes, interacting with specific trigger objects (e.g., a recorder) initiates temporal events that cause evidence to appear, alter, or disappear within a limited time window. For instance, activating a recorder may reveal that a password will temporarily emerge on a wall for a limited time before vanishing permanently. This design enforces a strict coupling between perception, memory, and

timing: *agents must not only identify relevant interactions but also recognize the transient nature of the evidence and act before it expires.* By incorporating design elements such as passwords disappearing over short periods, the availability of evidence changes over time. These designs in the time-aware level go beyond temporal action ordering, and allow direct evaluation whether models can reason under irreversible and time-conditioned information loss.

3.4 MM-Escape4D Benchmark

Built on the "prop chain" framework (Wang et al., 2025b) extended with our auditory clue designs (Section 3.2), MM-Escape4D defines three standardized difficulty levels and two stress-test variants. By treating auditory clues as an additional reasoning hop, the basic levels form a clean progression in cross-modal demand:

- **Difficulty-1:** A single-hop level requiring no props; the agent escapes by directly identifying and interacting with the door, evaluating basic spatial grounding and navigation.
- **Difficulty-2:** A two-hop level adding an auditory hop on top of Difficulty-1. The agent must first locate a sound-emitting object, trigger it to obtain the spoken password, then use that password to unlock the door (Figure 2(a)).
- **Difficulty-3:** A three-hop level requiring both auditory clues and physical props (e.g., keys), explicitly evaluating how well a model synthesises visual and auditory evidence into a coherent plan.

On top of these, **Difficulty-2-M** and **Difficulty-3-M** preserve the same hop structure but introduce a parallel distractor sound source playing irrelevant numerical content (Figure 2(b)); a model that attempts to act on the distractor's content is considered misled. Full discussion of the misleading design is provided in Appendix A.

Time-aware level design Time-aware levels, denoted as **Difficulty-2-T**, extend Difficulty-2 with perishable, action-triggered clues: triggering a recorder makes the password to the exit appear at a specified position only briefly before permanently disappearing (Figure 1(c)). All agent actions carry time costs (linear/angular velocity for movement and rotation, fixed cost for grab/trigger), so the model must locate the password before it expires. Detailed time-cost configuration and the audio-type taxonomy are listed in Appendix A, Table 3.

4 Experiments

4.1 Evaluation Metrics

We follow (Wang et al., 2025b) and adopt Escape Rate (ER), Prop Gain, Grab Success Rate (GSR), Grab Ratio (R_{grab}), and Intent-Outcome Consistency (C_{IO}). On top of these, we introduce four metrics tailored to auditory signals, cross-modal active perception, and time-awareness: **A**udio **M**isguidance **R**ate (AMR), how often models are distracted by irrelevant audio; **T**ime-**C**onstrained **S**earch **S**core (TCSS), how early a model finds the target relative to the time limit (zero if not found in time); **T**rigger **R**atio (TR) and **T**rigger **S**uccess **R**atio (TSR). Formal definitions are in Appendix B.

4.2 Evaluation Setup

We evaluate proprietary models, GPT-4o (OpenAI, 2024) and Gemini-3-Pro (Team et al., 2025), and open-source Omni models, MGM-Omni-32B (Wang et al., 2025a), Qwen3-Omni-30B-A3B-Thinking and Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025). As the official GPT-4o API does not accept text, image and audio simultaneously, we build a GPT-4o Omni-Agent that pairs `gpt-4o` (VL) with `gpt4o-audio-preview` (audio). Maximum interaction steps follow (Wang et al., 2025b) (50/65/80 for Difficulty-1/2/3, identical for misleading and time-aware variants). Full templates and additional setup details are in Appendix C and Appendix G.

4.3 Experimental Results

Basic game settings. Table 1 reveals a substantial and widening gap as difficulty increases. GPT-4o Omni-Agent leads with average ER 72.73%, followed by Gemini-3-Pro (66.67%), both more than double the strongest open-source candidate Qwen3-Omni-Thinking (30.30%). At Difficulty-3 the gap becomes most pronounced: every open-source model drops to 0.00% while proprietary models still complete a non-trivial fraction (GPT-4o 45.45%, Gemini-3-Pro 27.27%). This indicates MM-Escape4D probes capabilities not well captured by conventional multimodal benchmarks: completing a level requires active 3D exploration, triggering object-conditioned auditory cues, and reasoning over transient and irreversible signals,

Models	Difficulty-1				Difficulty-2						Difficulty-3							AVG
	ER (%)↑	Steps↓	GSR (%)↑	GR (%)	ER (%)↑	Steps↓	GSR (%)↑	GR (%)	TSR (%)↑	TR (%)	ER (%)↑	Prop (%)↑	Steps↓	GSR (%)↑	GR (%)	TSR (%)↑	TR (%)	ER (%)↑
Agentic Approach																		
GPT-4o Omni-Agent	90.91	**16.82**	48.52	24.58	**81.82**	38.82	12.75	33.66	55.30	9.15	**45.45**	**45.45**	59.27	**8.39**	34.73	34.85	24.09	72.73
Omni Models																		
Gemini3-Pro-Preview	**100.00**	**16.82**	**59.25**	21.95	72.73	**27.73**	**37.12**	22.67	**81.82**	6.42	27.27	27.27	**50.36**	7.82	29.69	**55.11**	4.52	66.67
Qwen3-Omni-Thinking	63.64	24.00	24.80	54.97	27.27	56.73	5.50	29.97	39.86	22.24	0.00	0.00	81.00	0.00	27.27	42.29	19.53	30.30
Qwen3-Omni-Instruct	18.18	47.55	6.82	9.24	27.27	60.27	12.21	7.99	56.06	2.67	0.00	0.00	81.00	0.00	9.09	29.52	6.06	15.15
MGM-Omni-32B	27.27	39.55	10.91	17.14	0.00	66.00	0.00	17.08	14.55	1.10	0.00	0.00	81.00	0.00	17.51	18.69	2.81	9.09

Table 1: Results of basic game setting. Prop: Prop Gain; Steps: average steps used to complete the game; GSR: Grab Success Rate, the precision of grabbing; GR: Grab Ratio, the portion of grabbing actions regarding the total consumed steps. TSR: Trigger Success Rate. TR: Trigger Ratio. The max allowed steps are 50, 65, 80 for Difficulty-1, -2, -3 respectively. The best score of each metrics is **bolded** and the second is underlined. The GPT-4o Omni Agent refers to our agentic approach built on GPT-4o models, `gpt-4o` for vision and language, together with `gpt4o-audio-preview` for audio. We use Qwen3-Omni-30B-A3B version for Qwen3-Omni models in this table.

Models	**Time-Aware**	**Misleading Levels**					
	TCSS↑	**MAT↓**	**AMR↓**	**D2-M-ER↑**	**D3-M-ER↑**	**D2-M-Steps↓**	**D3-M-Steps↓**
Agentic Approach							
GPT-4o Omni-Agent	**67.95%**	63.64%	100.00%	**72.73%**	**63.64%**	40.55	60.09
Omni Models							
Qwen3-Omni-Thinking	55.87%	59.09%	23.08%	0.00%	0.00%	66.00	81.00
Gemini-3-Pro-Preview	47.04%	40.91%	22.22%	63.64%	**63.64%**	**28.82**	**53.82**
MGM-Omni-32B	0.00%	**13.64%**	**0.00%**	0.00%	0.00%	66.00	81.00
Qwen3-Omni-Instruct	0.00%	40.91%	**0.00%**	0.00%	0.00%	66.00	81.00

Table 2: Comprehensive Experimental Results on MM-Escape4D. (I) Performance on **Time-aware Levels** measured by TCSS. (II) Robustness against *Misleading Auditory* cues: **MAT** refers to ratio of scenes where **m**isleading **a**udio was **t**riggered; **AMR** represents the ratio of scenes where the agent was deceived given that the misleading cue was triggered, within 3 following steps. **Dn-M-ER** and **Dn-M-Step** denote the Escape Rate and average steps for Difficulty-n Misleading levels. The best and second-best results are highlighted in **bold** and underline, respectively.

demanding tight coupling between perception, reasoning and action.

Strong performance on static multimodal benchmarks does not necessarily transfer. Although recent open-source Omni models report competitive numbers on MMMU (Yue et al., 2024), MathVista (Lu et al., 2024), and AIME25[1], Qwen3-Omni-Thinking drops to 27.27% ER on Difficulty-2 and 0% on Difficulty-3, and MGM-Omni-32B fails on Difficulty-2 entirely, implying persistent limitations in spatial reasoning and cross-modal coordination once time-sensitive, modality-intensive interaction is required.

Beyond raw escape rates, trigger metrics suggest a structural advantage of unified multimodal modelling. Qwen3-Omni-Thinking and Gemini-3-Pro both reach higher TSR than the GPT-4o Agent on Difficulty-3 (42.29% and 55.11% vs. 34.85%), indicating that integrating modalities in a unified representation, rather than serialising them through transcription, better supports complex cross-modal reasoning ability.

Modality bias. A consistent pattern across difficulties is a visible *modality bias*: agents weight visual evidence far more readily than auditory evidence even when audio is the only path to success. On Difficulty-2, where the password requires triggering a sound source, open-source Omni models exhibit very low Trigger Ratios (Qwen3-Omni-Instruct TR 2.67%, MGM-Omni-32B TR 1.10%), almost never attempting the auditory channel. The agentic GPT-4o pipeline triggers more often (TR 9.15%) because audio is routed through a separate tool, isolating the failure point as the internal decision of *when* to attend to audio. The same bias persists on Difficulty-3 (TR 0% for both Qwen3-Omni-Instruct and MGM-Omni-32B). Current Omni models do not yet treat modalities as truly interchangeable channels.

Misleading and time-aware settings. Table 2 reports the corresponding metrics. *Misleading levels.* Robustness factorises into exposure (MAT) and susceptibility (AMR), and the two factors expose

[1] American Invitational Mathematics Examination (AIME) 2025, https://maa.org.

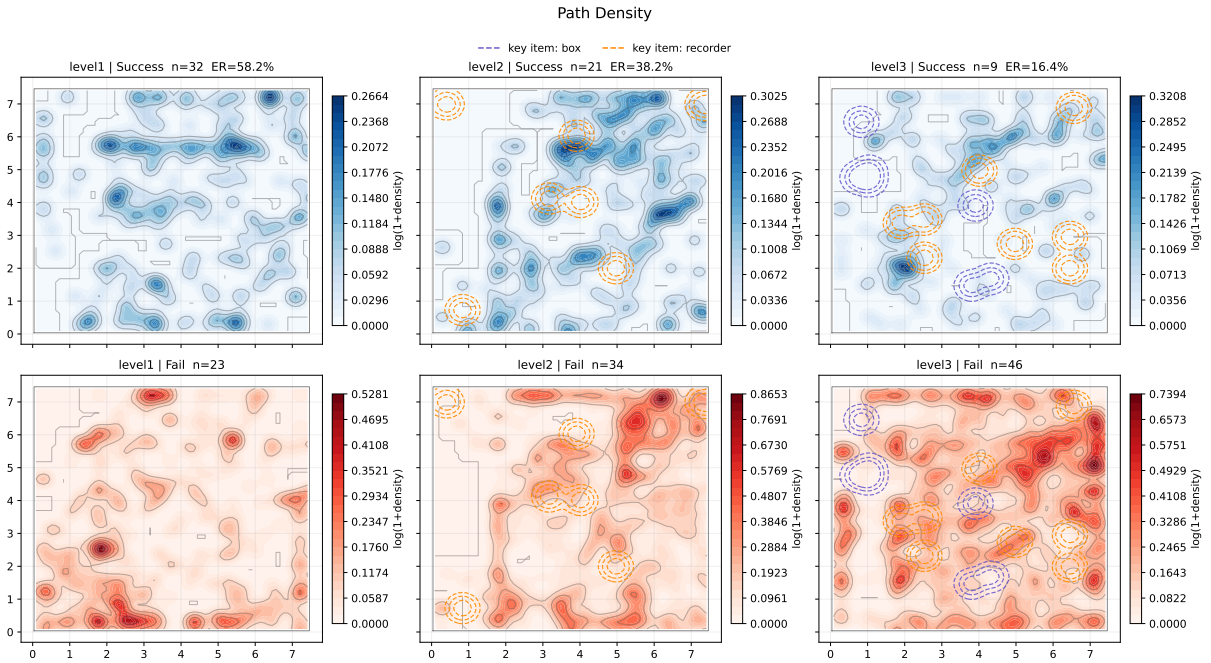

Figure 3: **Path density analysis for different levels in our MM-Escape4D benchmark.** Exits of different scenes are aligned as mentioned in Appendix F.1. The heatmaps represent the log-normalized density ($\log(1 + \text{density})$) of agent positions, with blue gradients for successful runs and red for failed ones. The dashed contours represent the spatial distribution of key items (e.g., recorders for auditory clues and boxes).

very different failure modes. GPT-4o is exposed often and fully susceptible (MAT 63.64%, AMR 100%), suggesting it actively interacts with audio sources but lacks any internal cross-checking once a plausible-sounding clue is heard. Qwen3-Omni-Thinking has comparable exposure but its reasoning chain partially filters the deception (AMR 23.08%); yet the underlying spatial reasoning is too weak to convert this resistance into success, so its misleading-level ER stays at 0%, showing that resistance to deception alone is not sufficient. Gemini-3-Pro achieves the best balance (MAT 40.91%, AMR 22.22%, D2/D3-M-ER 63.64%), being both selective about when to trigger an auditory source and able to discount its content when triggered. MGM/Qwen3-Instruct's near-zero MAT reflects interaction failure rather than genuine robustness; disambiguation never gets a chance to happen. *Time-aware levels.* TCSS spreads sharply: GPT-4o 67.95%, Qwen3-Omni-Thinking 55.87%, Gemini-3-Pro 47.04%, while MGM/Qwen3-Instruct fail to find the clue in time at all (0%). Higher TCSS is significantly associated with final success (Spearman $r = 0.433$, $p = 9.56 \times 10^{-4}$; permutation test on success vs. fail $p = 0.0016$). Together these reveal the gap that strong performance on static multimodal benchmarks does not directly transfer to embodied, time-variant cross-modal decision making and being *fast to* a transient clue is the bottleneck under irreversible time constraints.

Overall, we summarise the following takeaways: (i) Active perception under irreversible constraints exposes fundamental weaknesses. When agents must explore, trigger object-conditioned auditory cues, and reason over transient information, current Omni models exhibit pronounced degradation. (ii) 3D spatial reasoning and cross-modal temporal integration remain underdeveloped. Coordinating vision, language, and audio in dynamic settings is significantly more challenging than solving static or time-invariant multimodal reasoning tasks. (iii) Unified modality modeling improves coordination efficiency. Modality fusion demonstrates advantages in cross-modal interaction over modular, transcription-based agentic pipelines, indicating that structural integration of input modalities is beneficial for decision making in complex and dynamic environments.

5 Analysis

We further analyse the exploration behaviour of the agents and how auditory cues shape their trajectories from three complementary perspectives: exit-

aligned path density, an audio/no-audio ablation on Difficulty-1, and a paired audio-vs-non-audio comparison on Difficulty-2/3. Full details, hypothesis tests, and figures are in Appendix F.

Exploration patterns vary with difficulty (Sec. F.1). Exit-aligned path density heatmaps (Figure 3) reveal a clear progression. In Difficulty-1, successful runs concentrate tightly on the direct path from start to exit, while failed ones show sparse, near-uniform exploration suggesting an absence of basic spatial grounding. As complexity increases in Difficulty-2/3, success requires deviating from the exit-ward heading to interact with key items (recorders, boxes), and the heatmaps confirm successful runs exhibit higher density around those items than failed ones do. Failed runs are typically trapped near the centre or in corners and often fail to transition from exploration to goal-directed motion, with lower escape rates correlating with more fragmented path distributions. The discrepancy is most pronounced in Difficulty-3, where successful runs display a clean multi-hop pattern, first localising the recorder then proceeding to the exit, whereas failed ones miss the trigger zone or the timing window. The bottleneck thus appears less about raw perception and more about committing to a multi-hop plan that interleaves visual search with auditory interaction.

Ambient sound improves navigation (Sec. F.2). On Difficulty-1, comparing a spatialised ambient-sound condition with a silent baseline isolates the effect of audio on pure navigation. Across nine metrics, audio reduces all burden-related ones (*Steps*, *Time*, *PathLen*, *Turn*, *Frechet*), lifts efficiency ones (*ProgEff*, *PathEff*, *Mono*), and improves proximity to the exit (*MinDist*), winning in over 60% of trials. The non-audio group exhibits consistently higher Frechet values, reflecting more wandering and path distortion. Aligned distance-to-exit curves further show the audio group consistently closer to the exit through the central portion (0.2–0.8) with lower variance, so the gain comes from the middle of the trajectory. Paired Wilcoxon tests on AUC ($P^{\mathrm{AUC}}_{\mathrm{A\ vs\ NA}} = 0.042$) and terminal non-arrival ($P^{\mathrm{end}}_{\mathrm{A\ vs\ NA}} = 0.0125$), plus a point-label permutation test on path density ($P^{\mathrm{perm}}_{\mathrm{L1}} = P^{\mathrm{perm}}_{\mathrm{JSD}} = 0.00019$), confirm the effect is systematic. Spatialised audio thus acts as a continuous proximity signal rather than a destination beacon.

Auditory clues concentrate exploration (Sec. F.3). A normalised Herfindahl-Hirschman occupancy index (HII) over the trajectory shows audio-active segments are significantly more concentrated than non-audio segments across both outcomes (Difficulty-2: $p = 0.0004/0.0215$ for success/fail; Difficulty-3: $p = 0.0039/0.0028$): hearing a relevant clue prompts a tighter local search rather than continued wide exploration. A success-vs-fail gap emerges only in Difficulty-2 audio segments ($p = 0.0088$); in Difficulty-3 it disappears ($p = 0.4001$). A Mantel permutation test ($r = 0.1548$, $p = 0.0047$) attributes this to the additional visual key object, whose varying location dilutes the auditory contrast. Sign-flip permutation tests further confirm the auditory effect is robust ($p = 0.0003$). Omni agents thus exploit audio when present, but the same signal can be drowned out by a competing visual cue, indicating that cross-modal interference, not raw perception strength, drives failure when modalities compete. This is the most actionable finding: scaling either visual or auditory encoders in isolation will yield diminishing returns, and the next gains likely come from mechanisms that decide *when* each modality should dominate, dynamically reweighting attention as the agent moves between exploration, evidence collection, and time-bounded action.

6 Conclusion

We presented **EscapeCraft-4D**, a novel 4D multimodal environment for evaluating time-aware reasoning and active cross-modal perception in Omni models and agents. Our experiments across 66 scenes of different difficulties reveal that even state-of-the-art models struggle with modality bias, transient cues, and misleading information, highlighting substantial gaps in real-world multimodal reasoning. Incorporating auditory signals and temporal constraints significantly improves navigation efficiency, cue utilization, and multi-hop reasoning, as measured by TCSS, AMR, and HII metrics. Effective modality selection and time-sensitive decision-making remain challenging, which reveal the gap between the performance on static multimodal benchmarks and emobodied, time-variant and cross-modal decision making. Overall, EscapeCraft-4D provides a comprehensive platform to advance Omni models toward realistic, dynamic, and temporally-aware multimodal intelligence, supporting future research on embodied reasoning and proactive perception in complex and dynamic 4D environments.

Limitations

EscapeCraft-4D is a controlled simulated environment rather than a real-world deployment, so the conclusions drawn here should be interpreted within that scope. Auditory cues are restricted to spatialised ambient sound and short trigger speech, and the time-aware setting uses a single fixed window with deterministic action costs; richer acoustic conditions (overlapping speech, reverberation, non-speech semantic events) and nested or stochastic timers are left to future work. Our findings are based on a representative selection of current proprietary and open-source Omni models; while consistent across this slice, they should not be taken as a universal ordering over all Omni systems. Generalisation to physical embodied agents, where sensing noise and continuous control introduce further challenges, also remains an open direction. We view extending the environment to broader settings and evaluating a wider range of models as natural follow-ups that build on, rather than contradict, the findings reported in this paper.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: A Visual Language Model for Few-Shot Learning. *Preprint*, arXiv:2204.14198.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL Technical Report. *Preprint*, arXiv:2502.13923.

Ruzena Bajcsy. 1988. Active perception. *Proceedings of the IEEE*, 76(8):996–1005.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Mu Cai and 1 others. 2024. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. 2020. Learning to explore using active neural slam. In *International Conference on Learning Representations (ICLR)*.

Yijing Chen, Yihan Wu, Kaisi Guan, Yuchen Ren, Yuyue Wang, Ruihua Song, and Liyun Ru. 2025. Chronusomni: Improving time awareness of omni large language models. *arXiv preprint arXiv:2512.09841*.

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

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. Palm-e: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488. PMLR.

Thomas Eiter, Heikki Mannila, and 1 others. 1994. Computing discrete fréchet distance.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of large video-language models. *arXiv preprint arXiv:2405.21075*.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. 2025. VITA-1.5: Towards GPT-4o Level Real-Time Vision and Speech Interaction. *Preprint*, arXiv:2501.01957.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

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

A. Hirschman. 1964. The paternity of an index. *The American Economic Review*, 54:761–762.

Rongjie Huang, Mingze Zhang, Dongchao Li, Gin Meyer, Zhiyu Weng, Yu Zheng, Dongyu Deng, Yi Zhao, Yi Ren, Jinglin Peng, and 1 others. 2023. Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*.

Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Chao. 2024. MLLM-COMPBENCH: a comparative reasoning benchmark for multimodal LLMs. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 28798–28827.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *Preprint*, arXiv:2402.01831.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Yuying Zhang, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *Preprint*, arXiv:2201.12086.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1203–1213.

Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. 2024b. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18061–18070.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM computing surveys*, 56(10):1–42.

Seungwon Lim, Sungwoong Kim, Jihwan Yu, Sungjae Lee, Jiwan Chung, and Youngjae Yu. 2025. VisEscape: A benchmark for evaluating exploration-driven decision-making in virtual escape rooms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *Preprint*, arXiv:2304.08485.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-05-13.

OpenAI. 2025. OpenAI o3 and o4-mini System Card.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *Preprint*, arXiv:2103.00020.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Lu Weiqi, and Yuting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. In *Advances in Neural Information Processing Systems*, volume 36.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. *Preprint*, arXiv:1912.01734.

Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, C. Karen Liu, Silvio Savarese, Hyowon Gweon, Jiajun Wu, and Li Fei-Fei. 2021. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. *Preprint*, arXiv:2108.03332.

Barry E Stein and M Alex Meredith. 1993. *The merging of the senses*. MIT press.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. *Preprint*, arXiv:2310.13289.

Keda Tao, Wenjie Du, Bohan Yu, Weiqiang Wang, Jian Liu, and Huan Wang. 2026. Active perception agent for omnimodal audio-video understanding. *Preprint*, arXiv:2512.23646.

Chameleon Team. 2024a. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Gemini Team. 2024b. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gemini Team. 2025a. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, and 99 others. 2025. Gemini robotics:

Bringing ai into the physical world. *Preprint*, arXiv:2503.20020.

Meituan LongCat Team. 2025b. LongCat-Flash-Omni Technical Report. *Preprint*, arXiv:2511.00279.

OpenAI Team. 2024c. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Qwen Team. 2025c. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Chengyao Wang, Zhisheng Zhong, Bohao Peng, Senqiao Yang, Yuqi Liu, Haokun Gui, Bin Xia, Jingyao Li, Bei Yu, and Jiaya Jia. 2025a. MGM-Omni: Scaling Omni LLMs to Personalized Long-Horizon Speech. *Preprint*, arXiv:2509.25131.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415, Bangkok, Thailand. Association for Computational Linguistics.

Ziyue Wang, Yurui Dong, Fuwen Luo, Minyuan Ruan, Zhili Cheng, Chi Chen, Peng Li, and Yang Liu. 2025b. How do multimodal large language models handle complex multimodal reasoning? placing them in an extensible escape game. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4807–4817.

Ryan Wilson and 1 others. 2025. Pop-gs: Next best view in 3d-gaussian splatting with p-optimality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. In *arXiv preprint arXiv:2303.04671*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *Preprint*, arXiv:2408.01800.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Xiaoyu Zhan, Wenxuan Huang, Hao Sun, Xinyu Fu, Changfeng Ma, Shaosheng Cao, Bohan Jia, Shaohui Lin, Zhenfei Yin, LEI BAI, Wanli Ouyang, Yuanqi Li, Jie Guo, and Yanwen Guo. 2025. Actial: Activate spatial reasoning ability of multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *Preprint*, arXiv:2305.11000.

Lingfeng Zhang, Yuening Wang, Hongjian Gu, Atia Hamidizadeh, Zhanguang Zhang, Yuecheng Liu, Yutong Wang, David Gamaliel Arcos Bravo, Junyi Dong, Shunbo Zhou, Tongtong Cao, Xingyue Quan, Yuzheng Zhuang, Yingxue Zhang, and Jianye Hao. 2025a. Et-plan-bench: Embodied task-level planning benchmark towards spatial-temporal cognition with foundation models. *Preprint*, arXiv:2410.14682.

Lingfeng Zhang, Yuening Wang, Hongjian Gu, Atia Hamidizadeh, Zhanguang Zhang, Yuecheng Liu, Yutong Wang, David Gamaliel Arcos Bravo, Junyi Dong, Shunbo Zhou, and 1 others. 2024. Et-plan-bench: Embodied task-level planning benchmark towards spatial-temporal cognition with foundation models. *arXiv preprint arXiv:2410.14682*.

Shaolei Zhang, Shoutao Guo, Qingkai Fang, Yan Zhou, and Yang Feng. 2025b. Stream-omni: Simultaneous multimodal interactions with large language-vision-speech model. *Preprint*, arXiv:2506.13642.

Ziqiang Zhang, Sanyuan Chen, Long Zhou, Yu Wu, Shuo Ren, Shujie Liu, Zhuoyuan Yao, Xun Gong, Lirong Dai, Jinyu Li, and Furu Wei. 2023b. SpeechLM: Enhanced Speech Pre-Training with Unpaired Textual Data. *Preprint*, arXiv:2209.15329.

Kaiwen Zhou, Kwonjoon Lee, Teruhisa Misu, and Xin Wang. 2024a. Vicor: Bridging visual understanding and commonsense reasoning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10783–10795.

Qinhong Zhou, Sunli Chen, Yisong Wang, Haozhe Xu, Weihua Du, Hongxin Zhang, Yilun Du, Joshua B. Tenenbaum, and Chuang Gan. 2024b. Hazard challenge: Embodied decision making in dynamically changing environments. *Preprint*, arXiv:2401.12975.

A Level Design

This section provides the full level design that is briefly summarised in Section 3.4.

Basic game design By incorporating the auditory clue designs described in Section 3.2, we leverage the "prop chain" framework following MM-Escape (Wang et al., 2025b) to facilitate the batch generation of scenarios containing auditory modalities. By treating auditory clues as an additional reasoning hop, we define three standardized difficulty levels for each room:

- **Difficulty-1:** The single-hop reasoning path, requiring no visual props to escape. Models can exit by identifying and interacting directly with the door.
- **Difficulty-2:** This level adds one additional reasoning hop compared to Difficulty-1. The model must additionally locate a specific sound-emitting object and trigger the audio to obtain the password, which is then used to unlock the door (example in Figure 2(a)).
- **Difficulty-3:** A reasoning path with one additional hop beyond Difficulty-2 is required. It asks for both auditory clues and physical items (e.g., keys or passwords). This level integrates visual and auditory cues to evaluate the cross-modal synthesis capabilities of models under complex environments.

Misleading level design To evaluate the robustness of models against interference, we particularly design "misleading" versions for Difficulty-2 and Difficulty-3, named **Difficulty-2-M** and **Difficulty-3-M**. While maintaining the same hop-wise structure, we introduce a parallel distractor sound source to the hop containing auditory information. This distractor plays irrelevant audio containing numerical data intended to confuse the judgement of the model, as displayed by Figure 2(b). Misleading is measured by the ability of model to disregard this irrelevant information; if the model attempts to utilize the content of the distractor, it is considered misled by irrelevant modal.

Time-aware level design Time-aware levels, denoted as **Difficulty-2-T**, extend Difficulty-2 by introducing perishable clues, which are time-variant and action-related. Specifically, the first hop in this level involves an object, such as a recorder, that provides audio instructions guiding the agent to find the password within the subsequent constrained durations. The second hop is time constrained, requiring the model to be aware of this limited duration. As illustrated in Figure 1(c), after the recorder is triggered, the password to the exit appears at a specified position only temporarily before permanently disappearing. To realistically simulate the passage of time within our environment, we assign specific time costs and corresponding speeds to all actions: linear velocity for movement, angular velocity for rotations (turning or looking up/down), and fixed time intervals for grabbing and triggering interactions. Detailed configurations and examples are shown in Table 3.

Category	**Specification / Value**
Forward Speed	2.0 m/s
Rotation Speed	60°/s
Grab/Trigger	0.5 s (fixed cost)
Audio Types	**Examples**
Ambient Sound	Wind sound from exit
Auditory Clues	Passwords or prompts
Misleading Audio	Distractors

Table 3: Time-cost Configuration and Audio Categories.

B Evaluation Metrics

Follow (Wang et al., 2025b), we also use average escape rate (ER) as the indicator of game completion, and five more metrics for measuring intermediate interactions, including prop gain, average steps, grab count, grab success rate, and grab ratio. Denoting the total steps as S, amount of succeeded grabbing action as N_{grab}^{TP}, grab success rate as GSR, grab ratio as R_{grab} we have,

$$\text{Prop Gain} = \frac{N_{grab}^{TP}}{\sum \text{Prop count}}, \quad (1)$$

$$\text{GSR} = \frac{N_{grab}^{TP}}{\sum \text{Grabbing actions}}, \quad (2)$$

$$\mathrm{R}_{grab} = \frac{\sum \text{Grabbing actions}}{S}, \quad (3)$$

We also follow (Wang et al., 2025b) and use the metric *Intent-outcome Consistence*, denoted by C_{IO}, ranging [0,1] to evaluate how often the successful interaction was triggered by accident rather than the origin intention of the model. For example, a model aims at trying to open the refrigerator but opening a box instead. We use GPT-4o for auto-evaluation, and the prompt template was shown in Appendix G. We calculated the C_{IO} for

Difficulty-3 across GPT-4o Omni-Agent, Gemini3-Pro-Preview and Qwen3-Omni-Thinking. Results in Table 4 implies that many completed sub-goals are triggered by accident, where Gemini3 achieves a best C_{IO} with 53.75%, while GPT-4o achieves only 37.08% despite a best GSR of 8.39%, indicating the gap between reasoning and perception.

Models	GPT-4o Omni Agent	Gemini3-Pro-Preview	Qwen3-Omni-Thinking
Scores: $C_{IO}(GSR)$	37.08 (8.39)	53.75 (7.82)	27.55 (0.00)

Table 4: Consistency compared with GSR on Difficulty-3.

B.1 New Metrics for Auditory and Time-aware Settings

In addition to the metrics adopted from (Wang et al., 2025b), we design four metrics dedicated to auditory signals, cross-modal active perception, and time-awareness, as referenced in the main text. Specifically, we introduce (i) **A**udio **M**isguidance **R**ate (AMR) to quantify how often models are distracted by irrelevant modality, (ii) **T**ime-**C**onstrained **S**earch **S**core (TCSS) to measure how early a model finds the target relative to the time limit, with zero score assigned if the target is not found within the time budget, (iii) **T**rigger **r**atio (TR) and (iv) **T**rigger **s**uccess **r**atio (TSR). Denoting the total steps as S, the number of scenes in which the model is misled by irrelevant audio cues as $N_{\text{misguided}}$, amount of succeeded trigger action as N^{TP}_{tigger}, trigger success rate as TSR, trigger ratio as $\text{R}_{trigger}$, and N_{total} the number of scenes where misleading cue was triggered. Let t_{found} denote the consumed time duration when the target is found, and T_{lim} the time limit for each episode. We have,

$$\text{TSR} = \frac{N^{TP}_{trigger}}{\sum \text{Triggering actions}}, \tag{4}$$

$$\text{R}_{trigger} = \frac{\sum \text{Triggering actions}}{S}, \tag{5}$$

$$\text{AMR} = \frac{N_{\text{misguided}}}{N_{\text{total}}}, \tag{6}$$

$$\text{TCSS} = \begin{cases} 1 - \dfrac{t_{\text{found}}}{T_{lim}}, & \text{if found within } T_{\text{lim}}, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

C Evaluation Setup Details

We evaluated a range of open-source and proprietary models of varying scales, utilizing the tiered methodology described in Section 3.4 to assess their multimodal reasoning capabilities. Open-source models include MGM-Omni-32B (Wang et al., 2025a), Qwen3-Omni-30B-A3B-Thinking and Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025); proprietary models include GPT-4o (OpenAI, 2024) and Gemini-3-Pro (Team et al., 2025). For GPT-4o, we implement an agentic approach to enable Omni ability, as the officially provided APIs do not support simultaneously inputting all three modalities, text, image, and audio. Specifically, we employ `gpt-4o` as the VL backbone, and equip it with `gpt4o-audio-preview` to process any auditory information when required.

Detailed evaluation templates are provided in Appendix G. Following the experimental design in (Wang et al., 2025b), we set the maximum number of interaction steps to 50, 65, and 80 for Difficulty-1, Difficulty-2, and Difficulty-3, respectively, including their misleading and time-aware variants. This configuration ensures that models are granted sufficient opportunities to explore and interact with the environment, while preventing excessively long trajectories that could obscure performance differences across difficulty levels.

D Measurement Foundations

D.1 Discrete Fréchet Distance

To quantify the global structural similarity between the traversed trajectory P of the agent and the optimal reference "beeline" path Q (which is exactly the line connecting the start point of the agent and the coordinate where the exit is located), we employ the discrete Fréchet distance (d_F). Unlike the standard Euclidean distance between endpoints, d_F accounts for the specific sequencing and spatial flow of the movement. Formally, for two polygonal curves $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_m)$, the distance is defined as:

$$d_F(P, Q) = \inf_{\alpha,\beta} \max_{t\in[0,1]} \|P(\alpha(t)) - Q(\beta(t))\| \tag{8}$$

where α and β are continuous, non-decreasing reparameterizations. A lower d_F value indicates that the agent's navigation strategy closely adheres to the geometric optimum. Detailed description for discrete Fréchet can be found in (Eiter et al., 1994).

D.2 HHI-based Occupancy Index (HII)

We measure the spatial focus of the navigation of the agent using a normalized Herfindahl-Hirschman Index (HHI) applied to a discretized occupancy grid. The environment is partitioned into N equal-sized cells, where p_i represents the proportion of total episode time the agent occupies cell i. The raw HII is defined as:

$$HII_{raw} = \sum_{i=1}^{N} p_i^2 \tag{9}$$

To ensure comparability across different room scales, we normalize the index which is broadly used in calculating hii with in the grid range such that $HII \in [0, 1]$, which is defined as:

$$\mathrm{HII}_{\mathrm{norm}} = \frac{\mathrm{HII}_{\mathrm{raw}} - 1/C}{1 - 1/C} \in [0, 1]. \tag{10}$$

We interpret a higher HII as the evidence of signifying that the agent has developed a consistent, low-entropy navigation policy. Conversely, a lower HII indicates a high-entropy search pattern, often seen in agents lacking sufficient sensory guidance. However, it is also possible that the model is actually stuck in a certain area, causing an abnormally high HII. Therefore, we subsequently used various hypothesis tests to rule out the influence of extreme cases on the analysis results. Detailed description for the original definition of HII index can be found in (Hirschman, 1964).

D.3 Path Efficiency

Path efficiency (η_{path}) assesses the economy of movement by normalizing the theoretical shortest navigable distance L_{min} by the actual distance L_{actual} traversed by the agent:

$$\eta_{path} = \frac{L_{min}}{L_{actual}} \tag{11}$$

A value of 1.0 represents a strictly optimal trajectory. This metric isolates the ability of the agent to translate sensory inputs into direct kinetic progress.

D.4 Monotonic Progress Ratio

The Monotonic Progress Ratio (R_{mono}) evaluates the degree of hesitation or backtracking during an episode. Let $D(t)$ be the Euclidean distance to the exit at timestep t. We define the progress increment as $\Delta D_t = D(t-1) - D(t)$. The ratio is the proportion of the episode duration T where the agent successfully reduces the distance to the goal:

$$R_{mono} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}(\Delta D_t > 0) \tag{12}$$

where $\mathbb{I}$ is the indicator function. A high R_{mono} indicates a persistent, goal-oriented heading.

D.5 Survival Analysis of Exit Proximity

To evaluate the temporal efficiency of navigation, we model the probability of an agent not reaching the exit zone using a survival analysis framework. An agent is considered to have survived (failed to escape) at normalized time t if its distance to the exit remains above a critical threshold τ (e.g., $1.5m$). The survival function $S(t)$ is given by:

$$S(t) = P(D(t) > \tau) \tag{13}$$

By analyzing the decay rate of $S(t)$ and the resulting Area Under the Curve (AUC), we can statistically determine—via Wilcoxon signed-rank tests—whether auditory guidance significantly accelerates the arrival of the agent at the exit zone compared to baseline conditions.

D.6 Calculation of distance between trajectories

Jensen-Shannon Distance To quantify the dissimilarity between two trajectories P and Q of arbitrary lengths, we first discretize the room into a grid of $M \times M$ cells. Each trajectory is transformed into a discrete probability distribution, $\mathcal{P}$ and $\mathcal{Q}$, where the value of each cell represents the normalized residence steps of the agent within that spatial bin.

The divergence between these distributions is measured using the Jensen-Shannon (JS) divergence, which is a symmetrized and smoothed version of the Kullback-Leibler (KL) divergence. It is defined as:

$$D_{JS}(\mathcal{P} \| \mathcal{Q}) = \frac{1}{2} D_{KL}(\mathcal{P} \| \mathcal{M}) + \frac{1}{2} D_{KL}(\mathcal{Q} \| \mathcal{M}) \tag{14}$$

where $\mathcal{M} = \frac{1}{2}(\mathcal{P} + \mathcal{Q})$ is the average distribution. The D_{KL} term represents the relative entropy:

$$D_{KL}(\mathcal{P} \| \mathcal{M}) = \sum_{i=1}^{M^2} \mathcal{P}_i \log \left(\frac{\mathcal{P}_i}{\mathcal{M}_i} \right) \tag{15}$$

In our analysis, we utilize the **Jensen-Shannon distance**, defined as $d_{JS} = \sqrt{D_{JS}}$, which satisfies the

triangle inequality. Unlike simple Euclidean distance among discrete points, d_{JS} captures the complete topological difference between paths, **making it highly sensitive to detours, wall-following behaviors, and local search patterns**.

L1 Distance Use the same definition of discretizing the room into a grid of $M \times M$ cells, and denote p_c and q_c as the probability the trajectory of the agent falls into each cell. The L_1 distance is then calculated as follows:

$$L_1(\mathcal{P}, \mathcal{Q}) = \sum_c |p_c - q_c| \tag{16}$$

E Hypothesis tests

We performed a series of hypothesis tests to examine differences across various conditions (e.g., success vs. failure, audio vs. non-audio) and metrics (e.g., Hii_{norm}, path density). The following sections provide a detailed description of the specific tests employed. A comprehensive mapping of the hypothesis tests is presented in Table 6.

E.1 Mann-Whitney U Test

The Mann-Whitney U test (also known as the Wilcoxon rank-sum test) is a non-parametric statistical method used to determine whether there is a significant difference between the distributions of two independent groups. Unlike the t-test, it does not assume a normal distribution of the data. The calculation involves the following steps:

1. **Ranking:** Combine all observations from both groups and rank them by ascending order, assigning average ranks in the case of ties.

2. **Test Statistic (U):** The U statistic is calculated based on the sum of ranks for each group. For a group with size n_1 and rank sum R_1, the statistic is defined as:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{17}$$

3. **Evaluation:** The smaller of the U values from the two groups is compared against a critical value table or used to calculate a p-value. A small p-value suggests that one distribution is stochastically greater than the other.

E.2 Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a non-parametric alternative to the paired t-test, designed to compare two related samples or repeated measurements on a single sample to assess whether their population mean ranks differ. The procedure is as follows:

1. **Difference Calculation:** For each pair i, calculate the difference $d_i = x_{1,i} - x_{2,i}$. Pairs with $d_i = 0$ are typically excluded based on the "Wilcox" zero-handling method.

2. **Ranking of Absolute Differences:** Rank the absolute values $|d_i|$ from smallest to largest.

3. **Signed Ranks:** Attach the sign of the original difference to each rank. Let W^+ be the sum of ranks where $d_i > 0$, and W^- be the sum of ranks where $d_i < 0$.

4. **Test Statistic (W):** The test statistic W is the smaller of the two sums calculated before: $W = \min(W^+, W^-)$. The p-value is then derived from the distribution of W, indicating whether the median of the differences significantly deviates from zero.

E.3 Permutation-based Resampling Test

When standard rank-based tests are unavailable, a permutation-based resampling approach is employed as a robust estimation of the p-value. This method does not rely on any specific distribution parameters. The process involves:

1. **Observed Statistic:** Calculate the observed difference in means $\Delta\mu_{obs} = |\bar{x}_a - \bar{x}_b|$.

2. **Shuffling:** Under the null hypothesis that the two groups are identical, the group labels are randomly permuted $N = 5000$ times. For each permutation, a synthetic difference $\Delta\mu_{perm}$ is computed.

3. **Probability Estimation:** The p-value is estimated as the proportion of permutations where $\Delta\mu_{perm} \geq \Delta\mu_{obs}$:

$$p = \frac{\sum \mathbb{I}(\Delta\mu_{perm} \geq \Delta\mu_{obs}) + 1}{N + 1} \tag{18}$$

where $\mathbb{I}$ is the indicator function. The "+1" serves as a pseudo-count to prevent a p-value of zero.

E.4 Group-level Permutation Test

To determine if the observed spatial distribution difference between two independent cohorts (e.g., successful vs. failed runs) is statistically significant,

we perform a non-parametric permutation test. Let $\bar{\mathcal{P}}_A$ and $\bar{\mathcal{P}}_B$ be the mean occupancy probability maps of the two groups. The observed test statistic is defined as the JS distance between these mean distributions: $S_{obs} = d_{JS}(\bar{\mathcal{P}}_A, \bar{\mathcal{P}}_B)$.

Under the null hypothesis that group membership has no effect on spatial behavior, we pool all runs and randomly redistribute them into two surrogate groups of sizes n_A and n_B. This process is repeated for $N = 3000$ iterations to construct a null distribution of the test statistic S_{perm}. The p-value is calculated as the proportion of permutations where $S_{perm} \geq S_{obs}$:

$$p = \frac{\sum_{i=1}^{N} \mathbb{I}(S_{perm,i} \geq S_{obs}) + 1}{N + 1} \tag{19}$$

where $\mathbb{I}(\cdot)$ is the indicator function. This way to construct the test statistic ensures that the significance is derived from the collective spatial pattern of the group rather than being biased by individual outlier trajectories.

E.5 Paired Sign-flip Test for Stimulus Response

To isolate the direct impact of acoustic stimuli while controlling for individual agent variance, we employ a paired sign-flip test. For each run r, we compute the paired difference in occupancy: $\Delta_r = \mathcal{P}_{audio,r} - \mathcal{P}_{non_audio,r}$. This represents the spatial shift caused by the audio signal within run.

The null hypothesis states that the audio trigger has no systematic effect on the path of the agent, implying that the direction of the shift Δ_r is purely stochastic. To test this, we generate $N = 3000$ randomized samples where each difference map is assigned a random sign $s_r \in \{+1, -1\}$:

$$\bar{\Delta}_{perm} = \frac{1}{R} \sum_{r=1}^{R} s_r \cdot \Delta_r \tag{20}$$

The significance is determined by comparing the observed aggregate JS distance between the audio and non-audio conditions against the distribution of distances generated under random sign assignments. This paired approach is highly sensitive to subtle, consistent behavioral transitions that might be masked in a standard group-wise comparison.

E.6 Mantel Test

The Mantel test is a statistical procedure used to evaluate the correlation between two symmetric distance (or dissimilarity) matrices, D_X and D_Y, derived from the same set of n samples. In the context of spatial navigation, this test determines whether the dissimilarity in agent trajectories correlates with the dissimilarity in environmental configurations (e.g., target prop positions). The procedure is defined as follows:

1. **Matrix Construction:** Two matrices are constructed such that $D_{X(i,j)}$ represents the distance between the i-th and j-th agent trajectories (i.e. JS distance of occupancy probabilities), and $D_{Y(i,j)}$ represents the Euclidean distance between the corresponding target object positions for those runs.

2. **Observed Statistic (r_M):** The Mantel correlation coefficient is computed using the standardized Pearson or Spearman correlation between the corresponding off-diagonal elements of the two matrices:

$$r_M = \frac{1}{m-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{std}(D_{X(i,j)}) \cdot \text{std}(D_{Y(i,j)}) \tag{21}$$

   where $m = n(n-1)/2$ is the number of unique pairs.

3. **Permutation Procedure:** Since elements within distance matrices are not independent, a permutation-based approach is required for significance testing. The rows and columns of one matrix (e.g., D_X) are randomly shuffled $N = 3000$ times while keeping the other matrix D_Y fixed. For each shuffle k, a permuted statistic $r_{M,k}$ is calculated.

4. **Significance Evaluation:** The p-value is defined as the proportion of permutations where the absolute value of the permuted statistic equals or exceeds the observed statistic:

$$p = \frac{\sum \mathbb{I}(|r_{M,k}| \geq |r_{M,obs}|) + 1}{N + 1} \tag{22}$$

   A significant p-value indicates that the spatial distribution of the trajectories is non-randomly associated with the positioning of environmental landmarks.

F Detailed Analysis

This section provides the full analysis condensed in the main-text Analysis section.

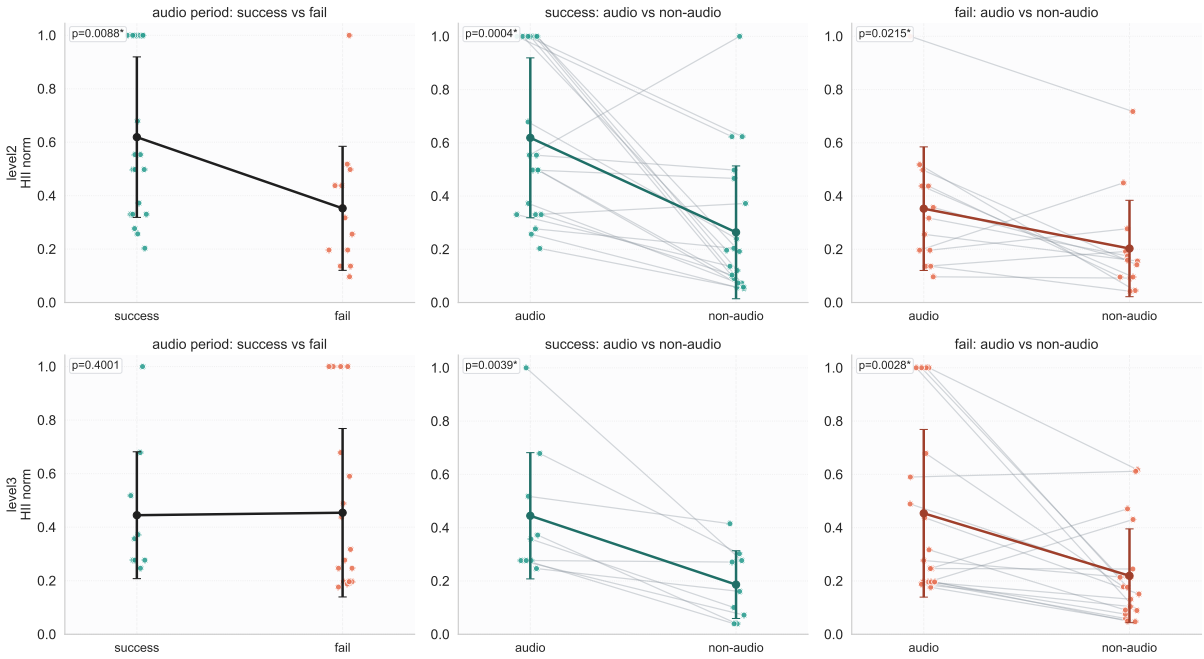

Figure 4: **Paired HII comparison under auditory conditioning for Difficulty-2 and Difficulty-3.** Each row corresponds to one difficulty level. Left column: unpaired success vs fail comparison in audio-active segments. Middle and right columns: paired audio vs non-audio comparison within success and fail runs. Each dot is one run-level $\mathrm{HII}_{\mathrm{norm}}$; lines connect paired points from the same run; error bars show mean $\pm$ standard deviation. The figure shows consistent concentration increase during audio-active periods, while success vs fail separation in audio periods is significant in Difficulty-2 and not significant in Difficulty-3.

Test	HII				JS Distance		Visual Effects
Difficulty	$P^{s}_{A \text{ vs } NA}$	$P^{f}_{A \text{ vs } NA}$	$P^{s/f}_{aud}$	$P^{s/f}_{non-aud}$	$P^{\mathrm{sf}}_{\mathrm{L}_1,\,\mathrm{A \text{ vs } NA},\,S}$	$P^{\mathrm{sf}}_{\mathrm{JSD},\,\mathrm{A \text{ vs } NA},\,F}$	$P^{\mathrm{Mantel}}_{\mathrm{box}\leftrightarrow\mathrm{path}}$
Difficulty-1	–	–	–	0.3612	–	–	–
Difficulty-2	**0.0004***	**0.0215***	**0.0088***	0.2388	**0.00033***	**0.00033***	–
Difficulty-3	**0.0039***	**0.0028***	0.4001	0.3568	**0.00033***	**0.00033***	**0.0047***

* Note: $P < 0.05$ indicates a statistically significant difference.

Table 5: **P-values for the hypothesis tests used**.

F.1 Analysis of Exploration Behavior and Path

To investigate the underlying exploration behavior, navigation strategies, and the impact of cross-modal cues on decision-making, we visualize the movement patterns using exit-aligned path density heatmaps in Figure 3. We rotate each trajectory by 90° increments so that the target exit is consistently aligned at the top edge for easier comparison, and then aggregate the spatial distributions of successful and failed runs across different level families.

In Difficulty-1, successful escape exhibit a direct and concentrated path towards the exit, whereas failed runs show sparse exploration, suggesting a lack of basic spatial grounding. As complexity increases in Difficulty-2 and -3, the density maps reveal a strategic shift, where models must deviate from the exit-ward path to interact with key items (e.g., recorders or boxes), as indicated by the dashed contour overlays. Note that successful runs in these levels show higher density around these key items compared to failed runs, indicating more effective active perception and target-oriented exploration. The lower ER in more complex levels correlate with more fragmented and exploratory path distributions.

A significant observation in failed runs for Difficulty-2 and -3 is that these runs are often trapped near the center or far-flung corners, where models often wander aimlessly or fail to transition from auditory-based exploration to final escape. The discrepancy between success and failure is

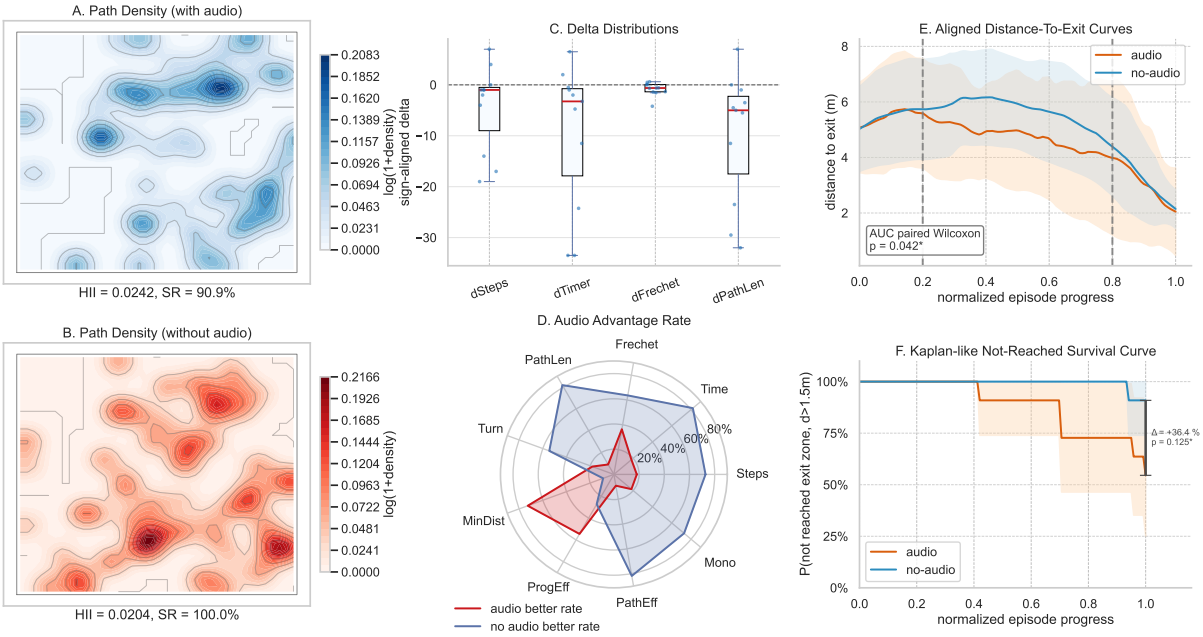

Figure 5: **Analysis of Ambient Sound via Audio and No-Audio settings.** (A-B) Path density heatmaps showing trajectory concentrations, using the same spatial aligning method as in Figure 3. (C) Delta distributions for sign-aligned metrics (*Steps, Timer, Frechet, PathLen*); the red median lines below zero indicate the superiority of the audio condition. (D) The winning rate of audio guidance across nine navigation metrics. Metrics where lower values are better: *Steps*, *Time*, *PathLen*, *Turn*, and *Frechet*. Efficiency metrics where higher values are preferred: *ProgEff*, *PathEff*, and *Mono*. (E) Mean distance-to-exit over normalized episode progress; the shaded area represents ± 1 standard deviation. (F) Kaplan-like survival plot showing the percentage of models that have not yet reached the 1.5m exit threshold. Statistical significance for AUC and endpoint success is determined via paired Wilcoxon tests ($*p < 0.05$ indicates a statistically significant difference).

most pronounced in Difficulty-3, where successful runs demonstrate a clear multi-hop traversal pattern, e.g., first localizing the auditory clue (recorder) and then proceeding to the exit, while failed ones often miss the crucial temporal or spatial trigger zones of these objects.

F.2 Ablation Study of Ambient Sound

To analyze the specific impact of ambient sound on active perception, we conduct an ablation study using Difficulty-1, which is uniquely suited for the analysis as it requires no complex prop chains or specific clues (e.g., passwords) to escape, relying solely on the ability to navigate toward the exit. By comparing the standard Difficulty-1 setting with a spatialized ambient sound source at the exit with a silent version, the effect of auditory cues on spatial reasoning and decision-making under uncertainty can be isolated.

We conduct analysis across nine metrics, including burden-related measures such as *Steps*, *Time*, *PathLen* (the total actual path length), *Turn* (turning back toward the opposite direction), and *Frechet*, as well as efficiency indicators including *ProgEff* (Progress Efficiency), *PathEff* (Path Efficiency), *Mono* (Monotonicity Ratio of trajectory), and *MinDist* (the closest proximity achieved to the exit). Together, these metrics characterize exploration behavior, path optimality, and progress toward the goal. Detailed explanations on these measurements can be found in Appendix D.

Results and Analysis Figure 5 compares the results between the Audio and No-Audio conditions in Difficulty-1. It shows that spatialized ambient sound significantly improves active perception and navigational efficiency. As illustrated in the Delta Distributions (Figure 5 C), auditory cues lead to a substantial reduction in *Steps*, *Time*, *PathLen*. Path quality is further evaluated using the *Frechet*. The Non-Audio group exhibits consistently higher Frechet values compared to the Audio group, reflecting significantly more wandering and path distortion. Correspondingly, the median deltas (Δ = Audio − No-Audio) for all these burden-related metrics fall well below the zero baseline, confirming that auditory feedback helps bypass redundant search behaviors and act

with greater decisiveness and efficiency.

The performance gap is further highlighted by the Audio Advantage Radar Chart (Figure 5 D), which shows the win rate of audio-guided settings across nine metrics. In this chart, the audio group presents lower burden in burden-related metrics *Steps*, *Time*, *PathLen*, *Turn*, and *Frechet*, and achieves higher efficiency in *ProgEff*, *PathEff*, and *Mono*, suggesting a strong audio benefit. The high better-rate in *MinDist* confirms that audio-guided agents not only move efficiently but also approach the exit more successfully. Overall, audio improves performance in over 60% of trials across nearly all dimensions, demonstrating robust multi-faceted advantages for navigation.

Temporal behavior analysis using Aligned Distance-to-Exit curves (Figure 5 E) shows the Audio group consistently closer to the exit between the central portion of the episode (0.2–0.8), with lower variance than the No-Audio group. A Paired Wilcoxon Signed-Rank Test on the AUC confirms the statistical significance ($P^{\text{AUC}}_{\text{A vs NA}} = 0.042^* < 0.05$), indicating that models actively leverage auditory cues rather than random variation. Similar trends appear in Kaplan-like Not-Reached survival curves in Figure 5 F, where the Audio curve decays faster, implying earlier entry into the exit zone. Another Paired Wilcoxon Test on terminal non-arrival rates ($P^{\text{end}}_{\text{A vs NA}} = 0.0125^* < 0.05$) further confirms that our ambient sound accelerates the transition from exploration to goal-oriented navigation with significance.

To test the comprehensive behavior through the whole path, a point-label permutation test was conducted on the path density. By dividing the room into a $M \times M$ grid, we get the path density distribution $\mathcal{P}_{aud}$ and $\mathcal{P}_{non-aud}$. Since the levels of audio and non-audio are exactly the same, we can use both L_1 and JS distance to calculate the gap between the path density of each level. The point-label permutation test reveals significant difference between the path density ($P^{\text{perm}}_{\text{L1}} = 0.00019^* < 0.05, P^{\text{perm}}_{\text{JSD}} = 0.00019^* < 0.05$), which confirms that the ambient sounds had a significant impact on the choice of agent path.

F.3 Influence of Auditory Clues

This subsection investigates how auditory clues such as password emitted by recorder would affect the trajectory of models. In this analysis, the trajectory is processed using the same steps in Section F.1. This normalization makes trajectories comparable across levels with different settings and runs under a shared coordinate frame.

The concentration of exploration is quantified by the normalized Herfindahl-Hirschman Index (HII), adapted from the Herfindahl concentration index in economics. We divide the room into $B \times B$ occupancy grid, and let p_c denote the probability mass of points on trajectories in cell c, with $\sum_c p_c = 1$ and $C = B^2$. The raw index is $\text{HII}_{\text{raw}} = \sum_{c=1}^{C} p_c^2$, and the normalized form is $\text{HII}_{\text{norm}} = \frac{\text{HII}_{\text{raw}} - 1/C}{1 - 1/C} \in [0, 1]$. A larger HII_{norm} indicates that trajectory occupancy is concentrated in fewer regions, while a smaller value indicates more spatially dispersed exploration. A detailed description for HII can be found in Appendix D.

Two complementary hypothesis tests are designed to separate *within-run auditory effects* from *between-outcome differences*. (i) A vs NA (paired, within run): each run provides two matched observations, the audio-active segment and the non-audio segment. The test target is whether the typical within-run difference ($\Delta_{\text{run}} = \text{HII}_{\text{audio}} - \text{HII}_{\text{non-audio}}$) is centered at zero. A paired two-sided Wilcoxon signed-rank test is used because the two observations are dependent and come from the same trajectory context. This design controls run-level heterogeneity and directly tests the effect of auditory condition on concentration. (ii) S vs F (unpaired, between runs): success and fail runs are independent groups, so that run-level HII distributions are compared with a two-sided Mann-Whitney U test.

The test is performed under three segment definitions (full steps, steps with audio playing, steps with no audio) to identify where outcome separation appears. This setup tests whether exploration concentration differs by final outcome, and whether that difference is specifically amplified during audio-active periods rather than a global trajectory difference.

Results from Table 5 show that **audio-active segments have significantly higher HII than non-audio segments across outcome classes, indicating more concentrated exploration with auditory cues**, indicating more concentrated exploration under auditory clues (Difficulty-2: $p = 0.0004^*$ for success and $p = 0.0215^*$ for fail; Difficulty-3: $p = 0.0039^*$ for success and $p = 0.0028^*$ for fail). The detailed run-level scatter-and-paired-line distributions are provided in Figure 4 and can further prove the consistency differ-

Test Symbol	Test Type	What Is Being Tested
$P^{\text{AUC}}_{\text{A vs NA}}$	Paired Wilcoxon signed-rank Appendix E.2	Whether audio changes the mid-trajectory distance-to-exit behavior between Difficulty-1 audio and non-audio settings (AUC on normalized progress window).
$P^{\text{end}}_{\text{A vs NA}}$	Paired Wilcoxon signed-rank Appendix E.2	Whether audio changes endpoint not-reached status between Difficulty-1 audio and non-audio settings (Kaplan-like endpoint comparison).
$P^{\text{perm}}_{\text{L1}}$, $P^{\text{perm}}_{\text{JSD}}$	Point-label permutation test Appendix E.3	Whether the density between difficulty-1 audio/non-audio settings changes significantly using L_1 and JS distance.
$P^{\text{A vs NA}}_{\text{S/F}}$	Paired Wilcoxon signed-rank Appendix E.2	Within successful/failed runs, whether HII behaviors change among audio/non-audio segments.
$P^{s/f}_{aud}/P^{s/f}_{non\text{-}aud}$	Mann-Whitney U test Appendix E.1	Whether audio-active-segment/non-audio-segment HII differs between success and fail runs.
$P^{\text{sf}}_{\text{JSD, A vs NA}, S/F}$	Paired sign-flip permutation Appendix E.5	Whether density between audio and non-audio segments changes significantly within successful/failed runs using JS distance.
$P^{\text{Mantel}}_{\text{box}\leftrightarrow\text{path}}$	Mantel permutation test Appendix E.6	Whether run-to-run box-position distance is associated with run-to-run path-distribution distance.

Table 6: Tests used for analysis.

ences in path selection strategies. For S vs F, a significant gap appears only in audio segments of Difficulty-2 ($p = 0.0088^*$), with success runs more concentrated than fail runs. However, Difficulty-3 shows no significant S vs F gap in audio segments ($p = 0.4001$). A plausible explanation is the additional visual key object in Difficulty-3, which dilutes isolated auditory contrast.

To verify this interpretation, a dedicated Difficulty-3 box-position analysis is conducted with a Mantel permutation test. This test is to answer if box location changes across runs, does agent path pattern change with it. The analysis compares path differences and box-location differences across runs, and also checks whether path centroids move with box coordinates. For each run, full-path distribution distance is computed by Jensen-Shannon distance between run histograms, and box-position distance is computed by Euclidean distance between aligned box coordinates. Additional robustness tests evaluate whether the audio-related shift is stable rather than random noise. The result is significant ($r = 0.1548$, $p = 0.0047^*$), showing that visual clue position is associated with path variation and affects Difficulty-3 auditory comparisons. This evidence supports strong cross-modal coupling in EscapeCraft4D.

Further, to confirm that **audio cues systematically affect behavior rather than reflecting random variation**, trajectory-level density distributions were analyzed with permutation tests. Jensen-Shannon distance and L_1 effect size were used, and paired audio vs non-audio comparisons applied sign-flip permutations on within-run density differences. Validation against thousands of randomized permutations shows that observed shifts are highly unlikely by chance, with $P^{s}_{A \text{ vs } NA}$ and $P^{f}_{A \text{ vs } NA}$ reaching the permutation lower bound 0.0003^*, demonstrating a strong and robust auditory impact on exploration strategies. All the tests we have used for analysis are listed in Table 6. For a detailed account of the statistical tests, see Appendix E.

G Prompt Template

We follow (Wang et al., 2025b) and define the following prompts.

System Prompt Compare to (Wang et al., 2025b), we added one more action space for auditory modality to the system prompt. The complete prompt is shown in table 7.

Instruction Prompt

You find yourself locked inside a room, and your ultimate goal is to escape the room, i.e., the room escape game.
You can explore the room, interact with objects, inspect items, and resolve puzzles. You can adopt the following actions to explore the room and interact with objects:

Operation Prompt

- move_forward: float, ranged between [-10, 10]. This is the number of meters you want to move forward. Negative value means moving backward.
- rotate_right: float, ranged between [-180, 180]. This is the number of degrees you want to turn right. Negative value means turning left.
- rotate_down: float, ranged between [-90, 90]. This is the angle you want to adjust your view vertically. Positive value means looking downward, while a negative value means looking upward. Angle 0 means looking straight ahead.
- jump: bool, whether you want to jump. This can be used together with moving forward, e.g., True represents the action "to jump".
- look_at: list[x: float, y: float], the range of x and y is [0, 1]. This parameter is the coordinates of the point in the image you want to look at. For reference, the coordinates of the upper left corner of the scene are (0, 0), and the coordinates of the lower right corner are (1, 1). Also note that there are no clues on the ceiling.
- grab: bool, whether you require to interact with the object located exactly at the center of the scene, marked by a red dot. For example, to grab the key or to interact with/open a box at the center of the scene, set grab=True. The red dot assists in locating the object you require to interact with. You might need to adjust the view or move closer to ensure the red dot is on your target object through the rotate_right, rotate_down, and move_forward actions. To successfully grab an object, you should center the object via the red dot and be within a certain distance to it. If the grabbing fails, try moving closer toward the object. If it fails multiple times at the same position, you should be aware that not all objects are interactable; do not get stuck in an uninteractable position.
- trigger: bool, whether you want to attempt to trigger the sound of a sound-emitting object, e.g., recorder/radiogram/musicbox, at the center of the scene. This ONLY works when you are close enough and facing it; otherwise, it will fail and you should move closer or adjust your view.
- interactions: dict:{"use_item_id": str, this is the item_id you require to view or use. When used together with grab=True, it means to use this item to interact with the target object you want to grab, e.g., using the item_id of the key to open the door in the scene. "input": str, this is the message you want to input when interacting with the center object}.
- read: str, this is the item_id that you want to get detailed information from your bag.
- rationale: str, represents the rationale of your action. This should explain your decision-making process and help the agent understand your thinking process.

You need to return data in the following format of JSON_string to interact with the scene:

```
{
    "move_forward": float,
    "rotate_right": float,
    "rotate_down": float,
    "jump": bool,
    "look_at": [x: float, y: float],
    "grab": bool,
    "trigger": bool,
    "interactions": {
        "use_item_id": str,
```

Continued on next page

The System Prompt continued

```
        "input": str
    },
    "read": str,
    "rationale": str
}
```

All of the above operations are optional. If no value is passed in, the interactive operation will not be performed.
You must follow the above instructions and do not say anything else except for the JSON_string of operations.

Table 7: The System Prompt

Step Prompt The complete prompt is shown in table 8.

Interaction Result
{interaction_result}
===
Inventory
The items in your bag usable include:
{bag_desc}
===
Step Prompt
Please determine the next action(s) that could help you observe the room or obtain useful tools or clues.
If you find yourself stuck in a corner, try turn around by passing rotate_right.
You need to return data in the following format of JSON_string to interact with the scene and don't say anything else:
`{`
`“move_forward”: float,`
`“rotate_right”: float,`
`“rotate_down”: float,`
`“jump”: bool,`
`“look_at”: [x: float, y: float],`
`“grab”: bool,`
`“trigger”: bool,`
`“interactions”: {`
`“use_item_id”: str,`
`“input”: str`
`},`
`“read”: str,`
`“rationale”: str`
`}`

Table 8: The Step Prompt

Prompt for Consistency Evaluation Compared to (Wang et al., 2025b), we found that prevailing models seem to prefer using more rationale to describe how to adjust the perspective, thus interfering with the judgement of the scoring model. Therefore, we added an additional special case to accommodate this situation. The complete prompt is shown in table 9.

Consistency Evaluation Prompt

You are a reasoning consistency evaluator for a multimodal agent benchmark. Your task is to determine whether the agent's intention described in the "rationale" matches the actual interaction result described in the "response".
Specifically, you are given:

- The agent's **rationale** for why it attempted an interaction, describing its goal or belief about the current environment.
- The **response** from the environment after the interaction, which includes the actual result, e.g., what item was interacted with and what was obtained.

Your goal is to determine whether the object the agent intended to interact with **matches** the object that was actually interacted with according to the response.
If the agent tried to interact with object A, e.g., a microwave, but the response shows interaction with object B, e.g., a box, and B was not the intended target, this is considered an **inconsistent interaction**, i.e., accidental success or misaligned action.
Special Cases: if the response is "Escaped successfully!", you should check whether the agent's rationale explicitly or implicitly indicates the intention to escape, e.g., trying to open the door to leave. In another special case, if the agent is trying to adjust its view to align with the door, and it has triggered the information of the door, this is considered consistent. If not, treat it as inconsistent.
You must give your feedback in the following JSON-string format and **DON'T** say anything else:

```
{
    "Consistency": 1 | 0
}
```

Where:
- `1` means the rationale and interaction are consistent, i.e., aligned.
- `0` means the interaction appears to be accidental, mismatched, or unintended.

Respond **only** with `Consistency: 1` or `Consistency: 0`.

—

Example 1:
rationale: *Moving closer to the microwave to try interacting with it one last time using '1264'. If this doesn't work, I'll need to explore other areas.*
response: *You used the correct password to unlock the box... You did not interact with any objects in the last step.*
Expected output: `Consistency: 0`

Example 2:
rationale: *I'll try using the key I just picked up on the door. Let's see if I can escape now.*
response: *Escaped successfully!*
Expected output: `Consistency: 1`

Example 3:
rationale: *I want to check if the small box has any useful items inside.*
response: *You opened the box and found a screwdriver.*
Expected output: `Consistency: 1`

Please score the following interaction:
rationale: `{rationale}`
response(s): `{response}`

Table 9: The Consistency Evaluation Prompt