TRACEAV-BENCH: BENCHMARKING MULTI-HOP TRAJECTORY REASONING OVER LONG AUDIO-VISUAL VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

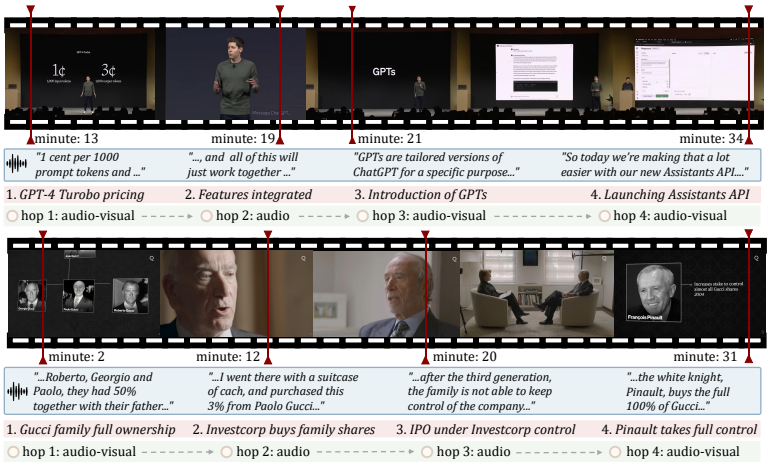
Real-world audio-visual understanding requires chaining evidence that is sparse, temporally dispersed, and split across the visual and auditory streams, whereas existing benchmarks largely fail to evaluate this capability. They restrict videos to short clips, isolate modalities, or reduce questions to one-hop perception. We introduce **TraceAV-Bench**, the first benchmark to jointly evaluate ***multi-hop reasoning over long audio-visual trajectories*** and ***multimodal hallucination robustness***. TraceAV-Bench comprises 2,200 rigorously validated multiple-choice questions over 578 long videos, totaling 339.5 hours, spanning 4 evaluation dimensions and 15 sub-tasks. Each question is grounded in an explicit multi-hop evidence trajectory that averages 3.68 hops across a 15.1-minute temporal span. The dataset is built by a three-step pipeline followed by a strict quality assurance process. Evaluation of multiple representative OmniLLMs on TraceAV-Bench reveals that the benchmark poses a persistent challenge across all models, with the strongest closed-source model (Gemini 3.1 Pro) reaching only 68.29% on general tasks, and the best open-source model under direct inference (Ming-Flash-Omni-2.0) reaching 51.70%, leaving substantial headroom. Further analysis shows that current models struggle to gather and combine evidence across long audio-visual videos, with performance varying across modalities and reasoning tasks. We hope TraceAV-Bench will facilitate progress toward reliable long-form audio-visual reasoning in OmniLLMs. The code and data are available at `https://anonymous.4open.science/r/TraceAV-Bench-Submission-5FD4`.

1 INTRODUCTION

The rapid advancement of Multimodal Large Language Models (MLLMs) Bai et al. (2025); Yao et al. (2024); Yin et al. (2024); Team et al. (2025a) has broadened the perceptual horizon of language models, enabling them to process visual Zhu et al. (2025); Steiner et al. (2024) and auditory Team (2026a); Shi et al. (2026) information far beyond the boundaries of language alone. Building upon this foundation, Omnimodal Large Language Models (OmniLLMs) Team (2026b); AI et al. (2025); Zhao et al. (2025); Team et al. (2025b); Tang et al. (2025); Li et al. (2024a); Comanici et al. (2025) have recently emerged, designed to jointly perceive and reason over real-world text, vision, and audio within a unified framework, achieving competitive performance across tasks such as audio-visual understanding and cross-modal reasoning Han et al. (2024); Wang et al. (2025f); Li et al. (2026b). This capability is fundamental for interacting with the physical world in a human-like manner, in which information from different modalities is simultaneously integrated.

However, despite these promising results, bringing OmniLLMs into genuine real-world utility reveals several unresolved challenges. ***1) Long-form multi-hop trajectory reasoning*** poses the most critical bottleneck Tao et al. (2026); Wang et al. (2025c); Fu et al. (2025); Chen et al. (2026b). While OmniLLMs can handle simple, short-clip perception tasks, real-world scenarios often demand chaining clues scattered across tens of minutes of continuous audio-visual content. This requires reasoning from a *trajectory* of temporally dispersed evidence spanning multiple events, rather than a single moment. ***2) Cross-modal information fusion*** remains equally unresolved Zhou et al. (2025); Li et al. (2025). A model that can process audio and visual streams does not necessarily know how to jointly leverage their complementary information under complex conditions, particularly when

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Audio-Visual: Temporal Sequencing
Question 1: Across the talk, the Speakers makes four major announcements involving GPT-4 Turbo pricing cuts, an upgraded ChatGPT with integrated features, the introduction of GPTs, and the launch of the Assistants API. Which option shows the correct chronological order in which he announces these four items in the video?"
Answer: C: "GPT-4 Turbo pricing cuts→ upgraded ChatGPT with integrated features→ introduction of GPTs→ launch of the Assistants API"

Hallunciation: Temporal Splicing Fallacy
Question 2: Consider the following claim based on the video: "After Maurizio Gucci's murder in 1995, the family regained control after Investcorp, continued through the IPO, and remained effectively in control even after François Pinault's acquisition." Based on the actual timeline shown in the video, how does this claim compare to what really happens?
Answer: A: "It is incorrect: by the time of Maurizio Gucci's murder and throughout the later IPO and Pinault's 100% acquisition, control had already passed out of Gucci family hands to Investcorp and then to Pinault, and it never re-consolidated as long-term family control."

Figure 1: Illustrative examples from TraceAV-Bench, each grounded in an explicit multi-hop evidence trajectory whose hops are tagged with their source modality. The first example (*Temporal Sequencing*, AVR) requires chronologically ordering four events by chaining speech and on-screen cues; the second (*Temporal Splicing Fallacy*, MH) requires rejecting a fabricated narrative that splices temporally isolated events into a false ownership timeline. Both showcase the main idea of TraceAV-Bench: composing evidence from temporally dispersed, cross-modal clues over long videos.

evidence in one modality is only interpretable in the context of the other. ***3) Multimodal hallucination*** further compounds these difficulties Bai et al. (2024); Sung-Bin et al. (2024); Xing et al. (2026a). When jointly processing mixed audio-visual inputs, OmniLLMs sometimes generate responses that are inconsistent with the input or not grounded in the observed content, raising serious reliability concerns. Critically, these challenges remain not only largely unsolved, but also insufficiently studied. Existing benchmarks mostly operate on short video clips of at most a few minutes, evaluate visual and audio modalities in isolation, or restrict questions to shallow single-hop inference, failing to comprehensively evaluate these challenges Li et al. (2024b); Gong et al. (2024).

To this end, we introduce **TraceAV-Bench**, a comprehensive benchmark designed to evaluate OmniLLMs on multi-hop reasoning over long audio-visual trajectories. TraceAV-Bench is built upon 578 long videos totaling 339.5 hours, spanning diverse genres and multiple languages, and it is, to the best of our knowledge, the ***first*** benchmark to simultaneously require multi-hop trajectory reasoning and assess multimodal hallucination robustness in the context of long audio-visual content. Via a three-step data synthesis pipeline, we construct 2,200 rigorously validated multiple-choice questions (MCQs) across 4 evaluation dimensions and 15 sub-tasks, covering ***Audio-Visual Joint Reasoning***, ***Visual-Centric Reasoning***, ***Audio-Centric Reasoning***, and ***Multimodal Hallucination***. Every question in TraceAV-Bench is grounded in an explicitly annotated multi-hop evidence trajectory (as showcased in Figure 1), where each reasoning chain spans at least two temporally non-adjacent events, and averages 3.68 hops across a temporal span of 15.1 minutes.

Our main contributions can be summarized as follows:

- ***The TraceAV-Bench Benchmark.*** We present a large-scale, carefully curated benchmark that extends the evaluation frontier for OmniLLMs over long multi-hop trajectories, providing the research community with a tool to rigorously diagnose model capabilities beyond existing benchmarks.
- ***A Data Construction Pipeline.*** We propose a three-step pipeline that decouples visual captioning, asynchronous audio-visual fusion, and LLM-based question generation to organize asynchronous evidence from long videos into logically coherent, trajectory-grounded MCQs at scale. A rigorous multi-stage quality assurance process further ensures that every retained question is of high quality.
- ***Comprehensive Empirical Analysis with Key Findings.*** We benchmark state-of-the-art open-source and closed-source OmniLLMs and MLLMs, and find that all evaluated models struggle on TraceAV-Bench. When answering multi-hop questions over long videos, these models have difficulty gathering and combining the required audio and visual evidence from different parts of the video to reach the correct answer. Our analysis further shows that these limitations vary across modalities and reasoning operations.

Benchmark	Modality	Duration	Avg. Length	Multi-hop	Multi-domain	Hallucination	Anno.	#Q
Short-clip benchmarks (≤ 60 ***s***)								
AVQA Yang et al. (2022)	V+A	∼10 s	∼10 s	×	×	×	M	57,335
Music-AVQA Li et al. (2022)	V+A	∼60 s	∼60 s	×	×	×	M	45,867
AVTRUSTBENCH Chowdhury et al. (2025)	V+A	10–60 s	-	×	✓	✓	A&M	600 K
AV-Odyssey Gong et al. (2024)	I+A	/	/	×	✓	×	M	4,555
AVHBench Sung-Bin et al. (2024)	V+A	∼10 s	∼10 s	×	✓	✓	A&M	5,302
OmniBench Li et al. (2024b)	I+A	-	-	×	✓	×	M	1,142
Daily-Omni Zhou et al. (2025)	V+A	30–60 s	42.8 s	×	×	×	A&M	1,197
Long-form video benchmarks (> 60 ***s***)								
WorldSense Hong et al. (2025)	V+A	15–656 s	141.1 s	×	✓	×	M	3,172
JointAVBench Chao et al. (2025)	V+A	-	97.2 s	×	✓	×	A&M	2,853
OmniVideoBench Li et al. (2025)	V+A	4–1955 s	384.2 s	×	✓	×	M	1,500
LVOmniBench Tao et al. (2026)	V+A	613–5482 s	2,070 s	×	✓	×	M	1,014
TraceAV-Bench (Ours)	**V+A**	**606–8394 s** (Ultra-long)	**2,112 s**	✓	✓	✓	**A&M**	**2,200**

Table 1: Comparison of TraceAV-Bench with existing audio-visual benchmarks. V, I, A: video, image, audio. Anno.: M = manual, A&M = automatic + manual.

2 Related Work

2.1 Omnimodal Large Language Models

Real-world perception is inherently multi-sensory, and closing the gap between human and machine understanding requires models that can handle vision, audio, and language. The research community has broadened the perceptual scope of LLMs, giving rise to Vision-Language Models (VLMs) Wang et al. (2025d); Bai et al. (2025); Team (2026b); Marafioti et al. (2025); Guo et al. (2025a); Liu et al. (2025a); Xiaomi & Team (2025), Audio-Language Models (ALMs) Tseng et al. (2025); Ding et al. (2025); Ghosh et al. (2025); Chu et al. (2024); Arora et al. (2025), and more recently OmniLLMs Liu et al. (2025b); Luo et al. (2025); Fu et al. (2024); Guo et al. (2025b); Team (2026b); Tian et al. (2026); Ding et al. (2026) capable of processing text, images, videos, and audio within a single framework. Both open-source models Ye et al. (2025); AI et al. (2025); Zhao et al. (2025); Tang et al. (2025); Xu et al. (2025b;a) and closed-source models such as the Gemini series Comanici et al. (2025), which combine multimodal pre-training with large context windows, have demonstrated strong performance across audio and visual tasks. As these capabilities mature, models are increasingly positioned to handle extended audio-visual inputs. However, rigorous evaluation in this regime remains scarce.

2.2 MLLM Benchmarks

MLLM benchmarks have evolved with model capabilities, progressing from modality-isolated suites for image Liu et al. (2024); Yue et al. (2024); Yu et al. (2023) and audio Sakshi et al. (2024); Wang et al. (2025a;b;e) understanding to specialized axes such as fine-grained perception Ghosh et al. (2026b); Qiang et al. (2025); Weng et al. (2026) and multimodal reasoning Ying et al. (2024); Yao et al. (2025); Du et al. (2025); Li et al. (2026a); Qian et al. (2025); Fu et al. (2026); Rawal et al. (2024). The rise of OmniLLMs has spurred a rich set of audio-visual benchmarks Zhou et al. (2025); Li et al. (2024b); Chen et al. (2026b); Cheng et al. (2025); Chen et al. (2025a); Gong et al. (2024); Zhang et al. (2025). These benchmarks have made important progress in broadening evaluation from isolated perception to joint audio-visual understanding. They provide a strong foundation for studying the capabilities of OmniLLMs. Several recent benchmarks have also extended evaluation to long-form videos Tao et al. (2026); Li et al. (2025); Han et al. (2025). These studies demonstrate the value of realistic extended content and reveal the difficulty of maintaining reliable multimodal understanding over time. Building on these contributions, TraceAV-Bench focuses on a complementary setting in which questions require multi-hop reasoning across temporally distant audio-visual evidence. It also evaluates multimodal hallucination robustness under the long-form setting, as summarized in Table **??**.

3 TraceAV-Bench

3.1 Dataset Collection

To evaluate the multi-hop reasoning capabilities of OmniLLMs over long trajectories in an audio-visual context, we define eligible long videos as those exceeding **10 minutes**, possessing semantically

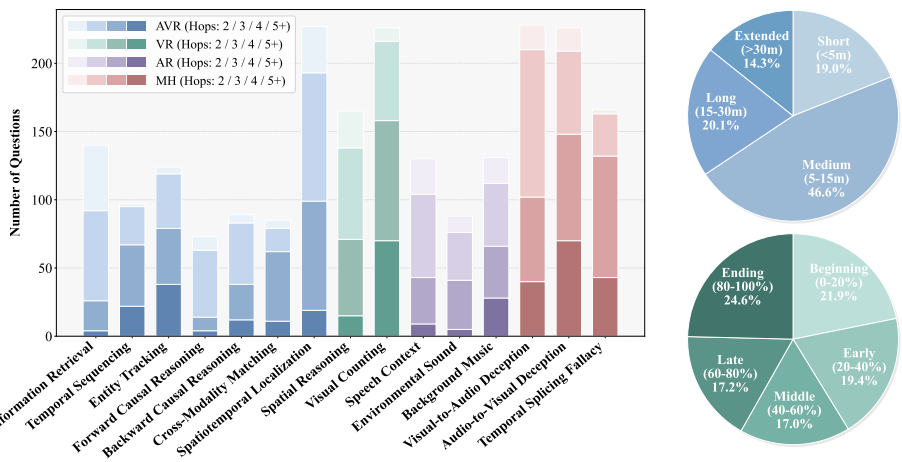


Figure 2: Statistics analysis of TraceAV-Bench. *left*: per-sub-task question counts by evaluation dimension (AVR, VR, AR, MH), stacked by hop count ($2/3/4/5+$). *right top*: distribution of temporal spans of multi-hop evidence trajectories. *right bottom*: distribution of the position for all evidence hops along the video timeline.

diverse and rich scene transitions, and containing aligned, meaningful information in both visual and audio modalities.

Video Collection and Deduplication. Our initial video corpus is mainly sourced from three publicly available video understanding benchmarks: OmniVideoBench Li et al. (2025), LVBench Wang et al. (2025c) and VideoMME Fu et al. (2025), where YouTube is their primary source, there is a natural risk of data overlap. To ensure the uniqueness of our benchmark, we apply a rigorous deduplication process based on video metadata and content hashing, yielding a diverse initial pool of candidate videos.



Figure 3: Video category distribution.

Video Quality Filtering. We further process the candidate videos through a strict filtering pipeline with the following criteria with a half-automated manner:

- ***Visual Dynamics and Event Density:*** Videos must exhibit sufficient visual dynamics, with frequent scene transitions and a steady stream of distinguishable events. Therefore, videos with fewer than 3 scene transitions are directly discarded.
- ***Modality Completeness:*** TraceAV-Bench explicitly evaluates the joint understanding of audio and visual streams. Thus, videos without an audio track or with a mean volume below -50 dB (indicating silence or negligible background noise) are discarded.
- ***Multi-hop Reasoning Potential:*** Our annotators manually review the videos to assess their potential for extracting multi-hop questions. Specifically, videos are retained only if they exhibit high intrinsic complexity, such as multiple interacting entities and linked events.

After executing this filtering process, we finalized a high-quality collection of **578** videos, yielding a total of **339.5** hours of continuous audio-visual content ready for complex multi-hop construction. As shown in Figure 3, the retained videos span a wide range of genres, ensuring broad topical diversity. Notably, the corpus is also multilingual, covering English, Chinese, and other major languages.

3.2 Task Definition

We meticulously design 15 distinct sub-tasks (with detailed definitions in Appendix C). These tasks are categorized into four core evaluation dimensions, briefly described as follows:

- ***Audio-Visual Joint Reasoning (AVR):*** This category forms the core of TraceAV-Bench, evaluating the model's ability to chain clues across both visual and auditory streams over long temporal spans. It includes seven sub-tasks: *Information Retrieval (IR)*, *Temporal Sequencing (TS)*, *Entity Tracking*

(ET), *Forward Causal Reasoning (FCR)*, *Backward Causal Reasoning (BCR)*, *Cross-Modality Matching (CMM)*, and *Spatiotemporal Localization (SL)*.

- ***Visual-Centric Reasoning (VR):*** Rather than joint reasoning, this dimension isolates the visual modality to evaluate how well the model maintains its visual reasoning capabilities when faced with long video inputs. It comprises *Spatial Reasoning (SR)* and *Visual Counting (VC)*.
- ***Audio-Centric Reasoning (AR):*** Complementary to VR, this dimension aims to assess the model's auditory capabilities under long-video settings. It encompasses examination across three distinct aspects: *Speech Context (SC)*, *Environmental Sound (ES)*, and *Background Music (BM)*.
- ***Multimodal Hallucination (MH):*** This category specifically investigates whether the model exhibits hallucinations in audio-visual contexts. It tests model robustness through *Visual-to-Audio Deception (V2A)*, *Audio-to-Visual Deception (A2V)*, and *Temporal Splicing Fallacy (TSF)*.

3.3 Data Construction Pipeline

To construct high-quality multi-hop questions over long audio-visual videos at scale, we develop a decoupled three-step pipeline that converts each video into a dense textual event catalog for trajectory-grounded question generation.

Step 1: Minute-Level Visual Captioning. To obtain fine-grained visual descriptions throughout long videos, we divide each video into 1-minute clips and sequentially process them with Qwen3-VL-32B-Instruct Bai et al. (2025). At this stage, we isolate the visual stream to avoid cross-modal interference and hallucination during initial content extraction Xie et al. (2026); Sung-Bin et al. (2024). To preserve identity consistency across temporally segmented clips, we maintain an *Entity Cache* that records previously identified persons, objects, locations, and their evolving states.

Step 2: Audio-Visual Fusion. Building on the visual captions and Entity Cache from Step 1, we adopt an *asynchronous fusion* strategy to align audio evidence with the existing visual narrative. Gemini-2.5-Flash Comanici et al. (2025) processes each corresponding 1-minute audio segment conditioned on both sources, producing a unified audio-visual description and selectively updating entity representations with new audio evidence. This yields a chronologically aligned, high-density bimodal event catalog for question generation.

Step 3: Agentic Question Generation. Although the fused captions capture rich content, they do not explicitly organize event boundaries or long-range dependencies, making direct generation prone to shallow or weakly grounded questions. We therefore employ a three-stage workflow powered by GPT-5.1 Singh et al. (2025): an *Event Segmentation Agent* groups adjacent captions into coherent event blocks, a *Trajectory Proposal Agent* identifies temporally dispersed and entity-linked evidence chains, and a *Question Generation Agent* converts validated trajectories into four-option MCQs, following three strict and hierarchical generation principles:

- ***Anti-Shortcut Formulation:*** The question stem must not be answerable via general world knowledge or common sense. It must require direct observation of the content named in the trajectory.
- ***Trajectory-Grounded Distractors:*** All incorrect options must be semantically plausible and grounded in actual events from the same video, preventing elimination by surface-level dissimilarity.
- ***Stylistic Uniformity:*** All options must be comparable in length, syntactic structure, and wording style to prevent guessing via surface-level cues.

3.4 Quality Assurance

Despite the strict generation constraints applied in Step 3, the model can still produce questions with latent flaws. To ensure the collected questions are high-quality and genuinely multi-hop, we process the generated candidates through a rigorous quality assurance stage led by the following principles:

- ***Rule-based Validation:*** We verify that all questions strictly adhere to their predefined tasks. For example, we enforce that AVR tasks genuinely necessitate evidence from both modalities and that unimodal tasks contain no cross-modal leakage. We additionally verify that each evidence hop is anchored to a specific and verifiable minute-level timestamp and that the MCQ format is strictly adhered to. Questions that fail any of these checks are automatically discarded.

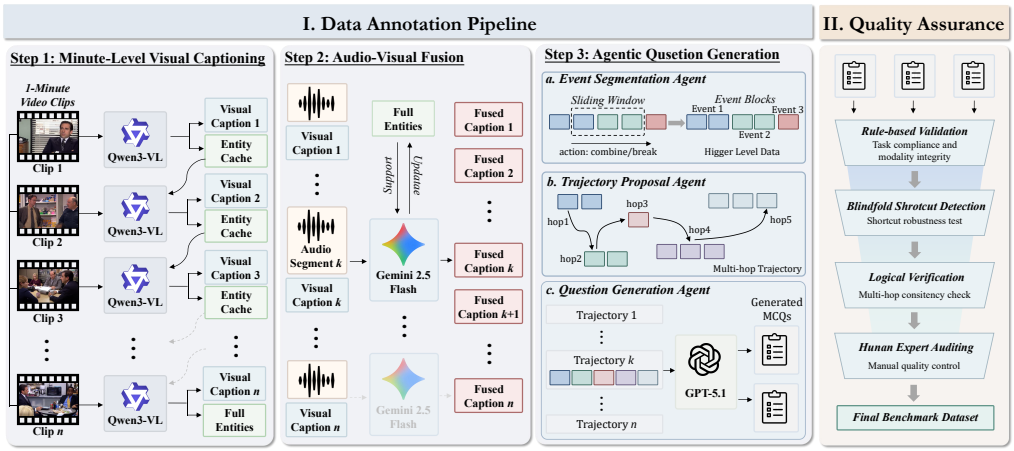


Figure 4: Overview of the TraceAV-Bench data construction pipeline.

- ***Logical Verification:*** A separate LLM conducts a deep audit of each surviving item's multi-hop integrity. It explicitly filters out "pseudo multi-hop" questions, checks for answer leakage within the question stem, and ensures that all distractors are semantically relevant and plausibly confusable, avoiding obvious stylistic or length disparities.
- ***Blindfold Shortcut Detection:*** Even with anti-shortcut prompting during generation, some questions may remain answerable through linguistic biases or world knowledge without viewing the video. We therefore submit each item to Gemini 2 Flash with only the question stem and options as text input, without all video content, and record whether the model selects the correct answer. Any item answered correctly by this "blindfolded" solver is removed from the pool.
- ***Human Expert Auditing:*** As a final quality gate, human annotators manually review a 15% sample from each generation batch, correcting unclear or weak items and verifying multi-hop integration across the referenced evidence. Any batch whose inspected items exceed a 5% error or ambiguity rate is discarded. Further auditing details are provided in Appendix G.

3.5 TraceAV-Bench Statistics

Video statistics. The benchmark is built upon **578** long videos with a total duration of **339.5 hours**. Video lengths range from 10.1 to 139.9 minutes, with a mean of 35.2 minutes. The majority (57.6%) fall in the 30–60 minute range, and 43 videos (7.4%) exceed one hour, underscoring the ultra-long-form nature of the benchmark. All videos feature resolutions up to 4K, with 73.7% at HD (720p) or above, and each contains a stereo audio track encompassing speech, sounds, and music.

Question statistics. TraceAV-Bench contains **2,200** questions across four dimensions: Audio-Visual Joint Reasoning (AVR, 7 tasks, 835 questions, 38.0%), Visual-Centric Reasoning (VR, 2 tasks, 391 questions, 17.8%), Audio-Centric Reasoning (AR, 3 tasks, 349 questions, 15.9%), and Multimodal Hallucination (MH, 3 tasks, 625 questions, 28.4%). The benchmark comprises 1,848 single-choice (84.0%) and 352 multi-choice questions (16.0%).

Multi-hop trajectory statistics. Each question is grounded in a multi-hop reasoning chain spanning specific minutes of the video. On average, each question requires **3.68** evidence hops, with a temporal span averaging **15.1 minutes** across the trajectory.

4 Experiments

4.1 Experimental Settings

Evaluated Models. We benchmark models in four categories: ***1) Closed-source OmniLLMs***: Gemini 2, Gemini 2.5 Comanici et al. (2025) and Gemini 3; ***2) Open-source OmniLLMs***: Qwen3-Omni Xu et al. (2025b), Qwen2.5-Omni Xu et al. (2025a), OmniVinci Ye et al. (2025), MiniCPM-o Yu et al. (2025), HumanOmni Zhao et al. (2025), Video-SALMONN 2 Tang et al. (2025), Baichuan-Omni-1.5 Li et al. (2024a), VideoLLaMA2.1 Cheng et al. (2024), Ming-Flash-Omni AI et al. (2025),

Model	AVR							VR		AR			Avg
	IR	TS	ET	FCR	BCR	CMM	SL	SR	VC	SC	ES	BM	
Closed-source OmniLLMs (With Visual and Audio)													
Gemini 3.1 Pro	**83.57**	60.82	**71.77**	**86.30**	**61.80**	49.41	**51.54**	**73.94**	**41.15**	**96.92**	63.64	**78.63**	**68.29**
Gemini 2.5 Pro	83.57	**63.92**	60.48	76.71	50.56	**54.12**	48.02	68.48	39.38	83.85	**65.91**	67.18	63.52
Gemini 3 Flash	82.14	53.61	65.32	83.56	59.55	49.41	29.07	73.33	36.73	86.92	61.36	66.41	62.28
Gemini 2.5 Flash	75.00	58.76	62.10	75.34	58.43	40.00	29.07	66.06	39.38	81.54	60.23	62.60	59.04
Gemini 2 Flash	66.43	53.61	58.06	64.38	43.82	41.18	25.11	54.55	27.43	70.00	55.68	58.78	51.59
Open-source OmniLLMs (With Visual and Audio)													
Ming-Flash-Omni-2.0	56.43	53.61	47.58	57.53	40.45	44.71	31.28	65.45	39.38	63.85	56.82	63.36	51.70
Qwen3-Omni-30B-A3B	47.14	51.55	35.48	43.84	50.56	40.00	32.60	58.18	38.50	63.85	59.09	60.31	48.43
OmniVinci-9B	49.29	44.33	38.71	57.53	34.83	35.29	33.48	55.15	34.51	65.38	**65.91**	54.20	47.38
MiniCPM-o 4.5	45.71	36.08	37.90	28.77	26.97	41.18	37.44	60.61	38.50	61.54	59.09	64.12	44.83
Qwen2.5-Omni-7B	46.43	32.99	37.10	30.14	35.96	37.65	37.44	49.70	35.40	60.00	52.27	49.62	42.06
Gemma 4-E4B	39.29	38.14	37.10	36.99	29.21	36.47	16.74	55.15	34.07	55.38	54.55	58.02	40.93
Video-SALMONN 2	42.14	41.24	29.03	30.14	29.21	32.94	31.28	48.48	39.38	47.69	47.73	44.27	38.63
HumanOmni-7B	37.86	31.96	29.84	31.51	31.46	25.88	35.68	55.15	34.96	52.31	51.14	44.27	38.50
VideoLLaMA2.1-AV-7B	36.43	29.90	25.81	17.81	16.85	25.88	22.91	38.79	38.94	36.15	39.77	35.88	30.43
Baichuan-Omni-1.5	37.14	14.43	20.97	15.07	12.36	20.00	30.40	37.58	26.99	32.31	42.05	33.59	26.91
Open-source Single-Modality MLLMs													
Qwen3-VL-32B	44.29	39.18	32.26	38.36	38.20	34.12	16.30	67.27	39.38	46.15	48.86	48.09	41.04
Qwen3-VL-8B	34.29	28.87	29.84	26.03	24.72	24.71	17.62	59.39	32.30	45.38	42.05	46.56	34.31
Qwen2-Audio-7B	30.71	27.84	33.06	20.55	26.97	29.41	29.96	26.67	23.45	38.46	37.50	44.27	30.74
Open-source OmniLLMs (Visual Only)													
Ming-Flash-Omni-2.0	37.86	35.05	32.26	45.21	31.46	27.06	19.82	54.55	38.94	43.08	47.73	49.62	38.55
Qwen3-Omni-30B-A3B	37.86	34.02	37.90	32.88	34.83	24.71	30.84	53.94	38.94	38.46	40.91	43.51	37.40
Open-source OmniLLMs (Audio Only)													
Ming-Flash-Omni-2.0	32.86	24.74	23.39	39.73	22.47	18.82	19.38	29.09	25.66	44.62	42.05	46.56	30.78
Qwen3-Omni-30B-A3B	27.14	31.96	20.16	23.29	33.71	29.41	18.94	27.88	21.24	43.85	40.91	45.80	30.36
Random Guess	23.39	21.39	16.33	20.21	12.64	18.53	23.57	25.00	23.78	20.58	18.75	19.27	20.29

Table 2: Main results on TraceAV-Bench (%) across 12 general tasks. Sub-tasks grouped by three dimensions: **AVR** (*Audio-Visual Joint Reasoning*), **VR** (*Visual-Centric Reasoning*), **AR** (*Audio-Centric Reasoning*). Abbreviations: IR = *Information Retrieval*, TS = *Temporal Sequencing*, ET = *Entity Tracking*, FCR = *Forward Causal Reasoning*, BCR = *Backward Causal Reasoning*, CMM = *Cross-Modality Matching*, SL = *Spatiotemporal Localization*, SR = *Spatial Reasoning*, VC = *Visual Counting*, SC = *Speech Context*, ES = *Environmental Sound*, BM = *Background Music*. **Avg** is the macro-average over all tasks. Best result per task is **bolded**, and best open-source result is underlined.

and Gemma 4; ***3) Single-modality MLLMs***: video-only Qwen3-VL Bai et al. (2025) and audio-only Qwen2-Audio Chu et al. (2024); ***4) Visual-only ablations***: Ming-Flash-Omni and Qwen3-Omni with the audio stream removed. ***5) Audio-only ablations***: Ming-Flash-Omni and Qwen3-Omni with the visual stream removed.

Evaluation Protocol. For open-source models, we follow official inference configurations and sample as many frames as permitted by the context window to maximize performance. For closed-source models, we use the recommended sampling rate (1 frame per second). These settings are chosen to obtain the best performance from each evaluated model. A response is correct only if the predicted option(s) exactly match the ground-truth answer(s).

4.2 Experimental Findings

Finding 1: Current OmniLLMs still struggle with long-form audio-visual multi-hop reasoning. Table 2 shows that TraceAV-Bench remains challenging for all evaluated models. The strongest closed-source model, Gemini 3.1 Pro, achieves an average accuracy of 68.29%, while the best open-source model, Ming-Flash-Omni-2.0, reaches 51.70%. Closed-source Gemini series models demonstrate ***dominant performance***, consistently occupying the top of the rankings, Nevertheless, even the strongest model remains below 70%, with particularly limited performance on tasks that require cross-modal matching, spatiotemporal localization, and visual counting. These results show

Model	V2A	A2V	TSF	MH Avg	Gen. Avg	Rank (MH→Gen.)
Closed-source OmniLLMs						
Gemini 3.1 Pro	**89.57**	79.91	**84.34**	**84.61**	**68.29**	$1 \rightarrow 1$
Gemini 3 Flash	76.52	75.55	87.35	79.81	62.28	$2 \rightarrow 3$
Gemini 2 Flash	74.78	**81.66**	77.11	77.85	51.59	$3 \rightarrow 5$
Gemini 2.5 Pro	79.13	74.24	75.90	76.42	63.52	$4 \rightarrow 2$
Gemini 2.5 Flash	60.87	66.81	66.87	64.85	59.04	$5 \rightarrow 4$
Open-source OmniLLMs						
Qwen3-Omni-30B-A3B	65.65	69.87	<u>66.87</u>	<u>67.46</u>	48.43	$1 \rightarrow 2$
Ming-Flash-Omni-2.0	71.30	67.25	62.65	67.07	<u>51.70</u>	$2 \rightarrow 1$
Gemma 4-E4B	<u>74.35</u>	69.43	56.02	66.60	40.93	$3 \rightarrow 6$
MiniCPM-o 4.5	70.87	<u>72.05</u>	56.63	66.52	44.83	$4 \rightarrow 4$
Qwen2.5-Omni-7B	60.43	55.46	53.61	56.50	42.06	$5 \rightarrow 5$
OmniVinci-9B	42.17	44.10	42.17	42.81	47.38	$6 \rightarrow 3$
Video-SALMONN 2	45.65	39.30	37.95	40.97	38.63	$7 \rightarrow 7$
HumanOmni-7B	33.91	37.99	28.31	33.40	38.50	$8 \rightarrow 8$
Baichuan-Omni-1.5	28.26	41.48	22.29	30.68	26.91	$9 \rightarrow 10$
VideoLLaMA2.1-AV-7B	35.22	29.69	19.88	28.26	30.43	$10 \rightarrow 9$
Random Guess	21.74	22.27	17.17	20.39	20.29	–

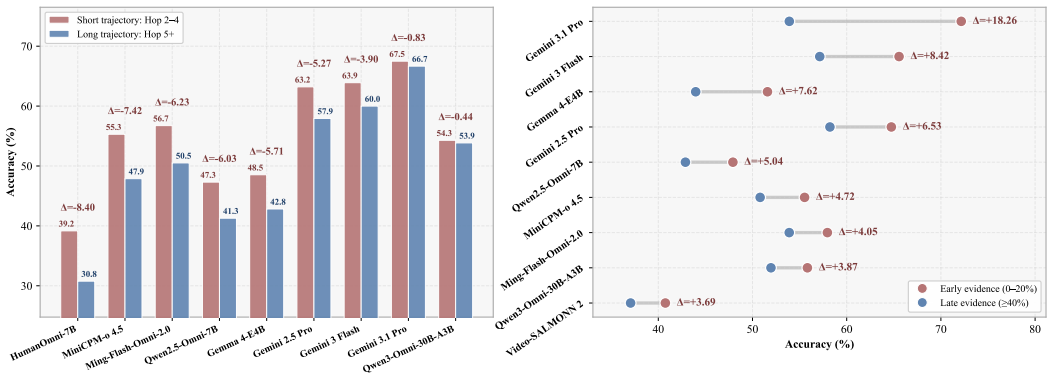
Table 3: Hallucination robustness on the MH dimension of TraceAV-Bench (%). V2A = *Visual-to-Audio Deception*, A2V = *Audio-to-Visual Deception*, TSF = *Temporal Splicing Fallacy*. **MH Avg** is the average over the three MH sub-tasks. **Gen. Avg** is the average over the 12 general sub-tasks (Table 2). Models are sorted by MH Avg within each group. **Rank (MH→Gen.)** shows each model's within-group rank by MH Avg and by Gen. Avg respectively. Best result per task is **bolded**, and best open-source result is <u>underlined</u>.

that ***current OmniLLMs have not yet mastered the integration of sparse audio-visual evidence across long videos***.

Finding 2: Audio-visual reasoning exhibits distinct performance profiles across tasks. The substantial variation across AVR sub-tasks shows that audio-visual reasoning is not a single and uniform capability. Models achieve higher observed accuracy when the required information can be directly retrieved or when an observed cause can be mapped to its consequence, while accuracy is lower on cross-modal binding, temporal localization, and backward causal reasoning. Gemini 3.1 Pro achieves 83.57% on Information Retrieval and 86.30% on Forward Causal Reasoning, while obtaining 49.41% on Cross-Modality Matching, 51.54% on Spatiotemporal Localization, and 61.80% on Backward Causal Reasoning. Ming-Flash-Omni-2.0 exhibits a similar pattern, reaching 56.43% on Information Retrieval and 57.53% on Forward Causal Reasoning, compared with 44.71% on Cross-Modality Matching, 31.28% on Spatiotemporal Localization, and 40.45% on Backward Causal Reasoning. Taken together, these results suggest that ***current models can often identify relevant evidence, but show lower observed accuracy when evidence must be aligned across modalities and composed into a coherent long-range inference***.

Finding 3: Fine-grained visual tracking and non-speech audio understanding remain modality-specific bottlenecks. The visual-centric results show that most models perform substantially worse on Visual Counting than on Spatial Reasoning. Gemini 3.1 Pro achieves 73.94% on Spatial Reasoning but only 41.15% on Visual Counting, while Ming-Flash-Omni-2.0 obtains 65.45% and 39.38% on the two tasks. A similar pattern appears for Qwen3-Omni-30B-A3B, whose performance decreases from 58.18% on Spatial Reasoning to 38.50% on Visual Counting. This result indicates that models can often recognize spatial information but struggle to track and accumulate repeated visual events across long videos. The audio-centric results reveal a related limitation for non-speech information. For example, Gemini 3.1 Pro achieves 96.92% on Speech Context, compared with 63.64% on Environmental Sound and 78.63% on Background Music. Across models, performance on Environmental Sound and Background Music is also more variable than performance on Speech Context. The results indicate that ***current models reason more reliably about spatial relations and speech content than tracking repeated visual events or interpreting sounds and music throughout long videos***.

Finding 4: General-task performance does not fully characterize hallucination robustness. Table 3 reveals substantial ranking changes between the general tasks and the Multimodal Hallucination dimension. Gemma 4 is a representative example, ranking third among open-source models



(a) Accuracy on questions grouped by reasoning Hop Count into short trajectories (Hop 2–4) and long trajectories (Hop 5+).

(b) Accuracy on questions whose first evidence hop falls in the early (0–20%) or late (≥40%) portion of the video.

Figure 5: Two diagnostic analyses of OmniLLMs on TraceAV-Bench.

on MH Avg with 66.60%, while ranking sixth on Gen. Avg with 40.93%. OmniVinci exhibits the opposite pattern, ranking third on Gen. Avg with 47.38% but sixth on MH Avg with 42.81%. A similar shift appears among closed-source models. Gemini 2 Flash ranks third on MH Avg with 77.85%, despite ranking fifth on Gen. Avg with 51.59%. Models also exhibit different strengths across Visual-to-Audio Deception, Audio-to-Visual Deception, and Temporal Splicing Fallacy. These results show that ***general audio-visual reasoning performance alone does not fully characterize a model's robustness to multimodal hallucinations***.

Finding 5: Longer annotated evidence trajectories are associated with lower accuracy. Figure 5(a) compares questions requiring 2 to 4 evidence hops with those requiring 5 or more hops. Most evaluated models achieve lower accuracy on the longer trajectories; this figure summarizes a descriptive grouping of benchmark items. MiniCPM-o 4.5 decreases from 55.3% on shorter trajectories to 47.9% on longer trajectories, while Ming-Flash-Omni-2.0 decreases from 56.7% to 50.5%. The same pattern appears in the Gemini family, with Gemini 3 Flash decreasing from 63.9% to 60.0% and Gemini 3.1 Pro decreasing from 67.5% to 66.7%. Although the magnitude varies across models, the observed trend is similar across different architectures. This descriptive pattern is consistent with an association between increased annotated hop count and lower accuracy on complex long-video questions.

Finding 6: Earlier evidence positions are associated with higher accuracy. Figure 5(b) compares questions whose first evidence hop appears in the first 20% of the video with questions whose first hop appears at or after 40%. Every evaluated model achieves higher accuracy when the first evidence hop appears near the beginning of the video. On average, accuracy is 6.91% higher for early-evidence questions, with model-specific differences of 3.69% to 18.26%. Gemini 3.1 Pro exhibits the largest difference of 18.26%, indicating that sensitivity to evidence position remains notable even in the strongest evaluated model. These results show that the position of the first annotated evidence hop remains an important factor in long-video performance.

5 Conclusion

In this paper, we present TraceAV-Bench, the first benchmark to jointly stress *multi-hop trajectory reasoning* and *multimodal hallucination robustness* in the long audio-visual video context. Built upon a decoupled three-step pipeline and a rigorous multi-stage quality assurance process, TraceAV-Bench comprises 2,200 trajectory-grounded multiple-choice questions over 578 long videos. Every question links temporally dispersed evidence, with AVR requiring both audio and vision and VR/AR deliberately isolating visual or auditory reasoning. We comprehensively evaluate representative OmniLLMs across the general and hallucination dimensions of TraceAV-Bench, and our analyses surface several insightful findings that we believe offer meaningful guidance for future model training and evaluation. We hope TraceAV-Bench serves as a rigorous testbed for the next generation of OmniLLMs that reason coherently and faithfully over long, evidence-sparse audio-visual content.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets and cleaning and reformatting the dataset. We have not used generative AI tools for proposing or refine hypotheses, designing or provide feedback on research methodology or experiments, implementing methods, assisting with translation, or interpreting results. The tasks concerning theoretical models, conceptual frameworks, mathematical claims, and proofs are not applicable to this empirical work. The remaining required disclosure tasks, including helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs, are not applicable to this work. Additionally, we used generative AI tools for editing a research paper to improve readability, sourcing/searching for information. We have reviewed all AI-assisted work. AI-generated benchmark intermediates and candidate questions were checked through the rule-based, model-based, and human quality-control procedures described in Section 3.3 and Appendices G and L. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

TraceAV-Bench is intended for non-commercial research on reliable audio-visual understanding, and its source videos are publicly accessible. We do not claim ownership of the source videos. Users should obtain video content from its original source and comply with the applicable licenses, platform terms, and copyright requirements. The released materials are designed to provide annotations, benchmark metadata, prompts, and evaluation resources without disclosing private annotator information. Human contributors performed paid annotation work, received task-specific training. The source material may reflect demographic, linguistic, cultural, and topical biases present on public video platforms. During data review, unsupported sensitive inferences and unsuitable content were removed, and generated annotations were checked against the original audio-visual evidence. Despite these safeguards, residual annotation errors, harmful content, or representational imbalance may remain. We encourage users to inspect the source context, consider these limitations when interpreting benchmark results, and report any identified issues so that future releases can be improved.

REPRODUCIBILITY STATEMENT

We support reproducibility through the detailed descriptions and released materials accompanying this paper. Section 3.3 presents the benchmark construction and quality-control workflow, while Appendix D.3 gives the agentic generation algorithm and Appendix L provides the prompts used at each construction and evaluation stage. Appendix E reports dataset and trajectory statistics, Appendix G documents the human auditing protocol and statistical reliability analysis, and Section 4.1 together with Appendix H specifies the evaluation protocol, model configurations, and deployment setting. The anonymous repository (from the abstract) provides the benchmark data, metadata, evaluation resources, and code needed to reproduce the reported evaluation.

REFERENCES

Inclusion AI, Bowen Ma, Cheng Zou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Chenyu Lian, Dandan Zheng, Fudong Wang, Furong Xu, et al. Ming-flash-omni: A sparse, unified architecture for multimodal perception and generation. *arXiv preprint arXiv:2510.24821*, 2025.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*, 2025.

Ali Asgarov, Kaushik Narasimhan, Najibul Haque Sarker, Hani Alomari, Chia-Wei Tang, Anushka Sivakumar, Zaber Ibn Abdul Hakim, Shaurya Mallampati, and Chris Thomas. Nest: Narrative event structures in time for long video understanding. *arXiv preprint arXiv:2606.19706*, 2026.

Kirolos Ataallah, Eslam Mohamed Bakr, Mahmoud Ahmed, Chenhui Gou, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. Infinibench: A benchmark for large multi-modal models in long-form movies and tv shows. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19496–19523, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Xinyue Cai, Chaoyou Fu, Yi-Fan Zhang, Ran He, and Caifeng Shan. Omnivideo-100k: A dataset for audio-visual reasoning through structured scripts and evidence chains. *arXiv preprint arXiv:2606.14702*, 2026.

Jianghan Chao, Jianzhang Gao, Wenhui Tan, Yuchong Sun, Ruihua Song, and Liyun Ru. Jointavbench: A benchmark for joint audio-visual reasoning evaluation. *arXiv preprint arXiv:2512.12772*, 2025.

Chen Chen, ZeYang Hu, Fengjiao Chen, Liya Ma, Jiaxing Liu, Xiaoyu Li, Ziwen Wang, Xuezhi Cao, and Xunliang Cai. Uno-bench: A unified benchmark for exploring the compositional law between uni-modal and omni-modal in omni models. *arXiv preprint arXiv:2510.18915*, 2025a.

Mingfei Chen, Zijun Cui, Ruoke Zhang, Hyeonggon Ryu, and Eli Shlizerman. Scenebind: Binding what and where across vision, audio and language. *arXiv preprint arXiv:2607.15265*, 2026a.

Qian Chen, Jinlan Fu, Changsong Li, See-Kiong Ng, and Xipeng Qiu. Futureomni: Evaluating future forecasting from omni-modal context for multimodal llms. *arXiv preprint arXiv:2601.13836*, 2026b.

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2159–2167, 2025b.

Xinlong Chen, Yue Ding, Weihong Lin, Jingyun Hua, Linli Yao, Yang Shi, Bozhou Li, Qiang Liu, Yuanxing Zhang, Pengfei Wan, et al. Avocado: An audiovisual video captioner driven by temporal orchestration. In *International Conference on Learning Representations*, volume 2026, pp. 24338–24362, 2026c.

Zixuan Chen, Depeng Wang, Hao Lin, Li Luo, Ke Xu, Ya Guo, Huijia Zhu, Tanfeng Sun, and Xinghao Jiang. Avid: A benchmark for omni-modal audio-visual inconsistency understanding via agent-driven construction. *arXiv preprint arXiv:2604.13593*, 2026d.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1590–1601, 2025.

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

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

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

Yue Ding, Yiyan Ji, Jungang Li, Xuyang Liu, Xinlong Chen, Junfei Wu, Bozhou Li, Bohan Zeng, Yang Shi, Yushuo Guan, et al. Omnisift: Modality-asymmetric token compression for efficient omni-modal large language models. *arXiv preprint arXiv:2602.04804*, 2026.

Hang Du, Jiayang Zhang, Guoshun Nan, Wendi Deng, Zhenyan Chen, Chenyang Zhang, Wang Xiao, Shan Huang, Yuqi Pan, Tao Qi, et al. From easy to hard: The mir benchmark for progressive interleaved multi-image reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 859–869, 2025.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

Archishman Ghosh, Abhinaba Roy, and Dorien Herremans. Karma-mv: A benchmark for causal question answering on music videos. *arXiv preprint arXiv:2605.08175*, 2026a.

Dhruba Ghosh, Yuhui Zhang, and Ludwig Schmidt. Understanding the fine-grained knowledge capabilities of vision-language models. *arXiv preprint arXiv:2602.17871*, 2026b.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*, 2025.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Siddharth Gururani, Hanrong Ye, Pritam Biswas, Yuanhang Su, Ehsan Hosseini-Asl, et al. Audio-visual flamingo: Open audio-visual intelligence for long and complex videos. *arXiv preprint arXiv:2607.16107*, 2026c.

Arushi Goel, Sreyan Ghosh, Vatsal Agarwal, Nishit Anand, Kaousheik Jayakumar, Lasha Koroshinadze, Yao Xu, Katie Lyons, James Case, Karan Sapra, et al. Mmou: A massive multi-task omni understanding and reasoning benchmark for long and complex real-world videos. *arXiv preprint arXiv:2603.14145*, 2026.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025a.

Qingpei Guo, Kaiyou Song, Zipeng Feng, Ziping Ma, Qinglong Zhang, Sirui Gao, Xuzheng Yu, Yunxiao Sun, Tai-Wei Chang, Jingdong Chen, et al. M2-omni: Advancing omni-mllm for comprehensive modality support with competitive performance. *arXiv preprint arXiv:2502.18778*, 2025b.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

ZhaoYang Han, Qihan Lin, Hao Liang, Bowen Chen, Zhou Liu, and Wentao Zhang. Longinsightbench: A comprehensive benchmark for evaluating omni-modal models on human-centric long-video understanding. *arXiv preprint arXiv:2510.17305*, 2025.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Wentao Wang, Zhenghao Song, Dingling Zhang, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*, 2025.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19108–19118, 2022.

Jiachun Li, Shaoping Huang, Zhuoran Jin, Chenlong Zhang, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Mmr-life: Piecing together real-life scenes for multimodal multi-image reasoning. *arXiv preprint arXiv:2603.02024*, 2026a.

Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, Shijian Wang, Guanting Dong, Jiajie Jin, Hao Wang, Yinuo Wang, Ji-Rong Wen, Yuan Lu, et al. Omnigaia: Towards native omni-modal ai agents. *arXiv preprint arXiv:2602.22897*, 2026b.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, et al. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 2024a.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4122–4134, 2025a.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025b.

Run Luo, Xiaobo Xia, Lu Wang, Longze Chen, Renke Shan, Jing Luo, Min Yang, and Tat-Seng Chua. Next-omni: Towards any-to-any omnimodal foundation models with discrete flow matching. *arXiv preprint arXiv:2510.13721*, 2025.

David Ma, Huaqing Yuan, Xingjian Wang, Qianbo Zang, Tianci Liu, Xinyang He, Yanbin Wei, Jiawei Guo, Zhenzhu Yang, Meng Cao, et al. Scalelong: A multi-timescale benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2026, pp. 154328–154347, 2026a.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. In *International Conference on Learning Representations*, volume 2026, pp. 132277–132304, 2026b.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Yusu Qian, Cheng Wan, Chao Jia, Yinfei Yang, Qingyu Zhao, and Zhe Gan. Prism-bench: A benchmark of puzzle-based visual tasks with cot error detection. *arXiv preprint arXiv:2510.23594*, 2025.

Chenhui Qiang, Zhaoyang Wei, Xumeng Han, Zipeng Wang, Siyao Li, Xiangyuan Lan, Jianbin Jiao, and Zhenjun Han. Ver-bench: Evaluating mllms on reasoning with fine-grained visual evidence. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12698–12705, 2025.

Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024.

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. Agentic very long video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46575–46602, 2026.

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

Shaden Shaar, Bradon Thymes, Sirawut Chaixanien, Claire Cardie, and Bharath Hariharan. Moviere-capsqa: A multimodal open-ended video question-answering benchmark. *arXiv preprint arXiv:2601.02536*, 2026.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-asr technical report, 2026. URL https://arxiv.org/abs/2601.21337.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*, 2024.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Changli Tang, Tianyi Wang, Fengyun Rao, Jing Lyu, and Chao Zhang. D-orca: Dialogue-centric optimization for robust audio-visual captioning. *arXiv preprint arXiv:2602.07960*, 2026.

Keda Tao, Wenjie Du, Bohan Yu, Weiqiang Wang, Jian Liu, and Huan Wang. Active perception agent for omnimodal audio-video understanding. *arXiv preprint arXiv:2512.23646*, 2025.

Keda Tao, Yuhua Zheng, Jia Xu, Wenjie Du, Kele Shao, Hesong Wang, Xueyi Chen, Xin Jin, Junhan Zhu, Bohan Yu, et al. Lvomnibench: Pioneering long audio-video understanding evaluation for omnimodal llms. *arXiv preprint arXiv:2603.19217*, 2026.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025a.

Meituan LongCat Team, Bairui Wang, Bin Xiao, Bo Zhang, Bolin Rong, Borun Chen, Chang Wan, Chao Zhang, Chen Huang, Chen Chen, et al. Longcat-flash-omni technical report. *arXiv preprint arXiv:2511.00279*, 2025b.

Qwen Team. Qwen3-asr technical report. 2026a.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026b.

Wenjie Tian, Zhixian Zhao, Jingbin Hu, Huakang Chen, Haohe Liu, Binshen Mu, and Lei Xie. Emoomni: Bridging emotional understanding and expression in omni-modal llms. *arXiv preprint arXiv:2602.21900*, 2026.

Liang-Hsuan Tseng, Yi-Chang Chen, Kuan-Yi Lee, Da-Shan Shiu, and Hung-yi Lee. Taste: Text-aligned speech tokenization and embedding for spoken language modeling. *arXiv preprint arXiv:2504.07053*, 2025.

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

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

Jiahao Wang, An Ping, Yanghai Wang, Yuanxing Zhang, Shihao Li, Hanyan Bian, Yichi Ren, Yize Zhang, Han Wang, Haowen Chen, et al. Omnicap-if: Benchmarking and improving instruction following abilities for omni-video captioning. *arXiv preprint arXiv:2606.08572*, 2026a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025c.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025d.

Yaoting Wang, Ziyi Zhang, Wenming Tu, Shaoxuan Xu, Wenjie Du, Cheng Liang, Weijun Wang, Yuanchao Li, Guangyao Li, Hao Fei, et al. Avi-bench: Toward human-like audio-visual intelligence of omni-mllms. *arXiv preprint arXiv:2606.07643*, 2026b.

Yingzhi Wang, Pooneh Mousavi, Artem Ploujnikov, and Mirco Ravanelli. What are they doing? joint audio-speech co-reasoning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025e.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18925–18935, 2025f.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. *arXiv preprint arXiv:2512.05774*, 2025g.

Tengjin Weng, Wenhao Jiang, Jingyi Wang, Ming Li, Lin Ma, and Zhong Ming. Oddgridbench: Exposing the lack of fine-grained visual discrepancy sensitivity in multimodal large language models. *arXiv preprint arXiv:2603.09326*, 2026.

LCT Xiaomi and Core Team. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 1(2):5, 2025.

Liuyue Xie, Avik Kuthiala, George Z Wei, Ce Zheng, Ananya Bal, Mosam Dabhi, Liting Wen, Taru Rustagi, Ethan Lai, Sushil Khyalia, et al. Maverix: Multimodal audio-visual evaluation and recognition index. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27090–27098, 2026.

Wenbin Xing, Quanxing Zha, Lizheng Zu, Mengran Li, Ming Li, and Junchi Yan. Learning to decode against compositional hallucination in video multimodal large language models. *arXiv preprint arXiv:2602.00559*, 2026a.

Zhenghao Xing, Ruiyang Xu, Yuxuan Wang, Jinzheng He, Ziyang Ma, Qize Yang, Yunfei Chu, Jin Xu, Junyang Lin, Chi-Wing Fu, et al. Native active perception as reasoning for omni-modal understanding. *arXiv preprint arXiv:2606.19341*, 2026b.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Ke Xu, Yuhao Wang, Ziyang Cheng, Hongcheng Liu, Yanfeng Wang, and Yu Wang. Agentic active omni-modal perception for multi-hop audio-visual reasoning. *arXiv preprint arXiv:2605.28192*, 2026.

Kaiying Yan, Luoyi Sun, Xiao Zhou, and Weidi Xie. Empowering long-form omni-modal understanding with robust audio perception. *arXiv preprint arXiv:2607.10299*, 2026.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 3480–3491, 2022.

Huanjin Yao, Jiaxing Huang, Yawen Qiu, Michael K Chen, Wenzheng Liu, Wei Zhang, Wenjie Zeng, Xikun Zhang, Jingyi Zhang, Yuxin Song, et al. Mmreason: An open-ended multi-modal multi-step reasoning benchmark for mllms toward agi. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 273–283, 2025.

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

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Xinlei Yin, Xiulian Peng, Xiao Li, Zhiwei Xiong, and Yan Lu. Hierarchical long video understanding with audiovisual entity cohesion and agentic search. *arXiv preprint arXiv:2601.13719*, 2026.

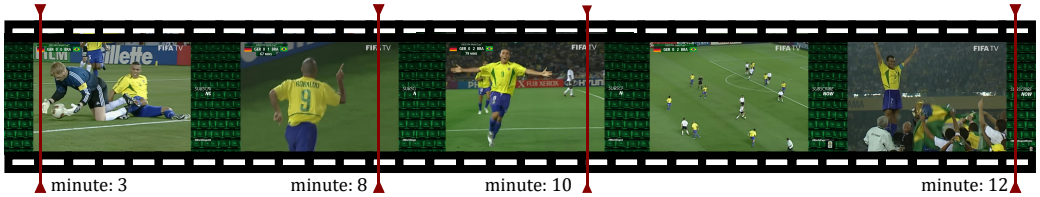
Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025. URL https://arxiv.org/abs/2509.18154.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 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*, pp. 9556–9567, 2024.

Chenshuang Zhang, Kyeong Seon Kim, Chengxin Liu, and Tae-Hyun Oh. Svhalluc: Benchmarking speech-vision hallucination in audio-visual large language models. *arXiv preprint arXiv:2606.02642*, 2026a.



Visual-to-Audio Deception

Question: Across Ronaldo's missed chances, both of his goals, and the final trophy celebrations, which of the following specific audio moments involving his name actually do NOT occur in the soundtrack?"

Options:

A: After Ronaldo's first goal (the 1–0 breakthrough), the crowd breaks into a clear, rhythmic "Ron-al-do, Ron-al-do" chant that rises above the general cheering.

B: During the immediate full-time scenes as Brazil are confirmed champions, the commentators lead into a repeated "Ronaldo, Ronaldo" chant that the crowd audibly joins.

C: Following Ronaldo's second goal to make it 2–0, the commentator repeats "Ronaldo" in excited narration, but there is no distinct, stadium-wide chant or song of his name—only generic crowd noise.

D: In the podium and trophy-lift sequence with confetti falling and an instrumental track playing, there is still no isolated or rhythmic crowd chant of "Ronaldo" that stands out from the music and general cheering.

✗ *Gemini 3.1-pro:* A B C

Error analysis: Gemini correctly locates Ronaldo's goals and the trophy celebration, but fails under Visual-to-Audio Deception. It infers a "Ronaldo" chant from the visually salient scoring and celebration scenes, confusing generic cheering and commentator mentions with an actual rhythmic crowd chant. This shows that the model relies on visual priors rather than verifying the aligned audio.

Figure 6: Case of a Visual-to-Audio Deception error. The model infers plausible crowd chants from the visual scenes without verifying whether they occur in the soundtrack.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *Advances in Neural Information Processing Systems*, 38:89863–89895, 2026b.

Yiman Zhang, Ziheng Luo, Qiangyu Yan, Wei He, Borui Jiang, Xinghao Chen, and Kai Han. Omnieval: A benchmark for evaluating omni-modal models with visual, auditory, and textual inputs. *arXiv preprint arXiv:2506.20960*, 2025.

Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Xihan Wei, Liefeng Bo, et al. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025.

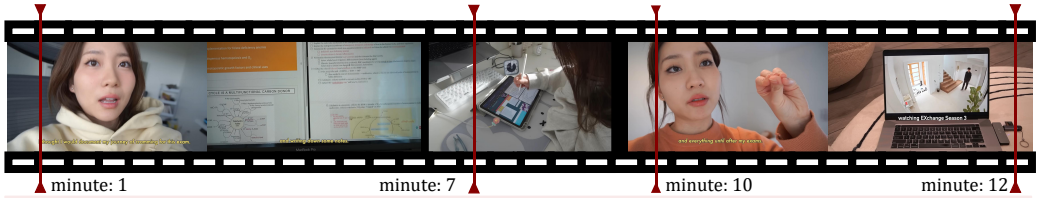
Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Error Analysis

We further examine representative errors to understand why long-context access does not necessarily translate into multi-hop reasoning. We select three cases from Gemini 3.1 Pro, the strongest model in our evaluation, as shown in Figures 6–8.

Common multi-hop failure. The three cases share a common structure: the model handles individual moments more successfully than it composes them into a complete trajectory. In Figure 6, the goals and celebrations are visually recognized, but generic cheering and commentator mentions are conflated with a distinct rhythmic crowd chant. The model consequently substitutes a cross-modal prior for verified audio evidence. In Figure 7, one valid antecedent is recovered, yet the later study behavior and examination completion are not jointly incorporated. In Figure 8, the local post-game scene is interpreted without maintaining a consistent identity-location representation across preceding



Backward Causal Reasoning

Question: Later in the video, the woman relaxes on her white sofa at night, eating a golden-brown fish-shaped snack and watching the K-drama 'Exchange Season 3' on her laptop. Based on the events shown earlier, which prior academic situation(s) ultimately led to her being able to relax like this?

Options:

A: She set up her study schedule at the start, explaining that a big cardiovascular/renal/respiratory exam in three days would heavily impact her grade and planning to cram for it over three full days.

B: She had just finished and done well on a major hematology-related exam, which she talks about while wearing an orange sweater and describing her post-exam cleaning and laundry plans.

C: She spent earlier days pushing through huge Anki decks—declaring she would get through 992 cards and stocking up on energy drinks—to prepare intensively for her upcoming exam."

D: She completed a separate assignment that was due in about five hours, working on cardiovascular and liver ultrasound content on her laptop late at night before the snack.

✗ *Gemini 3.1-pro:* A

Error analysis: Gemini captures the spoken exam setup in A but fails to integrate the later audio-visual evidence. It misses the visual study behaviors in C and the combined spoken and visual confirmation of exam success in B, preventing it from reconstructing the full causal chain leading to her relaxation. This reflects weak audio-visual grounding in causal reasoning.

Figure 7: Case of a Backward Causal Reasoning error. The model retrieves partial evidence but fails to reconstruct the complete causal chain across the video.



Spatial Reasoning

Question: Immediately after the 'JASON WINS GAME 2 7-1' graphic appears, where is Jason Rivera located relative to the playing court, based on how he is identified on the floor earlier and how he appears right after Game 2 ends?

Options:

A: Down by the opposite baseline near the crowd, standing in his game outfit with the ball at his side.

B: Off the playing floor near the sideline, standing and talking in a different white 'THE NEXT CHAPTER' shirt and cap.

C: Still on the court in his white Nike t-shirt and orange-trimmed shorts, standing behind the game helder.

D: Seated on the bench next to the Curly-Haired Man in the black 'A NEW WAVE NEXT CHAPTER' t-shirt along the near sideline.

✗ *Gemini 3.1-pro:* B

Error analysis: Gemini fails at multi-hop visual reasoning by incorrectly identifying the man in the different shirt and cap as Jason. It does not preserve Jason's earlier visual identity—white Nike shirt and orange-trimmed shorts—across the game sequence, and instead assumes that the person shown after the result graphic must be him. This reflects weak cross-shot identity tracking and spatial grounding rather than simple event localization.

Figure 8: Case of a Spatial Reasoning error. The model fails to preserve entity identity and spatial state across temporally separated shots.

shots. Before answering, the model does not reliably ensure that all required hops have been collected, correctly bound to their entities, times, and modalities, and jointly reconciled.

This behavior also illustrates the diagnostic advantage of TraceAV-Bench. Each question is grounded in an explicit, temporally localized, and modality-tagged evidence trajectory. A wrong answer can therefore be attributed to missing evidence, incorrect cross-hop binding, or failed aggregation, rather than being reduced to a single undifferentiated accuracy score. Moreover, the questions cannot be solved from one salient frame, one transcript sentence, or a short temporal neighborhood, and the model must reason over several dispersed events.

Failure modes across the four dimensions. Our broader manual audit reveals distinct and recurring failure patterns beyond the three illustrated examples. In **Audio-Visual Joint Reasoning (AVR)**, a common error is incomplete evidence coverage: models find one valid event and answer before checking the remaining hops, producing only a subset of the correct options. Even when all relevant events are recognized, they may be assigned the wrong temporal order, linked to the wrong entity, or matched across the wrong audio-visual boundary. Similar scenes create strong interference in localization, while causal questions expose poor episode segmentation: models select antecedents that are true and temporally nearby but not causally responsible for the queried outcome. In **Visual-Centric Reasoning (VR)**, models merge distinct but visually similar occurrences, count repeated views as new instances, or miss sparse events near the end of the video. Spatial errors often reflect stale state representations after a scene transition, identity confusion following changes in clothing or viewpoint, and reversal of camera-relative relations such as near versus far. In **Audio-Centric Reasoning (AR)**, models tend to preserve the most salient sound while losing less prominent instances distributed across the trajectory. They also compress a changing soundtrack into an imprecise global summary, confusing the genre, intensity, or narrative role of music across different stages of the video. In **Multimodal Hallucination (MH)**, the dominant behavior is schema-driven completion: models invent visually plausible object attributes, expected crowd chants, or conventional outro speech that never occurs. For temporal splicing, they often verify each fragment independently but fail to test whether the fragments actually form the claimed global chronology or causal narrative.

The persistence of these errors in the strongest evaluated models shows that the long video audio-visual multihop reasoning capability remains an open challenge.

B Evidence-Hop Masking Analysis

We use evidence-hop masking to examine the extent to which model predictions depend on the annotated evidence trajectories. This analysis has three expected patterns. First, masking one randomly selected hop (denoted as Mask-1) should reduce accuracy if individual hops contain task-relevant evidence. Second, masking all hops (denoted as Mask-All) should cause a further decline if the annotated hops provide complementary information. Third, performance under Mask-All should approach random guessing if the annotated trajectory captures a substantial share of the evidence used to solve each question.

Establishing strict necessity for every individual hop would require exhaustive conditional interventions over hop subsets and matched non-evidence segments, whose number grows combinatorially with trajectory length. We therefore use this analysis to evaluate trajectory-level evidence relevance and cumulative utility.

We evaluate these expectations under three input conditions. *Full* uses the original audio-visual video. For *Mask-1*, we randomly select one annotated evidence hop for each question, replace the corresponding video segment with black frames, and mute its audio. For *Mask-All*, we apply the same operation to every annotated hop, where each hop is mapped to its annotated temporal interval. When only a minute-level timestamp is available, we use the corresponding one-minute interval, and possible overlapping intervals are merged.

Evidence-hop masking probes dependence on annotated trajectories. The results show a consistent decline as more annotated evidence is masked. Averaged across the four evaluated models, Gen. Avg decreases from 43.84% under Full to 28.87% under Mask-1 and 22.02% under Mask-All. The Mask-All result is close to the 20.29% accuracy obtained by random guessing.

Masking one hop reduces accuracy in 47 of the 48 combinations of models and sub-tasks. This consistency indicates that a randomly selected annotated hop often contributes task-relevant information. More importantly, Mask-All underperforms Mask-1 in all 48 combinations. The remaining hops therefore contribute additional evidence after one hop has already been removed. This systematic second decline is consistent with the annotated trajectory providing cumulatively useful evidence.

Finally, Mask-All reduces the average result from 43.84% to 22.02%, which is close to random guessing. This convergence indicates that the annotated trajectories capture task-relevant evidence for these questions. Together, the results provide empirical evidence that the general-task trajectories contain complementary and cumulatively useful information.

	AVR							VR		AR			
Condition	**IR**	**TS**	**ET**	**FCR**	**BCR**	**CMM**	**SL**	**SR**	**VC**	**SC**	**ES**	**BM**	**Gen. Avg**
Qwen3-Omni-30B-A3B													
Full	47.14	51.55	35.48	43.84	50.56	40.00	32.60	58.18	38.50	63.85	59.09	60.31	48.43
Mask-1	26.43 (-20.71%)	27.84 (-23.71%)	22.58 (-12.90%)	23.29 (-20.55%)	22.47 (-28.09%)	18.82 (-21.18%)	23.79 (-8.81%)	39.39 (-18.79%)	23.89 (-14.61%)	44.62 (-19.23%)	44.32 (-14.77%)	41.98 (-18.33%)	29.95 (-18.48%)
Mask-All	20.00 (-27.14%)	23.71 (-27.84%)	15.32 (-20.16%)	15.07 (-28.77%)	12.36 (-38.20%)	14.12 (-25.88%)	20.70 (-11.90%)	29.09 (-29.09%)	20.80 (-17.70%)	37.69 (-26.16%)	34.09 (-25.00%)	35.88 (-24.43%)	23.24 (-25.19%)
OmniVinci-9B													
Full	49.29	44.33	38.71	57.53	34.83	35.29	33.48	55.15	34.51	65.38	65.91	54.20	47.38
Mask-1	44.29 (-5.00%)	29.90 (-14.43%)	20.16 (-18.55%)	35.62 (-21.91%)	21.35 (-13.48%)	20.00 (-15.29%)	23.79 (-9.69%)	41.82 (-13.33%)	23.89 (-10.62%)	46.92 (-18.46%)	50.00 (-15.91%)	39.69 (-14.51%)	33.12 (-14.26%)
Mask-All	29.29 (-20.00%)	17.53 (-26.80%)	12.90 (-25.81%)	21.92 (-35.61%)	12.36 (-22.47%)	14.12 (-21.17%)	20.70 (-12.78%)	32.12 (-23.03%)	20.80 (-13.71%)	38.46 (-26.92%)	40.91 (-25.00%)	33.59 (-20.61%)	24.56 (-22.82%)
Gemma 4-E4B													
Full	39.29	38.14	37.10	36.99	29.21	36.47	16.74	55.15	34.07	55.38	54.55	58.02	40.93
Mask-1	23.57 (-15.72%)	22.68 (-15.46%)	19.35 (-17.75%)	21.92 (-15.07%)	16.85 (-12.36%)	18.82 (-17.65%)	23.79 (+7.05%)	43.64 (-11.51%)	23.89 (-10.18%)	39.23 (-16.15%)	40.91 (-13.64%)	42.75 (-15.27%)	28.12 (-12.81%)
Mask-All	20.00 (-19.29%)	17.53 (-20.61%)	12.90 (-24.20%)	16.44 (-20.55%)	10.11 (-19.10%)	14.12 (-22.35%)	20.70 (+3.96%)	30.30 (-24.85%)	20.80 (-13.27%)	30.77 (-24.61%)	29.55 (-25.00%)	33.59 (-24.43%)	21.40 (-19.53%)
Video-SALMONN 2													
Full	42.14	41.24	29.03	30.14	29.21	32.94	31.28	48.48	39.38	47.69	47.73	44.27	38.63
Mask-1	25.71 (-16.43%)	22.68 (-18.56%)	16.13 (-12.90%)	20.55 (-9.59%)	12.36 (-16.85%)	18.82 (-14.12%)	23.79 (-7.49%)	33.33 (-15.15%)	23.89 (-15.49%)	35.38 (-12.31%)	31.82 (-15.91%)	26.72 (-17.55%)	24.27 (-14.36%)
Mask-All	20.00 (-22.14%)	17.53 (-23.71%)	12.90 (-16.13%)	15.07 (-15.07%)	8.99 (-20.22%)	14.12 (-18.82%)	20.70 (-10.58%)	24.85 (-23.63%)	20.80 (-18.58%)	28.46 (-19.23%)	23.86 (-23.87%)	19.08 (-25.19%)	18.86 (-19.77%)
Random guess	23.39	21.39	16.33	20.21	12.64	18.53	23.57	25.00	23.78	20.58	18.75	19.27	20.29

Table 4: Evidence-hop masking results on the 12 general sub-tasks of TraceAV-Bench (%). **Full** retains the original audio-visual input, **Mask-1** removes one randomly selected annotated evidence hop, and **Mask-All** removes all annotated evidence hops. Sub-task abbreviations follow Table 2. **Gen. Avg** is the unweighted macro-average over the 12 sub-tasks. Parenthesized values denote absolute accuracy changes from Full, with red and green indicating degradation and improvement, respectively. Models follow their order in Table 2.

Hallucination tasks show the same dependence on annotated trajectories. The MH results exhibit an even clearer progression. Averaged across the four models, MH Avg decreases from 54.46% under Full to 28.03% under Mask-1 and 20.96% under Mask-All. The Mask-All result nearly matches the 20.39% accuracy of random guessing. Moreover, all 12 combinations of models and sub-tasks follow the same ordering from Full to Mask-1 and then to Mask-All.

Together with the independent text-only evaluation in Appendix I, these results show that the evaluated models are sensitive to the source evidence supplied by the annotated trajectory. The consistent decline from Mask-1 to Mask-All further indicates that the annotated trajectory contributes cumulatively useful evidence for premise verification.

Taken together, Tables 4 and 5 show that performance is sensitive to the annotated evidence trajectory. Masking one annotated hop generally impairs prediction, while masking the full trajectory reduces performance to approximately random-guessing levels in both evaluation settings. These results provide empirical evidence that the annotated trajectories contain task-relevant, cumulatively useful evidence for solving the benchmark.

C Detailed Task Definitions

This section gives the full definition of each of the fifteen sub-tasks of TraceAV-Bench, grouped by the four evaluation dimensions introduced in the paper. For every sub-task we specify both *what the question asks* and, more importantly, *which capability of OmniLLMs the task is designed to probe*.

C.1 Audio-Visual Joint Reasoning (AVR)

The AVR dimension is the conceptual core of TraceAV-Bench. Every AVR sub-task is constructed so that the answer cannot be recovered from any single modality nor from any single moment in the video. Instead, the model must assemble evidence that is simultaneously dispersed across the visual

Condition	V2A	A2V	TSF	MH Avg
Qwen3-Omni-30B-A3B				
Full	65.65	69.87	66.87	67.46
Mask-1	30.87 (-34.78%)	32.31 (-37.56%)	27.71 (-39.16%)	30.30 (-37.16%)
Mask-All	23.48 (-42.17%)	24.89 (-44.98%)	20.48 (-46.39%)	22.95 (-44.51%)
OmniVinci-9B				
Full	42.17	44.10	42.17	42.81
Mask-1	28.70 (-13.47%)	30.57 (-13.53%)	30.12 (-12.05%)	29.79 (-13.02%)
Mask-All	19.57 (-22.60%)	23.58 (-20.52%)	20.48 (-21.69%)	21.21 (-21.60%)
Gemma 4-E4B				
Full	74.35	69.43	56.02	66.60
Mask-1	28.26 (-46.09%)	28.38 (-41.05%)	22.29 (-33.73%)	26.31 (-40.29%)
Mask-All	21.74 (-52.61%)	21.83 (-47.60%)	16.27 (-39.75%)	19.95 (-46.65%)
Video-SALMONN 2				
Full	45.65	39.30	37.95	40.97
Mask-1	32.17 (-13.48%)	24.45 (-14.85%)	20.48 (-17.47%)	25.70 (-15.27%)
Mask-All	23.91 (-21.74%)	19.65 (-19.65%)	15.66 (-22.29%)	19.74 (-21.23%)
Random guess	21.74	22.27	17.17	20.39

Table 5: Evidence-hop masking results on the Multimodal Hallucination dimension of TraceAV-Bench (%). V2A = *Visual-to-Audio Deception*, A2V = *Audio-to-Visual Deception*, and TSF = *Temporal Splicing Fallacy*. **MH Avg** is the unweighted macro-average over the three MH sub-tasks. Parenthesized values denote absolute accuracy degradation from Full. Models follow their order in Table 3.

and auditory channels and along the long temporal axis. The seven sub-tasks below probe orthogonal axes of cross-modal multi-hop reasoning.

- ***Information Retrieval (IR):*** IR probes whether a model can retrieve a specific fact (an on-screen number, a spoken name, an embedded text) when the retrieval requires fusing partial cues from at least two distant events, for instance, when the identity of an entity is established at one moment and a fact about that entity is uttered much later. The task therefore stresses long-context cross-modal grounding rather than simply fact recall.
- ***Temporal Sequencing (TS):*** Even when models recover the right set of facts, they often lose the timeline. TS asks the model to chronologically order a small set of events or entity-state changes drawn from non-adjacent positions in the video, isolating the question of whether internal representations preserve temporal order across long spans.
- ***Entity Tracking (ET):*** A persistent challenge for OmniLLMs is re-identifying the same entity after the visual surface changes (clothing, lighting, viewpoint) or the auditory surface changes (different speaker, ambient noise). ET targets the capability that the model must follow a single entity across multiple separate events and reason about its evolving role, relationships, or state.
- ***Forward Causal Reasoning (FCR):*** FCR begins with an early "cause" event and asks for its eventual consequence in a much later event. The reasoning chain must traverse intermediate events whose direct relevance is not immediately obvious, isolating the model's ability to project consequences forward across a long timeline.
- ***Backward Causal Reasoning (BCR):*** BCR operates in the reverse direction of FCR. The question is anchored to a late observable outcome and asks for its earlier root cause. We deliberately distinguish between forward and backward causality because the two directions exercise different inference patterns. Backward reasoning, in particular, requires the model to distinguish the true cause from spurious antecedents that merely co-occur along the timeline.
- ***Cross-Modality Matching (CMM):*** CMM probes audio-visual binding under temporal dispersion. The model must pair an auditory cue (a distinctive voice, a sound effect, a musical motif) with its correct visual referent among several visually plausible candidates that appear at different times.

The task is designed to expose models that handle each modality competently in isolation but fail to bind them when the evidence is temporally separated.

- ***Spatiotemporal Localization (SL):*** SL asks for the minute-level timestamp at which a specified target event occurs, where the timestamp itself must be triangulated by chaining clues from at least two different events. By forcing the model to commit to an explicit time, the task externalizes whether the model carries a faithful internal clock.

C.2 Visual-Centric Reasoning (VR)

The VR dimension isolates the visual modality by drawing all evidence from the visual stream alone. Its two sub-tasks correspond to the two visual capabilities most commonly degraded by long context:

- ***Spatial Reasoning (SR):*** SR asks about spatial relationships (relative positions, directions, layouts) that must hold consistently across multiple events. A correct answer requires the model to maintain a stable mental map of the scene over many minutes of video, rather than re-derive geometry from a single sampled frame.
- ***Visual Counting (VC):*** Open-ended counting is a well-known failure mode of MLLMs. VC further conditions the count on a constraint that is itself defined by another event (e.g., *"how many times does action X occur after event Y begins?"*), so that the counting target cannot even be located without first solving an additional reasoning hop.

C.3 Audio-Centric Reasoning (AR)

The AR dimension complements VR by isolating the auditory channel. Because audio in long videos arrives as a continuous stream rather than as discrete chunks, the three sub-tasks each target a distinct acoustic sub-stream: speech, non-speech environmental sound, and background music.

- ***Speech Context (SC):*** SC requires reasoning over spoken dialogue or narration spanning multiple events. To preempt visual leakage, the questions are constructed so that the answer is fully derivable from what is said and not from what is seen, providing a controlled probe of long-form speech comprehension that the visual stream cannot shortcut.
- ***Environmental Sound (ES):*** ES targets non-speech, non-musical sounds such as a siren, a closing door, or machinery. These cues are typically brief and easily missed, and threading them across distant events tests whether the model genuinely attends to acoustic events.
- ***Background Music (BM):*** BM asks about background music or ambient soundscape, including how musical shifts align with the unfolding narrative. The task probes a form of long-horizon audio understanding that carries strong narrative signal in real-world long videos.

C.4 Multimodal Hallucination (MH)

The MH dimension uses the same exact-match multiple-choice evaluation and measures whether a model can verify whether a cross-modal or temporal premise is supported or contradicted by the source video. Long audio-visual content is a fertile ground for hallucination, because evidence is sparse and the model must commit to an answer under uncertainty. A model that aces AVR but cannot reject false premises is not yet trustworthy.

- ***Visual-to-Audio Deception (V2A):*** The question is grounded in genuine visual evidence drawn from the video and asks whether a corresponding audio detail is supported within a specified event or time range. Items include both supported and contradicted premises, testing whether models verify rather than confabulate the audio half suggested by visual context.
- ***Audio-to-Visual Deception (A2V):*** A2V is the dual of V2A, where the question is anchored in genuine audio evidence and asks whether a corresponding visual detail is supported within a specified event or time range. Items include both supported and contradicted premises, testing whether models verify visual content rather than synthesizing details suggested by the audio context.
- ***Temporal Splicing Fallacy (TSF):*** TSF presents either a supported temporal narrative or a narrative that splices real but temporally isolated fragments into an explicitly contradicted sequence. The

model must verify the temporal relation from the observed timeline, which tests whether it reasons about the global temporal layout of evidence rather than verifying each fragment in isolation.

D More Detailed Related Work

D.1 Omnimodal Large Language Models

Real-world perception is inherently multi-sensory, and closing the gap between human and machine understanding requires models that can handle vision, audio, and language. The research community has broadened the perceptual scope of LLMs, giving rise to Vision-Language Models (VLMs) Wang et al. (2025d); Bai et al. (2025); Team (2026b); Marafioti et al. (2025); Guo et al. (2025a); Liu et al. (2025a); Xiaomi & Team (2025), Audio-Language Models (ALMs) Tseng et al. (2025); Ding et al. (2025); Ghosh et al. (2025); Chu et al. (2024); Arora et al. (2025), and OmniLLMs Liu et al. (2025b); Luo et al. (2025); Fu et al. (2024); Guo et al. (2025b); Team (2026b); Tian et al. (2026); Ding et al. (2026) that process text, images, videos, and audio within a unified framework. Both open-source models Ye et al. (2025); AI et al. (2025); Zhao et al. (2025); Tang et al. (2025); Xu et al. (2025b;a) and closed-source systems such as the Gemini series Comanici et al. (2025) have shown strong performance across audio and visual tasks.

Recent work further improves fine-grained audio-visual description, temporal alignment, dialogue grounding, robust audio perception, long-form reasoning, and semantic-spatial binding Ma et al. (2026b); Chen et al. (2026c); Tang et al. (2026); Yan et al. (2026); Ghosh et al. (2026c); Chen et al. (2026a). These advances provide stronger training data, representations, and model designs for processing extended audio-visual inputs. Agentic long-video systems complement this progress by using iterative search, active observation, structured memory, entity-aware indexing, and tool orchestration to gather task-relevant evidence Zhang et al. (2026b); Wang et al. (2025g); Rege et al. (2026); Yin et al. (2026); Tao et al. (2025); Xing et al. (2026b); Xu et al. (2026).

D.2 MLLM Benchmarks

MLLM benchmarks have evolved with model capabilities, progressing from modality-isolated suites for image Liu et al. (2024); Yue et al. (2024); Yu et al. (2023) and audio Sakshi et al. (2024); Wang et al. (2025a;b;e) understanding to specialized axes such as fine-grained perception Ghosh et al. (2026b); Qiang et al. (2025); Weng et al. (2026) and multimodal reasoning Ying et al. (2024); Yao et al. (2025); Du et al. (2025); Li et al. (2026a); Qian et al. (2025); Fu et al. (2026); Rawal et al. (2024). This broad benchmark ecosystem has played an important role in defining model capabilities and exposing systematic failure modes.

The rise of OmniLLMs has led to audio-visual benchmarks covering general perception, cross-modal understanding, reasoning, and model diagnosis Zhou et al. (2025); Li et al. (2024b); Chen et al. (2026b); Cheng et al. (2025); Chen et al. (2025a); Gong et al. (2024); Zhang et al. (2025). Recent efforts expand this scope through cognitively organized evaluation, long and complex real-world videos, instruction-following captioning, causal analysis of music videos, and evidence-chain data construction Wang et al. (2026b); Goel et al. (2026); Wang et al. (2026a); Ghosh et al. (2026a); Cai et al. (2026). Other benchmarks make valuable contributions to multimodal trustworthiness by testing audio-visual inconsistency and speech-vision hallucination Chen et al. (2026d); Zhang et al. (2026a). Together, these works provide broad task coverage and increasingly fine-grained evaluation of audio-visual intelligence.

Long-form video evaluation has also advanced rapidly. Existing benchmarks assess long omni-modal understanding Tao et al. (2026); Li et al. (2025); Han et al. (2025), grounded multi-hop reasoning in egocentric videos, movie and television understanding, open-ended movie question answering, multi-timescale comprehension, narrative event structure, and dispersed audio-visual evidence reasoning Chen et al. (2025b); Ataallah et al. (2025); Shaar et al. (2026); Ma et al. (2026a); Asgarov et al. (2026); Xu et al. (2026). These studies establish strong protocols for testing extended temporal context and show that long-video understanding contains several distinct challenges. Building on this foundation, TraceAV-Bench focuses on explicit, temporally dispersed, and modality-tagged multi-hop trajectories, with AVR requiring both audio and vision and VR/AR deliberately isolating one modality. It further places general reasoning and multimodal hallucination evaluation in the same long-form benchmark.

Algorithm 1: Agentic Question Generation Pipeline (Step 3).

Input : Bimodal event catalog $\mathcal{M} = [m_1, \ldots, m_N]$; task type τ; span threshold δ.
Output : Released MCQ set $\mathcal{Q}$.

Stage *1 – Event Segmentation Agent*:
- $\mathcal{E} \leftarrow \varnothing$; open event $e \leftarrow \{\text{captions: } [m_1],\ \text{summary: } m_1\}$
- **for** $t = 2$ **to** N **do**
  - $(a,\ s') \leftarrow \texttt{SegAgent}(e.\text{summary},\ m_t,\ m_{t+1},\ m_{t+2})$
  - **if** $a =$ CONTINUE **then**
    - extend e with m_t; $e.\text{summary} \leftarrow s'$
  - **else**
    - $\mathcal{E} \leftarrow \mathcal{E} \cup \{e\}$; open new event from m_t
- $\mathcal{E} \leftarrow \mathcal{E} \cup \{e\}$

Stage *2 – Trajectory Proposal Agent*:
- $\mathcal{C} \leftarrow \texttt{ProposeAgent}(\mathcal{E},\ \tau)$
- $\mathcal{T} \leftarrow \left\{\, c \in \mathcal{C} \,:\, \text{span}(c) \geq \delta \ \wedge\ \text{adjacent events of } c \text{ share a named entity} \right\}$

Stage *3 – Question Generation Agent*:
- $\mathcal{Q} \leftarrow \varnothing$
- **for** *each trajectory* $\mathbf{t} \in \mathcal{T}$ **do**
  - $q \leftarrow \texttt{QGenAgent}(\mathbf{t},\ \tau)$
  - refine each per-hop timestamp against the raw minute captions
  - randomly permute the four answer slots of q
  - $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{q\}$

return $\mathcal{Q}$

D.3 More details of the Agentic Question Generation Pipeline

Algorithm 1 consumes the per-minute bimodal catalog $\mathcal{M}$ produced by Steps 1 and 2 and a single span threshold δ that controls how temporally distant the evidence in a multi-hop chain must be, and it returns the set of trajectory-grounded MCQs $\mathcal{Q}$ that this video contributes to TraceAV-Bench. We now walk through the three stages in turn.

Stage 1: from minutes to events. Directly exposing downstream agents to the entire stream of minute-level captions would force them to reason over dozens of nearly-redundant entries, since adjacent minutes typically belong to the same scene. We therefore collapse $\mathcal{M}$ into discrete event blocks via a sliding-window segmentation agent. At each minute t, the agent jointly inspects the running summary of the currently open event, the current minute m_t, and a two-minute look-ahead (m_{t+1}, m_{t+2}), and decides whether m_t continues the event or marks a boundary. The look-ahead is what distinguishes our segmentation from a purely causal scan. When narrative threads briefly interleave (a cutaway, a flashback, a parallel scene), the agent can confirm whether the interruption is incidental or persistent before committing to a boundary, which prevents premature fragmentation of long coherent events. The output $\mathcal{E}$ is a partition of the timeline into self-contained event blocks, each carrying its constituent captions, a fused summary, and the union of its entity occurrences.

Stage 2: from events to multi-hop trajectories. Operating now on the much shorter sequence $\mathcal{E}$, a trajectory proposal agent generates candidate multi-hop chains together with a per-chain question-direction hint that records the focus, answer cue, distractor cue, and answer-mode preference of the question the chain is intended to support. To guarantee that the resulting chains exercise genuine multi-hop reasoning, every candidate is screened by two structural conditions. First, its temporal span must exceed δ event blocks, ruling out trivial chains over consecutive scenes. Second, every pair of adjacent events in the chain must share at least one named entity, so that each hop is bridged by explicit coreference rather than by thematic coincidence; this entity-bridge requirement is central to enforcing that no individual event in the chain is alone sufficient to answer the question. The chains that survive these conditions form $\mathcal{T}$ and are passed to question generation.

Stage 3: from trajectories to MCQs. For each retained trajectory, a question generation agent produces a four-option MCQ following the three generation principles stated in the main paper. An

Video Corpus		Question Pool		Reasoning Trajectories	
Total videos	578	Total questions	2,200	Evaluation dimensions	4
Total duration	339.5 hrs	Single-choice	1,848 (84.0%)	Sub-tasks	15
Duration range	10.1–139.9 min	Multi-choice	352 (16.0%)	Avg. hop count	3.68
Avg. duration	35.2 min	Avg. question length	43.6 words	Avg. temporal span	15.1 min
Resolution	73.7% HD or above	Avg. option length	22.1 words	Modality coverage	audio + visual

Table 6: Overview statistics of TraceAV-Bench, organized along three axes: the source video corpus, the released question pool, and the per-question multi-hop reasoning trajectories.

anti-shortcut formulation that forces the question to be answerable only by direct observation of the trajectory, trajectory-grounded distractors drawn from other events of the same video, and stylistic uniformity across the four options. The agent additionally annotates every hop with a modality label in $\{\text{audio}, \text{video}, \text{audio-video}\}$ and a candidate minute timestamp; the timestamp is then refined by mapping the cited evidence to its closest raw minute caption when it falls outside the event's time range, which yields the fine-grained per-hop grounding used by both the quality-assurance stage and the downstream evaluation. Finally, the four answer slots are randomly permuted on a per-question basis so that the position of the correct option is uncorrelated with the generation order, removing any positional bias that models might inherit from its prompting template.

E Statistics of TraceAV-Bench

This section provides the detailed statistics of TraceAV-Bench̊Table 6 reports the aggregate numbers under this three-axis view, and Table 7 breaks the question pool down further by sub-task.

Video corpus. The 578 videos in TraceAV-Bench amount to 339.5 hours of continuous audio-visual content, with an average length of 35.2 minutes and a broad range from just over 10 minutes to 2.3 hours. The visual quality of the collected videos is high, with 73.7% of videos at 720p or above and a non-trivial share at 4K. The corpus deliberately spans documentaries, vlogs, sports, music performances, and other genres so that no topical shortcut transfers across the benchmark.

Question pool. The 2,200 questions are uniformly four-option multiple-choice items but are not uniformly single-choice: 16.0% (352 items) are multi-choice, where two to four options can be simultaneously correct. Question stems average 43.6 words and options average 22.1 words.

Reasoning trajectories. Every question is grounded in an explicit multi-hop trajectory, which is the central structural property that distinguishes TraceAV-Bench from prior audio-visual benchmarks. A trajectory comprises 3.68 evidence hops on average and stretches across 15.1 minutes of video, so the typical question requires the model to integrate evidence from roughly four moments separated by a quarter of an hour of footage.

Per-sub-task structure. Table 7 breaks the 2,200 items down by the fifteen sub-tasks defined in Appendix C. Hop counts and temporal spans vary in interpretable ways across tasks. Visual Counting (4.24 hops) and Entity Tracking (4.19 hops) demand the deepest reasoning chains, since both require accumulating evidence over many events, while Information Retrieval (2.87 hops) is structurally the shallowest because the answer hinges on a single key fact assembled from a few cues. Spans are widest for Temporal Splicing Fallacy (19.8 min) and Cross-Modality Matching (19.3 min), both of which fundamentally rely on relating widely separated events.

F Quality Assurance

Quality assurance yield. The 2,200 questions retained in TraceAV-Bench are the result of an aggressive multi-stage quality filter applied to a much larger pool of candidate items emitted by the agentic generation pipeline of Section D.3. Table 8 reports, for each sub-task, the number of items produced before the quality assurance is applied, the number that ultimately enter TraceAV-Bench, and the corresponding retention rate. Across the fifteen sub-tasks, the overall retention rate is 50.5%. Roughly half of the candidate items are dropped during quality assurance. The task types that survive at slightly higher rates (localization, counting, and hallucination) are precisely those whose answer spaces are the most structurally constrained, such as integer minutes and integer counts, so a larger

Dim.	Sub-task (Abbrev.)	#Q	%	Hops	Span (min)	M-ch
AVR	Information Retrieval (IR)	140	6.4	2.87	13.5	9
	Temporal Sequencing (TS)	97	4.4	3.99	15.2	14
	Entity Tracking (ET)	124	5.6	4.19	18.3	43
	Forward Causal Reasoning (FCR)	73	3.3	3.11	10.1	14
	Backward Causal Reasoning (BCR)	89	4.0	3.53	14.1	44
	Cross-Modality Matching (CMM)	85	3.9	3.80	19.3	22
	Spatiotemporal Localization (SL)	227	10.3	3.40	12.1	13
VR	Spatial Reasoning (SR)	165	7.5	3.38	14.6	0
	Visual Counting (VC)	226	10.3	4.24	14.4	11
AR	Speech Context (SC)	130	5.9	3.22	15.7	23
	Environmental Sound (ES)	88	4.0	3.41	11.8	22
	Background Music (BM)	131	6.0	3.68	17.4	30
MH	Visual-to-Audio Deception (V2A)	230	10.5	3.60	14.2	30
	Audio-to-Visual Deception (A2V)	229	10.4	4.00	15.9	25
	Temporal Splicing Fallacy (TSF)	166	7.5	4.23	19.8	52
Total		**2,200**	**100.0**	**3.68**	**15.1**	**352**

Table 7: Per-sub-task statistics of TraceAV-Bench. *Hops*: mean number of evidence hops per question. *Span*: mean temporal gap, in minutes, between the first and last evidence hop. *M-ch*: number of multi-choice items.

Dim.	Sub-task (Abbrev.)	Pre-QA	Released	Dropped	Kept (%)
AVR	Information Retrieval (IR)	280	140	140	50.0
	Temporal Sequencing (TS)	215	97	118	45.1
	Entity Tracking (ET)	265	124	141	46.8
	Forward Causal Reasoning (FCR)	165	73	92	44.2
	Backward Causal Reasoning (BCR)	200	89	111	44.5
	Cross-Modality Matching (CMM)	195	85	110	43.6
	Spatiotemporal Localization (SL)	400	227	173	56.8
VR	Spatial Reasoning (SR)	340	165	175	48.5
	Visual Counting (VC)	400	226	174	56.5
AR	Speech Context (SC)	270	130	140	48.1
	Environmental Sound (ES)	195	88	107	45.1
	Background Music (BM)	280	131	149	46.8
MH	Visual-to-Audio Deception (V2A)	405	230	175	56.8
	Audio-to-Visual Deception (A2V)	400	229	171	57.3
	Temporal Splicing Fallacy (TSF)	350	166	184	47.4
Total		**4,360**	**2,200**	**2,160**	**50.5**

Table 8: Quality assurance yield per sub-task. *Pre-QA*: number of candidate items produced by the agentic generation pipeline before any quality filter is applied. *Released*: number of items retained in TraceAV-Bench. *Dropped*: items removed by the quality assurance stage. *Kept*: retention rate, computed as Released / Pre-QA.

fraction of generated candidates already complies with the rule-based portion of the filter and never reaches the more aggressive LLM-based audits.

Stage-wise attrition. Figure 9 visualizes how candidate items flow through the four sequential QA stages described in Section 3.3. Starting from a Pre-QA pool of 4,360 items, *Rule-based Validation* removes 312 format- or metadata-violating items, *Logical Verification* prunes the largest share (1,047 items) by filtering pseudo multi-hop questions and implausible distractors, *Blindfold Shortcut Detection* eliminates 578 items answerable without the video, and *Human Expert Auditing* discards a final 223 items through reviewer edits and batch rejection, yielding the 2,200 questions released in TraceAV-Bench.

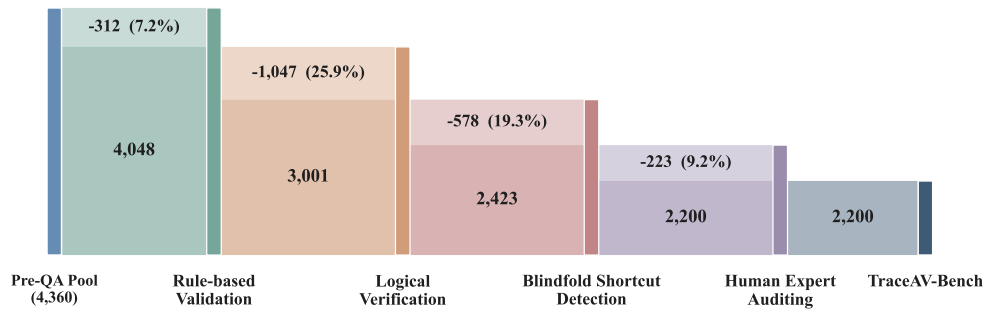


Figure 9: Flow of candidate items through the four-stage quality assurance pipeline.

G Human Auditing Protocol and Statistical Reliability

Human role, background, and training. Human experts intervened at source-video screening, review of textual intermediates before question generation, final construction-stage quality control, and release auditing. The release audit was conducted by six graduate-level annotators: four PhD students and two master's students in data science or artificial intelligence. All annotators had prior research experience in multimodal learning or benchmark construction. They were native Chinese speakers with professional working proficiency in English. Before formal auditing, annotators completed a training session, studied a written decision manual, and independently labeled 30 calibration items. Calibration disagreements were discussed against the source evidence before the formal audit began. Each sampled item was independently labeled by three annotators, and a senior PhD annotator who had not supplied an initial label adjudicated disagreements. Auditors were compensated at USD 15 per hour for training, annotation, and adjudication, consistent with institutional research-assistant rates.

Review of textual intermediates. The LLM-generated event catalog was treated as an intermediate representation rather than ground truth. For every retained video, an author replayed the original audio-video segments and checked the visual captions, fused captions, and entity cache for event coverage, speech and sound fidelity, cross-segment identity consistency, temporal alignment, language errors, and unsupported or sensitive inferences. Local defects were corrected before question generation, while systematically unreliable catalogs were returned for regeneration. This pre-generation gate limits error propagation because event segmentation, trajectory proposal, and MCQ generation all depend on these textual intermediates.

Source-grounded release-audit guideline. For each sampled item, all three auditors watched every cited segment with its original audio and video, expanding to adjacent context whenever event boundaries or causal relations were unclear. They formed a source-based judgment before consulting the generated event catalog, MCQ, and trajectory metadata. An *answer error* meant that the key was unsupported, incomplete, or non-unique; a *timestamp error* meant that the cited interval did not contain the claimed evidence or pointed to the wrong occurrence; and an *unnecessary-hop error* meant that a hop could be removed without making the answer underdetermined. A *modality-label error* recorded a mismatch between the assigned modality and the evidence actually required; *ambiguous wording* admitted a reasonable answer-changing interpretation; and an *invalid distractor* was duplicated, revealing, contradicted by the stem, or not plausibly confusable from the video. Reviewers independently assigned *pass*, *modify and retain*, or *reject*. Fleiss' κ was computed from the three pre-adjudication disposition labels, whereas the error and modification rates use adjudicated labels.

Sampling and reliability. All 2,200 released items have passed the construction pipeline, including its final human quality-control gate at either the inspected-item or batch level; this construction-stage screening is distinct from the release audit below. The stronger three-reviewer audit covers a fixed sample of 330 candidates from the provisional 2,200-item release pool, sampled without replacement within all 15 sub-tasks using proportional allocation with largest-remainder rounding. The dimension

totals were AVR/VR/AR/MH $= 126/59/52/93$. The task allocations, in the order of Table 7, were $21/15/19/11/13/13/34$, $25/34$, $19/13/20$, and $34/34/25$, respectively. For $N = 2{,}200$ and $n = 330$, the finite-population-corrected worst-case 95% margin is

$$1.96\sqrt{\frac{0.5(1-0.5)}{330}\,\frac{2{,}200-330}{2{,}200-1}} = 0.0497, \tag{1}$$

or $\pm 4.97\%$, and at a 5% prevalence it is ± 2.17 points. Because the proportional design is approximately self-weighting, hypergeometric inversion gives a finite-population 95% interval of 2.27–6.41% for the observed 13/330 item-level error rate (3.9%). For reference, zero observed errors would imply a one-sided 95% upper bound of 18 erroneous items in the population (0.82%), not proof that every unaudited item is error-free. The audit therefore quantifies residual-error uncertainty rather than certifying every item. Table 9 reports the adjudicated outcomes. The pre-adjudication Fleiss' κ of 0.82 indicates strong agreement among the three independent reviewers. The four rejected items were removed from the release pool and replaced with reserve items sampled from the same task strata. Each replacement underwent the same independent three-reviewer source-grounded audit and adjudication procedure before inclusion, so the final benchmark size remains 2,200. The rates in Table 9 use the original fixed 330-item sample; the replacements were not used to dilute the observed error rate.

Construction-stage rejection accounting. The preceding automatic stages removed 312/4,360 rule-violating items (7.2%), 1,047/4,048 logically invalid items (25.9%), and 578/3,001 blindfold-solvable items (19.3%). These conditional attrition rates characterize filter behavior, not residual ground-truth accuracy. A final-gate batch was a generation shard produced from a fixed source-video set and sub-task under one prompt/model version and generation run. During construction-stage screening, annotators inspected items sampled from every batch, covering 15% of the 2,423 candidates in total; a batch was rejected in full when more than 5% of its inspected items failed the audit rubric. Among the 112 batches entering the final human gate (median 21 items, range 12–32), nine exceeded this threshold, removing 171 items through whole-batch rejection. 52 additional candidates were removed through item-level decisions. Thus, $171 + 52 = 223$ candidates were removed at the final stage (9.2% of 2,423), leaving the 2,200-item release shown in Figure 9. These 223 construction-stage removals are distinct from the four items rejected in the subsequent audit of the provisional release pool.

H Additional Experimental Setup

All open-source models listed are deployed and evaluated on $8\times$NVIDIA H20 GPUs. We evaluate the closed-source models using their official APIs. We note that most current closed-source multimodal models, such as the GPT series, focus on the visual modality and do not natively accept audio input. Among the few that genuinely support joint audio-visual input, the Gemini series is the most representative. We therefore restrict our closed-source evaluation to the Gemini family. The specific versions used are as follows:

- Gemini-3.1-Pro-Preview: `gemini-3.1-pro-preview`
- Gemini-3-Flash-Preview: `gemini-3-flash-preview`
- Gemini-2.5-Pro: `gemini-2.5-pro`

Outcome/category	Count	Rate	Outcome/category	Count	Rate
Any item-level error	13	3.9%	Answer error	3	0.9%
Timestamp error	5	1.5%	Unnecessary-hop error	4	1.2%
Modality-label error	2	0.6%	Ambiguous wording	4	1.2%
Invalid distractor	6	1.8%	Modified and retained	9	2.7%
Rejected after audit	4	1.2%	Fleiss' κ	0.82	

Table 9: Adjudicated audit of the fixed 330-item provisional release sample. Error categories are non-exclusive; the item-level row counts each affected item once. The 13 affected items comprise nine corrected and retained items and four rejected items.

- Gemini-2.5-Flash: `gemini-2.5-flash`
- Gemini-2.0-Flash: `gemini-2.0-flash`
- GPT-5.1: `gpt-5.1-2025-11-13` (for data construction purpose only)

H.1 Long-video Input Configuration

For every video-capable model, we use the complete temporal span of each source video and do not crop inputs according to the annotated evidence locations. For the audio-only Qwen2-Audio baseline, we analogously use the complete source audio track. To obtain each model's strongest practical performance, we follow its officially supported inference configuration and provide as much of its supported modality input as its interface and context capacity permit.

Consequently, models may differ in their internal frame sampling or effective input density. Our objective is therefore not to isolate architectural reasoning ability under an identical information budget, but to evaluate the end-to-end long-video reasoning capability achievable under each model's practical supported configuration.

For every video-capable model, our inference pipeline submits the full source video rather than an evidence-centered clip. Within the officially supported interface of each model, we configure preprocessing and sampling to use the largest practical context allocation available to that implementation. Thus, any frame sampling or audio segmentation is allocated over the complete source timeline, not over a subclip selected using the annotated evidence locations. The resulting information density can differ across model interfaces; this is an inherent property of the end-to-end deployment setting considered here.

For direct-inference evaluation, each question is answered by the native inference path of the evaluated model from its supported source modality; we do not introduce an external captioner, retriever, voting procedure, or separate summarization model to combine segments before prediction. When an official processor internally samples or segments a long input, this operation remains part of that model's native preprocessing path.

I Additional Text-only Shortcut Analysis

To further assess potential linguistic shortcuts, we evaluate DeepSeek-V4.1-Flash, a strong model which was not used in data construction or quality filtering. During evaluation, we use only each question and its answer options, without video input. Table 10 reports its accuracy alongside the analytical uniform single-option guessing baseline used in the main results.

The text-only model obtains a 14.17% macro-average on the 12 general tasks, below the corresponding 20.29% baseline, and 20.13% on the three MH tasks, comparable to the 20.39% baseline. Thus, this independent strong language model does not reliably solve the released questions without access to the source video, providing an additional check against broad linguistic shortcuts.

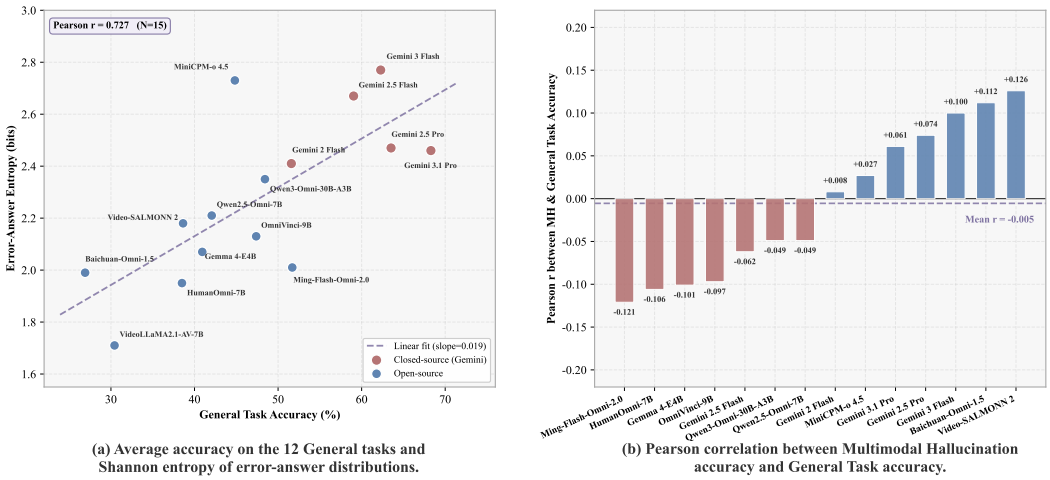
J More Experimental Findings

Stronger models make more *diverse* mistakes. Figure 10(a) plots each model's general-task accuracy against the Shannon entropy of its error-answer distribution, $H = -\sum_{i=1}^{k} p_i \log p_i$, where p_i is the frequency of the i-th error pattern (the sorted tuple of predicted options) over incorrect predictions. The two are strongly correlated (Pearson r=0.727), with weak models such as VideoLLaMA2.1 (30.4%, 1.71 bits) collapsing errors onto a few patterns and strong models such as Gemini 3 Flash (62.3%, 2.77 bits) spreading them broadly. This tight coupling between capability and error diversity suggests that error-answer entropy may reflect a model's potential, offering an indicator during evaluation throughout model training that complements what accuracy alone captures.

General-task performance is not a reliable indicator of hallucination robustness. Figure 10(b) shows, for each of the 14 models, the Pearson correlation between its Multimodal Hallucination (MH) accuracy and its general-task accuracy computed *across videos*. All 14 coefficients lie between -0.121 and +0.126, with a mean of r=-0.005. We use this video-level analysis as a descriptive diagnostic, since the number of questions available per video and dimension varies. The observed pattern is

Dim.	Sub-task	DeepSeek-V4.1-Flash	Random Guess
AVR	IR	17.86	23.39
	TS	27.84	21.39
	ET	7.26	16.33
	FCR	15.07	20.21
	BCR	13.48	12.64
	CMM	9.41	18.53
	SL	13.66	23.57
VR	SR	17.58	25.00
	VC	6.64	23.78
AR	SC	16.15	20.58
	ES	13.64	18.75
	BM	11.45	19.27
MH	V2A	20.43	21.74
	A2V	20.09	22.27
	TSF	19.88	17.17
General Avg. (12 tasks)		**14.17**	**20.29**
MH Avg. (3 tasks)		**20.13**	**20.39**

Table 10: Text-only evaluation of DeepSeek-V4.1-Flash on TraceAV-Bench (%). The model receives only the question and answer options. *Uniform Guess* denotes uniformly selecting one option from A–D; under exact-match scoring, this strategy receives zero expected credit on multi-choice items. Sub-task abbreviations follow Tables 2 and 3.



(a) Average accuracy on the 12 General tasks and Shannon entropy of error-answer distributions.

(b) Pearson correlation between Multimodal Hallucination accuracy and General Task accuracy.

Figure 10: Two diagnostic analyses of OmniLLMs on TraceAV-Bench.

consistent with general-task performance providing limited indication of MH performance on the same video. This mirrors the rank reversals in Table 3: models that perform better on general tasks are not consistently more robust to multimodal hallucinations.

K Evaluation of Agentic Audio-Visual Methods

Recent agentic methods use active observation, iterative evidence seeking, and tool-mediated memory to reduce the amount of long-video content processed at each reasoning step. We evaluate whether these designs improve multi-hop audio-visual reasoning on TraceAV-Bench. The evaluation covers three representative methods. *Active Video Perception* (AVP) follows a training-free plan, observe, and reflect workflow for query-conditioned evidence collection Wang et al. (2025g). *OmniAgent-RL-7B* is a native omni-modal agent trained with agentic supervised fine-tuning and reinforcement learning to alternate between frame, audio, and clip observations Xing et al. (2026b). *Audio-Guided OmniAgent*

Model or method	AVR							VR		AR			Avg
	IR	TS	ET	FCR	BCR	CMM	SL	SR	VC	SC	ES	BM	
Gemini 3.1 Pro	83.57	60.82	71.77	86.30	61.80	49.41	51.54	73.94	41.15	96.92	63.64	78.63	68.29
Ming-Flash-Omni-2.0	56.43	53.61	47.58	57.53	40.45	44.71	31.28	65.45	39.38	63.85	56.82	63.36	51.70
OmniVinci-9B	49.29	44.33	38.71	57.53	34.83	35.29	33.48	55.15	34.51	65.38	65.91	54.20	47.38
Gemma 4-E4B	39.29	38.14	37.10	36.99	29.21	36.47	16.74	55.15	34.07	55.38	54.55	58.02	40.93
Qwen3-Omni-30B-A3B backbone													
Direct	47.14	51.55	35.48	43.84	50.56	40.00	32.60	58.18	38.50	63.85	59.09	60.31	48.43
AVP	53.57	47.42	44.35	47.95	41.57	34.12	30.40	54.55	29.65	61.54	59.09	61.07	47.11
Audio-Guided	50.71	41.24	38.71	53.42	41.57	27.06	18.50	44.85	27.88	67.69	45.45	51.91	42.42
Qwen2.5-Omni-7B backbone													
Direct	46.43	32.99	37.10	30.14	35.96	37.65	37.44	49.70	35.40	60.00	52.27	49.62	42.06
OmniAgent-RL-7B	72.86	46.39	54.84	69.86	61.80	47.06	27.31	59.39	29.20	72.31	65.91	52.67	54.97

Table 11: General-task results of agentic audio-visual methods on TraceAV-Bench (%). General Avg is the macro-average over the 12 sub-tasks. Anchor models repeat representative results from Table 2 for context. AVP and Audio-Guided use direct Qwen3-Omni as the baseline. OmniAgent uses direct Qwen2.5-Omni, its base model.

Model or method	V2A	A2V	TSF	MH Avg
Gemini 3.1 Pro	89.57	79.91	84.34	84.61
Ming-Flash-Omni-2.0	71.30	67.25	62.65	67.07
OmniVinci-9B	42.17	44.10	42.17	42.81
Gemma 4-E4B	74.35	69.43	56.02	66.60
Qwen3-Omni-30B-A3B backbone				
Direct	65.65	69.87	66.87	67.46
AVP	70.87	71.18	67.47	69.84
Audio-Guided	60.87	50.66	62.05	57.86
Qwen2.5-Omni-7B backbone				
Direct	60.43	55.46	53.61	56.50
OmniAgent-RL-7B	40.00	47.16	51.20	46.12

Table 12: Hallucination robustness of agentic audio-visual methods on TraceAV-Bench (%). MH Avg is the macro-average over V2A, A2V, and TSF. Anchor models repeat representative results from Table 3.

uses audio cues to locate candidate events before applying fine-grained visual inspection Tao et al. (2025).

Evaluation setup. AVP uses Qwen3-Omni-30B-A3B as its planner, observer, reflector, and answer synthesizer, with at most three rounds and a confidence threshold of 0.7. Audio-Guided OmniAgent also uses Qwen3-Omni-30B-A3B for its brain and all audio and video tools, with a maximum of 30 iterations. We compare both methods with direct Qwen3-Omni-30B-A3B inference. OmniAgent uses the official OmniAgent-RL-7B checkpoint based on Qwen2.5-Omni-7B, with a maximum of 32 steps. Its matched baseline is direct Qwen2.5-Omni-7B inference. Note that the evaluated configurations differ from the default settings of the original methods. AVP uses Qwen3-Omni instead of Gemini-2.5-Pro. Audio-Guided OmniAgent uses Qwen3-Omni for both the brain and all tools instead of heterogeneous specialized models.

Finding 1: Performance gains derived from agentic methods are task-specific and do not transfer uniformly to hallucination robustness. OmniAgent achieves the strongest General Avg among the evaluated agentic methods at 54.97%, improving its direct Qwen2.5-Omni baseline by 12.91%. The largest gains occur on Information Retrieval, Forward Causal Reasoning, Backward Causal Reasoning, and Entity Tracking. Its performance decreases on Spatiotemporal Localization and Visual Counting. OmniAgent reduces MH Avg by 10.38% relative to direct Qwen2.5-Omni, with a 20.43% reduction on Visual-to-Audio Deception.

Method	Tokens per item	Mean wall	P95 wall
AVP	39,023	37.43 s	86.94 s
OmniAgent-RL-7B	133,348	308.82 s	785.69 s
Audio-Guided	200,105	75.94 s	244.51 s

Table 13: Inference cost of the three agentic methods over TraceAV-Bench. Mean and P95 wall time are computed from per-sample inference latency.

AVP reaches 47.11% General Avg, which is 1.32% below direct Qwen3-Omni. It improves Information Retrieval, Entity Tracking, and Forward Causal Reasoning by 6.43%, 8.87%, and 4.11%. AVP is the only evaluated agentic method that improves all three MH sub-tasks. It reaches 69.84% MH Avg and gains 2.38% over its direct baseline. Most of this gain is concentrated in multi-choice questions. AVP increases MH multi-choice accuracy from 18.69% to 42.06%, and its MH single-choice accuracy changes from 77.41% to 75.87%. Audio-Guided OmniAgent reduces General Avg by 6.01% and MH Avg by 9.60% relative to direct Qwen3-Omni.

Agentic gains concentrate on tasks with sparse, identifiable evidence, including retrieval, entity tracking, and forward causal reasoning. The same methods remain weak on global counting, precise temporal localization, cross-modal alignment, and premise verification.

Finding 2: Active evidence routing can both fix and introduce errors. Compared with direct Qwen3-Omni, AVP corrects 295 previously incorrect answers and introduces 306 new errors. Its overall accuracy therefore remains nearly unchanged. Audio-Guided OmniAgent corrects 346 errors and introduces 518 new errors, resulting in a clear overall decline. These results show that active routing is useful only when it selects the right evidence. Searching the wrong video segments leaves the model with incomplete information for the final answer.

Finding 3: Longer agent trajectories are associated with higher cost and lower success rates. Table 13 shows that AVP has the lowest recorded token count and per-sample latency among the three agentic methods. Audio-Guided OmniAgent records 5.13 times as many tokens per item as AVP under the same Qwen3-Omni serving stack. Errors also consume more tokens than correct predictions. The increases are 11.8% for AVP, 49.7% for OmniAgent, and 17.3% for Audio-Guided OmniAgent.

The same pattern appears when episodes are grouped by interaction length. AVP items that stop after one round achieve 54.49% accuracy with 33,178 tokens on average. Items reaching its three-round limit achieve 33.88% with 93,102 tokens. OmniAgent episodes using 2 to 6 steps achieve 61.10% with 35,243 tokens. Its 105 episodes reaching the 32-step limit achieve 26.67% with 458,798 tokens. Audio-Guided episodes using at most two high-level actions achieve 50.44% with 56,782 tokens. Its 202 episodes reaching the 30-step limit achieve 36.14% with 823,793 tokens. The capped episodes combine high inference cost with low accuracy. In Audio-Guided OmniAgent, 34 capped episodes repeatedly call the same audio QA tool after one global audio caption, showing that repeated tool calls remain a concrete failure mode.

L Prompt Templates

This section collects every LLM prompt used by the TraceAV-Bench pipeline, grouped by the stage in which it is invoked. Placeholder variables filled in at runtime are rendered as <variable_name>.

L.1 Pipeline Stage 1: Per-Minute Visual Captioning

The visual captioning stage processes each one-minute segment of a video while propagating a running entity cache to keep names and identities consistent across the entire video.

L.2 Pipeline Stage 2: Audio-Visual Caption Fusion

The fusion stage takes the per-minute visual caption together with the corresponding raw audio track and produces a single audio-visual integrated caption while updating entity descriptions with newly observed audio evidence.

Stage 1: Visual Captioning with Entity Tracking

Task: Analyze this 1-minute video segment as a continuous narrative.
[Identity Consistency Context]
Refer to these previously identified entities to maintain naming consistency:
<entity_cache>
[Analysis Requirements]

- Narrative Flow: describe the events as a dynamic sequence, focusing on how actions evolve from start to finish (<minute_summary>).
- Entity Tracking: (1) identify which entities from the Context above are actively visible or interacting in this specific minute (<present_entities>); (2) if any known entity has changed appearance, state, or role, update its description in <present_entities> to reflect the latest state; (3) identify any NEW prominent people or objects not listed in the Context (<new_entities>).
- Detail Capture: capture subtle gestures, clothing details, and any on-screen text.
- Environmental Shift: note changes in setting, lighting, or atmosphere if applicable.
- For every entity in <present_entities> and <new_entities>, always return a non-empty distinguishing description.

[Strict Constraints]
DO NOT list any entities from the Context that are NOT visible in this segment.
[Output Format]
Your response must be a valid JSON object with the keys <minute_summary>, <present_entities>, and <new_entities>.

Figure 11: Prompt used in Stage 1 to generate per-minute visual captions while propagating an entity cache for cross-minute identity consistency.

Stage 2: Audio-Visual Caption Fusion

```
You are an expert video analyst. Your goal is to synthesize
a 1-minute audio-visual integrated caption from two sources:
(1) [Visual Context], a summary of what is seen, and (2) [Audio
Track], the sounds and speech from the same segment.
[Visual Context]: <visual_summary>
[Visual Entities Present]: <visual_entities>
[Analysis Requirements]
    1. Information Preservation: the integrated caption MUST
       encapsulate all information from the visual context.
    2. Combined Narrative: integrate the visual description with
       the actual audio from this segment.
    3. Entity Alignment and Evolution: match voices and sounds to
       the visual entities above when a logical connection exists.
       If the audio reveals new details about an entity (a spoken
       name, a distinct voice, a characteristic sound), update its
       description. If an entity has no relevant audio in this
       minute, omit it from <updated_entities>.
    4. Audio Details: describe background music, environmental
       sounds, and the dialogue or speech context.
    5. Final Output: a single, high-density paragraph that
       captures both what is seen and what is heard, maintaining
       chronological and identity consistency.
[Output Format]
A valid JSON object with keys <minute>, <integrated_caption>, and
<updated_entities>.
```

Figure 12: Prompt used in Stage 2 to fuse the per-minute visual caption with the corresponding audio track into a single audio-visual caption.

L.3 Pipeline Stage 3a: Event Segmentation

To convert minute-level captions into discrete events, we slide an LLM judge across the timeline. At minute T, the model is shown the running summary of the ongoing event, the current minute, and a two-minute look-ahead, and decides whether the current minute continues, transitions out of, or hard-cuts away from the active event.

Stage 3a: Event Boundary Segmentation

```
Task: Analyze the narrative continuity between a past event
summary and the current video segment.
[Identity Consistency Context]
Active entities in the current minute: <current_entities>
[Past State Memory (Ongoing Event)]: <past_summary>
[Future Vision (Look-ahead Context)]: T+1 = <future_1>, T+2 =
<future_2>
[Analysis Requirements]
Determine the relationship of the Current Minute (Minute T) to the
Past State Memory:
    • CONTINUE: direct continuation of narrative and characters.
    • OVERLAP_TRANSITION: Minute T contains the resolution of the
      past and the start of the future.
    • HARD_CUT: total break from the past, aligning instead with
      the new context in the Future Vision.
[Output Format]
A valid JSON object with keys <action>, <reason>, and
<updated_summary>.
```

Figure 13: Prompt used in Stage 3a to slide an LLM judge across the timeline and decide CONTINUE / OVERLAP_TRANSITION / HARD_CUT at each minute, aggregating contiguous minutes into discrete events.

L.4 Pipeline Stage 3b: Trajectory Proposal

Given a video's event blocks, the trajectory proposer selects multi-hop chains of events that can support a high-quality question for the target task type τ. All fifteen task-specific prompts share the generic scaffold shown below. Note that only the task-specific *Core Principle* differs across tasks, and is listed individually for each task in Figures 16–30.

The fifteen task-specific instantiations (Core Principle for Stage 3b, Question Design and <task_specific_key> for Stage 3c) are listed individually in Figures 16–30 after the Stage 3c scaffold below.

L.5 Pipeline Stage 3c: Multiple-Choice Question Generation

Given a selected trajectory, the question generator produces a four-option multiple-choice question, assigns each evidence step a modality label, and writes a per-event minute timestamp. The fifteen task-specific prompts again share a common scaffold. The variable parts are the *Question Design* block and the <task_specific_key> pair, both listed per task in Figures 16–30.

Per-Task Instantiations We now list, for each of the fifteen sub-tasks defined in Appendix C, the three pieces that instantiate the generic Stage 3b/3c scaffolds for that task. The *Core Principle* that fills the Stage 3b proposer (Figure 14), the *Question Design* body that fills the Stage 3c generator (Figure 15), and the <task_specific_key> pair that the Stage 3c output is required to emit. The fifteen figures are ordered to follow the same task taxonomy as Appendix C (AVR $\rightarrow$ VR $\rightarrow$ AR $\rightarrow$ MH).

Stage 3b: Trajectory Candidate Proposer (Generic Scaffold)

Task: Select candidate multi-hop trajectories for <task_name> QA.
[Core Principle]
<task_specific_core_principle> *(per-task body in Figures 16-30)*
[Trajectory Quality]
Select only HIGH-QUALITY trajectories. A trajectory is high quality when ALL of the following hold:

1. STRICT Multi-hop: the intended question MUST be impossible to answer using any single event in the chain, and each hop must be a necessary piece of the puzzle.
2. No Common Sense / No Well-Known Facts: the answer MUST NOT be inferable from general world knowledge or widely known facts, and it must be uniquely tied to specific, non-obvious details of THIS video.
3. Challenging by design: the trajectory must support a question that would genuinely challenge an attentive viewer.
4. Strong evidence: every event in the chain contains concrete, specific evidence (named entity, exact number, quoted speech, visible action).
5. Clear entity bridge: adjacent events share at least one named entity whose state, role, or identity changes meaningfully between them.
6. Reject weak chains: do NOT include trajectories where the connection between events is only thematic or where the answer is guessable without watching the video.

[Question Direction]
For each candidate, propose a concrete and UNIQUE question direction. Across ALL candidates in this response, no two <focus> values may ask about the same fact, entity, or event. The fields <focus>, <answer_hint>, <distractor_hint>, and <answer_mode_hint> (single or multiple, respecting the per-task option-selection mode) must all be supplied.
[Output Format]
A valid JSON object containing a list <trajectory_candidates> of items, each with <trajectory_id>, <event_ids>, <why_multihop>, <range_type> (short or long), <bridge_type> (entity, semantic, temporal, or hybrid), and <question_direction>.

Figure 14: Generic scaffold for the Stage 3b trajectory proposer. The same scaffold is instantiated fifteen times, once per task type, by slotting in the task-specific Core Principle listed individually in Figures 16–30.

Stage 3c: Multiple-Choice Question Generator (Generic Scaffold)

Task: Generate one multiple-choice <task_name> question.
[Question Design]
<task_specific_question_design> *(per-task body in Figures 16-30)*
[Question Quality Priority]
When generating the question, enforce ALL of the following:

- CHALLENGING: genuinely difficult for an attentive viewer; not obvious after a single viewing.
- NO Common Sense / NO Well-Known Facts: a person who has never seen the video must NOT be able to guess the answer.
- STRICT Multi-hop: removing even ONE evidence step must make the question unanswerable; every hop must be load-bearing.
- Strong grounding: every option must be grounded in trajectory evidence; no invented facts.
- Balanced options: all four options must be plausible on the surface.

[Modality Labeling Rule]
For each evidence step, assign <label> based on what is ACTUALLY USED to solve that hop: audio (only auditory cues), video (only visual cues), or audio-video (both modalities jointly required).
[Distractor Design Rules]
Distractors MUST satisfy ALL of: (1) factually grounded in entities, events, numbers, or phrases that actually appear in the video; (2) logically and commonsensically valid; (3) superficially plausible to someone who partially watched the video; (4) length parity with the correct option; (5) no giveaway language such as "none of the above"; (6) no surface-level tells in tone or grammar.
[Option Selection Constraint]
Receives a per-task constraint <option_selection_constraint>: if single_choice_only, set <question_type>="single" with one item in <correct_options>; if single_or_multiple, either single-choice or multiple-choice is allowed (2 to 4 items in <correct_options> for the latter). Always keep exactly four options A/B/C/D.
[Timestamp Assignment]
For each node in <final_trajectory>, assign <timestamp_minute> by selecting the single minute within the event's [<start_minute>, <end_minute>] range that best matches the evidence text, using the per-minute <raw_segments> for grounding.
[task_specific_key]
<task_specific_key_pair> *(per-task value in Figures 16-30)*
[Output Format]
A valid JSON object with keys <final_trajectory> (per-hop <event_id>, <evidence>, <label>, <reason>, <timestamp_minute>), <task_specific_key>, <question_type>, <question>, <options> (A/B/C/D), <correct_options>, and <answer_text>.

Figure 15: Generic scaffold for the Stage 3c question generator. The Question Design block and <task_specific_key> pair are slotted in per task type (Figures 16–30).

Task: Information Retrieval (IR)

```
[Stage 3b Core Principle]
Identify events where a specific key fact (on-screen text, number,
spoken name, person identity) is embedded and can only be fully
retrieved by chaining evidence from multiple events.  Vary the
answer position across the trajectory (beginning, middle, or end).
Prefer events whose entities carry the attribute "audio-visual"
so that both modalities contribute to retrieval.  Include both
short-range and long-range chains.
[Stage 3c Question Design]
Identify a specific key fact (on-screen text, number, spoken name,
or person identity) as the retrieval target.  The question must
require chaining at least two evidence steps; the target fact must
NOT be directly answerable from a single event.  Vary where the
answer sits in the chain.  Distractors are plausible in-domain
values (similar numbers, similar names) drawn from elsewhere in
the trajectory.
[task_specific_key]
{<key_name>:  retrieval_target_type, <key_value>:
text | number | speech | person}
```

Figure 16: Per-task instantiation for **Information Retrieval (IR)**.

Task: Temporal Sequencing (TS)

```
[Stage 3b Core Principle]
Identify events that form a non-trivial temporal order or
entity-state evolution that cannot be inferred from any single
event.  Prefer events where an entity's appearance, role, or
state visibly or audibly changes between events.  Include both
short-range and long-range chains.  Avoid trajectories where the
ordering is trivially recoverable from event IDs alone.
[Stage 3c Question Design]
Ask the model to order events, identify what happened before/after
a given event, or describe how an entity's state changed across the
trajectory.  The correct ordering must require evidence from all
events.  Distractors should be plausible alternative orderings or
state sequences.
[task_specific_key]
{<key_name>:  sequence_length, <key_value>:  <number_of_ordered_items>}
```

Figure 17: Per-task instantiation for **Temporal Sequencing (TS)**.

Task: Entity Tracking (ET)

```
[Stage 3b Core Principle]
Centre each trajectory on a specific entity (person, object, or
sound source) that appears in at least two non-adjacent events with
meaningful state, role, or relationship changes. Leverage entities
whose attribute is "audio-visual" when both modalities are needed
to track the entity. Include both nearby transitions and far-apart
isolated transitions. Ensure the question cannot be answered from
a single event mention.
[Stage 3c Question Design]
Centre the question on a specific entity from the entity pool. Ask
about the entity's identity, role, relationship, or state at a
specific point in the trajectory, requiring cross-event linking.
Distractors should be other entities or plausible but wrong state
descriptions drawn from the entity pool or trajectory context.
[task_specific_key]
{<key_name>: tracked_entity, <key_value>: <entity_name>}
```

Figure 18: Per-task instantiation for **Entity Tracking (ET)**.

Task: Forward Causal Reasoning (FCR)

```
[Stage 3b Core Principle]
Identify event pairs or chains where an earlier event is a clear
prerequisite or cause for something that happens in a later event.
The causal link must require evidence from both events; neither
alone is sufficient to answer "what happens as a result of X".
Include both local causal chains and distant isolated causal
chains.
[Stage 3c Question Design]
Frame the question as: "Given that [cause from early event], what
happens / what is the outcome in [later event]?". The causal
link must be non-trivial and require evidence from both events.
Distractors should be plausible but causally unrelated or reversed
outcomes.
[task_specific_key]
{<key_name>: causal_direction, <key_value>: forward}
```

Figure 19: Per-task instantiation for **Forward Causal Reasoning (FCR)**.

Task: Backward Causal Reasoning (BCR)

```
[Stage 3b Core Principle]
Identify event pairs or chains where a later event shows an effect
or outcome that can only be explained by tracing back to an earlier
cause event.  The backward causal link must require evidence from
both events; neither alone is sufficient to answer "what caused
X". Include both local causal chains and distant isolated causal
chains.
[Stage 3c Question Design]
Frame the question as:  "Given that [effect / outcome in later
event], what was the cause / prerequisite in [earlier event]?".
The backward causal link must be non-trivial and require evidence
from both events.  Distractors should be plausible but causally
unrelated or forward-direction answers.
[task_specific_key]
{<key_name>:  causal_direction, <key_value>:  backward}
```

Figure 20: Per-task instantiation for **Backward Causal Reasoning (BCR)**.

Task: Cross-Modality Matching (CMM)

```
[Stage 3b Core Principle]
Identify events where a specific audio cue (speech phrase, sound,
music) and a specific visual element co-occur or are explicitly
linked across events.  Prefer events containing entities with
attribute "audio-visual".  Also identify events where an audio
or visual element appears WITHOUT its expected counterpart, since
these support plausible-but-wrong distractors.  Include both
short-range and long-range chains.
[Stage 3c Question Design]
Ask the model to correctly match an audio cue to its visual context
(or vice versa) by chaining evidence across events.  One option is
the correct match; the other three are plausible mismatches (audio
from a different event, visually similar but wrong scene).  All
evidence steps for this task type are labeled audio-video.
[task_specific_key]
{<key_name>:  match_direction, <key_value>:  audio->visual|visual->audio}
```

Figure 21: Per-task instantiation for **Cross-Modality Matching (CMM)**.

Task: Spatiotemporal Localization (SL)

```
[Stage 3b Core Principle]
Identify trajectories where pinpointing the EXACT MINUTE of a
target event or entity state requires chaining clues from at least
two events.  Each event is provided with minute-annotated captions
<minute_captions> keyed by absolute minute.  Prefer multi-minute
events so that fine-grained minute-level disambiguation is possible.
The question's correct answer and all distractors MUST be concrete
minute integers drawn from the trajectory's time range.  Include
both short-range and long-range chains.
[Stage 3c Question Design]
Ask the model to identify the EXACT MINUTE at which a specific
event or entity state occurs, requiring evidence from at least
two events to narrow it down.  The question stem must describe the
target event clearly enough that only one minute is correct.  ALL
FOUR options (A/B/C/D) MUST be concrete integer minute values drawn
from the trajectory's time range, NOT descriptive phrases.  The
correct option is the minute confirmed by the trajectory evidence
chain.  Distractors should be other plausible minutes from the same
trajectory.
[task_specific_key]
{<key_name>:  target_minute, <key_value>:  <correct_minute_integer>}
```

Figure 22: Per-task instantiation for **Spatiotemporal Localization (SL)**.

Task: Spatial Reasoning (SR)

```
[Stage 3b Core Principle]
Identify events where spatial relationships (position, direction,
relative layout) of objects or persons can only be determined by
combining visual evidence from multiple events.  Prefer events
with rich visual entity descriptions.  Audio evidence should NOT
be required to answer the question.  Include both short-range and
long-range chains.
[Stage 3c Question Design]
Ask about the spatial relationship (position, direction, relative
layout) of objects or persons that can only be determined by
combining visual evidence from multiple events.  Audio evidence
must NOT be needed.  Distractors should be plausible but spatially
incorrect alternatives.  All evidence steps for this task type are
labeled video.
[task_specific_key]
{<key_name>:  spatial_relation_type,
<key_value>:  position | direction | layout | relative}
```

Figure 23: Per-task instantiation for **Spatial Reasoning (SR)**.

Task: Visual Counting (VC)

```
[Stage 3b Core Principle]
Identify trajectories where counting actions, state changes, or
objects requires a condition established in one event and the
counting target distributed across other events.  Example pattern:
"How many times does X do Y after Z happens?".  Visual evidence
only; audio should not be required.  Include both short-range and
long-range chains.
[Stage 3c Question Design]
Ask how many times an action occurs, how many objects exist, or
how many state changes happen, under a condition established in
one event and counted across other events.  Produce four distinct
integer-count options.  Distractors should be nearby integers
(±1, ±2) or counts derived from wrong conditions.  Audio evidence
must NOT be needed.  All evidence steps for this task type are
labeled video.
[task_specific_key]
{<key_name>:  count_target, <key_value>:  <what_is_being_counted>}
```

Figure 24: Per-task instantiation for **Visual Counting (VC)**.

Task: Speech Context (SC)

```
[Stage 3b Core Principle]
Identify events where spoken dialogue or narration across multiple
events must be combined to answer a question about what was said.
Prefer events containing entities with attribute "audio-visual"
where the speech component is the primary evidence.  Visual
information should NOT be required.  Include both short-range and
long-range chains.
[Stage 3c Question Design]
Ask about what was said (spoken dialogue or narration) across
multiple events, requiring the model to chain speech evidence
from at least two events.  Visual information must NOT be needed.
Distractors should be plausible but incorrect speech content from
the same trajectory or nearby events.  All evidence steps for this
task type are labeled audio.
[task_specific_key]
{<key_name>:  speech_type, <key_value>:  dialogue | narration | announcement | other}
```

Figure 25: Per-task instantiation for **Speech Context (SC)**.

Task: Environmental Sound (ES)

```
[Stage 3b Core Principle]
Identify events where non-human environmental sounds (rain,
machinery, crowd, nature sounds) across multiple events must be
combined to answer a question.  The question must be answerable
from audio alone without visual evidence.  Include both short-range
and long-range chains.
[Stage 3c Question Design]
Ask about non-human environmental sounds (rain, machinery, crowd,
nature) that must be identified or reasoned about across multiple
events.  Visual information must NOT be needed.  Distractors
should be other plausible environmental sounds from the video.  All
evidence steps for this task type are labeled audio.
[task_specific_key]
{<key_name>:  sound_category, <key_value>:  <type_of_environmental_sound>}
```

Figure 26: Per-task instantiation for **Environmental Sound (ES)**.

Task: Background Music (BM)

```
[Stage 3b Core Principle]
Identify events where background music, ambient sound, or
non-foreground audio cues across multiple events must be combined
to answer a question.  The question must be answerable from audio
alone without visual evidence.  Include both short-range and
long-range chains.
[Stage 3c Question Design]
Ask about background music, ambient sound, or non-foreground audio
cues that must be reasoned about across multiple events.  Visual
information must NOT be needed.  Distractors should be other
plausible background audio descriptions.  All evidence steps for
this task type are labeled audio.
[task_specific_key]
{<key_name>:  background_audio_type, <key_value>:
music | ambient | other}
```

Figure 27: Per-task instantiation for **Background Music (BM)**.

Task: Visual-to-Audio Deception (V2A)

```
[Stage 3b Core Principle]
Identify events with rich visual content (entities with
attribute "visual" or "audio-visual") that make a candidate
audio detail plausible within a specified event or time range.
Construct a balanced set of premise_status=supported and
premise_status=fabricated items.  For supported items, the
candidate audio detail must genuinely occur in the specified range
and be verifiable from the soundtrack.  For fabricated items,
the visual scene should plausibly suggest the candidate audio
detail, but the detail must be contradicted by audio evidence
in the specified range.  Reject cases whose truth cannot be
determined because of inaudibility, off-screen activity, occlusion,
or insufficient coverage.  In both cases, the trajectory should
require source-grounded verification rather than local lexical
matching.  Include both short-range and long-range chains.
[Stage 3c Question Design]
Present real visual evidence from the trajectory and ask which
statement or statements about the corresponding audio are best
supported within the specified event or time range.  Construct
a balanced option set containing plausible acceptance and
rejection alternatives.  Both types of option must be concrete,
comparably detailed, and grounded in distinct candidate evidence
relations.  Follow the benchmark's single-/multi-choice allocation:
single-choice items have exactly one correct option, while
multi-choice items have a non-empty correct answer set.  Balance
the number and polarity composition of correct options within this
sub-task.  For supported items, the correct answer set affirms
the observed audio detail; for fabricated items, it rejects the
proposed detail by citing the contradictory audio evidence.  The
remaining options must include plausible but incorrect acceptance
and rejection rationales.  The stem must not reveal the premise
status or that the item is a hallucination test; phrase it as
a genuine audio inquiry.  Across the released set, ensure that
neither option polarity, negation words, option length, nor
explanation specificity predicts the correct answer set.  Label
the visual cue that motivates the candidate premise as video, and
label the evidence that verifies or refutes it according to its
actual modality (typically audio or audio-visual).
[task_specific_key]
{<key_name>:  halluc_type, <key_value>:  visual->audio},
{<key_name>:  premise_status, <key_value>:  supported|fabricated}
```

Figure 28: Per-task instantiation for **Visual-to-Audio Deception (V2A)**.

Task: Audio-to-Visual Deception (A2V)

```
[Stage 3b Core Principle]
Identify events with rich audio content (entities with attribute
"audio-visual") that make a candidate visual detail plausible
within a specified event or time range.  Construct a balanced set
of premise_status=supported and premise_status=fabricated items.
For supported items, the candidate visual detail must genuinely
appear in the specified range and be verifiable from the visual
stream.  For fabricated items, the audio should plausibly suggest
the candidate visual detail, but the detail must be contradicted
by visual evidence in the specified range.  Reject cases whose
truth cannot be determined because of inaudibility, off-screen
activity, occlusion, or insufficient coverage.  In both cases, the
trajectory should require source-grounded verification rather than
local lexical matching.  Include both short-range and long-range
chains.
[Stage 3c Question Design]
Present real audio evidence from the trajectory and ask which
statement or statements about the corresponding visual content
are best supported within the specified event or time range.
Construct a balanced option set containing plausible acceptance
and rejection alternatives.  Both types of option must be concrete,
comparably detailed, and grounded in distinct candidate evidence
relations.  Follow the benchmark's single-/multi-choice allocation:
single-choice items have exactly one correct option, while
multi-choice items have a non-empty correct answer set.  Balance
the number and polarity composition of correct options within this
sub-task.  For supported items, the correct answer set affirms
the observed visual detail; for fabricated items, it rejects the
proposed detail by citing the contradictory visual evidence.  The
remaining options must include plausible but incorrect acceptance
and rejection rationales.  The stem must not reveal the premise
status or that the item is a hallucination test; phrase it as
a genuine visual inquiry.  Across the released set, ensure that
neither option polarity, negation words, option length, nor
explanation specificity predicts the correct answer set.  Label the
audio cue that motivates the candidate premise as audio, and label
the evidence that verifies or refutes it according to its actual
modality (typically video or audio-visual).
[task_specific_key]
{<key_name>:  halluc_type, <key_value>:  audio->visual},
{<key_name>:  premise_status, <key_value>:  supported|fabricated}
```

Figure 29: Per-task instantiation for **Audio-to-Visual Deception (A2V)**.

Task: Temporal Splicing Fallacy (TSF)

```
[Stage 3b Core Principle]
Identify 2 to 4 events from temporally separated points in the
video and construct a candidate narrative about their ordering or
relation.  Construct a balanced set of premise_status=supported
and premise_status=fabricated items.  For supported items, the
narrative must faithfully describe a coherent sequence established
by the actual timeline.  For fabricated items, the narrative must
splice real fragments from non-adjacent time points into a sequence
explicitly contradicted by the observed temporal order.  Do not
treat missing evidence as evidence of impossibility.  In both cases,
the trajectory should require verification of the global temporal
layout rather than isolated recognition of individual fragments.
Prefer events with substantial temporal gaps.
[Stage 3c Question Design]
Present a candidate narrative grounded in real event fragments and
ask which statement or statements most accurately characterize its
relation to the video timeline.  Construct a balanced option set
containing plausible acceptance and rejection alternatives.  Both
types of option must be concrete, comparably detailed, and grounded
in distinct candidate temporal relations.  Follow the benchmark's
single-/multi-choice allocation:  single-choice items have exactly
one correct option, while multi-choice items have a non-empty
correct answer set.  Balance the number and polarity composition
of correct options within this sub-task.  For supported items, the
correct answer set affirms the narrative and identifies evidence
establishing its temporal coherence.  For fabricated items, it
rejects the narrative by identifying the observed ordering that
contradicts it.  The remaining options must include plausible
but incorrect acceptance and rejection rationales.  Use actual
timestamps from the trajectory to ground the decision, but do not
reveal the premise status in the stem.  Across the released set,
ensure that neither option polarity, negation words, option length,
nor explanation specificity predicts the correct answer set.
[task_specific_key]
{<key_name>:  halluc_type, <key_value>:  splice},
{<key_name>:  premise_status, <key_value>:  supported|fabricated}
```

Figure 30: Per-task instantiation for **Temporal Splicing Fallacy (TSF)**.

MH-specific balance and shortcut checks. For each MH sub-task separately, the final release must balance `supported` and `fabricated` items, the single-/multi-choice allocation, the number of correct options, and the polarity composition of correct answer sets after all quality filtering. Candidate items are rejected if option polarity, negation markers, option length, or explanation specificity predicts the gold answer set above the corresponding random baseline. We additionally construct a small, matched set of supported/fabricated premise pairs from comparable source evidence to audit whether the intended decision changes with the video-grounded premise status.

L.6 Quality Assurance Prompts

After generation, every candidate item passes through two LLM-based quality checks: a *text-only solver* that flags items whose answer can be guessed without the video, and a *verifier* that audits multi-hop integrity, distractor quality, and answer leakage.

QA-1: Text-only Direct Solver (System Prompt)

```
You are a strict multiple-choice solver. You only see question
text and options. You do NOT have access to the video. You are
NOT told whether this question is single-choice or multi-choice;
infer from wording only. If confidence is insufficient, abstain
instead of guessing. Return valid JSON only.
```

Figure 31: System prompt for the text-only solver, used to detect items whose answer can be guessed from textual cues alone.

QA-1: Text-only Direct Solver (User Prompt)

```
Solve the following benchmark question using text-only clues.
Question: <question>
Options: A. <option_A> B. <option_B> C. <option_C> D.
<option_D>
Return a JSON object with <predicted_options> (a list of unique
letters from A/B/C/D, possibly empty when abstaining), <confidence>
(a number in [0,1]), and <reason> (a brief justification).
Rules: do NOT assume single-choice by default; do NOT force a
guess when uncertain; abstain by returning an empty list.
```

Figure 32: User prompt paired with the text-only solver system prompt in Figure 31.

QA-2: Item Verifier (System Prompt)

```
You are a strict QA benchmark verifier. Given the question,
options, reference answer, and trajectory evidence, return JSON
only. Evaluate quality and potential flaws conservatively.
```

Figure 33: System prompt for the item-level verifier that audits multi-hop integrity, distractor quality, and answer leakage.

L.7 Evaluation Prompt

The same minimal prompt is issued to every candidate model on every TraceAV-Bench item, paired with the original video as input. We deliberately keep the wording short and free of chain-of-thought scaffolding The model is asked to return only the letter (or comma-separated letters for multi-choice)

QA-2: Item Verifier (User Prompt)

Verify this benchmark item.
Task type: <task_type> Question type: <question_type>
Question: <question>
Options: A/B/C/D = <options>
Reference correct_options: <correct_options> Reference answer_text: <answer_text>
Trajectory evidence: <trajectory_with_timestamps>
Please evaluate from the following angles:

1. Is the multi-hop chain real and correct?
2. Is this pseudo multi-hop (answer solvable by one hop or one local clue)?
3. Option quality: are there obvious superficial signals (e.g., one option much longer or stylistically different); are wrong options relevant rather than random noise; are wrong options plausibly confusing and semantically close?
4. Answer leakage: can the answer be directly inferred from wording in the question stem itself?

Return a JSON object with <is_real_multihop>, <multihop_issue>, <is_pseudo_multihop>, <pseudo_multihop_reason>, <option_quality> (with sub-fields <has_obvious_pattern>, <distractors_relevant>, <distractors_plausible>, <overall_quality> ∈ {very_poor, poor, fair, good, excellent}, <reason>), <answer_leakage_detected>, <answer_leakage_reason>, <should_drop>, <drop_reasons>, <severity> ∈ {none, low, medium, high}, and <summary>. The verifier sets <should_drop>=true for major quality flaws.

Figure 34: User prompt paired with the verifier system prompt in Figure 33; the structured JSON return is consumed by the dropping rule.

Evaluation: Model Inference

You are given a video. Base your answer on what you see and hear.
Directly provide the letter representing your choice (A/B/C/D) and nothing else. Do not include the full text of the option; do not provide any explanation. The problem could be a single-choice question or a multiple-choice question. If multiple options are correct, return letters joined by commas (example: A,C).
Question:
<question>
Options:
<options_text>

Figure 35: Unified evaluation prompt issued together with the source video to every candidate model on each TraceAV-Bench item.

of its selected option, which makes the response trivially parseable and avoids any prompt-side bias toward verbose explanations.