MS-OMNIBENCH: WHEN LISTENING ACROSS VIDEO STREAMS IS NECESSARY

Anonymous authors
Paper under double-blind review



Figure 1: **Audio-dependent reasoning across complementary video streams.** In this illustrative example, *Stream A* links the queried utterance to the person in green but does not reveal the pointed-to object. *Stream B* shows two pointing hands and the candidate phones but does not reveal which hand belongs to the target speaker. Combining *both streams* identifies the orange phone.

ABSTRACT

A camera may record the sound of an impact while its cause is visible only from another viewpoint. However, existing benchmarks for Omni-modal Large Language Models (Omni-LLMs) offer limited evaluation of reasoning that requires both sound and complementary information across streams. To address this gap, we introduce **MS-OmniBench**, a benchmark designed around questions that jointly require acoustic evidence and multiple streams. MS-OmniBench comprises 3,109 human-reviewed or human-authored questions across diverse real-world settings, organized into twelve subtasks spanning four families: *synchronous cross-stream understanding*, *asynchronous cross-segment reasoning*, *audio-anchored retrieval*, and *cross-modal consistency verification*. Answer-relevant evidence is distributed across camera viewpoints, source-specific audio tracks, and non-adjacent temporal segments, with annotated evidence chains identifying the supporting streams and time ranges. We evaluate seventeen open-weight and proprietary Omni-LLMs and four existing agentic systems. The best proprietary and open-weight end-to-end models achieve 61.9% and 52.0% accuracy, respectively, while the strongest existing agentic system reaches 49.8%. We further introduce **MESA**, a simple *M*ulti-stream *E*vidence *S*earch *A*gent that organizes observations into per-stream event timelines and uses them to guide targeted media inspection and cross-stream answer verification. With Gemini 3.8 Flash, MESA reaches 65.6% accuracy, improving by 3.7 percentage points over the same backbone in the end-to-end setting. MS-OmniBench provides a testbed for studying how models use sound to connect complementary observations across streams. The benchmark will be released at MS-OmniBench.

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.

1 INTRODUCTION

Understanding real-world events often requires linking acoustic and visual evidence (Arandjelovic & Zisserman, 2017; Zhao et al., 2018). Different camera views and audio channels can provide complementary observations (Sun et al., 2026), while acoustic cues can help identify which visual evidence is relevant (Tian et al., 2018). As shown in Figure 1, identifying the object under discussion requires linking a spoken utterance to a visible participant and then following that participant's gesture in another view. Neither recognizing the words nor detecting the objects independently resolves the reference. Answering such questions requires integrating acoustic evidence with complementary observations distributed across multiple streams.

However, existing benchmarks for Omni-modal Large Language Models (Omni-LLMs) provide limited assessment of this joint requirement. Recent evaluations span general video understanding (Li et al., 2024; Fu et al., 2025), long-context retrieval and reasoning (Wu et al., 2024; Wang et al., 2025), and audio-centric or modality-dependent reasoning (Sakshi et al., 2025; Radevski et al., 2026; Li et al., 2026). Complementary benchmarks examine cross-view and multi-stream integration (He et al., 2026; Shah et al., 2026), while X-Stream targets online multi-stream understanding, with around 30% of its questions incorporating audio or speech information (Sun et al., 2026). Therefore, existing benchmarks still lack a systematic evaluation of questions for which acoustic evidence and multiple streams are jointly necessary.

To address this gap, we introduce **MS-OmniBench**, a benchmark for jointly audio-dependent and multi-stream reasoning. The benchmark contains 3,109 human-reviewed or human-authored questions from six public datasets (Grauman et al., 2022; 2024; Carletta et al., 2005; Li et al., 2018; Donley et al., 2021; Shimada et al., 2023), organized into twelve subtasks across four families covering *synchronous cross-stream understanding*, *asynchronous cross-segment reasoning*, *audio-anchored retrieval*, and *cross-modal consistency verification*. Three question formats support text-only, reference-audio, and reference-image queries, while answer-relevant evidence is distributed across camera views, source-specific audio tracks, and non-adjacent temporal segments. Acoustic evidence extends beyond speech to speaker cues, non-speech events, music and instrument-specific signals, and ambient and spatial acoustic cues. Construction begins with aligned cross-stream events, with evidence chains identifying the supporting streams and time ranges. Model-assisted generation is followed by shortcut screening and human verification, while direct human authoring further broadens the coverage of the benchmark.

We evaluate seventeen proprietary (Comanici et al., 2025; Team, 2026a; Seed, 2026) and open-weight Omni-LLMs (Team, 2026b; Deshmukh et al., 2026; Li et al., 2025) and four existing agentic systems (Tao et al., 2025; Zhang et al., 2026; Zhou et al., 2025; Kurpath et al., 2025). The best proprietary and open-weight end-to-end models achieve 61.9% and 52.0% accuracy, respectively, while the strongest existing agentic system reaches 49.8%. Fine-grained analysis reveals substantial variation across supplied stream counts, evidence-chain lengths, and question formats. Accuracy does not decrease monotonically with either stream count or evidence-chain length, revealing distinct patterns of distributed evidence use across models and tasks. We further introduce **MESA**, a simple Multi-stream Evidence Search Agent that uses model-generated event timelines to guide targeted media inspection and cross-stream answer verification without access to benchmark evidence annotations. With Gemini 3.8 Flash, MESA reaches 65.6% accuracy, improving by 3.7 percentage points over its end-to-end counterpart. Overall, the contributions of this work are threefold:

- We introduce **MS-OmniBench** for jointly audio-dependent and multi-stream reasoning, covering complementary visual, acoustic, and temporal observations across twelve subtasks.
- We develop an evidence-grounded construction pipeline with shortcut screening and human verification, while providing explicit evidence chains for the released questions.
- We evaluate diverse Omni-LLMs and agents, analyze performance across input structures and question formats, and provide **MESA** as a simple multi-stream evidence-search baseline.

2 RELATED WORK

2.1 OMNI-MODAL BENCHMARKS

Recent benchmarks span temporal video understanding (Li et al., 2024; Fu et al., 2025), audio understanding (Sakshi et al., 2025), and joint audio-visual reasoning (Zhong et al., 2026; Hong et al.,

2026; Zhou et al., 2025). A central concern in audio-visual evaluation is whether performance genuinely depends on both modalities rather than textual priors or unimodal shortcuts. This has motivated benchmarks that explicitly test modality dependence and shortcut resistance (Radevski et al., 2026; Chao et al., 2026; Yang et al., 2025). OmniVideoBench further filters questions that can be solved from a single modality and provides atomic multimodal reasoning traces (Li et al., 2026). Other benchmarks provide fine-grained or task-specific diagnostics of audio-visual reasoning (Galougah et al., 2025; Wang et al., 2026; Lu et al., 2026). Robustness-oriented evaluations additionally examine hallucination, input perturbations, and cross-modal inconsistency (Sung-Bin et al., 2025; Chowdhury et al., 2025; Nazi et al., 2026; Chen et al., 2026b; van Sprang et al., 2026). A separate line of work studies reasoning when relevant evidence is distributed across extended temporal contexts or multiple recordings. Long-video benchmarks evaluate retrieval and reasoning over temporally dispersed evidence (Mangalam et al., 2023; Wu et al., 2024; Wang et al., 2025; Zhou et al., 2024; Chen et al., 2025; Tao et al., 2026; Goel et al., 2026; Kurpath et al., 2025; Feng et al., 2026). Cross-view benchmarks examine associations across cameras (He et al., 2026; Shah et al., 2026), while X-Stream extends this setting to online multi-stream understanding, with around 30% of its questions incorporating audio or speech information (Sun et al., 2026). Despite these advances, the joint requirement of acoustic evidence and multiple streams remains underexplored. **MS-OmniBench** addresses this gap by systematically evaluating questions that require both.

2.2 Omni-modal Models and Agents

Joint audio-visual processing is increasingly supported by proprietary systems (Hurst et al., 2024; Comanici et al., 2025; Team, 2026a) and open-weight omni-modal models (Jin Xu, 2025; Xu et al., 2025; Li et al., 2025; AI et al., 2025). Recent work has advanced audio-visual training (Tang et al., 2025), real-time interaction (Fu et al., 2026; Cui et al., 2026), and efficient inference (Deshmukh et al., 2026). For long-form reasoning, relevant evidence can be temporally dispersed, motivating selective retrieval and local inspection. VideoAgent and Deep Video Discovery iteratively search for question-relevant evidence in long videos (Wang et al., 2024; Zhang et al., 2026). Related methods use document retrieval and query-guided frame selection to reduce irrelevant visual context (Ma et al., 2025; Liu et al., 2025). Omni-modal agents further incorporate audio-guided active perception and iterative multimodal evidence search (Tao et al., 2025; Xu et al., 2026; Kurpath et al., 2025). OmniReasoner learns when and where to inspect local audio-visual content through tool use (Chen et al., 2026a). Building on these evidence-search approaches, **MESA** organizes observations into model-generated per-stream event timelines and uses them to guide targeted media inspection and cross-stream verification.

3 MS-OmniBench

3.1 Task Formulation and Question Formats

Each item is represented as $x = (\mathcal{S}, r, q, \mathcal{O}, y)$, where $\mathcal{S} = \{s_1, \ldots, s_k\}$ contains $k \geq 2$ scene streams, r is optional reference media, q is the question, $\mathcal{O}$ contains five answer options, and y identifies the unique correct option. Here $\mathcal{S}$ denotes the streams supplied with the question, not necessarily every stream available in its source case. A stream comprises video and its audio when available. Streams may differ by camera viewpoint, audio channel, or temporal segment, while preserving their original temporal relationships. Figure 2 illustrates three input question formats. **Text-only questions** specify the target through language. **Text with reference audio** supplies a short sound excerpt that must be associated with an event or source in the scene. **Text with reference image** supplies a visual exemplar whose corresponding entity must be located and connected to the relevant acoustic evidence.

3.2 Benchmark Tasks

We organize the benchmark into four task families and twelve subtasks, covering synchronous cross-stream association, asynchronous reasoning across segments, audio-anchored retrieval, and cross-modal consistency verification.

Synchronous cross-stream understanding tasks integrate time-aligned observations to resolve source, identity, and reference ambiguities across streams.



Figure 2: **Three input question formats in MS-OmniBench.** All formats are paired with multi-stream scene inputs, while reference audio or reference images provide additional cues for identifying the target evidence.

- *Cross-View Sound-Source Disambiguation (**SrcDis**).* Identify the actual sound source among visually similar candidates. Timbre, onset, and rhythm must be associated with visual observations across streams rather than inferred from appearance alone.
- *Off-Screen Sound Causal Reasoning (**OffCau**).* Determine the action or interaction responsible for a sound that is audible but not visible in one stream by locating its visual cause in another.
- *Multi-Speaker Identity Binding (**SpkBind**).* Associate a voice or conversational turn with a visible participant. The task requires linking vocal characteristics and turn timing to visual identity, which an utterance transcript alone does not establish.
- *Speech-to-Visual Grounding (**SpVis**).* Ground spoken references, requests, or comments in visual entities and actions. Beyond speaker identification, the answer concerns the referenced attribute, action, or spatial relation.

Asynchronous cross-segment reasoning tasks connect evidence across temporally separated or non-adjacent segments while preserving the relevant acoustic, visual, and cross-modal associations.

- *Cross-Segment Audio Memory (**MemA**).* Retain an earlier sound, voice, or musical theme and connect it to a subsequent event. Spoken-instruction following is included when an earlier instruction must be associated with its later execution.
- *Non-Speech Event Ordering (**EvOrd**).* Recover the sequence of impacts, door movements, equipment sounds, or other non-speech events. Acoustic timing and visual event identity must remain correctly associated across streams or segments.
- *Multi-Hop Cross-Segment Chains (**MHop**).* Resolve dependent inferences across modalities, views, and temporal segments. Each step identifies an intermediate entity or event required by a later step, making isolated clues insufficient for the complete answer.

Audio-anchored retrieval tasks use acoustic cues to specify what must be located, counted, or inferred from the scene observations.

- *Audio-Triggered Localization (**TrigLoc**).* Locate the visual moment, position, or participant associated with a specified sound. The model must distinguish its corresponding observation from plausible alternatives across streams.
- *Cross-Modal Counting (**XCount**).* Count events using complementary auditory and visual evidence. Audible but occluded and visible but silent events provide complementary observations. Correct counting must avoid omissions and duplicate counts across streams.
- *Ambient-Sound Scene Inference (**AmbSc**).* Infer spatial properties or scene transitions from reverberation, background sound sources, and their changes, using visual observations to corroborate the acoustic interpretation.

Cross-modal consistency verification tasks compare acoustic and visual evidence to identify agreement or conflict in natural recordings and controlled variants.

- *Natural Audio-Visual Inconsistency (**AVInc**).* Identify conflicts between acoustic and visual observations in unmodified recordings, including event-, text-, and affect-level discrepancies.
- *Controlled Manipulation & Shortcut Resistance (**Manip**).* Evaluate audio-visual correspondence under controlled muting, track swaps, contradictory subtitles, temporal shifts, and music replace-

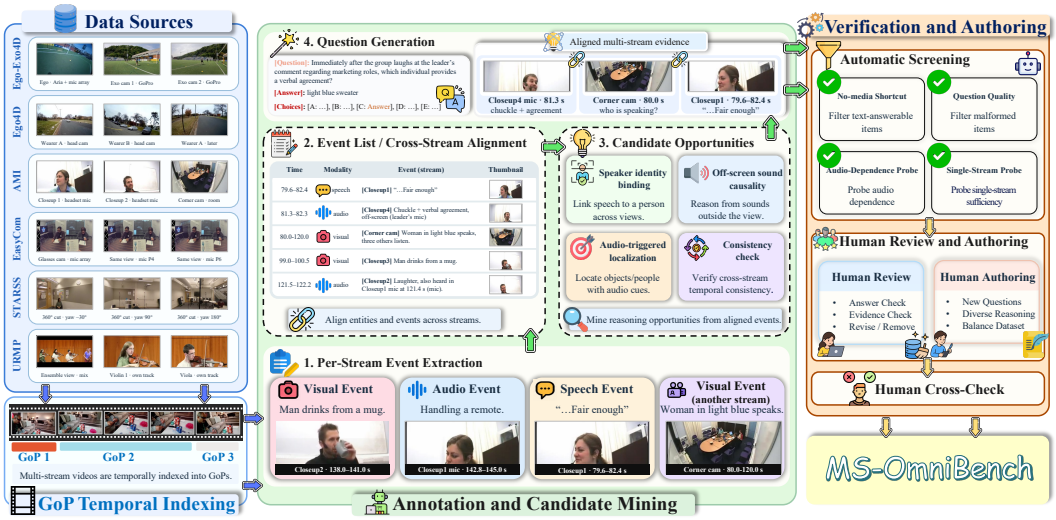


Figure 3: **Construction pipeline of MS-OmniBench.**

ment, together with naturally occurring subtitle conflicts. Decisions must be grounded in the observed evidence rather than textual shortcuts or expected audio-visual associations.

3.3 Data Sources and Construction

Figure 3 summarizes the construction pipeline. Gemini 3 Flash supports model-assisted event annotation, cross-stream alignment, question planning and writing, and shortcut screening.

Data collection. We select recordings from six public datasets to cover complementary visual, acoustic, and temporal observations. Ego-Exo4D supplies first- and third-person views of skilled activities (Grauman et al., 2024), while Ego4D provides egocentric activity recordings (Grauman et al., 2022). AMI and EasyCom contribute conversations with camera and microphone observations (Carletta et al., 2005; Donley et al., 2021). STARSS supplies spatial sound recordings with panoramic video (Shimada et al., 2023), and URMP provides aligned musical performances with individual instrument tracks (Li et al., 2018). We group selected media into cases using camera views, source-specific tracks, or non-adjacent segments. View crops and reused video paired with different audio are distinguished from separately captured camera recordings.

Temporal indexing and event annotation. Streams are re-encoded with scene-cut-aware key frames and indexed using group-of-pictures (GoP) intervals (Wiegand et al., 2003). GoP intervals, transcript times, and sound onsets provide temporal anchors for event annotation rather than predefined semantic boundaries (Wu et al., 2026; Shen et al., 2026). A per-stream model call extracts visual events, speech with speaker slots, non-speech sounds, and music, retaining stream identifiers and time intervals. It also links sounds to visible sources or marks them as off-screen.

Cross-stream alignment and candidate mining. A deterministic miner identifies recurring sounds or actions and off-screen sounds. A scene-level model call aligns the per-stream records, associates entities across views, and indexes candidate reasoning opportunities by task. This separates the discovery of useful evidence relations from the wording of a question.

Evidence-grounded question generation. Planning specifies the task, at least two supporting streams, event identifiers, distractors, and a multi-hop evidence chain. A writing call generates the question, five options, and answer. For model-assisted natural questions, distractors come from the same confusable event group, while stems avoid camera names, stream identifiers, timestamps, and verbatim speech quotations. Reference media are extracted from the scene. Annotators additionally construct controlled variants through muting, track swaps, subtitle injection, temporal shifts, and music replacement. The resulting questions and answers are verified against the modified media.

Automatic screening. Structural and question-quality checks identify malformed questions, prohibited wording, missing acoustic evidence, and invalid cross-stream evidence chains. A no-media shortcut test presents the question and options twice in different orders without scene media. Can-

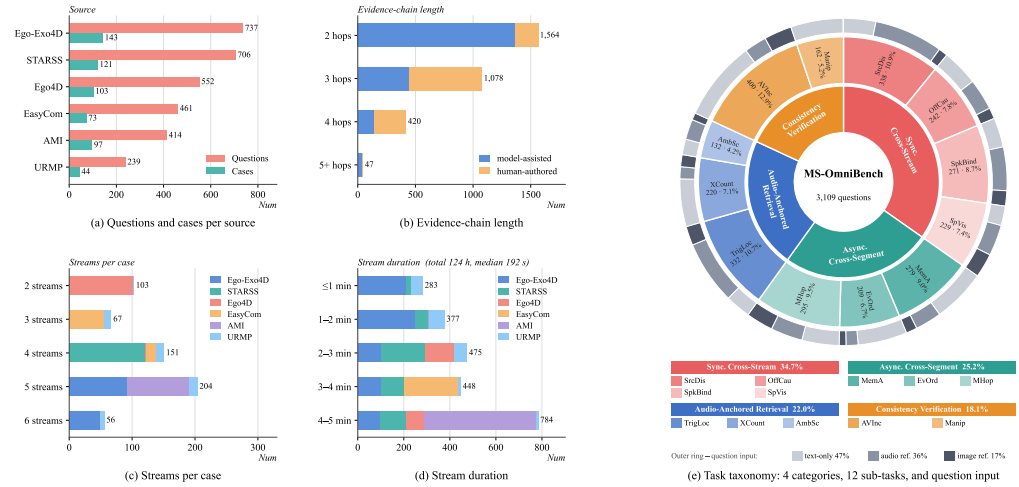


Figure 4: **Dataset statistics.** (a) Questions and cases by source. (b) Evidence-chain length, broken down by question origin. (c) Number of streams per case. (d) Distribution of individual stream durations by source. (e) Task families, subtasks, and question formats.

didates answered correctly in both trials are rejected or rewritten, with at most one rewrite allowed. Gemini 3 Flash additionally probes scene-muted and isolated-stream conditions to identify potential violations of audio dependence and multi-stream dependence. These construction-time probes guide candidate revision and removal before human review (Zhu et al., 2026; Sun et al., 2026).

Human verification and authoring. Fifteen English-proficient university students reviewed and revised model-generated candidates and authored additional questions. For each retained item, an annotator checked answer uniqueness and whether the answer required scene audio and evidence from at least two supplied streams. A randomly sampled subset was re-reviewed by the annotator team. Ambiguous items were referred to three additional reviewers outside the initial pool, who decided whether to revise or remove the item. Appendix C.2 details the inclusion criteria.

3.4 Benchmark Statistics

As shown in Figure 4, **MS-OmniBench** contains 3,109 questions from 581 cases and 2,367 streams across six source datasets. Cases contain two to six streams, with an average of 4.1 streams per case. Individual streams range from 25 to 300 seconds, with a median duration of 192 seconds. The total 124 hours sum stream durations rather than unique scene time. The benchmark covers twelve subtasks across four task families, with subtask sizes ranging from 132 to 400 questions. The three input question formats comprise 1,469 text-only, 1,115 reference-audio, and 525 reference-image questions. Evidence chains include 1,564 two-hop, 1,078 three-hop, 420 four-hop, and 47 five-or-more-hop questions. Additional details on the source datasets, data construction pipeline, representative benchmark examples, and dataset statistics are provided in the Appendix.

4 MESA: Multi-stream Evidence Search Agent

Relevant evidence in MS-OmniBench is distributed across streams and time. Motivated by selective evidence inspection in recent video and omni-modal agents (Wang et al., 2024; Zhang et al., 2026; Tao et al., 2025), we introduce **MESA**, a simple Multi-stream Evidence Search Agent. MESA organizes observations into per-stream event timelines to locate candidate evidence, then performs targeted media inspection and cross-stream verification before making a final decision. It uses the same backbone as its end-to-end counterpart for timeline construction and all perception calls.

Timeline construction. Before question answering, the backbone constructs a timeline for each stream from the raw media, using GoP intervals and audio onsets as temporal anchors. Records include timestamped visual events, speech, non-speech sounds, music, and within-stream audio-visual associations. Timelines are cached per case and shared across questions from the same case. They serve as retrieval aids rather than ground truth and may contain incomplete or incorrect ob-

servations. The agent does not access question-construction annotations, gold answers, annotated evidence chains, or task labels.

Evidence-guided inspection. Given the question, options, and optional reference media, the agent follows four stages of *planning, localization, discrimination, and decision*. Planning uses the cached timelines and stream summaries to identify candidate anchor windows and the observations needed to distinguish the options without directly inspecting scene media. For questions with reference media, MESA may additionally characterize the supplied reference before scene localization, with this call counted toward the same shared media-call budget. Candidate anchors are then verified against the original media, with coarse-to-fine search used when the timelines do not provide a reliable match. Targeted probes populate an option-by-evidence matrix. `compare_windows` jointly inspects candidate intervals to assess correspondences across streams or time. `ref_locate` matches supplied reference media against candidate scene windows and verifies the correspondence. All media-inspection tools use the backbone (Comanici et al., 2025; Xu et al., 2025) and return textual findings to the planner.

Answer selection. MESA performs question-time media inspection under a shared global budget of at most 12 evidence-search media calls. The agent revisits selected evidence windows and makes the final prediction from the accumulated option-specific evidence. Stage-wise call limits constrain localization, discrimination, and final re-inspection within this shared budget rather than defining independent additive budgets. The auxiliary end-to-end prediction is included in the reported question-time media-call and token totals. MESA outputs one option letter together with an evidence chain, while answer accuracy follows the same scoring protocol as end-to-end evaluation.

5 Experiments

5.1 Experimental Setup

Evaluated systems. We evaluate seventeen end-to-end systems, comprising eight proprietary and nine open-weight models. The proprietary group covers Gemini (Comanici et al., 2025), Qwen3.5-Omni (Team, 2026a), and Doubao-Seed-2.0 variants (Seed, 2026). The open-weight group includes Nemotron-3-Nano-Omni (Deshmukh et al., 2026), Qwen3-Omni (Xu et al., 2025), MiMo-V2.5 (Team, 2026b), Qwen2.5-Omni (Jin Xu, 2025), Baichuan-Omni-1.5 (Li et al., 2025), Ming-Lite-Omni-1.5 (AI et al., 2025), Phi-4-multimodal (Abouelenin et al., 2025), and MiniCPM-o 4.5 (Cui et al., 2026). We also evaluate four existing agentic systems, OmniAgent (Tao et al., 2025), Deep Video Discovery (Zhang et al., 2026), the Daily-Omni baseline (Zhou et al., 2025), and LongShOTAgent (Kurpath et al., 2025). Our **MESA** baseline is evaluated with Gemini 3.8 Flash and Qwen3-Omni-30B.

Inputs and metrics. Each item is evaluated with its specified scene streams, their available audio, and optional reference audio or image, rather than all recordings available in the source case. End-to-end models process these inputs directly, while agentic systems access them through their respective inspection pipelines. Performance is measured by whether the selected option matches the reference answer. We report subtask accuracy together with item-weighted category and overall accuracy.

Implementation details. End-to-end models use one frozen task prompt and a single run per item. Temperature is set to 0 when exposed by the corresponding inference interface. External agents retain their system-specific orchestration prompts while receiving the same benchmark questions and specified scene inputs, with final answers evaluated under the same scoring protocol. API-accessed systems are evaluated through their provider endpoints. The remaining open-weight models are deployed locally on A800-80GB GPUs using the authors' code. Further details on model versions, media preprocessing, input configurations, and prompts are provided in the Appendix.

5.2 Main Results

Table 1 presents the results across four task families and twelve subtasks. We summarize the principal observations below.

Aggregate rankings conceal substantial task specialization. Gemini 3.8 Flash achieves the highest end-to-end accuracy at 61.9%, while Nemotron-3-Nano-Omni leads the open-weight models

Table 1: **Main results on MS-OmniBench.** Subtask scores are aggregated into item-weighted category and overall accuracies. Bold and underline indicate the best and second-best results within the proprietary, open-weight, and agentic blocks, respectively.

Model	Overall	Sync. Cross-Stream					Async. Cross-Segment				Audio-Anchored Retrieval				Consistency Verification		
		SrcDis	*OffCau*	*SpkBind*	*SpVis*	**Avg**	*MemA*	*EvOrd*	*MHop*	**Avg**	*TrigLoc*	*XCount*	*AmbSc*	**Avg**	*AVInc*	*Manip*	**Avg**
Proprietary MLLMs																	
Gemini 3.8 Flash	**61.9**	**62.7**	64.5	65.7	68.1	65.0	**69.9**	57.4	70.2	**66.7**	62.0	38.6	67.4	55.6	60.5	49.4	**57.3**
Gemini 3.1 Pro	61.2	60.7	63.2	65.3	**70.3**	64.4	63.8	**65.6**	65.1	64.8	**64.2**	33.6	72.0	55.8	**61.2**	45.1	56.6
Gemini 3.5 Flash	61.0	61.2	**68.2**	**70.1**	69.4	**66.8**	66.7	59.8	65.4	64.4	61.1	36.4	**73.5**	55.6	56.8	40.1	52.0
Qwen3.5-Omni-Plus	59.9	58.6	63.6	64.6	64.2	62.4	63.8	57.9	64.7	62.6	60.2	40.5	72.0	**56.1**	58.0	**50.0**	55.7
Gemini 3 Flash	59.7	61.2	61.2	64.6	66.8	63.2	69.2	61.7	64.7	65.5	60.8	34.5	69.7	54.1	59.0	34.0	51.8
Doubao-Seed-2.0-lite	58.8	55.9	59.9	64.6	60.3	59.9	63.8	58.4	**71.2**	65.1	58.1	**45.5**	68.2	56.0	57.8	34.6	51.1
Qwen3.5-Omni-Flash	55.7	52.4	60.7	61.6	57.6	57.7	61.3	54.5	56.3	57.6	53.6	41.4	68.9	52.6	59.0	37.7	52.8
Doubao-Seed-2.0-mini	53.6	48.5	52.9	54.2	55.9	52.5	62.0	58.9	58.0	59.6	54.8	41.4	64.4	52.3	55.5	32.7	48.9
Open-weight MLLMs																	
Nemotron-3-Nano-Omni	**52.0**	**47.3**	**62.4**	49.4	**54.6**	**52.8**	**56.3**	**49.8**	**59.0**	**55.6**	**51.2**	37.3	54.5	47.4	53.0	**46.9**	**51.2**
Qwen3-Omni-30B	50.6	41.4	56.2	56.1	54.1	51.1	54.5	45.5	53.9	51.9	46.1	**43.6**	59.8	**48.0**	**53.8**	44.4	51.1
MiMo-V2.5	49.3	46.4	49.6	**56.5**	48.5	50.1	**56.3**	47.8	58.6	54.9	47.9	31.8	**62.1**	45.5	50.0	30.9	44.5
Qwen2.5-Omni-7B	42.7	35.5	50.0	45.0	44.1	43.0	46.2	45.9	48.8	47.1	41.0	31.4	53.0	40.2	43.5	27.2	38.8
Baichuan-Omni-1.5	40.9	35.5	43.4	42.4	38.9	39.7	45.9	45.5	43.7	45.0	42.5	31.4	45.5	39.5	45.8	22.8	39.1
Ming-Lite-Omni-1.5	39.0	37.6	49.6	42.4	39.3	41.9	41.9	42.6	39.7	41.3	36.7	29.5	42.4	35.5	41.8	17.9	34.9
Phi-4-multimodal	34.2	33.4	32.2	36.2	29.7	33.1	37.3	37.3	35.9	36.8	29.2	35.0	36.4	32.5	35.8	32.1	34.7
Qwen2.5-Omni-3B	31.8	30.8	36.8	36.2	29.7	33.2	33.3	28.2	35.9	33.0	31.3	22.3	37.9	29.7	34.8	18.5	30.1
MiniCPM-o 4.5	29.6	29.9	33.9	21.0	25.8	27.7	34.1	31.1	31.5	32.3	27.4	29.5	42.4	31.0	34.0	12.3	27.8
Agentic Systems																	
OmniAgent	49.8	39.6	44.6	42.1	69.4	47.7	54.8	60.8	45.8	53.0	44.3	40.0	63.6	46.6	52.5	**55.6**	53.4
Deep Video Discovery	46.2	47.3	36.8	55.7	47.2	47.0	57.7	51.7	44.4	51.1	40.4	22.7	66.7	39.8	55.8	19.8	45.4
Daily-Omni Agent	40.4	26.3	42.1	42.1	47.2	38.2	40.5	36.4	43.1	40.4	42.2	45.9	56.8	46.2	42.5	24.7	37.4
LongShOT Agent	37.6	26.3	52.5	28.0	58.5	39.4	31.2	36.4	32.9	33.2	34.6	37.3	47.0	37.9	44.3	28.4	39.7
MESA$_{[\text{Qwen3-Omni}]}$	52.1	43.5	61.2	53.9	56.3	52.8	51.6	48.8	60.0	54.0	43.1	46.4	65.2	48.4	54.0	49.4	52.7
MESA$_{[\text{Gemini 3.8}]}$	**65.6**	**66.0**	**72.7**	**66.4**	**73.8**	**69.3**	**68.1**	**63.2**	**76.9**	**70.1**	**59.9**	**46.8**	**77.3**	**59.1**	**62.5**	54.9	**60.3**

at 52.0%. However, no single end-to-end model dominates the individual reasoning skills. Five different proprietary models achieve the best result on at least one subtask, indicating substantial specialization across evidence operations. The proprietary–open-weight gap is also highly task dependent, ranging from only 1.9 percentage points on XCount to 13.6 points on SpkBind. Thus, a higher overall score does not imply a uniform advantage across the benchmark.

Counting and correspondence verification remain shared bottlenecks. No evaluated end-to-end model exceeds 45.5% on XCount, while the strongest Manip result reaches only 50.0%. For Gemini 3.8 Flash, XCount accuracy is 38.6%, compared with 62.0% on TrigLoc and 67.4% on AmbSc. This large variation within the same task family shows that locating or interpreting acoustic evidence does not imply reliable enumeration and de-duplication across streams. The proprietary–open-weight gaps on XCount and Manip are only 1.9 and 3.1 points, respectively, suggesting that these difficulties persist across model classes rather than being resolved by stronger overall models.

Existing agents provide selective advantages but limited aggregate gains. The strongest external agent, OmniAgent, reaches 49.8% overall, below the leading end-to-end models. Nevertheless, agentic systems can be competitive on selected operations. OmniAgent reaches 55.6% on Manip, exceeding the best end-to-end result of 50.0%, while Daily-Omni reaches 45.9% on XCount compared with the end-to-end maximum of 45.5%. Their limited aggregate advantage may partly reflect the bounded per-stream duration of MS-OmniBench, which reduces the extreme long-context retrieval burden targeted by many video agents. The results suggest that long-video search capabilities do not automatically translate into reliable association of acoustic and visual evidence across streams. These comparisons remain system-level because the external agents use different backbones, tools, and inference budgets.

MESA provides broad family-level gains while remaining subtask dependent. MESA improves Gemini 3.8 Flash from 61.9% to 65.6% and Qwen3-Omni from 50.6% to 52.1%. Both backbones improve across all four task families, with larger gains on tasks involving multi-step evidence combination and verification, such as MHop and Manip. In contrast, TrigLoc and MemA decrease for both backbones. These shared patterns suggest that explicit evidence search is more beneficial for complex evidence integration than for tasks with relatively direct evidence access. The results characterize the complete MESA configuration rather than isolating the contribution of timeline construction alone.

Table 2: **End-to-end accuracy (%) by supplied stream count, annotated evidence-chain length, and question format.** Avg-17 denotes the unweighted mean over all seventeen end-to-end models.

Group	n	Gemini 3.8 Flash					Doubao-Seed-2.0-mini					Qwen3-Omni-30B					Avg-17				
		Sync.	*Async.*	*Retr.*	*Verif.*	**Ovr.**	*Sync.*	*Async.*	*Retr.*	*Verif.*	**Ovr.**	*Sync.*	*Async.*	*Retr.*	*Verif.*	**Ovr.**	*Sync.*	*Async.*	*Retr.*	*Verif.*	**Ovr.**
Input streams																					
2	1,898	64.2	65.4	50.0	58.0	60.2	50.8	61.9	48.1	57.4	54.1	50.6	54.4	49.5	57.7	52.5	50.5	54.6	44.7	51.2	50.4
3	437	60.6	69.4	59.1	58.5	62.0	56.1	55.1	60.2	49.1	55.8	57.1	53.1	58.0	50.9	55.6	52.9	53.9	53.4	48.2	52.6
4	310	70.3	75.0	69.6	57.4	68.4	54.9	69.4	62.0	29.4	54.5	49.5	54.2	35.4	38.2	44.5	52.3	55.1	48.1	36.0	48.3
5	358	72.0	64.0	70.0	61.7	66.5	50.5	50.5	56.7	40.4	48.9	37.6	40.5	46.7	45.7	42.2	45.2	46.8	48.2	38.7	44.5
6	106	73.5	70.6	59.3	32.1	58.5	64.7	41.2	55.6	28.6	49.1	67.6	35.3	29.6	25.0	41.5	58.1	41.5	39.4	26.5	42.3
Annotated evidence-chain length																					
2 hops	1,564	64.7	64.7	48.9	54.1	58.8	51.8	60.0	46.3	49.0	51.0	50.1	54.2	49.2	53.2	51.2	50.3	53.9	45.0	45.2	48.4
3 hops	1,078	64.7	67.2	65.3	67.8	66.1	54.8	60.2	60.4	49.2	57.6	52.3	51.0	49.4	45.8	50.4	51.4	53.8	49.7	48.8	51.6
$\geq$4 hops	467	67.1	67.4	50.9	58.8	62.7	51.4	58.0	50.0	47.1	53.3	53.6	51.4	41.1	44.1	49.0	52.1	51.0	42.1	42.0	48.5
Question format																					
Text-only question	1,469	60.4	66.4	45.6	56.3	58.3	52.1	64.6	47.8	48.6	54.0	48.8	53.8	51.1	52.6	51.7	49.8	55.1	41.5	45.4	48.6
Text + reference audio	1,115	66.6	67.1	63.3	60.9	65.5	50.7	56.5	58.5	46.4	53.6	52.1	50.5	49.3	44.9	50.6	50.9	51.4	51.6	46.5	50.9
Text + reference image	525	69.4	66.7	59.3	59.3	64.8	58.5	49.3	47.5	52.3	52.6	53.0	47.8	37.3	48.8	47.4	52.7	49.7	44.2	47.1	49.1

5.3 Fine-Grained Analysis

Table 2 groups the end-to-end results by supplied stream count, annotated evidence-chain length, and question format. Evidence-necessity diagnostics are provided in Appendix C, with source-by-stream and origin-by-chain breakdowns in Appendix D.

The effect of additional streams depends on the evidence they contribute. The aggregate results are strongly non-monotonic. Gemini improves from 60.2% on two-stream questions to 66.5% on five-stream questions, whereas Qwen decreases from 52.5% to 42.2%. Within-source analysis clarifies this divergence. On EasyCom, where additional streams provide source-specific microphone observations, all three models improve as more streams are supplied. In contrast, additional viewpoints in STARSS and Ego-Exo4D do not produce consistent gains, particularly for the open-weight models. Stream count alone is therefore a poor proxy for difficulty; performance depends on the information contributed by the additional streams and the model's ability to integrate it.

The relation between chain length and difficulty depends on authoring origin. The seventeen-model mean is 48.4%, 51.6%, and 48.5% for two, three, and at least four hops, respectively. This apparently flat trend partly reflects differences in question origin. Within model-assisted questions, performance drops from three to four-or-more hops, with the seventeen-model mean decreasing from 51.0% to 43.2%. The corresponding human-authored three-hop and four-or-more-hop groups both achieve 52.1%. Since human-authored questions constitute a substantial fraction of the longest-chain group, the aggregate statistics obscure this interaction. Evidence-chain length should therefore be interpreted jointly with authoring origin and task rather than as a universal measure of difficulty.

Reference media change the grounding problem rather than uniformly changing difficulty. Gemini performs better on both reference-audio and reference-image questions than on text-only questions, while Qwen shows a marked drop on image-referenced questions. More detailed analysis shows that reference audio can provide a strong anchor when the sound is distinctive, whereas reference images become more challenging when the referenced entity must be re-identified across viewpoints. The effect of question format therefore depends on the grounding operation required by the reference, rather than on the presence of additional reference media alone.

6 Conclusion

We present **MS-OmniBench**, a benchmark for reasoning that jointly requires acoustic evidence and complementary information across multiple streams. It contains 3,109 human-reviewed or human-authored questions across four task families and twelve subtasks. Evaluation of seventeen Omni-LLMs and four existing agentic systems reveals substantial task specialization, with cross-modal counting and correspondence verification remaining particularly challenging. We further introduce **MESA**, a simple multi-stream evidence-search baseline that improves both evaluated backbones overall, although its gains vary across tasks. These results show that access to multiple modalities and streams alone is insufficient. Effective reasoning also requires locating, associating, and verifying distributed evidence across streams. We hope **MS-OmniBench** supports progress toward more reliable omni-modal reasoning across recordings.

Ethics Statement

MS-OmniBench is constructed from six existing public research datasets and does not introduce new recordings of human participants. We use the source media solely for research purposes and follow the access and licensing conditions of the corresponding datasets. The benchmark release will not redistribute the original video or audio; instead, we release newly created questions, annotations, evidence metadata, and reconstruction and evaluation tools, while the source media must be obtained through their official channels. Because several source datasets contain identifiable faces and voices, privacy and redistribution constraints follow the terms of the original datasets. Controlled variants modify audio tracks or on-screen text only for benchmark evaluation and do not synthesize new identities. Model-assisted question construction and screening are followed by human verification of the answer and supporting audio-visual evidence. We will document source-specific licenses, data provenance, and release conditions together with the benchmark.

Reproducibility Statement

We provide the benchmark formulation, task taxonomy, construction pipeline, MESA design, and evaluation protocol in the main paper. The supplementary material further documents the source-to-stream construction, media preprocessing and temporal indexing, benchmark construction and screening prompts, human verification protocol, system-specific input configurations, additional diagnostic results, and the implementation and computational cost of MESA. The benchmark will be released at MS-OmniBench.

AI Use Statement

Generative AI tools were used for two purposes in this work. First, Gemini 3 Flash was used to generate model-assisted benchmark content during the construction of MS-OmniBench, including event annotations, cross-stream alignments, question plans, question wording, and shortcut-screening outputs. All benchmark questions produced through this model-assisted pipeline subsequently underwent the automatic checks and human verification procedures described in the paper. Second, generative AI assistants were used only for language polishing and editorial refinement of the manuscript. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final manuscript.

References

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

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

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 609–617. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.73. URL https://doi.org/10.1109/ICCV.2017.73.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pp. 28–39. Springer, 2005.

Jianghan Chao, Wenhui Tan, Yuchong Sun, Ruihua Song, Liyun Ru, et al. Jointavbench: A benchmark for joint audio-visual reasoning evaluation. In *International Conference on Learning Representations*, volume 2026, pp. 84035–84062, 2026.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025.

Yu Chen, Caorui Li, Ziyu Xiong, Yidong Wang, Mingqi Gao, Shuman Liu, Biao Liu, Chunfeng Yang, Anxiang Zeng, Haibo Zhang, et al. Omnireasoner: Thinking with long audio-video via native tool use. *arXiv preprint arXiv:2607.19339*, 2026a.

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*, 2026b.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1590–1601. IEEE, 2025.

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.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, et al. Nemotron 3 nano omni: efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. *arXiv preprint arXiv:2107.04174*, 2021.

Hengyi Feng, Hao Liang, Mingrui Chen, Bohan Zeng, Meiyi Qiang, Zhengyang Zhao, Zimo Meng, Zeang Sheng, and Wentao Zhang. Traceav-bench: Benchmarking multi-hop trajectory reasoning over long audio-visual videos. *arXiv preprint arXiv:2605.07593*, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long MA, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *Advances in Neural Information Processing Systems*, 38:75300–75320, 2026.

Siminfar Samakoush Galougah, Rishie Raj, Sanjoy Chowdhury, Sayan Nag, and Ramani Duraiswami. Aura: A fine-grained benchmark and decomposed metric for audio-visual reasoning. *arXiv preprint arXiv:2508.07470*, 2025.

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.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

Yuping He, Yifei Huang, Guo Chen, Baoqi Pei, Jilan Xu, Tong Lu, and Jiangmiao Pang. Egoexobench: A benchmark for first-and third-person view video understanding in mllms. *Advances in Neural Information Processing Systems*, 38, 2026.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 52423–52443, 2026.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Mohammed Irfan Kurpath, Jaseel Muhammad Kaithakkodan, Jinxing Zhou, Sahal Shaji Mullappilly, Mohammad Almansoori, Noor Ahsan, Beknur Kalmakhanbet, Sambal Shikhar, Rishabh Lalla, Jean Lahoud, et al. A benchmark for omni-modal reasoning in long videos. *arXiv preprint arXiv:2512.16978*, 2025.

Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21(2):522–535, 2018.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Zhenghao Song, Dingling Zhang, Ying He, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. In *International Conference on Learning Representations*, volume 2026, pp. 138214–138236, 2026.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22195–22206, 2024.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3318–3327. IEEE, 2025.

Lidong Lu, Guo Chen, Zhu Wei, Zhiqi Li, Yicheng Liu, and Tong Lu. Av-reasoner: Improving and benchmarking clue-grounded audio-visual counting for mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33477–33487, 2026.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18936–18946. IEEE, 2025.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Zabir Al Nazi, Shubhashis Roy Dipta, and Md Rizwan Parvez. Omni-modal dissonance benchmark: Systematically breaking modality consensus to probe robustness and calibrated abstention. *arXiv preprint arXiv:2603.27187*, 2026.

Gorjan Radevski, Teodora Popordanoska, Matthew Blaschko, and Tinne Tuytelaars. Dave: Diagnostic benchmark for audio visual evaluation. *Advances in Neural Information Processing Systems*, 38, 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. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

Sahil Shah, SP Sharan, Harsh Goel, Manvik Pasula, Adithya Hebbalae, Minkyu Choi, and Sandeep P Chinchali. Crossview: Can vision-language models reason across cameras? In *European Conference on Computer Vision*, pp. 266–285. Springer, 2026.

Xin Shen, Rui Zhu, Lei Shen, Zhuojie Wu, Xinyu Wang, Kaihao Zhang, Tianqing Zhu, Shuchen Wu, Chenxi Miao, Weikang Li, et al. Fingercap: Fine-grained finger-level hand motion captioning. In *European Conference on Computer Vision*, pp. 531–550. Springer, 2026.

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

Peiwen Sun, Xudong Lu, Huadai Liu, Yang Bo, Dongming Wu, Huankang Guan, Minghong Cai, Jinpeng Chen, Xintong Guo, Shuhan Li, et al. X-stream: Exploring mllms as multiplexers for multi-stream understanding. *arXiv preprint arXiv:2606.02482*, 2026.

Kim Sung-Bin, Oh Hyun-Bin, Lee Jung-Mok, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *International Conference on Learning Representations*, volume 2025, pp. 24244–24271, 2025.

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.

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.

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

Xiaomi MiMo Team. Mimo-v2. 5, 2026b.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European conference on computer vision*, pp. 252–268. Springer, 2018.

Angela van Sprang, Laurens Samson, Ana Lucic, Erman Acar, Sennay Ghebreab, and Yuki M Asano. Same content, different answers: Cross-modal inconsistency in mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8781–8790, 2026.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967. IEEE, 2025.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024.

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*, 2026.

Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7): 560–576, 2003.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Zhuojie Wu, Shijie Wang, and Xin Yu. Metom: Metadata-guided token merging for efficient video llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10441–10450, 2026.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

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.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6580–6598, 2025.

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, 2026.

Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *European conference on computer vision*, pp. 587–604. Springer, 2018.

Muyan Zhong, Erfei Cui, Sen Xing, Weiyun Wang, Wen Wu, Yuchen Hu, Yanting Zhang, Xiaowei Hu, Wenhai Wang, Chao Zhang, et al. Have-bench: Hierarchical audio-visual evaluation from perception to interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8801–8812, 2026.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2(5):6, 2024.

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.

Rui Zhu, Xin Shen, Shuchen Wu, Chenxi Miao, Xin Yu, Yang Li, Weikang Li, Deguo Xia, and Jizhou Huang. Video-msr: Benchmarking multi-hop spatial reasoning capabilities of mllms. *arXiv preprint arXiv:2601.09430*, 2026.

This appendix is organized as follows:

- LIMITATION AND FUTURE WORK (Section A).
- BENCHMARK CONSTRUCTION AND PROMPT DESIGN (Section B).
- QUALITY CONTROL AND BENCHMARK VALIDATION (Section C).
- EVALUATION DETAILS AND ADDITIONAL ANALYSIS (Section D).
- MESA IMPLEMENTATION AND ANALYSIS (Section E).
- REPRESENTATIVE BENCHMARK CASES AND RELEASE CONTENTS (Section F).

A Limitation and Future Work

MS-OmniBench is designed to isolate audio-dependent reasoning over complementary multi-stream evidence while keeping the evaluation practical for current Omni-LLMs. We nevertheless recognize limitations of the current benchmark and MESA, which also motivate directions for future work.

Longer multi-stream context. Each individual stream in MS-OmniBench is limited to at most five minutes, which reduces the confounding effect of extreme context length and places greater emphasis on locating and associating evidence across streams. This design also takes into account the practical input and inference constraints of current Omni-LLMs. A multi-stream case may still contain substantially more total media than any individual stream, but the benchmark does not directly evaluate hour-scale memory, continual monitoring, or long-horizon online reasoning. Future versions can extend the current construction pipeline to longer synchronized and asynchronous recordings, enabling evaluation of how models retain, retrieve, and integrate acoustic and visual evidence over substantially longer multi-stream contexts.

Benchmark coverage and evaluation format. MS-OmniBench is constructed from six existing research datasets, so its coverage is bounded by the scenes, activities, acoustic sources, and sensor configurations available in these sources. The five-option format enables consistent evaluation across heterogeneous models, but answer accuracy does not fully capture open-ended reasoning quality or the faithfulness of model-generated evidence. Model-assisted construction may also influence the resulting question distribution despite shortcut screening, human verification, and direct human authoring. Future work can use the current evidence-grounded construction pipeline to expand MS-OmniBench with additional domains, sensing conditions, and reasoning patterns. Open-ended answers and generated evidence chains could further be evaluated with carefully validated LLM-based judges, extending the benchmark beyond multiple-choice accuracy.

MESA and evidence-search reliability. MESA is intended as a simple evidence-search baseline rather than a complete solution to multi-stream reasoning. Its gains are demonstrated with two backbones, and the current experiments do not establish that the same improvements generalize across a broader range of models and agent architectures. The model-generated timelines and subsequent tool calls may omit relevant events, introduce incorrect associations, or provide misleading evidence to later reasoning stages. Repeated media inspection also introduces additional computation relative to direct end-to-end answering. Future work can evaluate evidence-search strategies across more backbones and matched computational budgets, while improving timeline construction, tool reliability, and evidence verification to reduce error propagation during multi-stream reasoning.

B Benchmark Construction and Prompt Design

This section provides additional details on the construction pipeline described in the main paper. We focus on source-to-stream construction, temporal event annotation, cross-stream alignment, evidence-grounded question generation, and verification. Gemini 3 Flash is used throughout the model-assisted stages with frozen stage-specific prompts. For readability, we present shortened versions of the prompts that retain the task-defining instructions and omit implementation-specific formatting. The complete prompts and construction code will be released with MS-OmniBench.

Table 3: **Source-to-stream construction.** Audio structure means the acoustic evidence exposed through the constructed streams.

Source	Cases	Questions	Stream structure	Audio structure
Ego-Exo4D	143	737	One ego view with several synchronized exocentric views	Audio is retained on the ego stream, while the exocentric streams are video-only in our construction.
STARSS	121	706	Directional views extracted from panoramic video	Spatial audio is down-mixed to stereo and associated with the primary directional stream.
Ego4D	103	552	Concurrent egocentric views or non-adjacent temporal segments	Native audio from the selected egocentric recordings or temporal segments.
EasyCom	73	461	Glasses-camera video paired with different microphone observations	Array-microphone and participant-specific close-microphone tracks provide distinct acoustic viewpoints.
AMI	97	414	Participant close-up views together with a room view	Close-up streams are paired with participant headset microphones, while the room-view stream is video-only.
URMP	44	239	Ensemble view together with player-specific visual streams	The ensemble mix and isolated instrument tracks provide source-specific acoustic evidence.

Table 4: **Example event records used for cross-stream alignment.**

Time	Modality	Stream	Event
79.6–82.4	Speech	Closeup 1	"Fair enough"
81.3–82.3	Audio	Closeup 4	Chuckle and verbal agreement
80.0–120.0	Visual	Corner	Woman in light blue speaks
99.0–100.5	Visual	Closeup 3	Man drinks from a mug

B.1 Source-to-Stream Construction

Table 3 summarizes how the six source datasets are converted into the stream structures used by MS-OmniBench. A *stream* is a media unit supplied to the model as part of the scene input. A stream may correspond to a camera viewpoint, a source-specific acoustic observation paired with visual content, or a non-adjacent temporal segment. The streams supplied with a question need not contain every recording available in the corresponding source case. Reference audio and reference images belong to the question specification and are not counted as scene streams.

Data access and licensing. All six sources are publicly available research datasets used under their original licenses or access agreements. Ego4D and Ego-Exo4D require users to accept their official dataset license agreements before accessing the corresponding media. AMI is publicly distributed under the Creative Commons Attribution 4.0 International license. EasyCom is distributed under the Creative Commons Attribution-NonCommercial 4.0 International license. STARSS is publicly distributed under the MIT License. URMP is distributed through the Dryad repository, where the released dataset is available under CC0. We use these datasets for research under their respective terms and do not redistribute their original video or audio. The MS-OmniBench release instead provides newly created questions, annotations, evidence metadata, reconstruction information, and evaluation tools.

B.2 Event Annotation and Cross-Stream Alignment

Each stream is temporally indexed before model-assisted annotation. Group-of-pictures intervals provide visual anchors, while transcript timestamps and sound onsets provide additional temporal anchors when available. These anchors support event annotation without imposing fixed semantic boundaries.

A per-stream Gemini 3 Flash call extracts timestamped visual events, speech, non-speech sounds, and music. Visual and acoustic events are represented separately so that audio-visual correspondence is not assumed during extraction. Each event retains its stream identifier and temporal range. Visible sound sources are linked to visual events only when supported by the media. Sounds without a visible source are explicitly retained as off-screen or background observations. Table 4 shows a compact example of the intermediate event representation.

A deterministic mining stage identifies recurring events, confusable observations, off-screen sounds, and temporally separated event pairs. These mined anchors provide candidate evidence relations for

Prompt Template for Cross-Stream Alignment

System:

You are an expert in aligning audio-visual evidence across multiple streams.

User:

You are given per-stream event records, synchronization information, audio-bearing streams, and automatically mined difficulty anchors.
Your task is to build a cross-stream relation graph and identify useful opportunities for multi-stream reasoning.
Requirements:

- Align recurring people, objects, sounds, scenes, and actions across streams.
- Preserve the supporting stream identifiers, event identifiers, and temporal ranges.
- Identify cross-stream relations involving identity, temporal correspondence, visible sound sources, off-screen sound sources, and plausible distractors.
- Prefer opportunities supported by confusable events, off-screen acoustic evidence, or temporally separated observations.
- Keep uncertain cross-stream correspondences unresolved rather than forcing a match.
- Avoid opportunities centered on a unique visual event that can be solved without acoustic evidence.

Organize the resulting opportunities according to the four MS-OmniBench task families.

subsequent question construction. A scene-level alignment call then associates recurring entities and events across streams while preserving their stream and temporal provenance. The resulting relations include cross-view identity, temporal correspondence, visible sound-source candidates, off-screen source candidates, and potential visual or acoustic distractors. Uncertain correspondences remain unresolved rather than being forced into a single match.

B.3 Evidence-Grounded Question Generation

Question generation is separated into planning and natural-language writing. The planning stage specifies the required evidence relation before the final question wording is produced. Each plan identifies the target task, question format, supporting streams, audio and visual evidence, distractors, and an evidence chain. This separation allows the intended reasoning structure to be checked independently of the final natural-language realization.

Question planning. One Gemini 3 Flash planning call is performed for each case. Each call produces up to eight candidate plans. A plan contains two to five evidence hops and requires at least two supporting streams. Each hop is associated with a cross-modal, cross-view, or cross-time relation. The plan additionally specifies why acoustic evidence and multiple streams are required for resolving the question. For asynchronous questions, the supporting evidence is separated across non-adjacent temporal regions. Distractors are preferentially selected from the same confusable event group as the answer.

Question writing. Each accepted plan is converted into a five-option multiple-choice question by a multimodal Gemini 3 Flash call. The writer receives the approved plan together with the required scene streams. The source media are used to verify the details specified by the plan before producing the final question. For model-assisted natural questions, distractors are taken from the planned confusable event group whenever possible. Reference audio and reference images are extracted from the source media according to the generated reference specification.

B.4 Screening and Human Verification

Generated candidates undergo structural validation, shortcut screening, dependence probes, and human verification before inclusion in MS-OmniBench. Structural checks verify the task label, required streams, evidence identifiers, evidence-chain structure, question format, reference specifi-

Prompt Template for Question Planning

System:

You are the question planner for a challenging audio-dependent multi-stream benchmark.

User:

Your task is to construct question plans from the aligned cross-stream evidence.

Produce plans only and do not write the final question.

Requirements:

- Select one MS-OmniBench task family and subtask.
- Select a text-only, reference-audio, or reference-image question format.
- Require at least two supporting streams that contribute necessary evidence.
- Identify the required acoustic and visual evidence.
- Construct an evidence chain containing two to five hops.
- Include cross-modal or cross-view relations in the evidence chain.
- Make each hop necessary for resolving an ambiguity in the question.
- Prefer confusable, off-screen, or temporally separated evidence identified during candidate mining.
- Select distractors from the same confusable event group when possible.
- Ensure that the planned answer cannot be justified from muted visual input alone.
- Ensure that the intended evidence relation cannot be resolved from a single designated stream.
- For asynchronous questions, use evidence from temporally separated events.

Record the required streams, evidence identifiers, distractors, and evidence chain for the subsequent writing stage.

cation, answer options, and gold answer. Referenced evidence must resolve to valid events in the corresponding streams and temporal ranges. Question-quality checks additionally reject wording that exposes camera names, stream identifiers, timestamps, solution procedures, or other unintended shortcuts. Candidates that fail the structural or question-quality checks can be rewritten once before removal.

No-media shortcut screening. The no-media test presents only the question and answer options to Gemini 3 Flash. Each candidate is evaluated twice using different option orders. Predicted option letters are mapped back to the canonical ordering. Candidates answered correctly in both trials are rejected or rewritten. This test reduces questions that can be solved primarily from textual priors or option wording.

Audio-dependence probe. Gemini 3 Flash additionally evaluates candidates under a scene-muted condition to identify potential violations of audio dependence. The question and non-audio inputs

Prompt Template for Question Generation

System:

You are an expert in writing evidence-grounded questions for multi-stream audio-visual reasoning.

User:

Write one five-option multiple-choice question that faithfully realizes the provided evidence plan.

The question should describe the underlying event rather than the recording process.

Requirements:

- Preserve the required evidence relation and evidence chain from the approved plan.
- Do not mention cameras, views, streams, recordings, devices, or timestamps.
- Do not describe the procedure that a model should use to solve the question.
- Avoid verbatim quotations that directly reveal the answer.
- For reference-media questions, refer only to "the reference audio" or "the reference image".
- Generate five mutually exclusive answer options with comparable grammatical form and length.
- Ensure that every answer option corresponds to a real entity, event, location, order, or count from the source case.
- Use confusable events from the plan as distractors rather than generic commonsense alternatives.
- Preserve every necessary evidence hop from the plan.
- Revise the question if muted visual input or one designated stream is sufficient to resolve the intended evidence relation.

Return the question, five options, gold answer, supporting evidence, evidence chain, and reference specification.

are preserved while scene audio required by the original item is removed. Candidates that remain readily answerable under this condition are flagged for revision or removal before human review. The probe is used as a construction-time screening signal rather than as a formal proof of acoustic necessity.

Single-stream probe. Gemini 3 Flash also evaluates candidates under isolated-stream conditions to identify potential violations of multi-stream dependence. The question is evaluated with individual scene streams separated from the complete multi-stream input. Candidates that appear solvable from an isolated stream are flagged for revision or removal. The probe complements the planning constraints that require multiple supporting streams and explicit cross-stream evidence relations.

Human verification and authoring. Human verification comprises an item-level first pass, a randomly sampled second-pass audit, and separate adjudication of ambiguous items. Annotators inspect the question, answer options, supplied streams, and supporting evidence windows. Review covers answer correctness, acoustic dependence, cross-stream dependence, distractor validity, and reference-media validity. Candidates are accepted, revised, or removed, and items that remain ambiguous after review are discarded. Annotators also construct questions directly from the media and create controlled variants to diversify phrasing and improve coverage of sparse task types.

Table 5: **Summary of benchmark construction and human authoring.** Question-generation counts include rewrite attempts and are therefore not directly comparable with the distinct-item counts used in the subsequent review stages.

Stage	Items
Candidate question plans	4,009
Question-generation attempts (incl. rewrites)	4,158
Distinct candidates entering review-stage QC	2,391
Automatically removed	193
Removed during human review	50
Retained from review-stage QC	2,148
Model-assisted questions	1,999
Human-authored controlled variants	149
Direct human-authored candidates	1,032
Not retained after verification	71
Retained direct human-authored questions	961
Final model-assisted (MA)	1,999
Final human-authored (HA)	1,110
Final release	3,109

The final benchmark contains 1,999 model-assisted (MA) questions and 1,110 human-authored (HA) questions. Controlled variants are human-authored and are therefore included in the HA group. The two origin groups form a disjoint partition of all 3,109 released questions.

C Quality Control and Benchmark Validation

This section provides additional details on the quality-control procedures used during benchmark construction and the diagnostics used to examine the intended evidence requirements. Automatic screening is used to identify likely construction failures and shortcut solutions, while final inclusion relies on human inspection of the complete media and supporting evidence.

C.1 Construction Funnel

MS-OmniBench combines model-assisted question construction with direct human annotation and authoring. The model-assisted pipeline produces structured question plans before natural-language question generation, followed by structural validation, shortcut screening, dependence probing, and review-stage quality control. A total of 2,391 distinct candidate questions enter the review-stage quality-control process. During this stage, 193 candidates are removed by automatic benchmark-validity and quality-control checks, and an additional 50 candidates are removed during human review. This leaves 2,148 questions retained from the review-stage pool.

The final benchmark is organized into two question-origin groups used consistently throughout our analysis. The model-assisted (MA) group contains 1,999 questions generated through the model-assisted construction pipeline. The human-authored (HA) group contains 1,110 questions produced through human annotation, including controlled variants constructed during the annotation process. The two groups form a disjoint partition of all 3,109 released questions. Figure 4(b) further shows how these two origins are distributed across annotated evidence-chain lengths.

Human verification and authoring. Human verification comprises an item-level first pass, a randomly sampled second-pass audit, and separate adjudication of ambiguous items. Annotators inspect the question, answer options, supplied streams, and supporting evidence windows. Review covers answer correctness, acoustic dependence, cross-stream dependence, distractor validity, and reference-media validity. Candidates are accepted, revised, or removed, and items that remain ambiguous after review are discarded. Annotators also construct questions directly from the media and create controlled variants to diversify phrasing and improve coverage of sparse task types.

The final benchmark contains 1,999 model-assisted (MA) questions and 1,110 human-authored (HA) questions. Controlled variants constructed during human annotation are included in the HA group. These variants cover muting, audio-track swaps, subtitle injection, temporal shifts, and music replacement. The corresponding questions and answers are verified against the modified media before

inclusion in the benchmark. Together, the two origin groups form a disjoint partition of all 3,109 released questions.

C.2 Human Verification of Joint Evidence Necessity

The defining property of MS-OmniBench is that answering a question requires both acoustic evidence and complementary evidence distributed across multiple streams. We therefore perform a dedicated human necessity verification on *every question* retained in the final benchmark, rather than estimating this property from a sampled subset. A pool of 15 English-proficient university students reviewed and revised model-generated candidates and wrote additional questions. Each retained item received a first-pass check against the inclusion criteria below. Members of the annotator pool subsequently re-checked a randomly sampled subset of retained items. Ambiguous items were escalated to three reviewers outside the initial pool for a revise-or-remove decision. The sampled second pass should not be interpreted as a second independent review of every item.

For each candidate item, reviewers inspect the complete supplied input, including all scene streams, the question and answer options, and the annotated supporting evidence. An item is retained only if all of the following conditions are satisfied:

1. **Full-input sufficiency.** The complete provided input contains sufficient evidence to uniquely support the gold answer.
2. **Acoustic necessity.** After removing the answer-relevant scene audio, the remaining evidence is insufficient to fully justify the gold answer.
3. **Multi-stream necessity.** No individual supplied scene stream, when considered in isolation, contains sufficient evidence to fully justify the gold answer.

Reviewers verify these conditions against the underlying media rather than relying on model predictions. Questions that violate any condition, remain ambiguous, or contain an unintended shortcut are revised and re-checked or removed from the benchmark.

Consequently, all 3,109 released questions satisfy these human-verified inclusion criteria for joint acoustic and multi-stream dependence. This item-level verification is distinct from the reduced-input model diagnostics in Section C.3. Human verification determines whether the required evidence is necessary to justify the answer, whereas model ablations measure how current systems behave when parts of that evidence are removed. A model may occasionally select the correct option from incomplete evidence through residual cues, option elimination, or chance; such behavior does not imply that the removed evidence is sufficient to justify the answer.

C.3 Reduced-Input Diagnostics

We complement the item-level human necessity verification with controlled reduced-input evaluations on two representative Omni-LLMs. These experiments do not determine whether an item satisfies the benchmark inclusion criteria; that property is verified manually for every released question as described in Section C.2. Instead, they quantify how model performance changes when different components of the intended evidence are removed or when additional streams are supplied.

We consider five input conditions. FULL uses exactly the scene streams and reference media specified by the benchmark item. ALL-STREAMS provides all available scene streams from the corresponding source case, rather than only the benchmark-specified streams. MUTED preserves the visual inputs while removing scene audio. SINGLE evaluates each required scene stream independently and reports the average accuracy across these isolated-stream evaluations; it therefore does not privilege a particular stream or depend on stream ordering. ASR-ONLY replaces the original audio-visual scene evidence with speech transcripts. Table 6 reports overall accuracy and performance across the four task families. Removing scene audio causes substantial degradation for both models, while isolated-stream evaluation reduces accuracy to approximately the 20% chance level (21.4% for Gemini and 19.7% for Qwen). In contrast, supplying all available streams provides no consistent benefit, indicating that additional observations can also introduce irrelevant or distracting evidence. The item-level inclusion decision follows the human evidence-sufficiency checks in Section C.2. The reduced-input scores describe average model behavior under missing evidence; in

Table 6: **Reduced-input diagnostics on MS-OmniBench.** FULL uses the benchmark-specified scene streams. ALL-STREAMS additionally supplies all available scene streams from the corresponding case. MUTED removes scene audio. SINGLE evaluates each required scene stream independently and reports the average accuracy across isolated-stream evaluations. ASR-ONLY replaces the original audio-visual scene evidence with speech transcripts. Δ denotes the change in overall accuracy relative to FULL. Chance accuracy is 20% in the five-option setting.

Condition	Δ	Overall	Sync.	Async.	Retr.	Verif.
Gemini 3.8 Flash						
FULL	–	61.9	65.0	66.7	55.6	57.3
ALL-STREAMS	+0.6	62.5	65.6	67.3	55.8	57.8
MUTED	-25.5	36.4	41.4	41.8	29.1	28.5
SINGLE	-40.5	21.4	20.2	23.1	21.3	21.2
ASR-ONLY	-24.6	37.3	36.2	51.7	31.3	26.5
Qwen3-Omni-30B						
FULL	–	50.6	51.1	51.9	48.0	51.1
ALL-STREAMS	-5.0	45.6	44.8	46.7	44.2	47.2
MUTED	-26.4	24.2	27.1	27.3	18.9	20.8
SINGLE	-30.9	19.7	18.9	20.8	20.2	19.2
ASR-ONLY	-25.1	25.5	23.8	36.8	19.9	20.1

particular, SINGLE is not an estimate of the best individual stream for each item and does not by itself establish item-level necessity.

D EVALUATION DETAILS AND ADDITIONAL ANALYSIS

This section provides additional details on the frozen evaluation protocol used for the main end-to-end comparison. All seventeen end-to-end Omni-LLMs receive the same system message, question text, canonical answer options, and output instruction. No few-shot examples or model-specific prompt variants are used. Each item is evaluated once with temperature set to 0 when the corresponding inference interface exposes this parameter. The only system-specific differences are how the supplied video, audio, and reference media are delivered through the supported model interfaces.

D.1 FROZEN END-TO-END EVALUATION PROMPT

Each model receives the scene streams in the order specified by the benchmark item. Before every media file, a text label identifies the stream index, stream identifier, and whether the corresponding video contains an audio track. Reference audio or reference images are appended after the scene streams when required by the question. The answer options are always provided in their canonical benchmark order.

The media files are then supplied in sequence using labels of the following form:

Media type	Label
Scene stream	`[STREAM i] id=stream_xx · video+audio`
Scene stream without audio	`[STREAM i] id=stream_xx · video only`
Reference audio	`[REFERENCE AUDIO]`
Reference image	`[REFERENCE IMAGE]`

For all questions, the stream order follows the required streams field when it is available and otherwise follows the complete stream order of the case. The labels `video+audio` and `video only` reflect whether the supplied media file actually contains an audio track. The prompt requests only the final option in a fixed JSON format and does not request an explicit reasoning trace. This frozen text prompt is shared by all seventeen end-to-end models. Differences across systems therefore arise from the model and its supported media interface rather than from model-specific question prompting.

System Prompt for End-to-End Evaluation

System:

You are answering a multiple-choice question about a real-world scene that was recorded by one or more synchronized streams.
A stream is a video from one camera, possibly with its own audio track.
Some streams share the same view but carry different audio, for example one microphone per speaker or one track per instrument.
Some streams are two non-adjacent segments of the same recording.
Streams are labeled STREAM 0, STREAM 1, ... in the order they are given.
If a REFERENCE AUDIO or REFERENCE IMAGE is attached, the question refers to it as "the reference audio" or "the reference image".
Use ONLY the provided evidence: what you can see and hear across ALL streams.
Think silently, then output your final answer as a single JSON object on the last line.
Do not output anything after the JSON.

User Prompt for End-to-End Evaluation

User:

[EVIDENCE]
(media attached above, in order)
[QUESTION]
{question}
[OPTIONS]
A. {A} B. {B} C. {C} D. {D} E. {E}
[INSTRUCTIONS]
Choose exactly one option.
Base your answer only on the evidence above.
Output strictly: {"answer":"<A-E>"}

Table 7: **Media-delivery configurations for the evaluated systems.** The seventeen end-to-end models receive the same frozen text prompt. Differences reflect the media interfaces and context constraints supported by each system.

System family	Access	Video input	Sampling	Audio input	Additional setting
Gemini series	API	Source resolution	Provider default, approximately 1 fps	Interleaved with video	Thinking enabled
Qwen3.5-Omni	DashScope batch	180-pixel copy	Provider allocation based on video duration and resolution	Audio retained in video after loudness normalization	196k input limit, thinking disabled
Doubao-Seed-2.0	Ark batch	180-pixel copy	1 fps with provider video-token limits	Audio retained in video	224k context, provider-default thinking
MiMo-V2.5	API	180-pixel copy	Provider 1 fps processing	Audio retained in video	Provider-default reasoning configuration
Local open-weight models	Local inference	180-pixel copy	32–64 frames per stream within the supported visual budget	Separate 16-kHz audio track for each audio-bearing stream	A800-80GB GPUs using the corresponding implementation
External agents	Mixed	System-specific	System-specific	System-specific	Same benchmark questions and answer-scoring protocol

D.2 System Input Configurations

Although the text prompt is identical, current Omni-LLMs expose different media interfaces and context-allocation policies. The main leaderboard therefore compares complete systems under their supported input configurations rather than claiming an identical perceptual budget across providers. Table 7 summarizes the frozen media-delivery configurations.

Media preprocessing is fixed before evaluation for each system family and is not adapted on a per-question basis. When an interface cannot accommodate the full default media allocation, the

Table 8: **Accuracy (%) by source dataset and task family.** Each cell reports Gemini 3.8 Flash / Doubao-Seed-2.0-mini / Qwen3-Omni-30B. Sync., Async., Retr., and Verif. denote synchronous cross-stream understanding, asynchronous cross-segment reasoning, audio-anchored retrieval, and cross-modal consistency verification.

Source	Sync.	Async.	Retr.	Verif.	Overall
Ego-Exo4D	67/48/43	60/54/49	53/51/43	53/47/46	59/50/45
STARSS	65/48/52	68/62/44	53/42/36	62/46/47	62/49/45
Ego4D	64/48/48	63/57/52	59/49/45	60/52/65	62/51/51
EasyCom	63/60/60	83/73/71	70/80/78	63/63/52	70/67/65
AMI	67/56/48	66/55/51	54/43/49	60/52/49	64/54/49
URMP	61/66/65	62/57/41	46/64/63	42/34/51	53/56/56

Table 9: **Accuracy (%) by total supplied media duration.** Results are descriptive because duration groups differ in source and task composition.

Duration	Gemini	Doubao	Qwen	Avg-17
$\leq$ 5 min	57.7	50.6	49.1	47.9
5–10 min	62.0	55.4	51.0	50.6
10–15 min	63.7	56.5	54.8	50.8
15–20 min	74.3	52.6	55.9	50.6
$>$ 20 min	69.5	50.5	42.5	46.4

predefined overflow policy reduces visual sampling rather than changing the question or answer options. External agents retain their original orchestration and media-inspection procedures, so their results are interpreted as system-level comparisons rather than controlled backbone comparisons.

D.3 Source-Wise Performance

Table 8 reports performance across the six source datasets for three representative end-to-end models. The analysis complements the aggregate leaderboard by showing variation across recording environments, sensor configurations, and acoustic structures.

Performance varies substantially across source datasets and model families. For example, all three representative models perform relatively strongly on EasyCom, while their relative behavior differs more substantially on Ego-Exo4D, STARSS, and URMP. These differences indicate that the aggregate benchmark score combines multiple forms of cross-stream evidence rather than reflecting performance on a single recording configuration.

D.4 Total Supplied Media Duration

Each individual stream in MS-OmniBench lasts at most five minutes, but a question may contain several streams. The total supplied media duration is therefore defined as the sum of the durations of all scene streams provided with a question. For synchronized streams, this quantity can substantially exceed the elapsed duration of the underlying scene.

Accuracy does not decrease monotonically as total supplied media duration increases. The seventeen-model mean remains similar from 5 to 20 minutes and decreases for the longest-duration group. Because the duration bins also differ in source, task, and stream composition, these results should not be interpreted as a causal estimate of context-length difficulty. They instead show that the benchmark includes substantial total multi-stream context even though each individual stream is bounded to five minutes.

D.5 Authoring-Origin Analysis

Because Gemini 3 Flash is used during the model-assisted construction pipeline, we additionally examine whether model-assisted questions exhibit a systematic advantage for Gemini-family models relative to human-authored questions. For this analysis, the released benchmark is partitioned into 1,999 model-assisted (MA) questions and 1,110 human-authored (HA) questions. The two groups form a disjoint partition of all 3,109 released questions. Table 10 reports this comparison for the eight proprietary models in the main evaluation.

Table 10: **Accuracy (%) by question origin for proprietary models.** MA denotes the 1,999 model-assisted questions, while HA denotes the 1,110 human-authored questions. The two groups form a disjoint partition of all 3,109 released questions.

Model	MA n=1999	HA n=1110	$\Delta_{\text{HA}-\text{MA}}$	Overall n=3109
Gemini 3.8 Flash	59.2	66.9	+7.7	61.9
Gemini 3.1 Pro	58.9	65.3	+6.4	61.2
Gemini 3.5 Flash	58.2	66.1	+7.9	61.0
Qwen3.5-Omni-Plus	59.3	60.9	+1.6	59.9
Gemini 3 Flash	59.7	59.7	+0.0	59.7
Doubao-Seed-2.0-lite	58.6	59.1	+0.5	58.8
Qwen3.5-Omni-Flash	56.3	54.6	-1.7	55.7
Doubao-Seed-2.0-mini	53.9	53.2	-0.7	53.6

Table 11: **Accuracy (%) by source dataset and supplied stream count.** Entries separated by slashes correspond to the stream-count groups shown in the second column. Only groups containing at least 15 questions are reported. Mean-17 denotes the unweighted mean over all seventeen end-to-end models.

Source	Streams	n	Gemini 3.8	Doubao mini	Qwen3-Omni	Mean-17
AMI	2 / 3 / 5	212 / 45 / 157	59.9 / 55.6 / 72.0	55.2 / 51.1 / 52.2	51.9 / 48.9 / 45.9	48.8 / 45.9 / 48.3
EasyCom	2 / 3 / 4	243 / 179 / 39	65.4 / 72.1 / 84.6	63.8 / 70.4 / 76.9	60.1 / 70.4 / 74.4	55.8 / 62.7 / 67.7
Ego4D	2	552	62.0	51.4	51.4	48.6
Ego-Exo4D	2 / 3 / 5 / 6	382 / 92 / 165 / 90	58.1 / 53.3 / 64.2 / 57.8	54.2 / 41.3 / 44.8 / 48.9	51.6 / 38.0 / 36.4 / 40.0	51.1 / 43.8 / 41.0 / 42.8
STARSS	2 / 3 / 4	400 / 76 / 230	60.2 / 55.3 / 66.5	51.0 / 44.7 / 47.0	49.0 / 50.0 / 37.4	49.5 / 45.8 / 43.5
URMP	2 / 3 / 4 / 5 / 6	109 / 45 / 33 / 36 / 16	47.7 / 57.8 / 60.6 / 52.8 / 62.5	55.0 / 51.1 / 75.8 / 52.8 / 50.0	58.7 / 48.9 / 60.6 / 52.8 / 50.0	50.4 / 49.0 / 58.5 / 44.0 / 39.7

The results do not show a systematic advantage for Gemini-family models on the model-assisted subset. Gemini 3.8 Flash, Gemini 3.1 Pro, and Gemini 3.5 Flash perform 7.7, 6.4, and 7.9 percentage points better on human-authored questions than on model-assisted questions, respectively. Gemini 3 Flash performs nearly identically across the two subsets. The remaining proprietary models show smaller differences in both directions, with $\Delta_{\text{HA}-\text{MA}}$ ranging from -1.7 to $+1.6$ points. Averaged across the eight proprietary models, accuracy is 58.0% on model-assisted questions and 60.7% on human-authored questions.

These results do not support the interpretation that Gemini-family performance is primarily explained by an advantage on questions produced by the Gemini-assisted construction pipeline. We do not interpret the raw difference between the two subsets as a direct measure of intrinsic difficulty because their task and input-format compositions are not identical. The item-weighted combination of the two subsets reproduces the Overall results in Table 1, providing a consistency check between this analysis and the main leaderboard.

D.6 Cross-Tabulated Analysis

The aggregate analyses in Table 2 mix questions from different sources, stream structures, and authoring origins. We therefore provide two cross-tabulated analyses to examine the interactions underlying the aggregate trends discussed in Section 5.3. These comparisons are descriptive because the groups differ in question composition and should not be interpreted as controlled interventions in which only stream count or evidence-chain length changes.

Table 11 shows that increasing stream count does not have a uniform relationship with accuracy across sources. EasyCom exhibits a monotonic increase for all three representative models and for Mean-17, which rises from 55.8% to 62.7% to 67.7% across the two-, three-, and four-stream groups. The corresponding group sizes are 243, 179, and 39, so the four-stream result should be interpreted cautiously because of its smaller sample size. This pattern differs from STARSS and Ego-Exo4D, where additional supplied viewpoints do not consistently improve performance and the Mean-17 accuracy decreases from 49.5% to 43.5% and from 51.1% to 42.8% between the lowest and highest reported stream-count groups, respectively. Together with the source structures in Table 3, these results suggest that the information contributed by additional streams is more informative than stream count alone when characterizing difficulty.

Table 12 further shows that the association between evidence-chain length and accuracy differs by question origin. For model-assisted questions, Mean-17 is 50.2% for two-or-fewer hops, 51.0%

Table 12: **Accuracy (%) by question origin and annotated evidence-chain length.** MA and HA denote the 1,999 model-assisted and 1,110 human-authored questions, respectively. The HA group includes the human-authored controlled variants. Mean-17 denotes the unweighted mean over all seventeen end-to-end models.

Origin	Hops	n	Gemini 3.8	Doubao mini	Qwen3-Omni	Mean-17
MA	≤ 2	1,364	60.0	52.9	52.0	50.2
MA	3	448	58.9	58.9	53.6	51.0
MA	≥ 4	187	54.0	48.7	41.7	43.2
HA	≤ 2	200	51.0	37.5	46.0	36.1
HA	3	630	71.3	56.7	48.1	52.1
HA	≥ 4	280	68.6	56.4	53.9	52.1

for three hops, and 43.2% for four or more hops. The decrease from three to four-or-more hops is 7.8 percentage points, with the corresponding group size decreasing from 448 to 187 questions. For human-authored questions, the two-or-fewer-hop group is substantially more difficult at 36.1%, whereas both the three-hop and four-or-more-hop groups achieve 52.1%. The human-authored subset also constitutes 280 of the 467 questions in the four-or-more-hop group. Consequently, the nearly flat aggregate relationship between hop count and accuracy in Table 2 partly reflects differences in question-origin composition rather than a uniform effect of evidence-chain length. These results reinforce that annotated chain length should be interpreted jointly with authoring origin and task composition.

E MESA Implementation and Analysis

This section provides additional implementation details for MESA. MESA separates reusable case-level evidence indexing from question-time targeted media inspection. The same Omni-LLM backbone is used for timeline construction and subsequent media-perception calls. The agent does not access benchmark gold answers, annotated evidence chains, or task labels.

E.1 Algorithm and Tool Interface

Before question answering, MESA constructs or loads a model-generated event timeline for each supplied stream. Each timeline contains timestamped visual events, speech, non-speech sounds, music, and within-stream audio-visual associations. The timelines are cached for reuse across questions from the same case and are treated as retrieval aids rather than ground truth.

At question time, MESA uses a shared global budget of at most 12 evidence-search media calls. The workflow contains four reasoning stages: Plan, Localize, Discriminate, and Decide. For reference-media questions, an optional reference-characterization call may be performed between planning and scene localization, and this call counts toward the same global media budget.

Plan. The agent queries cached stream summaries and event timelines to identify candidate anchor events, temporal windows, and observations that could support or falsify the leading answer options. This stage does not directly inspect scene media.

Localize. Candidate anchors are verified against the original media using targeted perception calls. When the timelines do not provide a reliable candidate, MESA performs a coarse-to-fine temporal search rather than scanning the full stream sequentially. The localization stage uses up to four media calls, subject to the remaining global budget.

Discriminate. The agent probes the strongest competing options using the localized evidence. Each observation is recorded as supporting, opposing, or unresolved in an option-by-evidence matrix. The discrimination stage uses up to six media calls, subject to the remaining global budget.

Decide. MESA re-inspects the most informative evidence windows and selects the final answer from the accumulated observations. The decision stage uses up to two media calls, subject to the remaining global budget. Because all stages share the same 12-call cap, the individual stage ceilings are not additive and need not all be reached on a given item. Table 13 summarizes the principal tools used during evidence search.

Table 13: **Core MESA tools used during evidence search.**

Tool	Purpose
`summary`	Read a compact summary of one stream.
`timeline_query`	Retrieve candidate events and temporal windows from the model-generated timeline.
`view_clip`	Inspect visual evidence in a local temporal window.
`listen_clip`	Inspect acoustic evidence in a local temporal window.
`compare_streams`	Jointly inspect aligned evidence across multiple streams.
`compare_windows`	Compare candidate windows across streams or time.
`ref_locate`	Match supplied reference media against candidate scene windows.

E.2 MESA Evidence-Search Prompt

MESA uses one frozen system prompt together with stage-specific instructions. The prompt explicitly warns that both timeline entries and media-tool outputs may be incomplete or incorrect. The planner is therefore instructed to verify important claims against the original media and to test competing hypotheses rather than simply confirm an initial prediction. All evidence-search media calls, including optional reference characterization, share the same global budget of at most 12 calls. The auxiliary end-to-end prediction is outside this search budget.

Prompt Template for MESA Evidence Search

System:
You are a verification agent answering a multiple-choice question over several audio-visual streams.
You have access to model-generated per-stream event timelines and media tools for inspecting short temporal windows.
Timeline entries and media-tool outputs may be incomplete or incorrect.
Use the timelines to locate candidate evidence and verify important claims against the original media.
Stage A: Plan
Identify the anchor events required by the question.
Query cached stream summaries and timelines for candidate temporal windows.
Specify observations that could support or falsify the leading answer options.
Do not directly inspect scene media during this stage.
If reference media are provided, they may be characterized before localization, with the call counted toward the shared global media budget.
Stage B: Localize
Verify candidate anchor events using targeted media inspection.
Use at most four media calls, subject to the remaining global budget.
When no reliable candidate is available, perform a coarse-to-fine temporal search rather than scanning the stream sequentially.
Localize the required event before testing the answer options.
Stage C: Discriminate
Use the localized evidence to test the strongest competing options.
Use at most six media calls, subject to the remaining global budget.
Prefer open-ended probes and attempt to falsify each candidate.
Record each observation as supporting, opposing, or unresolved.
Select the decisive temporal windows required for the final decision.
Stage D: Decide
Re-inspect the decisive evidence using at most two media calls, subject to the remaining global budget.
Select the answer best supported by the accumulated evidence.

For identity and correspondence questions, MESA can compare relevant windows jointly rather than treating observations independently. For reference-media questions, candidate matches are verified against the supplied reference and the original scene media. The direct end-to-end prediction from the same backbone is retained only as an auxiliary signal during the final decision.

E.3 Representative MESA Trace

Figure 5 visualizes a complete MESA run on a reference-audio speaker-binding question. The direct end-to-end prediction identifies the correct speaker attribute but selects the wrong hand configura-

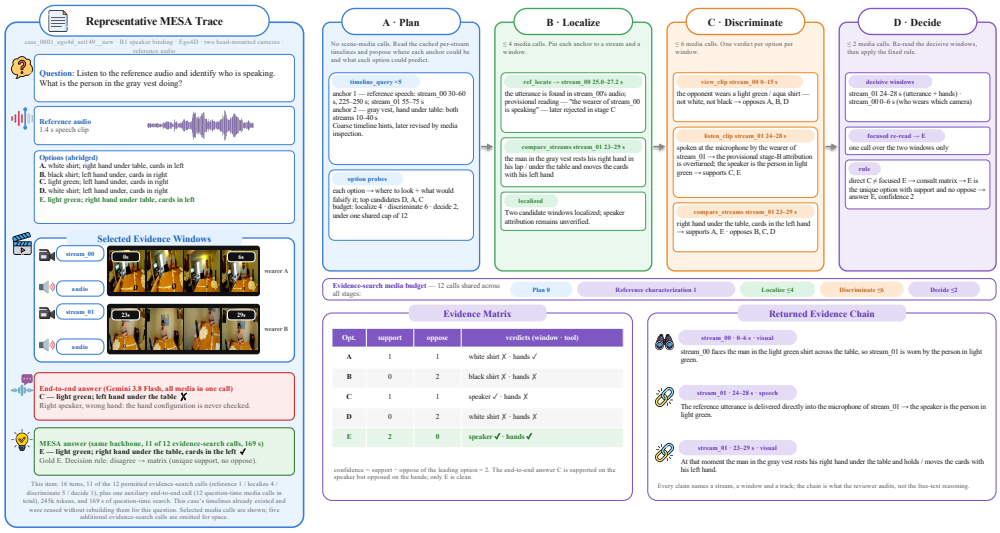


Figure 5: **Representative MESA evidence-search trace.** Planning operates over cached per-stream timelines without directly inspecting scene media. Media seconds include inspected evidence windows but exclude the auxiliary full-input call.

Table 14: **MESA question-time computational cost.** Media-call counts include optional reference characterization when used. Media seconds sum all temporal windows inspected during question-time evidence search. Reusable case-level timeline construction is excluded from these question-time averages.

Backbone	Media calls/item	Media seconds/item	Tokens/item
Gemini 3.8 Flash	9.1	170	144k
Qwen3-Omni-30B	10.1	180	162k

tion. MESA first uses the cached timelines to propose candidate windows and then verifies the relevant anchors against the original media. An initial speaker attribution produced during localization is explicitly treated as provisional and is subsequently overturned when microphone evidence is checked during discrimination. The agent then evaluates the competing answer options using an evidence matrix and re-inspects the decisive windows before selecting the final answer. In this example, MESA uses 11 of the permitted 12 evidence-search calls, including one reference-characterization call. The returned evidence chain records the stream, temporal window, modality, and claim associated with each decisive observation.

E.4 Computational Cost

Timeline construction is a reusable case-level preprocessing step, while question-time computation is dominated by targeted media inspection. Cached timelines are shared across questions from the same case and are therefore not rebuilt for every question. Table 14 reports the average question-time computational cost. The reported media-call counts include optional reference characterization when it is used. A comparison call that examines multiple windows contributes all inspected windows to the reported media duration. MESA introduces additional inference cost relative to direct end-to-end answering because it performs iterative evidence search and verification. This additional computation should therefore be considered together with the corresponding accuracy gains. The reported question-time costs are separated from reusable timeline construction so that the two sources of computation are not conflated. These figures describe question-time inference and exclude reusable case-level timeline construction. The reported accuracy gains characterize the complete MESA configuration rather than the isolated effect of timeline construction. The auxiliary end-to-end prediction is included in the reported question-time media-call and token totals.



Figure 6: Representative MS-OmniBench questions from the four task families. The displayed time intervals are answer-relevant excerpts; evaluated models receive the full specified scene streams.

F Representative Benchmark Cases and Release Contents

F.1 Representative Benchmark Cases

Figure 6 presents representative examples from the four MS-OmniBench task families. Each example shows the supplied scene streams, the relevant acoustic and visual observations, a compact evidence chain, and the verified answer. Together, these examples illustrate how answer-relevant evidence can be distributed across synchronized viewpoints, temporally separated observations, acoustic anchors, and cross-modal consistency relations.

The four examples emphasize different forms of evidence complementarity. The synchronous cross-stream example links sound, action, and participant identity across synchronized views. The asynchronous cross-segment example maintains an acoustically grounded identity across temporally separated observations. The audio-anchored retrieval example uses sound to identify the relevant moment before retrieving complementary visual evidence. The consistency-verification example reconciles acoustic and visual event counts rather than relying on either modality alone.

F.2 Release Contents

The MS-OmniBench release contains the 3,109-question benchmark manifest, question and option text, gold answers, source and case identifiers, supplied stream sets, reference-media specifications, evidence annotations, and construction metadata. We additionally release the construction prompts, reconstruction scripts, evaluation code, and prediction records used to generate the reported results. The release also includes the supporting stream identifiers and temporal evidence windows required to audit benchmark questions and reproduce the reported analyses. Original source video and audio are not redistributed. They must be obtained from the corresponding source datasets under their original licenses or access agreements.