EVAOMNI: AN EVIDENCE-GROUNDED BENCHMARK FOR AUDIO-VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-visual reasoning requires models to locate key audio and visual evidence and integrate information across modalities. However, existing benchmarks that rely primarily on multiple-choice questions typically score models solely on answer correctness, making it difficult to assess whether they can identify the evidence supporting their answers. Moreover, textual or unimodal shortcuts may allow models to answer correctly without analyzing the necessary audio-visual evidence, further weakening these benchmarks' ability to assess audio-visual reasoning. To address these issues, we introduce **EvaOmni**, an evidence-grounded benchmark comprising 420 videos and 1,200 question-answer pairs across 15 capabilities and three levels: retrieval and grounding, basic understanding, and complex reasoning. Experts annotate audio and visual evidence and construct video-grounded distractors; text-only, audio-only, and visual-only checks filter questions that can be answered without both modalities. Models must return both answers and supporting evidence intervals. We propose the Coverage Grounded Score (C.G.) to measure how much of the annotated evidence a model identifies, and the Bottleneck Grounded Score (B.G.) to penalize poor temporal grounding in either modality; both incorporate answer correctness. Across 7 open-weight and 15 proprietary Omni LLMs, the strongest model achieves 80.0% answer accuracy but only 35.8% B.G., compared with 94.3% for human experts. At Level 3, its 79.1% accuracy contrasts with 12.4% B.G. These results show why audio-visual reasoning should be evaluated through both answers and the supporting evidence. We will publicly release EvaOmni to support the continued evolution of evaluation systems and further advance model development.

1 INTRODUCTION

Omni-modal large language models (Omni LLMs) (Xu et al., 2025a;b; Team, 2026; Deshmukh et al., 2026) can jointly process text, audio, images, and video, integrating cross-modal cues to interpret multimodal inputs and reason over the evidence they provide. Audio-visual reasoning is particularly useful for assessing these capabilities because it requires models to connect auditory and visual signals over time. This involves identifying relevant evidence in each modality and integrating cues from different moments, abilities that support applications such as sports event understanding and the analysis of films and television programs. Evaluating the capabilities of existing Omni LLMs provides valuable guidance for model development and selection.

Existing benchmarks for joint audio-visual reasoning (Zhou et al., 2025b; Hong et al., 2026; Li et al., 2026; Chao et al., 2026; Tao et al., 2026; Xie et al., 2026) cover everyday activities, real-world scenarios, and diverse tasks, providing valuable ways to assess Omni LLMs. However, our analysis reveals a key limitation of multiple-choice question (MCQ) evaluations: **a correct answer does not necessarily demonstrate joint use of audio and visual evidence**. As illustrated in Fig. 1, some questions can be answered using a single modality or even textual cues alone. This ambiguity has two sources. First, answer-only scoring cannot distinguish evidence-supported answers from lucky guesses. Second, question and answer-option design can introduce unimodal or logical shortcuts. For example, the correct option may be inferred from common knowledge, audio-only input, or vision-only input. Additional examples from existing benchmarks are provided in the Appendix.

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.

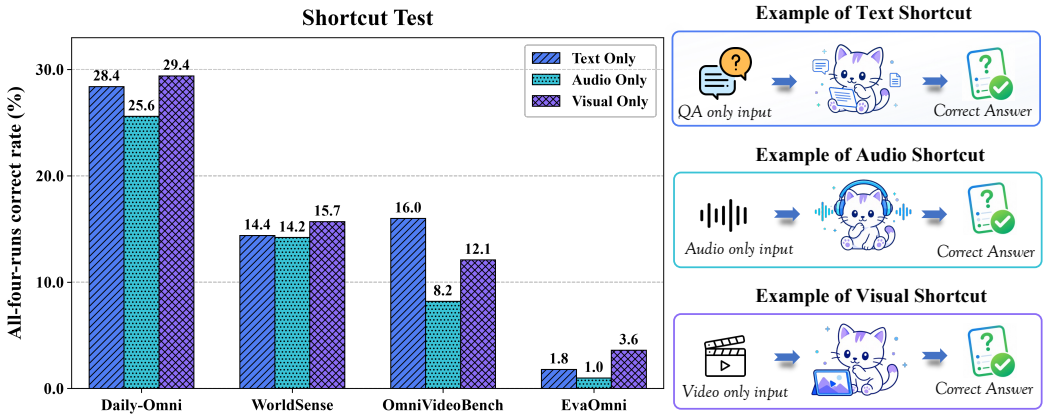


Figure 1: EvaOmni reduces shortcuts through a rigorous annotation pipeline. Its evidence-grounded evaluation limits the credit awarded to correct answers that lack supporting evidence.

Together, these limitations motivate an evaluation protocol that requires models to identify the audio and visual evidence supporting each answer, enabling a more verifiable evaluation.

Motivated by these observations, we introduce **EvaOmni**, an **ev**idence-grounded benchmark for **a**udio-visual reasoning. It comprises **420 videos and 1,200 question-answer pairs** spanning **15 fine-grained capabilities** at three levels of reasoning and evidence complexity: **Level 1-Retrieval & Grounding** requires locating relevant audio and visual evidence; **Level 2-Basic Understanding** requires interpreting and combining complementary cues; and **Level 3-Complex Reasoning** requires integrating temporally dispersed evidence or performing multi-step cross-modal reasoning. Experts annotate the audio and visual intervals supporting each answer, and models must return both an answer and its supporting intervals. Systematic video selection, expert annotation, cross-checking, and model-assisted filtering help reduce logical and unimodal shortcuts. Beyond answer accuracy (Acc.), we report the Coverage Grounded Score (C.G.) to measure coverage of annotated evidence and the Bottleneck Grounded Score (B.G.) to penalize poor temporal localization in either modality. Both incorporate answer correctness; more grounded metrics are reported in the Appendix.

We evaluate 7 open-weight and 15 proprietary Omni LLMs on EvaOmni. Performance varies substantially within both groups. Although the strongest proprietary model outperforms the strongest open-weight model, some open-weight models outperform several proprietary ones. More notably, both groups show substantial gaps between answer accuracy and evidence-grounded scores. For example, the strongest model overall, Gemini3.7-Flash, achieves 80.0% accuracy and 71.7% C.G., but only 35.8% B.G.; the corresponding human expert scores are 96.8%, 96.0%, and 94.3%. These findings motivate models that both answer correctly and precisely locate supporting evidence, allowing their conclusions to be verified against the audio-visual input. We will publicly release EvaOmni, including its data, annotations, and evaluation framework, to support **reliable, interpretable, and hierarchical** evaluation of audio-visual reasoning.

2 Related Work

Omni LLMs. Omni LLMs process text, images, video, and audio, enabling cross-modal understanding and reasoning (Fu et al., 2024; 2026; Liu et al., 2025). ImageBind (Girdhar et al., 2023) learned shared representations across text, audio, depth, and images; PandaGPT (Su et al., 2023) connected such representations to an LLM; and NExT-GPT (Wu et al., 2023) extended omni-modal systems to image, audio, and video generation. GPT-4o (Hurst et al., 2024) introduced streaming speech generation for natural interaction. The Gemini (Team et al., 2023; 2024; Comanici et al., 2025) and Qwen (Xu et al., 2025a;b; Team, 2026) families are leading proprietary and open-weight Omni LLMs, respectively, advancing cross-modal understanding and reasoning. SeedRealtime (ByteDance Seed, 2026) further integrates audio, video, and text in a full-duplex architecture for continuous scene understanding, interaction-timing awareness, and responsive behavior. Together, these systems mark a shift toward joint reasoning and natural interaction.

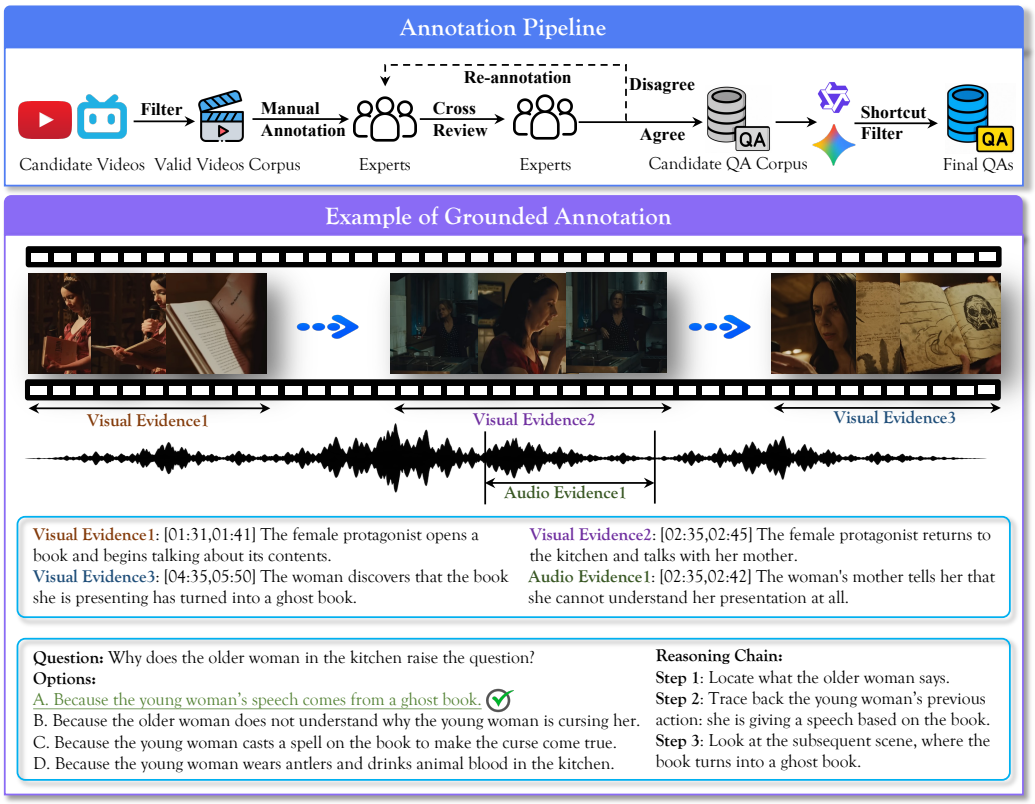


Figure 2: EvaOmni annotation pipeline and an example of evidence-grounded annotation.

Benchmarks for Omni LLMs. Unlike benchmarks of individual image, audio, or video understanding (Sakshi et al., 2025; Wang et al., 2026; Chen et al., 2024; Yue et al., 2024; Wang et al., 2025; Zhou et al., 2025a), Omni LLM benchmarks assess joint cross-modal reasoning. DailyOmni (Zhou et al., 2025b) covers everyday scenarios; WorldSense (Hong et al., 2026) emphasizes audio-visual understanding in real-world videos; and OmniVideoBench (Li et al., 2026) requires reasoning over both modalities. Yet textual and unimodal shortcuts can undermine answer-only scores, and existing benchmarks do not jointly score answer correctness and the localization of supporting audio-visual evidence. They also rarely organize capabilities by reasoning and evidence complexity.

3 EvaOmni: A New Evidence-Driven Paradigm for Omni-Modal Evaluation

3.1 Rethinking Audio-Visual Reasoning Benchmarks

We first examine how existing audio-visual reasoning benchmarks are constructed and evaluated. Table 1 compares seven representative benchmarks.

In the shortcut-checks column, T/A/V/I/Tr denote text-, audio-, video-, image-, and transcript-based tests; U denotes an unspecified unimodal test, D a dependency and logic check, and H human shortcut review. In the taxonomy column, cat., dims., diff., and caps. denote categories, dimensions, difficulty tiers, and capabilities. OmniBench uses static images, while SocialOmni focuses on social interaction rather than video question answering. Most existing benchmarks use multiple-choice questions (MCQs) and score responses by answer correctness. This raises **RQ1: Can answer-only MCQ accuracy establish whether a correct answer is grounded in both audio and visual evidence?** It cannot: the same credit is assigned to evidence-supported answers and unsupported correct choices. For four-option questions, uniform random guessing yields an expected accuracy of 25%, illustrating one way to obtain correct answers without using either modality. Answer accuracy also provides no measure of whether a model can locate the evidence supporting its prediction.

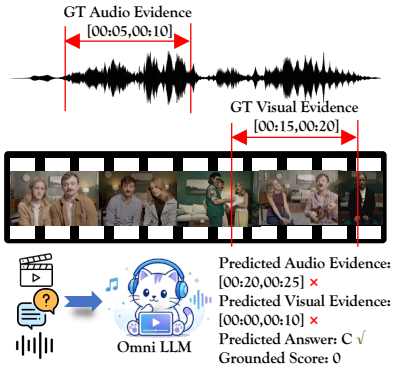


Figure 3: Grounded evaluation.

Table 1: **Comparison of omni-modal benchmark designs.** We compare shortcut checks, evidence annotation, grounded evaluation, and task taxonomies. Grounded Ann. requires question-level audio and visual evidence intervals; Grounded Eval. scores answer correctness and evidence grounding.

Benchmark	Reliability Control		Evaluation Design	
	Shortcut checks	Grounded Ann.	Grounded Eval.	Taxonomy
OmniBench	T/I/Tr/H	No	No	3 cat./8 tasks
Daily-Omni	T/H	No	No	6 tasks
WorldSense	V/H	No	No	3 levels/26 tasks
OmniVideoBench	T/U/H	No	No	13 tasks
JointAVBench	D/H	No	No	5 dims./15 tasks
LVOmniBench	T/A/V/H	No	No	3 diff./9 types
SocialOmni	–	No	No	3 tasks
EvaOmni (ours)	**T/A/V/H**	**Yes**	**Yes**	**3 levels/15 caps.**

We also ask **RQ2: To what extent can existing questions be answered through textual, logical, or unimodal shortcuts?** We evaluate three representative video question-answering benchmarks—Daily-Omni, WorldSense, and OmniVideoBench—under text-, audio-, and video-only inputs. Figure 1 illustrates shortcut cases; the full protocol and aggregate results appear in Appendix D.1. Some questions are answered correctly across repeated runs without full audio-visual input. This indicates potential shortcut susceptibility under our test conditions, although it does not establish how models arrive at their answers when given both modalities.

These observations motivate **RQ3: How can a benchmark limit shortcuts while evaluating answers and their supporting audio-visual evidence across levels of reasoning?** We address this question with EvaOmni, introduced next.

3.2 Introducing EvaOmni

To address RQ3, we introduce **EvaOmni**, an evidence-grounded benchmark for audio-visual reasoning. For each question, expert annotators identify the audio and visual intervals supporting the correct answer. At evaluation time, models must provide both an answer and timestamped evidence intervals for each modality. EvaOmni evaluates answer correctness together with evidence coverage and temporal grounding, thereby limiting the credit awarded to correct answers that lack matching evidence. Text-only, audio-only, and vision-only checks during question construction further reduce shortcut susceptibility, as described below. EvaOmni organizes 15 fine-grained capabilities into three levels of reasoning and evidence complexity:

Level 1: Retrieval and Grounding. Tests temporal alignment, event binidng, and entity binding, such as identifying the speaker of an utterance or linking a sound to its visible source.

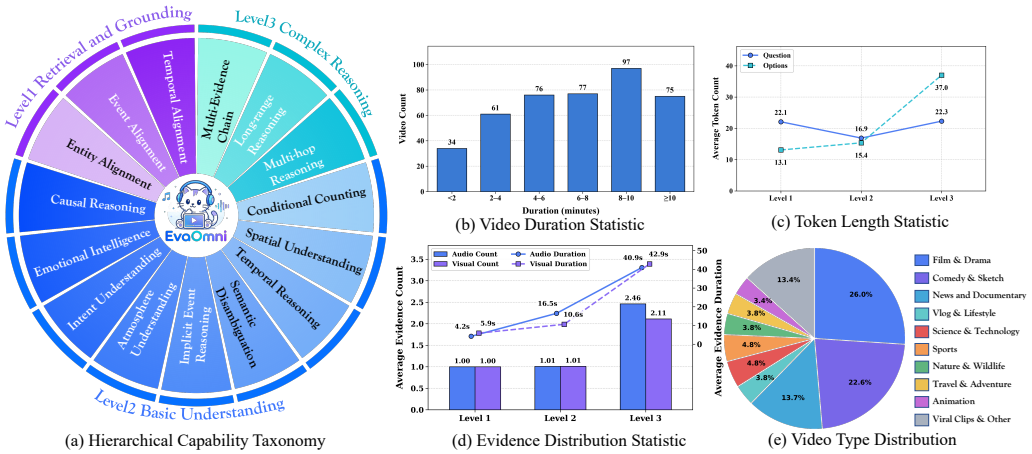
Level 2: Basic Understanding. Requires models to interpret and integrate complementary audio-visual cues to understand events, scene context, and their relationships. The nine fine-grained capabilities are detailed in Appendix A.5.

Level 3: Complex Reasoning. Tests long-range reasoning, multi-evidence chains, and multi-hop reasoning. These tasks require models to integrate evidence from different moments, perform multiple inference steps, or both.

The three levels contain three, nine, and three capabilities, respectively, as summarized in Fig. 4(a). The detailed capability distribution is provided in the Appendix.

3.3 Data Collection

To mitigate potential training data contamination, we collect recent videos from open platforms such as YouTube and Bilibili. The selected 1–12-minute videos span diverse real-world scenarios and contain informative visual events and audio cues. We exclude videos with hard-coded subtitles or prominent textual overlays to limit OCR-based shortcuts. Detailed selection and screening criteria are provided in Appendix A.



(a) Hierarchical Capability Taxonomy (b) Video Duration Statistic (c) Token Length Statistic (d) Evidence Distribution Statistic (e) Video Type Distribution

Figure 4: **Overview and statistics of EvaOmni.** (a) Hierarchical capability taxonomy. (b) Video duration statistics. (c) Mean question and answer-option lengths in tokens by level. (d) Mean counts and durations of annotated audio and visual evidence segments. (e) Video type distribution.

3.4 Annotation Pipeline

Expert annotators construct questions requiring joint use of audio and visual information and mark the temporal evidence for each modality. Each sample then undergoes human cross-validation for capability alignment, evidence quality, and potential textual or unimodal shortcuts. Disputed pairs receive independent review by two additional experts and are discarded unless both approve them. Model-assisted screening under text-only, audio-only, and vision-only inputs identifies additional shortcuts; pairs flagged at this stage are revised and re-evaluated. Appendix A provides the full annotation and quality-control procedure.

3.5 Evaluation Metrics

EvaOmni evaluates both answer correctness and supporting audio-visual evidence. For each question, the model returns an answer and audio and visual evidence intervals specified by start and end timestamps. We report two grounded scores that combine the answer score s ($s = 1$ for a correct option and $s = 0$ otherwise) with evidence localization quality:

Coverage Grounded Score (C.G.). This metric measures how much annotated audio and visual evidence is covered by the predictions. Let audio_I and visual_I be the numbers of human-annotated audio and visual evidence segments that have a positive-duration temporal overlap with at least one predicted segment in the same modality. Each annotated segment is counted at most once. Let $\mathrm{audio}_{\mathrm{all}}$ and $\mathrm{visual}_{\mathrm{all}}$ denote the total numbers of annotated segments in the respective modalities. We define

$$\mathrm{C.G.} = \frac{s}{2}\left(\frac{\mathrm{audio}_I}{\mathrm{audio}_{\mathrm{all}}} + \frac{\mathrm{visual}_I}{\mathrm{visual}_{\mathrm{all}}}\right). \tag{1}$$

C.G. averages the two modality-specific evidence hit rates and weights them by answer correctness. We interpret C.G. as a lenient evidence hit-rate metric, not a measure of precise temporal localization: any positive-duration overlap counts as a hit, and overly broad predictions are not directly penalized for imprecise interval boundaries.

Bottleneck Grounded Score (B.G.). Let $\mathrm{IoU}_{\mathrm{audio}}$ and $\mathrm{IoU}_{\mathrm{visual}}$ denote the temporal intersection over union between predicted and human-annotated evidence for the audio and visual modalities, respectively. For each modality, we first take the union of all predicted intervals and the union of all annotated intervals; their IoU is the duration of the intersection of these two sets divided by the duration of their union. We define

$$\mathrm{B.G.} = s \times \min(\mathrm{IoU}_{\mathrm{audio}}, \mathrm{IoU}_{\mathrm{visual}}). \tag{2}$$

Table 2: **Overall performance on EvaOmni.** We report Acc., C.G., and B.G. for each capability level and overall. All values are in %.

Model	Level1 R.&G.			Level2 B.U.			Level3 C.R.			Overall		
	Acc.	C.G.	B.G.	Acc.	C.G.	B.G.	Acc.	C.G.	B.G.	Acc.	C.G.	B.G.
Open-Weight Omni LLMs												
Qwen2.5-Omni-7B	16.6	1.2	0.6	17.4	0.5	0.0	13.7	0.3	0.0	16.4	0.8	0.3
Qwen3-Omni-Instruct-30B-A3B	28.6	5.1	0.6	22.4	3.4	0.4	22.2	2.4	0.1	25.2	4.0	0.4
Qwen3-Omni-Thinking-30B-A3B	26.3	11.4	4.8	24.4	6.2	1.2	21.6	4.1	0.6	24.8	8.3	2.8
Nemotron3-Nano-Omni-30B-A3B	30.7	15.0	5.3	22.1	13.1	3.9	10.5	4.7	0.7	24.2	12.6	4.0
MiMo-V2.5-310B-A15B	**65.4**	**61.2**	**30.7**	**51.5**	**41.2**	**14.5**	45.1	**29.7**	**6.7**	**56.8**	**48.5**	**20.7**
MiMo-V2.6-Flash-310B-A15B	61.2	55.2	29.0	47.1	39.4	13.9	45.8	28.7	5.9	53.3	44.9	19.5
MiMo-V2.6-Pro-1T-A42B	60.1	51.2	28.7	43.1	34.0	12.3	**49.7**	26.7	4.0	52.0	40.7	18.6
Proprietary Omni LLMs												
Doubao-Seed-2.1-Lite	54.5	36.4	10.7	51.5	31.0	7.8	62.8	32.2	5.9	54.5	33.7	8.9
Qwen3.5-Omni-Flash	47.3	41.8	15.5	33.3	26.1	8.3	29.4	15.8	2.7	39.1	31.7	10.7
Qwen3.5-Omni-Plus	58.0	52.1	22.2	42.9	24.6	7.6	42.5	19.9	3.3	49.7	36.4	13.6
Qwen3.8-Omni-Flash	64.9	53.7	19.4	51.3	39.6	11.1	56.2	36.2	7.9	58.3	45.5	14.4
Gemini2.5-Flash-Lite	52.9	47.1	22.1	32.8	23.8	7.8	27.5	14.7	2.2	41.2	32.9	13.5
Gemini2.5-Flash	59.1	55.1	26.8	41.2	33.0	12.2	41.8	24.7	4.0	49.5	41.8	17.5
Gemini2.5-Pro	76.4	73.0	**49.2**	64.7	54.3	**30.1**	72.6	46.1	9.5	71.3	62.0	29.4
Gemini3.1-Flash-Lite	59.4	52.5	24.3	38.4	32.5	10.9	32.0	19.1	3.0	47.0	39.5	15.8
Gemini3.1-Flash-Lite-Preview	59.6	53.6	25.1	41.7	33.2	11.0	38.6	14.7	3.3	49.4	41.0	16.2
Gemini3.1-Pro-Preview	74.6	72.8	41.4	67.2	58.3	23.8	72.1	48.7	11.2	72.1	63.4	29.8
Gemini3.5-Flash-Lite	60.5	55.5	26.5	39.5	29.1	11.4	42.5	24.3	3.5	49.6	40.5	17.0
Gemini3.5-Flash	64.0	59.4	28.7	68.6	59.9	26.9	77.8	48.1	11.3	68.0	57.7	25.0
Gemini3.6-Flash	80.8	77.9	44.7	68.6	61.6	26.7	73.2	44.4	10.7	75.0	66.3	32.4
Gemini3.7-Flash	**85.2**	**84.1**	49.0	**74.0**	**66.2**	29.9	**79.1**	**49.4**	**12.4**	**80.0**	**71.7**	**35.8**
Gemini3.8-Flash	83.4	82.0	48.4	**74.0**	63.3	27.7	74.5	43.2	9.5	78.4	68.6	34.2
Human Expert	97.8	97.8	95.8	96.5	94.8	93.2	94.4	93.8	93.0	96.8	96.0	94.3

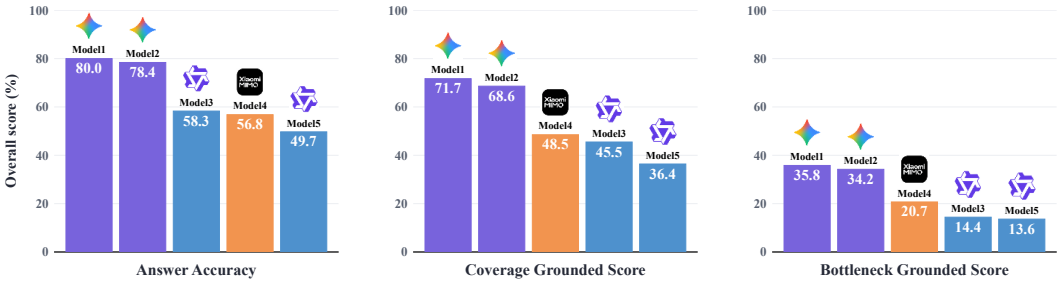


Figure 5: **Overall performance rankings for selected models.** We select the top two Gemini models, top two Qwen models, and top MiMo model by overall Acc. in Table 2, then rank the same five models independently by Acc., C.G., and B.G. Model1: Gemini3.7-Flash; Model2: Gemini3.8-Flash; Model3: Qwen3.8-Omni-Flash; Model4: MiMo-V2.5-310B-A15B; Model5: Qwen3.5-Omni-Plus. Model IDs remain fixed across panels, which share the same score scale.

Taking the minimum makes B.G. sensitive to the less accurately localized modality: strong localization in one modality cannot compensate for poor localization in the other. Thus, C.G. measures broad evidence coverage, whereas B.G. requires accurate localization in both modalities. Additional grounded metrics are provided in Appendix D.2.

4 Experiments

4.1 Evaluated Models

We evaluate 22 Omni LLMs on EvaOmni, comprising 7 open-weight and 15 proprietary models. The open-weight models include Qwen2.5-Omni-7B (Xu et al., 2025a), Qwen3-Omni-Instruct and Qwen3-Omni-Thinking (Xu et al., 2025b), Nemotron3-Nano-Omni (Deshmukh et al., 2026),

MiMo-V2.5-310B-A15B, MiMo-V2.6-Flash-310B-A15B, and MiMo-V2.6-Pro-1T-A42B. The proprietary models include Qwen3.5-Omni-Flash and Qwen3.5-Omni-Plus (Team, 2026), Qwen3.8-Omni-Flash (Qwen Team, 2026), Doubao-Seed-2.1-Lite (Volcano Engine, 2026), and 11 Gemini variants spanning the Pro, Flash, and Flash-Lite series (Comanici et al., 2025; Google Deepmind, 2026a;c;d;b). The complete model list is provided in Table 2. We also report human expert performance as a reference for comparing answer correctness and evidence grounding; its protocol is described in Appendix B.2.

4.2 Experimental Settings

We evaluate all 22 models on 1,200 question-answer pairs drawn from 420 videos, using the same prompt and preprocessed audio-visual content. Video frames are uniformly sampled at 1.0 FPS and resized so that their longer side is 480 pixels while preserving the aspect ratio. Open-weight models outside the MiMo family are evaluated on a single server equipped with eight NVIDIA H20 GPUs, using the inference frameworks and settings recommended in their model cards. The open-weight MiMo models and all proprietary models are evaluated through their providers' official APIs. The prompt, input preprocessing procedures, and further evaluation details are provided in the Appendix.

4.3 Main Findings

Table 2 reports performance across capability levels and overall, while Figure 6 compares two representative models with human performance. We summarize the key findings below.

Overall Performance Analysis. (1) *A substantial gap remains between the strongest evaluated open-weight and proprietary models.* MiMo-V2.5-310B-A15B achieves overall Acc., C.G., and B.G. scores of 56.8%, 48.5%, and 20.7%, respectively, whereas Gemini3.7-Flash reaches 80.0%, 71.7%, and 35.8%. The respective gaps of 23.2, 23.2, and 15.1 percentage points span answer correctness, evidence coverage, and precise dual-modal localization. (2) *The gap to human experts is especially pronounced in precise grounding.* Human experts achieve 96.8% Acc., 96.0% C.G., and 94.3% B.G., exceeding Gemini3.7-Flash by 16.8, 24.3, and 58.5 percentage points, respectively. The model's 80.0% accuracy and 71.7% C.G. indicate strong answering performance and substantial evidence coverage, yet its 35.8% B.G. reveals persistent difficulty in precisely localizing evidence in both modalities. Although grounded scores also incorporate answer correctness, the much smaller decline from Acc. to B.G. for humans (2.5 points, versus 44.2 for Gemini3.7-Flash) highlights the remaining grounding challenge. (3) *Grounding limitations are more severe among the four open-weight models with fewer than 100B total parameters.* Their overall C.G. ranges from 0.8% to 12.6% and B.G. from 0.3% to 4.0%, compared with 31.7%–71.7% and 8.9%–35.8% for the proprietary models. For example, Qwen2.5-Omni-7B achieves 16.4% accuracy but only 0.8% C.G. and 0.3% B.G. Jointly requiring correct answers and grounded evidence thus exposes limitations that answer-only evaluation can obscure. These models leave substantial room for improvement in consistently satisfying both dual-modal localization and answer correctness. In Figure 5, we compare the rankings of five models. Model rankings by grounding score broadly follow those by Acc., although a few shifts in position highlight differences in evidence localization among models with similar answer accuracy.

> ***Finding 1:*** *Grounded evaluation shows that even leading models trail humans in evidence localization, while open-weight models under 100B parameters lag behind proprietary models in evidence perception and localization.*

Analysis Across Capability Levels. Table 2 shows that answer accuracy decreases from Level 1 to Level 2 for 20 of the 22 models, but increases from Level 2 to Level 3 for 11 models. For example, Gemini3.7-Flash improves from 74.0% to 79.1% accuracy between Levels 2 and 3, while its C.G. drops from 66.2% to 49.4% and B.G. from 29.9% to 12.4%. This divergence is consistent with the design of Level 3, which emphasizes long-range reasoning, combinations of multiple evidence segments, and multi-hop inference. These tasks place greater demands on assembling and localizing supporting evidence, without necessarily producing a monotonic decline in multiple-choice accuracy. One plausible explanation is that partial evidence can help eliminate distractors and narrow the candidate set, allowing a model to select the correct answer without fully localizing the supporting evidence. The aggregate results are compatible with this explanation, but do not directly establish

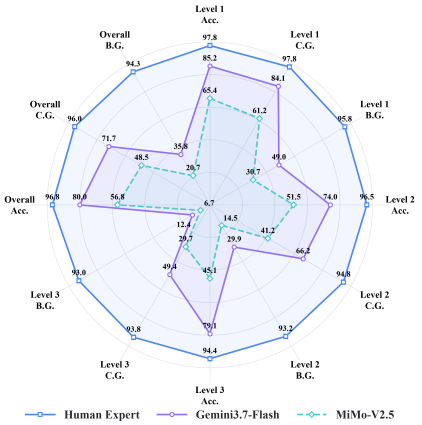


Figure 6: Overall performance.

Table 3: Cost efficiency performance. Cost efficiency is USD per expected correct answer (↓), using Overall Acc. in Table 2.

Model	Cost (USD)	Cost Efficiency
Qwen3.8-Omni-Flash	13.48	0.0193
Doubao-Seed-2.1-Lite	18.05	0.0276
Gemini3.1-Flash-Lite	21.82	0.0387
Gemini3.7-Flash	51.89	0.0541
Qwen3.5-Omni-Flash	48.96	0.1043
Gemini3.5-Flash	103.90	0.1273
Gemini2.5-Flash	79.09	0.1331
Gemini3.1-Pro-Preview	138.54	0.1601
Qwen3.5-Omni-Plus	152.55	0.2558
Gemini2.5-Pro	262.29	0.3061

the models' answer-selection strategies. In contrast, Level 3 yields the lowest B.G. for all 22 models, including one tie, and the lowest C.G. for 21 models. The C.G. exceptions is Doubao-Seed-2.1-Lite. This broad trend across open-weight and proprietary models highlights the value of grounded evaluation: C.G. rewards evidence coverage alongside correct answers, whereas B.G. rewards accurate temporal localization in both modalities alongside correct answers. Reporting Acc. together with these two metrics therefore provides a more complete assessment of answering, evidence coverage, and precise grounding across capability levels.

Finding 2: *As the task level increases, evidence-grounding quality declines, while answer accuracy does not follow the same trend.*

Cost-Efficiency Analysis. Table 3 compares the estimated evaluation costs of 10 selected proprietary models over 1,200 audio-video samples. We measure cost efficiency as USD per expected correct answer, using Overall Acc. in Table 2; lower values indicate better efficiency. The resulting ranking differs substantially from the accuracy ranking. Gemini3.7-Flash achieves the highest accuracy among these models (80.0%), but ranks fourth in cost efficiency at $0.0541 per correct answer. In contrast, Qwen3.8-Omni-Flash, despite its lower accuracy of 58.3%, ranks first at $0.0193 per correct answer, approximately 64.3% lower than Gemini3.7-Flash. Its total estimated cost is also lower ($13.48 versus $51.89). However, higher accuracy does not necessarily require higher spending: Gemini3.7-Flash outperforms Gemini2.5-Pro (71.4% accuracy), while reducing cost per correct answer from $0.3061 to $0.0541, an approximately 82.3% reduction. These comparisons suggest that model selection should account for both accuracy requirements and inference budgets. Qwen3.8-Omni-Flash is an economical option when its accuracy meets the application's requirements, whereas Gemini3.7-Flash offers higher accuracy at a lower cost than several less accurate alternatives. This cost metric captures answer correctness rather than evidence localization; applications requiring reliable grounding should additionally consider C.G. and B.G.

Finding 3: *Higher accuracy does not necessarily imply better cost efficiency. Model selection should balance task-specific performance requirements with budget constraints.*

4.4 Further Analysis

Input Modality Ablation. In Table 4 and the following analysis, we abbreviate MiMo-V2.5-310B-A15B, Qwen3VL-8B-Instruct, and Qwen3VL-Instruct-30B-A3B as MiMo-V2.5, Qwen3VL-8B, and Qwen3VL-A3B, respectively. Table 4 compares three models under different input conditions, yielding three observations. (1) *Native audio outperforms ASR transcripts in the evaluated MiMo-V2.5 settings.* With audio alone, MiMo-V2.5 achieves 30.0% Acc. and 10.0% C.G., compared with 23.8% and 7.1% using ASR alone, respectively. The advantage persists when video is available: Video+Audio achieves 56.8% Acc., 48.5% C.G., and 20.7% B.G., exceeding Video+ASR by 8.1, 10.4, and 9.1 percentage points. Thus, ASR transcripts do not fully substitute for audio in these

Table 4: **Input modality ablation.** Bold marks the best input setting for each model and metric.

Model	Input	Acc.	C.G.	B.G.
MiMo-V2.5	Audio	30.0	10.0	0.0
	ASR	23.8	7.1	0.0
	Video	26.8	9.3	0.0
	Video+ASR	48.7	38.1	11.6
	Video+Audio	**56.8**	**48.5**	**20.7**
Qwen3VL-8B	Video	13.8	0.3	0.0
	Video+ASR	**23.2**	**9.7**	**1.9**
Qwen3VL-A3B	Video	8.5	0.7	0.0
	Video+ASR	**19.3**	**7.0**	**0.8**

Table 5: Answer–grounding diagnostics. A/V: unweighted audio/visual IoU; C: correct answer; G: both IoUs ≥ 0.5. Here $\neg$ means "not" and a comma means "and".

Model	A	V	C, G	$C, \neg G$	$\neg C, G$	$\neg C, \neg G$
MiMo-V2.5	41.1	39.3	21.5	35.3	8.3	34.7
Qwen3.5-Omni-Plus	30.5	29.8	13.1	36.6	5.5	44.7
Qwen3.8-Omni-Flash	27.4	34.3	11.8	46.5	4.4	37.3
Gemini2.5-Flash	40.2	36.7	18.2	31.3	8.8	41.5
Gemini2.5-Pro	47.8	46.2	32.4	38.9	7.2	21.5
Gemini3.5-Flash	45.6	41.9	27.0	41.0	6.8	25.2
Gemini3.7-Flash	52.6	51.1	41.8	38.2	5.8	14.3

settings. Non-speech sounds, prosody, and transcription or timestamp errors may contribute to this gap. (2) *Removing either modality substantially degrades answering and grounding.* Relative to Video+Audio, MiMo-V2.5's accuracy drops by 26.8 percentage points with Audio alone and 30.0 points with Video alone. C.G. similarly falls from 48.5% to 10.0% and 9.3%, respectively, while B.G. drops from 20.7% to 0.0% in both cases. These declines highlight the value of complementary audio-visual evidence. (3) *ASR improves vision-language models, but substantial grounding limitations remain.* Adding ASR to video improves Qwen3VL-8B's accuracy from 13.8% to 23.2% and Qwen3VL-A3B's from 8.5% to 19.3%, gains of 9.4 and 10.8 percentage points. Their C.G. scores also increase from 0.3% to 9.7% and from 0.7% to 7.0%, respectively. Nevertheless, their B.G. scores remain at only 1.9% and 0.8%, compared with MiMo-V2.5's 11.6% under the same Video+ASR condition. Transcripts therefore provide useful additional evidence, but do not by themselves ensure effective evidence integration and precise localization.

Finding 4: *Both modalities improve answer accuracy. Native audio outperforms ASR, while ASR improves the evaluated VLMs but leaves their precise grounding performance low.*

Error Analysis. Table 5 decomposes answering and grounding outcomes for seven models; details are in Appendix D.3. Audio IoU exceeds visual IoU by 0.7–3.7 points for six of the seven models; Qwen3.8-Omni-Flash is the only exception (27.4% versus 34.3%). The two failure modes are asymmetric. Requiring both modality IoUs to reach 0.5, 47.8%–79.8% of correct answers fail the criterion $(C, \neg G)$: Gemini3.7-Flash answers 80.0% correctly but satisfies both criteria on only 41.8%, and Qwen3.8-Omni-Flash drops from 58.3% to 11.8%. Conversely, 12.2%–32.6% of grounding-passing predictions still choose the wrong option $(\neg C, G)$, and $\neg C, G$ never exceeds 8.8% of all questions, so accurate dual-modal localization is rarely reached where answers are wrong. Therefore, neither answer correctness nor temporal overlap alone is sufficient to establish reliable evidence-grounded answering.

Finding 5: *Answer correctness does not guarantee accurate dual-modal grounding, and grounding success does not guarantee a correct answer.*

5 Conclusion

We introduce EvaOmni, a new evidence-grounded benchmark for audio-visual reasoning. Rather than evaluating only whether a model selects the correct answer, EvaOmni also requires the model to identify the supporting evidence necessary to justify its prediction. Through a series of rigorous annotation and quality-control procedures, EvaOmni comprises 1,200 questions organized into three hierarchical levels, providing diverse scenarios and rich data for comprehensive evaluation. We further introduce multiple Grounded Score metrics to evaluate the accuracy of evidence localization. Among them, the Coverage Grounded Score rewards evidence coverage, while the Bottleneck Grounded Score prevents strong localization in one modality from masking failure in the other. Additional Grounded Score variants are provided in the Appendix. Large-scale evaluation shows that combining grounded evaluation with accuracy reveals insights missed by accuracy alone. EvaOmni provides a reproducible, interpretable basis for measuring progress and identifies precise temporal grounding as the capability current Omni LLMs most clearly lack. We will release the data, annotations, and evaluation framework in full to further advance model development and evaluation.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and translation. We have not used generative AI tools to generate synthetic datasets, and providing mathematical proofs is not applicable to this work. Additionally, we used generative AI tools to create and modify figures and tables. We have reviewed all AI-assisted work. Through a manual literature review, we verified that the references identified by large language models correspond to existing publications. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our research does not raise any concerns under the Code of Ethics. EvaOmni is built from 420 publicly accessible videos collected from open platforms such as YouTube and Bilibili. We do not claim ownership of the source media and do not redistribute it: our release consists of video identifiers, time spans, questions, answer options, expert-annotated evidence intervals, and evaluation code, so that copyright remains with the original uploaders and takedown requests can be honored by removing the corresponding entries.

REPRODUCIBILITY STATEMENT

Our aim is to make every numerical result reported in this paper reproducible. We will release the full EvaOmni specification, including video identifiers and time spans, 1,200 question-answer pairs with capability categories and level labels, and the corresponding audio and visual evidence intervals. We will also release scripts for downloading and preprocessing the videos, scoring model outputs, and reproducing the tables and figures, including the response-parsing rules used for malformed outputs. Appendix A describes video selection, the annotation protocol, human cross-validation, and model-assisted shortcut screening. Appendix D.3 explains the exclusion of malformed reference intervals from the diagnostic analysis and how failed or malformed model outputs are scored. The shared prompt and required JSON output format appear in Appendix C; video sampling at 1.0 FPS and resizing the longer side to 480 pixels are specified in Appendix B.1. The two main Grounded Score metrics are defined in Section 3.5, and additional metrics and results across IoU thresholds are reported in Appendix D.2.

REFERENCES

ByteDance Seed. Seedrealtime audio-visual full-duplex llm released: Toward omni-modal natural interaction. `https://seed.bytedance.com/en/blog/seedrealtime-audio-visual-full-duplex-llm-released-toward-omni-modal-natural-interaction`, August 2026.

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.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

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

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.

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

Chaoyou Fu, 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.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

Google Deepmind. Gemini 3.1 Pro . https://deepmind.google/models/gemini/pro/, 2026a.

Google Deepmind. Gemini 3.6 Flash . https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-6-flash-3-5-flash-lite-3-5-flash-cyber/, 2026b.

Google Deepmind. Introducing Gemini 3.7 Flash. https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-3-7-flash/, 2026c.

Google Deepmind. Introducing Gemini 3.8 Flash and 3.8 Flash Cyber. https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/, 2026d.

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.

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.

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

Qwen Team. Qwen3.8-omni-flash: Omni senses. agentic delivery., September 2026. URL https://qwen.ai/blog?id=qwen3.8-omni-flash.

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.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the Era of Interactive Assistants!*, pp. 11–23, 2023.

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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

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

Volcano Engine. Coding plan personal model release record, September 2026. URL https://docs.volcengine.com/docs/ark/coding-plan-personal-model-release?lang=zh.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2026, pp. 31374–31410, 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.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

Tianyu Xie, Jinfa Huang, Yuexiao Ma, Rongfang Luo, Yan Yang, Qingchuan Ma, Wang Chen, Yuhui Zeng, Yixuan Zou, Zhiqiang Lu, et al. Socialomni: Benchmarking audio-visual social interactivity in omni models. *arXiv preprint arXiv:2603.16859*, 2026.

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

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*, 2025b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025a.

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*, 2025b.

APPENDIX CONTENTS

A Dataset Construction and Annotation Details **13**

A.1 Video Collection and Filtering 13

A.2 Annotation Principles and Schema 14

A.3 Human Cross-Validation 14

A.4 Model-Assisted Shortcut Screening 14

A.5 EvaOmni Structure 15

B Experimental Details **15**

B.1 Input Preprocessing and Inference Setup 15

B.2 Human Expert Baseline 16

C Prompt Templates **16**

D Additional Quantitative Results and Analysis **19**

D.1 Shortcut Evaluation and Case Studies 19

D.2 Additional Grounded Evaluation Results 29

D.3 Details of Failure Mode Analysis 30

D.4 Analysis Across Fine-Grained Capabilities 30

E Limitations and Future Work **31**

E.1 Limitations 31

E.2 Future Work 31

A DATASET CONSTRUCTION AND ANNOTATION DETAILS

A.1 VIDEO COLLECTION AND FILTERING

We select candidate videos according to four criteria, aiming to reduce potential training data contamination and textual shortcuts while providing sufficient cues for constructing audio-visual reasoning questions.

Temporal freshness and potential contamination screening. We consider only videos uploaded after January 2025. As an additional screening step, we randomly sample three frames from each video and provide them to Gemini3.1-Pro-Preview, asking it to describe the video content. Videos for which the model can reconstruct the full content from these frames are flagged as potential contamination risks and excluded from the candidate pool. This procedure is a heuristic: neither the upload date nor the model's ability to reconstruct the content establishes whether a video was present in its training data.

Avoiding subtitle-based textual shortcuts. We exclude videos with hard-coded subtitles to prevent models from substituting on-screen text recognition for audio understanding. This criterion supports the intended requirement that questions depend jointly on audio and visual information.

Manageable video duration. We select videos lasting 1–12 minutes to balance temporal scope, model input budgets, and annotation feasibility. At a sampling rate of 1.0 FPS, this range corresponds to approximately 60–720 frames, helping control the visual input size while retaining tem-

porally structured event sequences. Extending the collection to longer videos is a direction for future work on audio-visual reasoning over longer time spans.

Sufficient diversity of entities and events. We prioritize videos with varied participants, actions, or event developments. Videos with highly repetitive content, such as a single person continuously performing one action, may provide too few cues for constructing plausible distractors. We manually exclude videos that do not support adequate distractor construction, reducing opportunities to answer questions through unimodal cues or simple elimination.

A.2 Annotation Principles and Schema

Five senior PhD students with research backgrounds in computer science and specialized knowledge of linguistics participate in annotation. Each sample records the video URL, question, answer options, capability level, and fine-grained capability category. Annotators separately mark the visual and audio evidence supporting the answer, with each evidence segment containing start and end timestamps in `mm:ss` format and a content description. For Level 3 multi-hop reasoning questions, the annotation additionally includes a step-by-step reasoning chain.

Question and distractor design. Each question must match its assigned capability category and level and be designed to require joint use of audio and visual information. Distractors draw on visual and audio cues present in the video so that a single modality is insufficient to reliably distinguish the correct option from the alternatives. They also represent logically plausible alternative explanations, limiting opportunities to eliminate incorrect options solely through wording or logical cues in the question and answer choices.

A.3 Human Cross-Validation

After initial annotation, each expert's question-answer pairs are randomly divided into four parts, which are randomly assigned, one to each of the other four experts, for cross-validation. Each expert therefore reviews approximately as many pairs as they initially annotated. The reviewer checks three dimensions:

1. **Capability alignment:** whether the question matches its assigned fine-grained capability and capability level.
2. **Shortcut susceptibility:** whether the question can be reliably answered using only the question and options, audio alone, or vision alone through logical or unimodal cues.
3. **Evidence quality:** whether the audio and visual evidence intervals and descriptions are accurate and support the annotated answer.

Approximately 78% of the initially submitted pairs pass this first review. The remaining disputed pairs are independently reviewed by two additional experts, neither of whom is the original annotator or the first reviewer. A pair is discarded if either additional expert identifies a problem; it passes only if both approve it. Overall, approximately 82% of the initially submitted pairs pass human cross-validation, while the remaining 18% are discarded at this stage.

A.4 Model-Assisted Shortcut Screening

Following human validation, we use Gemini3.1-Pro-Preview and Qwen3-Omni-Instruct-30B-A3B for additional shortcut screening under text-only, audio-only, and vision-only conditions. We inspect both the answers and their supporting rationales, as well as the evidence from the available modality in unimodal settings. If a model can reliably answer a question correctly and provide a plausible supporting rationale despite the absence of one or both modalities, the sample is returned for revision. Revised samples must pass model-assisted screening again before inclusion.

The resulting EvaOmni benchmark contains 1,200 question-answer pairs curated through this annotation and quality-control process.

A.5 EvaOmni Structure

EvaOmni contains 420 videos: 180 at Level 1, 150 at Level 2, and 90 at Level 3. These videos yield 540, 480, and 180 question-answer pairs at the respective levels, for a total of 1,200 pairs. The main text defines the three levels; here we define the fine-grained capabilities within each one.

Level 1 (540 question-answer pairs).

1. **Temporal Alignment (180):** Identify the main content of one modality using information from the other modality.
2. **Event Binding (180):** Summarize the main event in one modality using information from the other modality.
3. **Entity Binding (180):** Link the main content of one modality to the corresponding content in the other modality.

Level 2 (480 question-answer pairs).

1. **Causal Reasoning (65):** Infer the causes of events.
2. **Emotional Intelligence (60):** Interpret emotional states from audio-visual cues.
3. **Intent Understanding (60):** Infer a person's intent from audio-visual cues.
4. **Atmosphere Understanding (60):** Interpret the atmosphere of a scene from audio-visual cues.
5. **Implicit Event Reasoning (60):** Infer unstated events, such as an omitted or likely future event.
6. **Semantic Disambiguation (60):** Resolve the specific meaning of ambiguous speech or behavior using context from both modalities.
7. **Temporal Understanding (60):** Determine the order, overlap, or duration relationships between audio and visual events.
8. **Spatial Understanding (35):** Reason about spatial locations and relationships.
9. **Conditional Counting (20):** Count entities or events that satisfy specified conditions.

Level 3 (180 question-answer pairs).

1. **Multi-hop Reasoning (60):** Answer questions that require multiple linked inference steps.
2. **Long-range Reasoning (60):** Integrate dispersed evidence whose most widely separated start times are at least four minutes apart.
3. **Multi-Evidence Chain (60):** Reason over at least five temporally separated pieces of evidence.

B Experimental Details

B.1 Input Preprocessing and Inference Setup

Input preprocessing. For evaluation, all videos are resized to 480 pixels on the longer side while preserving their original aspect ratios. Frames are uniformly sampled at 1.0 FPS and provided to the Omni LLMs together with the corresponding audio. For the 1–12-minute videos in EvaOmni, this configuration keeps the audio-visual inputs within the context windows of the evaluated models.

Inference setup. All open-weight models except those in the MiMo family are evaluated on a single server equipped with eight NVIDIA H20 GPUs, using the inference frameworks and parameter settings recommended in their respective model cards. The three MiMo models are evaluated through MiMo's official API platform with the same input preprocessing settings described above. Since the weights of these three models have been publicly released, we classify them as open-weight models.

B.2 Human Expert Baseline

The human expert baseline is obtained from test participants who are independent of the annotators who constructed the questions and marked the reference evidence. They answer the same questions using the same test videos as the models and report supporting audio and visual intervals; their responses are scored using the same Acc., C.G., and B.G. metrics. Participants may pause and replay the videos when answering and locating evidence. This allowance helps compensate for human memory, which is much more limited than a model's access to the supplied video context, but it also makes the viewing protocol different from model inference. Even with replay, the need to retain and revisit details across a video may disadvantage humans, especially for long or temporally dispersed evidence. We therefore treat the human scores as a reference baseline rather than a strictly matched comparison of memory or viewing conditions.

C Prompt Templates

Figures 7, 8, 9, 10, 11, and 12 show the prompt templates used for the six input settings.

Prompt for Audio-Only Inputs

You are answering a multiple-choice question based ONLY on the Audio provided.
Answer the multiple-choice question and identify the evidence that directly supports your answer.
Question:
{question}
Options:
{options}
Evidence requirements:
* Each evidence item MUST contain a precise timestamp range in the format "[mm:ss,mm:ss]".
* Include only evidence that directly supports the answer.
* Keep each timestamp range as precise as possible.
Strict output format:
Return ONLY a valid JSON object with the following fields:
{{
"answer": "<single option letter>",
"audio_evidence": [
{{"time": "[mm:ss,mm:ss]"}}
]
}}

Figure 7: Prompt used for audio-only inputs.

Prompt for Text-Only Inputs

You are answering a multiple-choice question. You DO NOT have access to any video or audio.
Answer using your general knowledge and reasoning only.
Question:
{question}
Options:
{options}
Output:
{{
"answer": "<single option letter>"
}}

Figure 8: Prompt used for text-only inputs.

Prompt for Video Only Input

You are answering a multiple-choice question based ONLY on the Video provided.
Answer the multiple-choice question and identify the evidence that directly supports your answer.
Question:
{question}
Options:
{options}
Evidence requirements:
* Each evidence item MUST contain a precise timestamp range in the format "[mm:ss,mm:ss]".
* Include only evidence that directly supports the answer.
* Keep each timestamp range as precise as possible.
Strict output format:
Return ONLY a valid JSON object with the following fields:
{{
"answer": "<single option letter>",
"visual_evidence": [
{{"time": "[mm:ss,mm:ss]"}}
]
}}

Figure 9: Prompt used for video-only inputs.

Prompt for Video+Audio Input

You are answering a multiple-choice question based on the Video and Audio provided.
Answer the multiple-choice question and identify the evidence that directly supports your answer.
Question:
{question}
Options:
{options}
Evidence requirements:
* Each evidence item MUST contain a precise timestamp range in the format "[mm:ss,mm:ss]".
* Include only evidence that directly supports the answer.
* Keep each timestamp range as precise and minimal as possible.
Strict output format:
Return ONLY a valid JSON object with the following fields:
{{
"answer": "<single option letter>",
"visual_evidence": [
{{"time": "[mm:ss,mm:ss]"}}
],
"audio_evidence": [
{{"time": "[mm:ss,mm:ss]"}}
]
}}

Figure 10: Prompt used for Video+Audio inputs.

Prompt for ASR Input

You are answering a multiple-choice question based ONLY on an automatic speech recognition (ASR) transcript of a video's audio track, with timestamps. Answer the multiple-choice question and identify the evidence that directly supports your answer.
Question:
{question}
Options:
{options}
Evidence requirements:
* Each evidence item MUST contain a precise timestamp range in the format "[mm:ss,mm:ss]".
* Include only evidence that directly supports the answer.
* Keep each timestamp range as precise and minimal as possible.
Strict output format:
Return ONLY a valid JSON object with the following fields:
{{
"answer": "<single option letter>",
"audio_evidence": [
{{"time": "[mm:ss,mm:ss]"}}
]
}}

Figure 11: Prompt used for ASR-only inputs.

Prompt for Video+ASR Input

You are answering a multiple-choice question based on the Video and automatic speech recognition (ASR) transcript of the video's audio track with timestamps provided.
Answer the multiple-choice question and identify the evidence that directly supports your answer.
Question:
{question}
Options:
{options}
Evidence requirements:
* Each evidence item MUST contain a precise timestamp range in the format "[mm:ss,mm:ss]".
* Include only evidence that directly supports the answer.
* Keep each timestamp range as precise and minimal as possible.
Strict output format:
Return ONLY a valid JSON object with the following fields:
{{
"answer": "<single option letter>",
"visual_evidence": [
{{"time": "[mm:ss,mm:ss]"}}
],
"audio_evidence": [
{{"time": "[mm:ss,mm:ss]"}}
]
}}

Figure 12: Prompt used for Video+ASR inputs.

Table 6: **Benchmark comparison under text-only, audio-only, and video-only inputs.** All values are percentages. Each model is evaluated twice, and an answer is considered correct only if both independent runs produce the correct answer.; Intersection denotes the proportion of evaluated questions answered correctly in all four runs under the corresponding input condition.

(a) Daily-Omni

Input	Model 1	Model 2	Intersection
Text Only	45.1	38.6	28.4
Audio Only	30.2	57.9	25.6
Video Only	41.2	45.2	29.4

(b) WorldSense

Input	Model 1	Model 2	Intersection
Text Only	24.7	24.5	14.4
Audio Only	20.7	46.0	14.2
Video Only	25.5	39.8	15.7

(c) OmniVideoBench

Input	Model 1	Model 2	Intersection
Text Only	29.2	25.8	16.0
Audio Only	13.6	30.0	8.2
Video Only	22.8	26.8	12.1

(d) EvaOmni

Input	Model 1	Model 2	Intersection
Text Only	**9.7**	**3.8**	**1.8**
Audio Only	**3.2**	**18.2**	**1.0**
Video Only	**7.3**	**12.5**	**3.6**

Model 1 / Model 2 by input condition:
Text Only: Gemini3.1-Pro-Preview / Qwen3.8-Flash.
Audio Only: Audio-FlamingGo-3 / Qwen3-Omni-Instruct-30B-A3B.
Video Only: Qwen3VL-Instruct-30B-A3B / Qwen3-Omni-Instruct-30B-A3B.

D Additional Quantitative Results and Analysis

D.1 Shortcut Evaluation and Case Studies

Figure 1 summarizes the shortcut evaluation results for three representative benchmarks for Omni LLMs, alongside EvaOmni. Here, we provide further details of the evaluation protocol and report the complete results in Table 6.

Evaluation protocol. For the visual-only condition, denoted as *Video Only* in the table, we use Qwen3-Omni-Instruct-30B-A3B and Qwen3VL-Instruct-30B-A3B. For the audio-only condition, we use Qwen3-Omni-Instruct-30B-A3B and Audio-FlamingGo-3. For the text-only condition, we use Gemini3.1-Pro-Preview and Qwen3.8-Flash, providing only the question and answer options. Under each input condition, we evaluate each model twice, yielding four sets of predictions, and take the intersection of the sets of correctly answered questions across all four runs. The *Intersection* column reports the percentage of all evaluated questions answered correctly in every run. This design aims to reduce the influence of chance successes in a single run and biases specific to an individual model, although it does not eliminate these factors.

Interpreting the results. These results serve as diagnostic indicators of potential shortcut susceptibility. Consistently correct answers from multiple models under unimodal or text-only inputs do not necessarily imply that a question is flawed, nor do they alone establish that the models exploit a shortcut. Instead, they indicate the extent to which correct answers can be repeatedly obtained without full audio-visual input under the tested models and settings, highlighting a potential vulnerability of answer-only evaluation. The reported rates should therefore not be interpreted directly as the proportion of flawed questions or the probability that a benchmark can be exploited.

Illustrative Shortcut Cases. Figures 13–18, 19–24, and 25–30 show two text-only, two visual-only, and two audio-only shortcut cases from each of WorldSense, OmniVideoBench, and Daily-Omni, respectively.

Case of Text Only Shortcut

QA ID: 124

Question:
What is the correct sequence for testing the fuel pump as presented in the video?
Options:
A. (b) Connect the power probe to the ground wire, (a) Identify the power wire and ground wire, (c) The power probe lights up green, (d) Check the water flow in the pipe.
B. (a) Identify the power wire and ground wire, (b) Connect the power probe to the ground wire, (c) The power probe lights up green, (d) Check the water flow in the pipe.
C. (a) Identify the power wire and ground wire, (b) Connect the power probe to the ground wire, (d) Check the water flow in the pipe, (c) The power probe lights up green.
D. (c) The power probe lights up green, (a) Identify the power wire and ground wire, (b) Connect the power probe to the ground wire, (d) Check the water flow in the pipe.
Text Shortcut Explanation: For option B (the correct option), it presents a reasonable testing sequence, so the correct answer can be reliably selected based on the text alone.

Figure 13: Text-only shortcut case 1 from WorldSense.

Case of Text Only Shortcut

QA ID: 60

Question:
What is David Popovici's recorded time from the previous 100m World Championship?
Options:
A. 48.77 seconds.
B. 47.3 seconds.
C. 49.37 seconds.
D. 47.56 seconds.
Text Shortcut Explanation: For option D (the correct option), David Popovici's recorded time is 47.56 seconds (2022), which matches option D.

Figure 14: Text-only shortcut case 2 from WorldSense.

Case of Visual Only Shortcut

QA ID: 0

Visual Evidence : [00:04, 00:11] The metal rolling shutter is clearly visible to the right of the woman in white.

Question:
What is the position of the metal roller door relative to the woman wearing white in the video?

Options:
A. To the left of the woman wearing white.
B. To the right of the woman wearing white.
C. In front of the woman wearing white.
D. Behind the woman wearing white.

Visual Shortcut Explanation: For option B (the correct option), the metal roller door is clearly visible to the right of the woman, making the correct answer reliably identifiable from the visual information alone.

Figure 15: Visual-only shortcut case 1 from WorldSense.

Case of Visual Only Shortcut

QA ID: 11

Visual Evidence : [00:08, 01:00] The video shows two people playing the violin.

Question:
How many instruments are in the video?

Options:
A. Three.
B. Two.
C. One.
D. Four.

Visual Shortcut Explanation: For option B (the correct option), the video clearly shows two people playing the violin, making the correct option reliably identifiable from visual evidence alone.

Figure 16: Visual-only shortcut case 2 from WorldSense.

Case of Audio Only Shortcut

QA ID: 9

Audio Evidence : [00:00, 01:00] The audio features a duet of the guzheng and piano.

Question:
What is the instrument on the right in the video?
Options:
A. Guitar.
B. Cello.
C. Banjo.
D. Guzheng.
Audio Shortcut Explanation: For option D (correct option), the guzheng clearly matches option D, making the answer identifiable from audio alone.

Figure 17: Audio-only shortcut case 1 from WorldSense.

Case of Audio Only Shortcut

QA ID: 14

Audio Evidence(ASR) : [00:04, 00:09] "they become so serious and so sinister that McMahon is no longer an active part of WWE"

Question:
Will McMahon return to WWE?
Options:
A. I am not sure.
B. Yes.
C. No.
D. None of the above.
Audio Shortcut Explanation: For option C (the correct option), the audio clearly states that McMahon will not return to WWE, so the correct answer can be reliably selected using the audio alone.

Figure 18: Audio-only shortcut case 2 from WorldSense.

Case of Text Only Shortcut

QA ID: 136

Question:
What would the long-haired man do if the guy hadn't sung out?

Options:
A. Cheer for him
B. Feel lost for him
C. Feel admiration for him
D. Feel disappointed in him

Text Shortcut Explanation: Obviously, disappointment is a natural emotional response to failing to accomplish something. Therefore, the correct answer to this question can be reliably inferred through common-sense reasoning.

Figure 19: Text-only shortcut case 1 from OmniVideoBench.

Case of Text Only Shortcut

QA ID: 556

Question:
Why does this video evoke a sense of peace and prosperity?

Options:
A. Because the lights and the song are both beautiful.
B. The song's beautiful meaning matches the visuals.
C. This song evokes a sense of peace and prosperity.
D. This visual evokes a sense of peace

Text Shortcut Explanation: For option B (the correct option), it involves both the music and the visual content, which allows options C and D, each involving only one modality, to be ruled out. In addition, option B aligns more closely with the wording of the question than option A. Therefore, the correct answer can be reliably selected through common-sense reasoning.

Figure 20: Text-only shortcut case 2 from OmniVideoBench.

Case of Visual Only Shortcut

QA ID: 45

Visual Evidence : [00:02, 00:14] The video shows that the hair on the right side of the girl's head is singed.

Question:
Why does the laughter sound effect appear at the beginning of the video?

Options:
A. Because the girl's left hair is scorched.
B. Because the girl's right hair is scorched.
C. Because all the girl's hair is scorched.
D. Because the girl's clothes are scorched.

Visual Shortcut Explanation: For option B (the correct option), the video clearly shows the distinctive scene of the hair on the right side of the girl's head being singed, so the correct option can be reliably selected based on the visual input alone.

Figure 21: Visual-only shortcut case 1 from OmniVideoBench.

Case of Visual Only Shortcut

QA ID: 85

Visual Evidence : [04:57, 05:04] The video shows the man smiling at the end and happily walking out of the bowling alley with his girlfriend.

Question:
What emotion did the man have on the bowling ally at last?

Options:
A. Delighted
B. Dislike
C. Indifferent
D. Angry

Visual Shortcut Explanation: For option A (the correct option), the end of the video clearly shows the man smiling, so the correct answer can be reliably selected without using the audio.

Figure 22: Visual-only shortcut case 2 from OmniVideoBench.

Case of Audio Only Shortcut

QA ID: 544

Audio Evidence : [00:00, 00:49] A piece of lighthearted and cheerful music.

Question:
How do I feel about this job?
Options:
A. It's boring.
B. It's challenging.
C. It's exciting.
D. It's comfortable.
Audio Shortcut Explanation: For option D (the correct option), the background music suggests a relaxed and cheerful mood, so the correct answer can be reliably selected based on the audio alone.

Figure 23: Audio-only shortcut case 1 from OmniVideoBench.

Case of Audio Only Shortcut

QA ID: 564

Audio Evidence : [00:00,01:08] The background music is monotonous and subdued.

Question:
How would I most likely feel about this kind of life?
Options:
A. A bit lonely.
B. Very interesting.
C. Very negative.
D. A bit noisy.
Audio Shortcut Explanation: For option A (the correct option), there is no dialogue throughout the audio, only a monotonous piece of music, which suggests that the person may be feeling somewhat lonely.

Figure 24: Audio-only shortcut case 2 from OmniVideoBench.

Case of Text Only Shortcut

QA ID: 00004_G_VTkkb34gw

Question:
Which audio event occurred simultaneously with the two female police officers engaging in target practice at the indoor shooting range?

Options:
A. Sharp aerodynamic swoosh
B. Foreign-language command phrases
C. Numerical countdown sequence
D. Sustained gunfire bursts

Text Shortcut Explanation: For option D (the correct option), because the two female officers are practicing at an indoor shooting range, the sustained gunfire directly matches the activity shown.

Figure 25: Text-only shortcut case 1 from Daily-Omni.

Case of Text Only Shortcut

QA ID: 00040_sNf0WG57RtA

Question:
Why did the presenter emphasize the '$(CH3COO)_3$' portion of the chemical formula while explaining the compound's name?

Options:
A. To indicate the three acetoxy groups corresponding to the formula's subscript
B. To highlight the sodium (Na) component in the formula
C. To demonstrate the borohydride (BH) portion's role as the reducing agent
D. To connect the acetoxy groups to the 'triacetoxy' prefix in the chemical name

Text Shortcut Explanation: For option D (the correct option), "Triacetoxy" means three acetoxy groups, corresponding directly to the $(CH_3COO)_3$ part of the formula.

Figure 26: Text-only shortcut case 2 from Daily-Omni.

Case of Visual Only Shortcut

QA ID: 00012_d6b4OmUFt7I

Visual Evidence : [00:09, 00:30] A man appears on screen with the words "BOTOX WITHOUT THE BOTOX" behind him, followed by a woman holding the product and explaining it.

Question:
What visual elements were displayed immediately after Dr. Rajani's 'BOTOX WITHOUT THE BOTOX' video concluded?

Options:
A. Still product bottle → Price text overlay
B. Facial treatment demonstration → Presenter holding product while explaining
C. Presenter's torso shot → Secondary screen activation
D. Bookshelf backdrop → Close-up of string lights

Visual Shortcut Explanation: For option B (correct option), immediately after the man appears, the video shows a facial-care demonstration while the host explains the product, making it possible to answer the question reliably using the video alone.

Figure 27: Visual-only shortcut case 1 from Daily-Omni.

Case of Visual Only Shortcut

QA ID: 00031_7yY1OWjT4zA

Visual Evidence : [00:00, 00:03] A hand appears on screen, accompanied by the word "Conservation" in white text.

Question:
Which event occurs FIRST in the video's timeline?

Options:
A. White fabric unrolled in a room with ambient music
B. Fabric being unrolled on a reflective surface with music
C. Liz Rose pushing a cart through a workshop doorway
D. ' Conservation' text overlay with blurred hands

Visual Shortcut Explanation: For option D (the correct option), the opening visual already closely matches the option description, so a visual-only model can answer it reliably.

Figure 28: Visual-only shortcut case 2 from Daily-Omni.

Case of Audio Only Shortcut

QA ID: 00008_G_VTkkb34gw

Audio Evidence : [00:09, 00:10] The audio contains an overlap of electronic music and gunfire.

Question:
Which audio events correspond to the indoor shooting range activities shown in the video?

Options:
A. Countdown sequence during target practice
B. Electronic music overlapping with gunshots
C. Hip-hop music with whoosh sounds
D. Wind sounds followed by human speech

Audio Shortcut Explanation: For option B (correct option), the overlap of electronic gunfire and electronic music in the audio closely matches the question, making it very easy to answer correctly using the audio alone.

Figure 29: Audio-only shortcut case 1 from Daily-Omni.

Case of Audio Only Shortcut

QA ID: 00000_Ec_lQgZ9wlg

Audio Evidence (ASR) : [00:19, 00:25] "We get our class time, go through all that, and then after we get to go on the equipment for a couple hours and get used to it"

Question:
What immediate task required Sara to review documents in the office before equipment operation?

Options:
A. To analyze soil composition charts
B. To finalize her course payment details
C. To prepare safety checklists for machinery
D. To complete classroom requirements stated in the audio

Audio Shortcut Explanation: For option D (correct option), The audio mentions that they have class time to learn about these devices, so the question can be answered reliably using the audio alone.

Figure 30: Audio-only shortcut case 2 from Daily-Omni.

Table 7: **Additional grounded evaluation results across capability levels.** Mean averages the answer score weighted by the mean audio and visual IoU. Thr@τ averages the answer score after gating it by whether that mean IoU reaches τ, for $\tau \in \{0.3, 0.5, 0.7\}$. All values are percentages; higher is better. Within each model group, **bold** and underlined values indicate the best and second-best distinct scores in each column, respectively; ties receive the same formatting.

Model	Level 1				Level 2				Level 3				Overall			
	Mean	Thr@0.3	Thr@0.5	Thr@0.7	Mean	Thr@0.3	Thr@0.5	Thr@0.7	Mean	Thr@0.3	Thr@0.5	Thr@0.7	Mean	Thr@0.3	Thr@0.5	Thr@0.7
Open-Weight Omni LLMs																
Qwen2.5-Omni-7B	0.7	0.9	0.9	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.0	0.0
Qwen3-Omni-Instruct-30B-A3B	2.8	5.3	2.5	0.0	1.2	1.1	0.6	0.0	0.5	0.0	0.0	0.0	1.8	2.9	1.4	0.0
Qwen3-Omni-Thinking-30B-A3B	6.5	10.2	7.4	2.8	2.5	3.4	1.1	0.6	1.0	0.7	0.0	0.0	4.1	6.0	3.8	1.5
Nemotron3-Nano-Omni-30B-A3B	7.7	11.3	7.6	3.2	5.9	9.5	5.3	1.7	1.7	2.0	0.0	0.0	6.1	9.1	5.5	2.1
MiMo-V2.5-310B-A15B	**37.1**	**56.6**	42.3	**22.2**	**19.9**	**30.0**	**19.1**	7.6	10.6	14.4	**2.0**	**1.3**	**26.3**	**39.7**	26.9	**13.3**
MiMo-V2.6-Flash-310B-A15B	35.2	52.2	**43.7**	20.1	19.8	**30.0**	**19.1**	**8.4**	**11.0**	**17.0**	**2.0**	0.0	25.4	38.1	**27.6**	12.4
MiMo-V2.6-Pro-1T-A42B	33.0	48.5	40.4	20.1	17.0	26.6	15.7	7.3	9.5	9.8	**2.0**	0.7	23.1	33.9	24.8	12.1
Proprietary Omni LLMs																
Doubao-Seed-2.1-Lite	17.7	30.0	15.2	4.4	14.4	22.7	10.9	4.8	12.6	17.0	2.6	0.7	15.6	25.1	11.6	3.9
Qwen3.5-Omni-Flash	20.7	34.2	18.5	6.9	11.3	16.5	10.1	3.9	4.9	3.9	1.3	0.7	14.6	22.6	12.5	4.8
Qwen3.5-Omni-Plus	28.5	46.0	30.0	11.8	11.2	17.1	9.5	3.1	6.8	7.8	2.0	0.7	18.4	28.8	17.7	6.7
Qwen3.8-Omni-Flash	26.7	44.6	27.5	7.2	17.4	28.0	13.2	3.6	13.9	17.0	5.9	1.3	21.1	33.8	18.6	4.9
Gemini2.5-Flash-Lite	27.3	41.6	30.7	13.2	10.7	15.4	8.4	5.0	4.3	2.6	0.7	0.0	17.3	25.4	17.4	8.0
Gemini2.5-Flash	33.3	51.5	38.8	17.3	17.2	25.5	16.0	9.2	7.5	7.2	1.3	0.0	23.0	34.5	24.1	11.5
Gemini2.5-Pro	49.2	70.2	60.5	34.2	30.1	45.1	31.1	14.3	17.8	22.9	7.2	1.3	36.9	53.0	40.7	21.3
Gemini3.1-Flash-Lite	31.1	47.8	36.3	15.7	15.8	25.2	13.7	6.2	5.4	3.9	0.7	0.7	21.1	32.1	22.0	9.7
Gemini3.1-Flash-Lite-Preview	32.0	50.4	37.4	16.6	16.0	24.4	13.2	6.4	6.5	6.5	0.0	0.0	21.8	33.4	22.2	10.1
Gemini3.1-Pro-Preview	48.5	68.8	60.7	31.4	31.5	45.9	29.7	15.7	20.1	28.8	5.2	2.6	37.5	53.7	40.0	20.8
Gemini3.5-Flash-Lite	33.5	51.5	38.6	18.5	15.6	24.4	14.0	8.1	8.3	9.8	0.7	0.7	22.6	34.5	23.1	11.7
Gemini3.5-Flash	36.7	55.9	44.1	21.5	34.4	50.7	35.3	19.1	20.6	27.5	9.8	1.3	33.3	49.3	35.2	17.3
Gemini3.6-Flash	52.4	75.1	66.1	37.0	35.1	52.7	34.5	19.6	20.6	31.4	8.5	2.0	40.7	59.5	44.8	24.7
Gemini3.7-Flash	**56.9**	**81.8**	**72.3**	40.4	**37.9**	**56.0**	**41.2**	**21.3**	**23.5**	**36.0**	**12.4**	**3.3**	**44.3**	**64.6**	**50.8**	27.2
Gemini3.8-Flash	56.4	79.2	69.3	**43.4**	36.3	52.7	37.0	20.7	18.3	26.1	5.9	2.0	42.7	60.6	46.8	**28.1**

D.2 Additional Grounded Evaluation Results

In addition to the Bottleneck Grounded Score (B.G.) and Coverage Grounded Score (C.G.) reported in the main text, we report a continuous score and threshold-based scores derived from the average temporal IoU across audio and visual evidence. These metrics provide complementary views of performance under different localization requirements.

For sample i, let s_i denote the answer score defined in the main text, and let a_i and v_i denote the temporal IoUs for audio and visual evidence, respectively. Their average is

$$m_i = \frac{a_i + v_i}{2}. \tag{3}$$

For an evaluation set of N samples, we define

$$\text{Mean} = \frac{1}{N} \sum_{i=1}^{N} s_i m_i, \tag{4}$$

$$\text{Thr@}\tau = \frac{1}{N} \sum_{i=1}^{N} s_i \, \mathbf{1}\{m_i \geq \tau\}, \qquad \tau \in \{0.3, 0.5, 0.7\}, \tag{5}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. All results are reported as percentages.

Mean continuously measures the joint performance of answering and average evidence localization across modalities; it is not a localization metric independent of answer correctness. For MCQs with binary answer scores, Thr@τ is the fraction of all samples that are answered correctly and have an average IoU of at least τ, rather than a pass rate computed only over correct answers. Higher thresholds impose stricter requirements on average localization. Unlike B.G., which is sensitive to the weaker modality, average IoU allows stronger localization in one modality to compensate for weaker localization in the other. Passing a threshold on average IoU therefore does not require both modality-specific IoUs to exceed that threshold.

Overall performance. Table 7 shows that Gemini3.7-Flash achieves the highest overall Mean (44.3%), followed by Gemini3.8-Flash (42.7%) and Gemini3.6-Flash (40.7%). Among the evaluated open-weight models, MiMo-V2.5 obtains the highest overall Mean at 26.3%, trailing Gemini3.7-Flash by 18.0 percentage points. Thus, a substantial gap between the leading open-weight and proprietary models persists when the weaker-modality IoU used in B.G. is replaced by average IoU.

Performance across capability levels. Mean decreases from Level 1 to Level 2 and again to Level 3 for all 22 evaluated models. For example, Gemini3.7-Flash scores 56.9%, 37.9%, and 23.5% across

the three levels, while MiMo-V2.5 scores 37.1%, 19.9%, and 10.6%. This pattern is consistent with the greater demands on evidence integration and temporal localization at higher levels. Level 3 remains challenging even when localization quality can be traded off across modalities. Because these scores also incorporate answer correctness, however, the decline cannot be attributed entirely to localization errors.

Sensitivity to localization thresholds. Stricter average-IoU thresholds substantially reduce scores. Gemini3.7-Flash obtains overall scores of 64.6%, 50.8%, and 27.2% at thresholds of 0.3, 0.5, and 0.7, respectively; the corresponding scores for MiMo-V2.5 are 39.7%, 26.9%, and 13.3%. Thus, many correctly answered samples satisfying a permissive localization criterion fail to meet a stricter one. Rankings also depend on the threshold: Gemini3.7-Flash leads in overall Mean, Thr@0.3, and Thr@0.5, whereas Gemini3.8-Flash has a slightly higher Thr@0.7 score (28.1% versus 27.2%). Continuous scores and results at multiple thresholds therefore provide complementary information about average performance and success under stricter localization requirements.

D.3 Details of Failure Mode Analysis

Scope and validation. We analyze the available predictions from Gemini3.7-Flash, Gemini3.5-Flash, Qwen3.5-Omni-Plus, Gemini2.5-Pro, Gemini2.5-Flash, Qwen3.8-Omni-Flash, and MiMo-V2.5-310B-A15B. Each model is evaluated on the same subset of the available annotations, matched by question ID. We also verify video IDs, questions, options, levels, and reference answers for every record.

Scoring protocol. Ground-truth evidence is read from the annotation file for every model. For each modality, we compute temporal IoU between the unions of the predicted and annotated intervals, without multiplying by answer correctness. Failed inference, response parsing, or truncated output receives zero answer and localization scores and remains in the denominator. Comma-based decimal timestamps are normalized numerically. These choices preserve a common evaluation set without granting localization credit to malformed predictions.

Let C denote answer correctness and $G_\tau = \mathbb{1}[\min(\mathrm{IoU}_{\mathrm{audio}}, \mathrm{IoU}_{\mathrm{visual}}) \geq \tau]$ denote successful dual-modal localization. Unlike the mean-IoU threshold metric above, this criterion requires each modality to meet the threshold. Table 5 reports the four outcomes at $\tau = 0.5$; each outcome is expressed as a percentage of all questions.

Modality-specific localization. Six of the seven models in Table 5 obtain higher mean audio than visual IoU overall. Among these six, the difference ranges from 0.7 percentage points for Qwen3.5-Omni-Plus to 3.7 points for Gemini3.5-Flash. Qwen3.8-Omni-Flash shows the opposite pattern, with 27.4% audio IoU and 34.3% visual IoU.

Answer–grounding disagreement. Across the seven models, 47.8%–79.8% of correct answers fail the dual-modal localization criterion at $\tau = 0.5$. For Gemini3.7-Flash, 47.8% of correct answers fail grounding, whereas 12.2% of grounding-passing predictions answer incorrectly. The converse error also occurs in the other models: for MiMo-V2.5, 27.9% of grounding-passing predictions answer incorrectly. These outcomes distinguish temporal localization from answer selection; temporal overlap alone establishes neither semantic understanding nor causal use of the reported evidence.

Threshold sensitivity. For Gemini3.7-Flash, the fraction of correct answers for which either audio or visual IoU falls below the grounding threshold is 32.9%, 47.8%, and 78.2% at thresholds of 0.3, 0.5, and 0.7, respectively. These proportions depend on the threshold and are not direct evidence that correct answers were guessed.

D.4 Analysis Across Fine-Grained Capabilities

Figure 31 compares the two representative models across 15 fine-grained capabilities. On the three Level 1 alignment tasks, Gemini3.7-Flash achieves 82.3%–88.8% accuracy and 47.5%–49.9% B.G., compared with 60.4%–68.8% accuracy and 28.1%–34.0% B.G. for MiMo-V2.5. Both models show a larger separation between answering and precise grounding on complex tasks. On multi-hop reasoning, Gemini3.7-Flash reaches 80.7% accuracy but only 10.6% B.G., while MiMo-V2.5 reaches 44.6% accuracy and 4.9% B.G.; on multi-evidence chains, the corresponding scores are

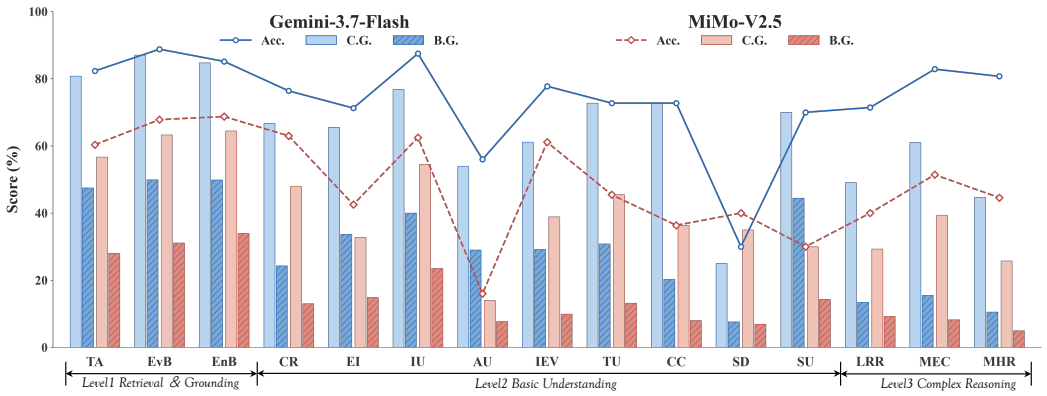


Figure 31: **Performance across fine-grained capabilities.** Lines show answer accuracy (Acc.); solid and hatched bars show Coverage Grounded Score (C.G.) and Bottleneck Grounded Score (B.G.), respectively. All values are percentages. Task abbreviations follow the capability taxonomy in Appendix A.5: Level 1: TA = Temporal Alignment, EvB = Event Binding, EnB = Entity Binding; Level 2: CR = Causal Reasoning, EI = Emotional Intelligence, IU = Intent Understanding, AU = Atmosphere Understanding, IEV = Implicit Event Reasoning, TU = Temporal Understanding, CC = Conditional Counting, SD = Semantic Disambiguation, SU = Spatial Understanding; Level 3: LRR = Long-range Reasoning, MEC = Multi-Evidence Chain, MHR = Multi-hop Reasoning.

82.9%/15.6% and 51.4%/8.2%. These comparisons suggest that selecting an answer is substantially easier for both models than locating the dispersed evidence needed to support it.

E Limitations and Future Work

E.1 Limitations

Dataset coverage. EvaOmni contains 1,200 question-answer pairs drawn from videos lasting 1–12 minutes with sufficient audio-visual cues. Excluding hard-coded subtitles and prioritizing varied entities and events support the construction of questions requiring both modalities, but also constrain the data distribution. The results therefore do not establish generalization to substantially longer videos, subtitle-rich content, scenes with sparse events, or all linguistic and cultural contexts. The three capability levels characterize different task requirements; they do not guarantee a strictly increasing difficulty ordering in answer accuracy across different question sets.

E.2 Future Work

Broader data and task coverage. Future extensions could include longer videos, a wider range of linguistic and cultural contexts, and scenes with sparser audio-visual cues while retaining checks for the necessity of both modalities. Subtitle-rich content would also enable explicit separation of text reading from audio understanding. Broader evaluation of open-ended question answering could test whether the current findings hold without cues from answer options.