



MAVERIX: Multimodal Audio-Visual Evaluation and Recognition IndeX

Liuyue Xie^{*1} Avik Kuthiala^{*1} George Z. Wei^{*1}
 Ce Zheng¹ Ananya Bal¹ Mosam Dabhi¹ Liting Wen¹
 Taru Rustagi¹ Ethan Lai¹ Sushil Khyalia¹ Rohan Choudhury¹
 Morteza Ziyadi² Xu Zhang² Hao Yang² László A. Jeni¹

¹Carnege Mellon University, ²Amazon, ^{*}Equal contribution

Abstract

We introduce MAVERIX (Multimodal audiovisual Evaluation and Recognition IndeX), a unified benchmark to probe the video understanding in multimodal LLMs, encompassing video, audio, text inputs with human performance baselines. Although recent advancements in models with vision and audio understanding capabilities have shown substantial progress, the field lacks a standardized evaluation framework to thoroughly assess their cross-modality comprehension performance. MAVERIX curates 2,556 questions from 700 videos, in the form of both multiple-choice and open-ended formats, explicitly designed to evaluate multimodal models through questions that necessitate tight integration of video and audio information, spanning a broad spectrum of agentic scenarios. MAVERIX uniquely provides models with audiovisual questions, closely mimicking the multimodal perceptual experiences available to humans during inference and decision-making processes. To our knowledge, MAVERIX is the first benchmark aimed explicitly at assessing comprehensive audiovisual integration in such granularity. Experiments with state-of-the-art models, including Qwen 2.5 Omni and Gemini 2.5 Flash-Lite, show performance around 64% accuracy, while human experts reach near-ceiling performance of 92.8%, exposing a substantial gap to human-level comprehension. With standardized evaluation protocols, a rigorously annotated pipeline, and a public toolkit, MAVERIX establishes a challenging testbed for advancing audiovisual multimodal intelligence.

Introduction

Human cognition seamlessly integrates visual and auditory information to reason, infer, and interact within dynamic environments. Replicating this ability in Multimodal Large Language Model (MLLM) systems remains a central challenge for AI, as autonomous agents must process complex audiovisual input to engage meaningfully with the world (Lin et al. 2023; Amirizaniani et al. 2024; tes 2024).

Recent progress in multimodal foundation models has brought us closer to this goal, but current benchmarks fall short in assessing their abilities to reason with multimodal inputs. Most benchmarks focus on static images (Chen et al. 2015; Agrawal et al. 2019; Li et al. 2025b), simple recognition, or questions that can be solved through unimodal cues,

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

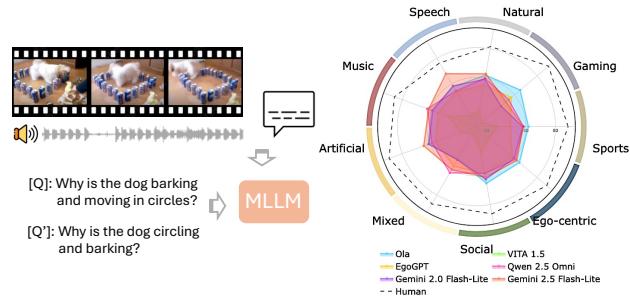


Figure 1: An illustration of our proposed benchmark, which includes highly audiovisual correlated questions and paraphrased questions, can be used to evaluate the model’s underlying comprehension abilities and their gaps to humans.

such as transcripts (Singh et al. 2019; Chen et al. 2024b). These benchmarks fail to probe the deeper, joint reasoning across modalities needed for real-world scenarios, such as interpreting social interactions or anticipating off-screen events (Chandrasegaran et al. 2024; Mangalam, Akshulakov, and Malik 2023).

A core obstacle in designing effective multimodal benchmarks is ensuring that the questions genuinely require multimodal understanding rather than allowing models to exploit unimodal shortcuts or common sense from the training data. For benchmarks designed to expose the model understanding for highly multimodal data, their common adoption of a 4-way multiple-choice question for evaluation provides limited insights into the underlying interpretations (Li et al. 2025b; Hong et al. 2025). Many existing video-language benchmarks reduce to visible-object recognition or dialogue parsing, bypassing the need to synthesize audiovisual dependencies (Patraucean et al. 2023; Keser et al. 2023; Li et al. 2024b).

To address this, we introduce **MAVERIX**, a benchmark designed to evaluate multimodal video-audio understanding through questions that have tight modality interdependence. MAVERIX features questions from challenging agentic scenario categories: factual recall, causal understanding, sentiment analysis, temporal recall, situational awareness, context understanding, social interaction understanding, and emotional synthesis, covering 700 videos and 2,556 care-

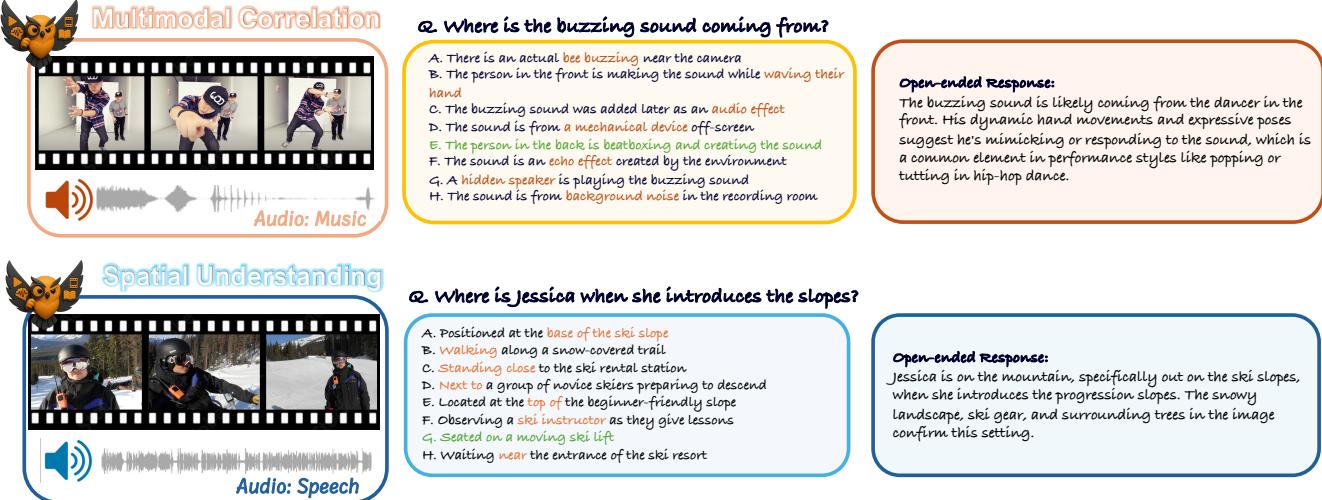


Figure 2: Example Agentic Categories and corresponding QAs in the MAVERIX benchmark.

fully designed questions. These are constructed through a hybrid human-AI pipeline to ensure that solving them requires intertwined audiovisual synthesis, revealing the underlying multimodal reasoning capabilities of models applied to the benchmark questions.

Evaluations of state-of-the-art proprietary and open-source models of different sizes, including Gemini 2.0/2.5 Flash-Lite(FL)(Team and Petko Georgiev 2024), GPT-4o (OpenAI and Aaron Hurst 2024), and Qwen 2.5 Omni (Xu et al. 2025), reveal significant gaps as shown in Fig. 1. Gemini 2.5 Flash-Lite, even with direct audio-video inputs, achieves 54.7% accuracy on multiple choice questions, significantly lower than human performance (92.8%). Open-ended responses further expose weaknesses in temporal reasoning and contextual understanding, with models averaging only 1.9/5 vs. human 2.79/5 in GPT-4o-judged scoring. Further, models that are not capable of processing raw audio and rely solely on transcripts perform even worse, highlighting the inadequacy of text-only proxies for rich audiovisual comprehension (Ning et al. 2023; Fu et al. 2024).

By providing a unified evaluation framework, high-quality human-validated questions, and an open-source toolkit, MAVERIX aims to advance research toward robust multimodal reasoning at the human level.

Benchmark Design and Construction

MAVERIX challenges multimodal large language models (MLLMs) to *integrate* audio and visual evidence under realistic conditions. This section elaborates on four key aspects: (i) the motivation behind our design, (ii) the dataset construction pipeline, (iii) the dual-format evaluation protocol, and (iv) dataset statistics. Fig. 3 visualizes the pipeline; Tab. 1 compares the benchmark components with the relevant works; Tab. 2 summarizes key dataset statistics.

Design Motivation and Principle

While previous video-understanding benchmarks curate multiple-choice question-answer pairs over different topics,

some still suffer unimodal shortcuts and are limited in exposing models’ underlying biases. For evaluating the models’ multimodal understanding abilities, we source videos that capture a wide range of temporal events, spatial motions, and audiovisual correlations. While sourcing the videos and constructing the multiple choice and open-ended question-answer pairs, we follow the following design principles.

Avoid Unimodal Shortcut. Many existing image and video based question answering benchmarks (e.g., TVQA, MSR-VTT Q&A) contain questions that can be answered from captions or a single salient frame, enabling unimodal shortcuts.

Wide range of evaluation dimensions. Deploying MLLMs in the real world requires that the models understand and handle scenarios with different skills. Single-skill benchmarks do not reflect the breadth of reasoning required in open-world settings. The curated benchmark measures the models in six evaluation dimensions, covering acoustic understanding, agentic skills, understanding by broad and sub-topic taxonomy, temporal understanding, and multimodal synthesis abilities.

Prevention of guess inflation. We design a hybrid of eight-way multiple-choice, and open-ended QA for the benchmark, such that by-passing the questions with model-inherent biases can be evidently exposed. The hybrid design evaluates the models’ actual abilities to interpret the input sources in different modality settings and provides a fair evaluation of their capabilities.

Dataset Generation Pipeline

Video Collection. We primarily source our video content from five datasets: YouTube-8M(Abu-El-Haija et al. 2016), MSR-VTT(Xu et al. 2016), UR-FUNNY-V2(Hasan et al. 2019), Ego4D(Grauman et al. 2021), and AudioSet(Gemmeke et al. 2017). YouTube-8M(Abu-El-Haija et al. 2016) and AudioSet(Gemmeke et al. 2017) are large-scale datasets covering a wide range of taxonomies, where the videos have strong audiovisual correlations. MSR-VTT

Benchmark	#Vid.	Med. Len. (s)	#Q	Mod.	MCQ.	#Div.	Diff.	Shortcut	Human	OE.
MSRVT-T-QA	2,990	15	72,821	V	4-MCQ	1	X	X	X	X
MSVD-QA	504	9	50,505	V	4-MCQ	1	X	X	X	X
ActivityNet-QA	5,800	15	58,000	V	4-MCQ	1	X	X	X	X
How2QA	1,517	11	71,812	V+S+A	4-MCQ	1	X	X	X	X
AutoEval-Video	327	32	450	V	4-MCQ	1	X	X	X	✓
TempCompass	410	11	1,540	V	4-MCQ	2	X	X	X	✓
Video-MME	900	1,072	2,700	V+S	4-MCQ	3	X	X	X	X
OmniBench	—	—	1142	I+A	4-MCQ	3	X	✓	X	X
WorldSense	1,662	141.1	3,172	V+S+A	4-MCQ	3	X	X	X	X
HourVideo	500	2,742	12,976	V+S	4-MCQ	2	X	X	X	X
MAVERIX	700	106	2,556	V+S+A	8-MCQ	7	✓	✓	✓	✓
MAVERIX-Long	700	345	2,556	V+S+A	8-MCQ	7	✓	✓	✓	✓

Table 1: **Comparison with prior video-question benchmarks.** **Mod.:** V (video), S (subtitles), A (audio). **Ans.:** 4-MCQ, “8-MCQ+OE” (eight-option plus open-ended). **Diff.:** crowdsourced or expert difficulty labels. **#Div.:** Number of division types. **Shortcut:** dataset validated against audio-only / video-only ablations. **Human:** ✓ if a benchmark reports any human baseline.

comprises of high quality video descriptions designed for video translation QAs. UR-FUNNY-V2 exhibits videos of different emotional states, challenging the models in their sentimental understanding. Lastly, Ego4D consists of ego-centric, long-duration videos to probe model’s understanding to daily interactions for agentic scenarios. The videos are selected by human annotators according to the principles described above, ensuring a thorough distribution across the topics, with different durations: *short* with < 1 min, *medium* with 1-10 mins, and *long* with 10-65 mins. Each video is processed and accompanied by subtitles generated with Whisper-v3 (Radford et al. 2022) to ensure a fair evaluation on video-text models without audio-support.

Initial Question Answering Annotation. A team of 8 expert annotators engaged in the initial question-answering pair curation. The annotators provided at least one question answer pair to each video to generate the initial ground truths. Then the same pair is expanded into eight-way multiple choice question with alternative distractive answers.

Shortcut Removal and Validation. Following the initial annotation, we use a semi-automated approach to validate difficulties with MLLMs and refine the questions to avoid any potential shortcuts. Each question undergoes three ablation tests with GPT-4o-mini and Gemini 2.0-FL: *text-only*, *video-only*, and *videos+subtitles*. If any ablation yields the correct answer for both models, the item is flagged and revised to reduce reliance on unimodal cues. All revisions are logged, and the final set is approved after a second expert pass. For example, a valid question might ask, “*Why did the mechanic abruptly stop speaking?*” requiring both visual cues (e.g., discovering a leak) and audio cues (e.g., sudden silence). This protocol ensures MAVERIX’s QA pairs demand genuine modality interdependence, preventing reliance on any single modality.

The difficulty labels were crowd-sourced through Amazon MTurk service (mturk 2005) with 219 unique participants for gauging common consensus, and are determined based on the subtlety of cross-modal cues, the depth of understanding required, and the ease of locating relevant information in

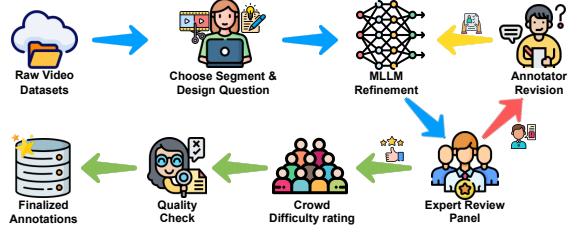


Figure 3: The framework to construct annotation sets with hybrid annotator and MLLM-as-judge quality assurance.

the video. The human performance evaluations were gathered through MTurk with 382 participants answering a 1/3 subset of the MCQs and open-ended questions.

Quality Assurance To ensure the reliability of MAVERIX’s videos and annotations, each QA pair undergoes four checks by an expert annotator, as illustrated in Fig. 3: (1) linguistic validity for clear and grammatical phrasing, (2) answerability of whether the question is resolvable via the video’s audiovisual content, (3) option integrity to ensure one correct answer with plausible distractors like semantically tangent or structurally identical options, and (4) modality interdependence, using cross-modal invalidation tests from Section (e.g., disabling audio or video to detect shortcuts). For open-ended questions, reviewers also confirm that rephrased variants preserve meaning without overlapping with the ground-truth wording.

Dual-Format Evaluation

Eight-Option MCQs. Each of the 2556 questions offers one ground-truth answer and seven carefully crafted distractors. Annotators design distractors that remain semantically consistent with the clip yet differ in key audiovisual details, forcing models to discriminate subtle cross-modal cues. Expanding to eight options lowers random accuracy to 12.5 %, yielding finer score resolution. Initial difficulty labels from

Statistic	Audio Type			Agentic Categories			Topic Domain		Overall
Sub-class of QA	Mixed Sound	Speech	Artificial Sound	Information Querying	Egocentric Agent	Sentiment Analysis	Humanity & Society	Business & Commerce	
<i>Agentic Abilities</i>									
Causal Relationship	66	111	15	75	27	33	27	21	201
Emotional Inference	51	57	18	66	21	27	27	12	129
Factual Recall	516	690	60	672	171	186	210	120	1311
Situational Understanding	27	27	10	33	9	12	12	6	70
Context Understanding	309	414	45	237	147	138	111	84	771
<i>QA Lengths</i>									
Question	11.26	11.63	9.06	10.31	10.41	10.00	10.09	10.51	11.28
Options	11.76	10.52	11.99	9.92	12.29	12.89	10.30	12.54	11.13
Open-ended answer	13.30	12.69	11.68	10.46	12.77	13.51	12.73	12.12	12.85
Subtitle length	440.79	682.24	419.87	485.08	1257.19	360.40	351.32	488.34	558.06
<i>Video–Audio Statistics</i>									
Media Length	319.98	381.24	327.25	315.77	1039.18	289.09	222.42	259.69	352.63
Media min. Length	6.15	10.03	5.57	5.57	10.04	5.57	5.57	10.15	5.57
Media max. Length	6620.63	4427.76	3205.50	6620.63	6620.63	3851.93	1800.17	511.23	6620.63
Media std.	527.50	681.63	482.65	603.49	1341.37	439.54	268.20	117.12	610.82

Table 2: Statistics from the included data. *Agentic Categories*: counts per category. *QA Lengths*: mean word counts for questions, mean per-option length (MCQs), and open-ended answers (computed only when options are absent). *Video–Audio Statistics*: duration in seconds (mean, min, max, std).

Gemini and GPT-4o are later re-calibrated by crowd workers to align machine estimates with human perception.

Open-Ended Generation. Every clip is paired with at least one semantically unique free-form prompt, ranging from causal explanation to future prediction, requiring natural-language output. For each unique question, we paraphrase the question and prompt the model again to test its robustness against paraphrasing. The open-ended responses are graded by GPT-4o on a five-factor rubric covering factual correctness, attention to detail, contextual grounding, temporal coherence, and paraphrase robustness.

Together, the MCQ and generation tracks marry *scalability*, through high-throughput accuracy metrics, with *depth*, by exposing weaknesses in explanation quality.

MAVERIX Statistics

MAVERIX comprises a diverse set of videos spanning 155 real-world scenarios across evaluation dimensions including agentic categories, topic domains, sub-topic domains, audio categories, duration, and difficulty. These are complemented by hierarchies over topics (e.g., travel, technology), video categories (e.g., documentaries, vlogs), and multimodal abilities (e.g., temporal reasoning, emotion recognition), supporting broad and balanced evaluation. Each question is also tagged with a difficulty level judged by human annotators. A detailed breakdown of these splits is provided in Tab. 2.

The dataset contains 105.8 hours of video footage, with durations ranging from 10 seconds to 63 minutes. Videos are distributed across three duration categories: 16.8% short clips (<1 minute) for rapid context-switching understanding, 75.7% medium-length videos (1-10 minutes) for sustained understanding, and 7.4% long-form content (10-65 minutes) for testing temporal coherence. The average video length is 352.63 seconds, with an std. of 610.82, suggesting that the included medias have a diverse distribution in lengths. Constructing a audiovisual benchmark with diverse

media lengths reflects real-world use cases for MLLMs and sufficiently challenges their ability to generalize.

Each video is paired with 3 to 4 questions on average with 2,556 in total, including 852 eight-option multiple-choice questions (MCQs) and 1704 open-ended prompts, with examples shown in Fig. 2. Questions span over the evaluation dimensions to ensure a thorough evaluation. To mitigate positional bias, answer labels are uniformly redistributed across options.

Experiments

Evaluation Protocol

MAVERIX adopts a dual evaluation framework to assess multimodal LLMs (MLLMs) through eight-option multiple-choice questions (MCQs) and open-ended response generation.

Evaluation is conducted under two settings: localized, where models access only the *timestamped video segment* relevant to each question, and global (MAVERIX-Long), where the full-length video is provided. The localized setting limits the context to the specific temporal window required for understanding, whereas providing the full videos demands the models to localize the required information from the haystack of frames.

For MCQs, we report both split-specific and overall accuracy, with answer choices uniformly distributed across positions (A-H) to reduce positional bias. Open-ended responses are evaluated using a GPT-4o scoring pipeline, adapted from Video ChatGPT (Maaz et al. 2024), which assesses the output in five dimensions on a scale of 0-5. The results are aggregated across modalities (Tab. 3), with separate analyses for easy, medium, and hard videos to diagnose comprehension limitations in Tab. 4. We also report the model’s open-ended response qualities and token counts in Tab. 6. Our proposed evaluation protocol ensures reproducibility while ad-

dressing modality interdependence and human performance baselines.

Baselines

We evaluate MAVERIX on a diverse suite of 17 MLLMs, encompassing both proprietary and open-source models, to assess their ability to reason over intertwined audiovisual modalities. Proprietary models include Gemini 2.0-FL (Team and Petko Georgiev 2025), Gemini 2.5-FL (Team and Petko Georgiev 2025), GPT-4o (OpenAI and Aaron Hurst 2024), Grok4 (xAI 2025), Claude Sonnet 3.5 (Anthropic 2024), Nova-Lite (Intelligence 2024), and Nova-Pro (Intelligence 2024). While open source representatives feature Ola (Liu et al. 2025b), EgoGPT (Yang et al. 2025), VITA 1.5 (Fu et al. 2025), Qwen 2.5 Omni (Xu et al. 2025), InternVL2 (Chen et al. 2024c), Qwen2.5-VL (Bai et al. 2025), LLaVA-OneVision (Li et al. 2024a), DeepSeek-VL2-Small (Wu et al. 2024). Among them, Ola, EgoGPT, VITA 1.5, Qwen 2.5 Omni and Gemini are equipped with native audiovisual processing, enabling direct ingestion of raw video-audio streams. For the tested models, we maximize temporal resolution by sampling frames at their maximum supported rates. However, most architectures, including GPT-4o and LLaVA-OneVision, require transcribed subtitles as text proxies for audio. To standardize inputs, we preprocess all videos using Whisper-v3 (Radford et al. 2022) to extract time-synced subtitles, with the timestamps provided to the evaluated models.

All models receive inputs in the unified format [video frames, subtitles, question], with frames uniformly sampled at their maximum supported context window. For audio-incapable models, subtitles replace raw audio tracks, while the audio-supported models additionally process synchronized audio-video pairs. We employ a standardized prompt template across models, ensuring fairness by eliminating instructional biases. This setup isolates modality interdependence as the critical challenge: models must synthesize potentially asynchronous audiovisual cues, such as the startled expression of a character with an auditory context like an off-screen crash to match human-like understanding.

Model analysis with MAVERIX

This section discusses model behaviour along five dimensions: multimodal gains across modal designs, training-recipe variation, agentic ability relative to humans, temporal-horizon sensitivity, and perceived question difficulty, using the benchmark set.

Multimodal Gains Across Model Capacity and Architecture. Tab. 3 summarizes performance across five architectural families and multiple model sizes. Best unimodal accuracies span $\sim 26\text{--}55\%$, and most models show gains when additional modalities are provided. Many systems follow an audiovisual encoder+MLP+LLM design; among these, Ola, EgoGPT, and Qwen 2.5 Omni, which incorporate Whisper v3 for audio, generally do not regress and often improve relative to their strongest unimodal scores, whereas VITA 1.5 shows a regression with audio-visual input. The lightweight Gemini 2.0-FL and 2.5-FL variants

also improve with multimodal inputs. For video-text models, adding a modality yields consistent gains; Qwen2-VL, Nova-Lite, Nova-Pro, and Claude 3.5 Sonnet improve by roughly $\sim 10\%$. Despite these gains, a sizable gap to human performance remains, suggesting that current models under-utilize cross-modal cues. This is evident when contrasting V+S and V+A: several models (e.g., Gemini 2.5-FL, EgoGPT, VITA 1.5, and Qwen 2.5 Omni) score lower with V+A than with V+S, indicating missed auditory details or limitations in audio-video fusion.

Training-Recipe Variants: SFT, RL, and Data Composition. We assess open-source recipes built on identical backbones to isolate curriculum effects. Omni-modal models such as Ola and Qwen 2.5 Omni use an image-text warmup followed by separate alignment for the audio modality, and they typically improve on audiovisual evaluations; reinforcement learning appears to further increase the multimodal gains. VITA 1.5, on the other hand, emphasizes alignment to the video modality during training, which may bias attention toward the visual stream and results in regression when subtitles and audio are added.

Video-text models follow a more streamlined path: initial pretraining on image-caption pairs to align images with text, then instruction tuning or long chain-of-thought data for fine-tuning. Aside from minor architectural differences, they vary primarily in data curation and sources. Qwen 2.5 VL uniquely includes chain-of-thought data during fine-tuning to encourage explicit reasoning and stronger multimodal synthesis. While its overall performance is strong, the relative gain from multimodal inputs appears similar to its counterpart without chain-of-thought fine-tuning, which may point to reward hacking during SFT and warrants further study.

Agentic Ability in Comparison with Humans. Tab. 5 indicates that humans perform best on social and egocentric questions, with slightly lower accuracy on gaming and sports that demand domain knowledge. Model behavior is less uniform. Gemini, Qwen 2.5 Omni, and VITA 1.5 tend to be weaker on egocentric videos and comparatively stronger on domain-specific categories, while the remaining models show different per-category strengths.

Across audio categories, human accuracy is largely stable. Models trained on broader multimodal corpora such as Qwen 2.5 Omni and the Gemini family exhibit smaller fluctuations across audio types, yet many systems underperform when music dominates. EgoGPT shows relatively strong auditory understanding, likely reflecting its use of a pretrained Whisper encoder. In contrast, Ola and VITA 1.5 that trained on smaller datasets, display larger variance across categories, with notable drops on the music category.

Taken together, these patterns suggest that coverage of first-person content and diverse audio during training may be as important as scale for robust agentic ability across domains.

Temporal Horizons: Long- versus Short-Clip Performance. As shown in Fig. 4, we evaluate models on short, pre-localized audiovisual clips in MAVERIX and on their full-length counterparts in MAVERIX-Long. Across models, localized clips yield higher accuracy. Among the agen-

Table 3: Multimodal gains across models on MCQs (measured in % accuracy). A, V, and S denote the Audio, Video, and Subtitle modalities, respectively. *Best-Uni* = $\max(A, S, V)$; *Best-Multi* = $\max(V+S, V+A)$; ΔMulti = *Best-Multi* – *Best-Uni*.

Model	Audio Support	Size	Arch	Recipe	Unimodal Analysis				Multimodal Performance			
					A	S	V	Best-Uni	V+S	V+A	Best-Multi	ΔMulti
Human	–	–	–	–	44.3	41.7	81.4	81.4	86.4	92.8	92.8	+11.4
EgoGPT-7B	✓	7B	Dual-Tower	SFT	29.9	43.0	45.2	45.2	55.0	45.2	55.0	+9.8
Ola-7B	✓	7B	Tri-Tower	SFT	49.4	43.9	37.6	49.5	49.6	53.1	53.1	+3.6
VITA 1.5	✓	7B	Dual-Tower	SFT	32.4	43.4	20.2	43.5	22.3	18.5	22.3	-21.2
Qwen 2.5 Omni	✓	7B	Dual-Tower	SFT+RL	46.5	41.4	35.4	46.5	57.9	49.5	57.9	+11.4
Qwen-2-VL	7B	ViT-LLM	SFT	–	43.0	48.0	48.0	57.5	–	57.5	57.5	+9.5
Qwen-2.5-VL	7B	ViT-LLM	SFT	–	40.3	46.9	46.9	55.3	–	55.3	55.3	+8.4
InternVL2	8B	ViT-LLM	SFT	–	24.1	26.3	26.3	33.1	–	33.1	33.1	+6.8
LLaVA-OneVision	7B	SigLIP-LLM	SFT	–	44.5	46.8	46.8	55.6	–	55.6	55.6	+8.8
DeepSeekVL2-small	2.8B	Hybrid Enc.-MoE	SFT	–	34.3	33.2	34.3	42.4	–	42.4	42.4	+8.1
Gemini 2.0-FL	✓	–	–	–	43.8	38.0	42.1	43.8	41.1	50.2	50.2	+6.4
Gemini 2.5-FL	✓	–	–	–	44.8	47.7	48.8	48.8	56.7	54.7	56.7	+7.9
Claude Sonnet 3.5	–	–	–	–	–	55.0	42.0	55.0	64.1	–	64.1	+9.1
GPT-4o	–	–	–	–	–	55.3	54.3	55.3	64.0	–	64.0	+8.7
Grok 4	–	–	–	–	–	41.8	54.5	54.5	59.4	–	59.4	+4.9
GPT-4o-mini	–	–	–	–	–	45.4	35.5	45.4	50.0	–	50.0	+4.6
NOVA-Lite	–	–	–	–	–	40.4	39.7	40.4	51.0	–	51.0	+10.6
NOVA-Pro	–	–	–	–	–	46.6	45.4	46.6	55.8	–	55.8	+9.2

Table 4: Difficulty-wise MCQ accuracy (%) for audio-enabled models. AV cells show Δ vs A.

Model	Easy			Medium			Hard					
	A	V	AV	A	V	AV	A	V	AV			
Human	46.4	84.7	93.4	47.0†	44.8	81.4	92.5	47.7†	38.5	73.9	92.1	53.6†
EgoGPT	29.4	50.2	50.2	20.8†	32.3	44.3	44.3	12.0†	25.2	36.4	36.4	11.2†
Ola-7B	54.1	36.9	57.1	3.0†	48.4	41.0	54.1	5.7†	41.7	30.5	41.1	0.6†
VITA 1.5	33.0	21.3	20.1	12.9†	34.8	19.8	18.5	16.3†	25.2	18.5	14.6	10.6†
Qwen-2-Omni	50.6	39.9	52.1	1.5†	44.6	34.8	48.8	4.2†	41.5	25.9	45.2	3.7†
Gemini 2.0-FL	43.8	45.3	57.7	13.9†	44.0	42.7	48.9	4.9†	32.5	33.1	36.4	3.9†
Gemini 2.5-FL	47.7	48.3	59.8	12.1†	44.8	53.5	54.9	10.1†	37.7	37.7	43.0	5.3†

tic categories, the questions that depend on immediate, synchronous audiovisual cues, such as those from factual recall and near-term causal inference, show the smallest degradation. When the relevant segment is pre-localized, models can more reliably extract the necessary information.

By contrast, social relationship, emotion, and situational understanding often rely on fine-grained and sometimes asynchronous cues distributed over time. Performance drops more on long videos, reflecting challenges in localizing these signals and integrating them over extended context. Overall, a gap to human performance remains, especially for longer videos and for recognizing subtle contextual cues. These trends suggest that current MLLMs are stronger at retrieving salient, object or event level signals than at integrating evolving context and social nuance over time.

Model and Human Perception of Difficulty. We analyze performance by difficulty and observe that multimodal inputs often help most on easy items, with smaller gains on hard ones, though trends vary by model. For the audio-enabled models where split statistics are available, Gemini 2.0-FL improves by +12.4% on easy items and +3.3% on hard items, and Gemini 2.5-FL improves by +11.5%

Table 5: Accuracy (%) on AV inputs across taxonomy and audio categories (single-column).

Models	Taxonomy (AV)			
	Social	Ego-centric	Sports	Gaming
Human	92.7	95.2	82.3	74.5
EgoGPT	34.3	45.8	42.9	40.8
Ola	52.5	57.6	52.9	46.6
VITA 1.5	21.2	6.8	16.8	12.6
Qwen 2.5 Omni	47.5	42.4	49.6	41.7
Gemini 2.0-FL	46.5	32.3	48.7	47.6
Gemini 2.5-FL	47.1	39.0	54.3	61.9

Models	Audio Category (AV)				
	Natural	Speech	Music	Artificial	Mixed
Human	88.4	94.1	90.3	89.6	92.8
EgoGPT	41.2	44.2	46.2	48.0	47.3
Ola	48.5	52.0	38.5	58.0	56.5
VITA 1.5	29.4	17.0	7.7	26.0	17.6
Qwen 2.5 Omni	52.9	47.8	53.8	48.0	51.5
Gemini 2.0-FL	47.1	51.1	38.5	54.0	49.6
Gemini 2.5-FL	51.5	56.7	50.0	52.0	53.1

on easy items and +5.3% on hard items. Owen 2.5 Omni shows a different pattern with substantial benefit on hard items as well. These mixed results suggest that current systems leverage straightforward cross-modal cues more reliably than they integrate sparse or subtle signals in harder cases.

Although MAVERIX is designed to elicit cross-modal reasoning, some models still achieve moderate scores with a single modality, likely because many real videos contain aligned audio and visual streams that allow plausible inferences from partial evidence. GPT-4o-mini is one such exam-

Table 6: Open-ended response correctness scores (out of 5) as judged by GPT, reported per modality. For models without native audio support, scores for Subtitle (S) and Subtitle+Video (SV) are shown instead of Audio (A) and Audio+Video (AV). Response token length statistics are analyzed separately.

Model	Audio Support	Judged Score			Token Length			
		A/S	V	AV/SV	Avg	Max	Std	
Human		1.7	2.6	3.4	0.8 [↑]	16.1	140	12.7
EgoGPT	✓	0.7	1.2	1.3	0.1 [↑]	9.4	112	17.8
Ola-7B	✓	1.4	0.9	1.5	0.1 [↑]	17.9	71	15.4
VITA 1.5	✓	0.8	0.7	0.5	0.3 [↓]	55.3	273	29.1
Qwen-2-Omni	✓	1.0	1.2	1.2	0.0 [↓]	47.0	178	26.5
Qwen-2-VL		1.2	1.4	1.6	0.2 [↑]	22.9	72	15.7
Qwen-2.5-VL		1.1	1.4	1.7	0.3 [↑]	52.9	86	17.7
InternVL2		0.9	0.9	1.1	0.2 [↑]	24.0	95	17.6
DeepSeek-VL2		1.2	1.0	1.4	0.2 [↑]	27.4	512	61.0
LLaVA-OneVision		1.3	1.4	1.6	0.2 [↑]	18.1	70	9.3
Gemini 2.0-FL	✓	1.4	1.6	1.9	0.3 [↑]	20.9	103	14.6
Gemini 2.5-FL	✓	1.4	1.4	1.9	0.5 [↑]	32.6	112	20.7
Claude Sonnet 3.5		1.6	1.7	2.2	0.5 [↑]	59.8	95	4.8
GPT-4o		1.6	1.4	2.2	0.6 [↑]	50.8	102	17.9
Grok-4		1.7	2.0	2.4	0.4 [↑]	131.7	11022	322.3
NOVA-Lite		1.1	1.0	1.2	0.1 [↑]	23.5	75	18.4
NOVA-Pro		1.1	1.2	1.5	0.3 [↑]	35.4	78	20.9

ple of respectable unimodal performance. Humans also benefit from aligned cues and strong priors, yet the jump from 81.4% with video-only to 92.8% with audiovisual highlights the value of genuine cross-modal understanding and sets a clear target for future modeling. We hope continued progress in cross-modal alignment will narrow this gap and eventually surpass the current human baseline.

Related Work

MLLM Benchmarks. Early vision-language benchmarks centered on images for grounding and recognition, using captions and QA pairs (Chen et al. 2015; Agrawal et al. 2019; Zhu et al. 2016; Plummer et al. 2015), followed by domain-specific and knowledge-intensive settings (Saikh

et al. 2022; Lu et al. 2023; Singh et al. 2019). A-OKVQA targets external-knowledge reasoning beyond visible content (Schwenk et al. 2022). More recent efforts, including MMMU and MMMU-Pro, broaden question diversity and reading-from-image skills (Yue et al. 2024a,b).

Image-only evaluation lacks temporal and acoustic context, motivating video benchmarks that probe motion, sequence, and temporal localization (Li et al. 2024b; Patraucean et al. 2023; Kesen et al. 2023; Huang et al. 2020; Song et al. 2024a; Maaz et al. 2024; Fang et al. 2025; Li et al. 2024c; Ning et al. 2023; Chen et al. 2024a; He et al. 2024; Mangalam, Akshulakov, and Malik 2023; Song et al. 2024b). However, most emphasize short clips and constrained domains, rely on MCQ-only protocols, and provide limited coverage of everyday social or situational reasoning. Video-MME and AV-Odyssey scale video duration but remain MCQ-only, omitting open-ended assessment (Fu et al. 2024; Gong et al. 2024). *MAVERIX* elevates audio as one of the primary signals and stresses cross-modal integration as it evaluates both 8-way MCQs and open-ended responses to measure multimodal synthesis under realistic audiovisual conditions.

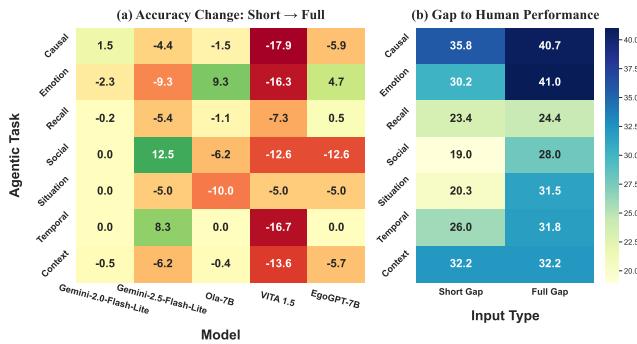
Video Understanding Models. Contrastive pretraining on image-text data yields transferable representations and has been adapted to spatio-temporal reasoning; post-training with instruction tuning and RLHF further aligns models with human preferences (Sun et al. 2023; Zhai et al. 2024; Lin et al. 2023; Wang et al. 2024; Bai et al. 2025; Zhang 2024). MoE-style routing improves scalability without linear cost growth (Wu et al. 2024; Lin et al. 2024; DeepSeek-AI and Aixin Liu 2025; Sun, Chen, and Yiqing Huang 2024; Cai et al. 2024; Liu et al. 2025a). Yet whether these advances enable human-comparable multimodal reasoning in real-world audiovisual settings remains open (tes 2024; Amirizaniani et al. 2024; Campbell et al. 2024; Kazemi et al. 2024). Our evaluations on *MAVERIX* show a substantial gap to human accuracy on MCQs, often on the order of several tens of percentage points, and highlight persistent challenges in integrating temporal, social, and auditory cues in models with different architecture and training recipes.

Conclusion

Agentic scenarios such as assisting collaborative work and navigating dynamic environments require strong audiovisual reasoning, yet these abilities remain under-assessed in recent MLLMs. We introduce **MAVERIX**, a benchmark for complex, real-world audiovisual understanding, comprising 700 videos and 2,556 carefully crafted, human-authored questions. The suite evaluates models with both 8-way multiple-choice and open-ended responses.

Our results indicate that multimodal inputs generally improve accuracy, but sizable gaps to human performance persist, especially for socially grounded or dynamic scenarios. Models benefit most when relevant segments are pre-localized and tend to struggle on longer videos that demand integrating subtle, asynchronous cues over time; robust audio integration also remains uneven across systems. We hope *MAVERIX* will guide progress toward stronger

Figure 4: Impact of Video Length on Agentic Category Performances. (a) Accuracy change (%) from short to full-length videos across models and questions. (b) Accuracy gap (%) to human performance for short and full inputs.



cross-modal alignment, better temporal reasoning, and more context-aware, socially intelligent models.

References

???? Hugging Face Hub. <https://huggingface.co/>. Accessed: 2025-08-04.

2005. Amazon Mechanical Turk. <https://www.mturk.com/>. Crowd-sourcing platform service.

2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7): 1285–1295. Publisher Copyright: © The Author(s) 2024.

Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; and Vijayanarasimhan, S. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. In *ArXiv*.

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8948–8957.

Amirizaniani, M.; Martin, E.; Sivachenko, M.; Mashhadni, A.; and Shah, C. 2024. Can LLMs Reason Like Humans? Assessing Theory of Mind Reasoning in LLMs for Open-Ended Questions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, 34–44. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704369.

Anthropic. 2024. Introducing Claude 3.5 Sonnet \ Anthropic. [Online; accessed 2025-03-06].

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Cai, R.; Muralidharan, S.; Heinrich, G.; Yin, H.; Wang, Z.; Kautz, J.; and Molchanov, P. 2024. Flextron: Many-in-One Flexible Large Language Model. *arXiv:2406.10260*.

Campbell, D.; Rane, S.; Gialanza, T.; Sabbata, N. D.; Ghods, K.; Joshi, A.; Ku, A.; Frankland, S. M.; Griffiths, T. L.; Cohen, J. D.; and Webb, T. W. 2024. Understanding the Limits of Vision Language Models Through the Lens of the Binding Problem. *arXiv:2411.00238*.

Chandrasegaran, K.; Gupta, A.; Hadzic, L. M.; Kota, T.; He, J.; Eyzaguirre, C.; Durante, Z.; Li, M.; Wu, J.; and Li, F.-F. 2024. HourVideo: 1-Hour Video-Language Understanding. In *Advances in Neural Information Processing Systems*, volume 37.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Chen, X.; Lin, Y.; Zhang, Y.; and Huang, W. 2024a. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*, 179–195. Springer.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024b. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *arXiv preprint arXiv:2404.16821*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

DeepSeek-AI; and Aixin Liu, e. a. 2025. DeepSeek-V3 Technical Report. *arXiv:2412.19437*.

Fang, X.; Mao, K.; Duan, H.; Zhao, X.; Li, Y.; Lin, D.; and Chen, K. 2025. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37: 89098–89124.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Fu, C.; Lin, H.; Wang, X.; Zhang, Y.-F.; Shen, Y.; Liu, X.; Cao, H.; Long, Z.; Gao, H.; Li, K.; Ma, L.; Zheng, X.; Ji, R.; Sun, X.; Shan, C.; and He, R. 2025. VITA-1.5: Towards GPT-4o Level Real-Time Vision and Speech Interaction. *arXiv:2501.01957*.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*. New Orleans, LA.

Gong, K.; Feng, K.; Li, B.; Wang, Y.; Cheng, M.; Yang, S.; Han, J.; Wang, B.; Bai, Y.; Yang, Z.; and Yue, X. 2024. AV-Odyssey Bench: Can Your Multimodal LLMs Really Understand Audio-Visual Information? *arXiv:2412.02611*.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; and Antonino Furnari, e. a. 2021. Ego4D: Around the World in 3, 000 Hours of Egocentric Video. *CoRR*, abs/2110.07058.

Hasan, M. K.; Rahman, W.; Bagher Zadeh, A.; Zhong, J.; Tanveer, M. I.; Morency, L.-P.; and Hoque, M. E. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2046–2056. Hong Kong, China: Association for Computational Linguistics.

He, X.; Feng, W.; Zheng, K.; Lu, Y.; Zhu, W.; Li, J.; Fan, Y.; Wang, J.; Li, L.; Yang, Z.; et al. 2024. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*.

Hong, J.; Yan, S.; Cai, J.; Jiang, X.; Hu, Y.; and Xie, W. 2025. WorldSense: Evaluating Real-world Omnimodal Understanding for Multimodal LLMs. *arXiv:2502.04326*.

Huang, Q.; Xiong, Y.; Rao, A.; Wang, J.; and Lin, D. 2020. MovieNet: A Holistic Dataset for Movie Understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Intelligence, A. A. G. 2024. The Amazon Nova family of models: Technical report and model card. *Amazon Technical Reports*.

Kazemi, M.; Dikkala, N.; Anand, A.; Devic, P.; Dasgupta, I.; Liu, F.; Fatemi, B.; Awasthi, P.; Guo, D.; Gollapudi, S.; and Qureshi, A. 2024. ReMI: A Dataset for Reasoning with Multiple Images. *arXiv:2406.09175*.

Kesen, I.; Pedrotti, A.; Dogan, M.; Cafagna, M.; Acikgoz, E. C.; Parcalabescu, L.; Calixto, I.; Frank, A.; Gatt, A.; Erdem, A.; et al. 2023. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. *arXiv preprint arXiv:2311.07022*.

Li*, B.; Zhang*, P.; Zhang*, K.; Pu*, F.; Du, X.; Dong, Y.; Liu, H.; Zhang, Y.; Zhang, G.; Li, C.; and Liu, Z. 2024. LMMs-Eval: Accelerating the Development of Large Multimoal Models.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024a. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*.

Li, D.; Sun, R.; Huang, Y.; Zhong, M.; Jiang, B.; Han, J.; Zhang, X.; Wang, W.; and Liu, H. 2025a. Preference Leakage: A Contamination Problem in LLM-as-a-judge. *arXiv:2502.01534*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Li, S.; Li, L.; Liu, Y.; Ren, S.; Liu, Y.; Gao, R.; Sun, X.; and Hou, L. 2024c. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, 331–348. Springer.

Li, Y.; Zhang, G.; Ma, Y.; Yuan, R.; Zhu, K.; Guo, H.; Liang, Y.; Liu, J.; Wang, Z.; Yang, J.; Wu, S.; Qu, X.; Shi, J.; Zhang, X.; Yang, Z.; Wang, X.; Zhang, Z.; Liu, Z.; Benetos, E.; Huang, W.; and Lin, C. 2025b. OmniBench: Towards The Future of Universal Omni-Language Models. arXiv:2409.15272.

Lin, B.; Tang, Z.; Ye, Y.; Cui, J.; Zhu, B.; Jin, P.; Zhang, J.; Ning, M.; and Yuan, L. 2024. MoE-LLaVA: Mixture of Experts for Large Vision-Language Models. *arXiv preprint arXiv:2401.15947*.

Lin, J.; Du, Y.; Watkins, O.; Hafner, D.; Abbeel, P.; Klein, D.; and Dragan, A. 2023. Learning to Model the World with Language.

Liu, J.; Tang, P.; Wang, W.; Ren, Y.; Hou, X.; Heng, P.-A.; Guo, M.; and Li, C. 2025a. A Survey on Inference Optimization Techniques for Mixture of Experts Models. arXiv:2412.14219.

Liu, Z.; Dong, Y.; Wang, J.; Liu, Z.; Hu, W.; Lu, J.; and Rao, Y. 2025b. Ola: Pushing the Frontiers of Omni-Modal Language Model. arXiv:2502.04328.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

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

Ning, M.; Zhu, B.; Xie, Y.; Lin, B.; Cui, J.; Yuan, L.; Chen, D.; and Yuan, L. 2023. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.

OpenAI; and Aaron Hurst, e. a. 2024. GPT-4o System Card. arXiv:2410.21276.

Patraucean, V.; Smaira, L.; Gupta, A.; Recasens, A.; Markeeva, L.; Banarse, D.; Koppula, S.; Malinowski, M.; Yang, Y.; Doersch, C.; et al. 2023. Perception test: A diagnostic benchmark for multi-modal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2641–2649.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavy, C.; and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint arXiv:2212.04356*.

Saikh, T.; Ghosal, T.; Mittal, A.; Ekbal, A.; and Bhattacharyya, P. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3): 289–301.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, 146–162. Springer.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; Lu, Y.; Hwang, J.-N.; and Wang, G. 2024a. MovieChat: From Dense Token to Sparse Memory for Long Video Understanding. arXiv:2307.16449.

Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. 2024b. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.

Sun, X.; Chen, Y.; and Yiqing Huang, e. a. 2024. Hunyuan-Large: An Open-Source MoE Model with 52 Billion Activated Parameters by Tencent. arXiv:2411.02265.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; Keutzer, K.; and Darrell, T. 2023. Aligning Large Multimodal Models with Factually Augmented RLHF.

Team, G.; and Petko Georgiev, e. a. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530.

Team, G.; and Petko Georgiev, e. a. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; Xie, Z.; Wu, Y.; Hu, K.; Wang, J.; Sun, Y.; Li, Y.; Piao, Y.; Guan, K.; Liu, A.; Xie, X.; You, Y.; Dong, K.; Yu, X.; Zhang, H.; Zhao, L.; Wang, Y.; and Ruan, C. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. arXiv:2412.10302.

xAI. 2025. Bringing Grok to Everyone. [Online; accessed 2025-07-30].

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; Zhang, B.; Wang, X.; Chu, Y.; and Lin, J. 2025. Qwen2.5-Omni Technical Report. arXiv:2503.20215.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5288–5296.

Yang, J.; Liu, S.; Guo, H.; Dong, Y.; Zhang, X.; Zhang, S.; Wang, P.; Zhou, Z.; Xie, B.; Wang, Z.; Ouyang, B.; Lin, Z.; Cominelli, M.; Cai, Z.; Zhang, Y.; Zhang, P.; Hong, F.; Widmer, J.; Gringoli, F.; Yang, L.; Li, B.; and Liu, Z. 2025. EgoLife: Towards Egocentric Life Assistant. arXiv:2503.03803.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024a. 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*, 9556–9567.

Yue, X.; Zheng, T.; Ni, Y.; Wang, Y.; Zhang, K.; Tong, S.; Sun, Y.; Yu, B.; Zhang, G.; Sun, H.; et al. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Zhai, Y.; Bai, H.; Lin, Z.; Pan, J.; Tong, S.; Zhou, Y.; Suhr, A.; Xie, S.; LeCun, Y.; Ma, Y.; and Levine, S. 2024. Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhang, G. 2024. VL-RLHF: A RLHF Infrastructure for Vision-Language Model. <https://github.com/TideDra/VL-RLHF>.

Zhang, K.; Li, B.; Zhang, P.; Pu, F.; Cahyono, J. A.; Hu, K.; Liu, S.; Zhang, Y.; Yang, J.; Li, C.; and Liu, Z. 2024. LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models. *arXiv:2407.12772*.

Zhu, Y.; Groth, O.; Bernstein, M.; and Fei-Fei, L. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4995–5004.

MAVERIX Supplementary Material

- Section A: MAVERIX Release v1.0
- Section B: Rationale and Limitations
- Section C: Additional Experiments
 - Section C.1: Standardized Prompt Design
 - Section C.2: GPT-4o as Judge Criteria
 - Section C.3: Human Performance by Categories
 - Section C.4: Quantitative GPT Evaluations
 - Section C.5: Performance by Categories
 - Section C.6: Common Error Modes
- Section D: Details on the Human Survey
- Section E: Dataset Distribution Continued

A. MAVERIX Release v1.0

We are releasing **MAVERIX v1.0**, our proposed benchmark dataset for multimodal video-language understanding, built with support to a common benchmark platform LMMs-Eval (Li* et al. 2024). The benchmark dataset is provided as a single JSON file for ease of use and straightforward integration with existing benchmarking pipelines. For each video, the dataset includes metadata and contains multiple-choice and open-ended questions covering multiple tasks from our proposed task suite. Each task is accompanied by a set of questions designed to evaluate visual reasoning, situational awareness, and cross-modal understanding. Relevant timestamps are provided to allow precise video trimming.

We will provide all 1400 video clips used in our benchmark through HuggingFace (hug), with both localized and full length videos, and users can download the corresponding videos from the Ego4D website after reviewing and accepting the Ego4D license agreement. We also provide sample videos with annotations from MAVERIX. All materials are available through the project page upon acceptance. The code snippets providing main input prompts, response aggregation, judge response generations are provided as part of the supplement materials, with unit test functions for testing data corruption included.

Structure: MAVERIX v1.0 release is organized as follows:

• **data/**

- MAVERIX_v1_0.json: Contains all 2556 questions in the benchmark dataset.
- vid_only/: Contains all muted videos for evaluations without audio modality.
- vid/: Contains all videos for evaluations with both video and audio modality.
- audio/: Contains all audio files for evaluations with audio modality.
- subtitles/: Contains all subtitles for evaluations requiring subtitle access.
- sample_annotations/: Given that **MAVERIX** is an evaluation benchmark, ground truth annotations are not released to the public. For review purposes, we provide ground truth annotations for select sample videos in an .json file.

• **src/**

- maverix_benchmark.py: A script for dataloading, processing, and evaluation functionalities.
- _default_template.yaml: A file to specify the default processing template for the evaluation. The data loading path, cache directory, and evaluation modes can be specified here.
- maverix_<mode>.yaml: Individual setting files specifying the test settings for different modes. The currently supported modes are: audio_only, video_only, sub_only, video_sub, and video_audio.

– **Remark:** Except for the open-sourced models listed in the paper or available through LMMs-Eval(Zhang et al. 2024; Li* et al. 2024), all other experiments require access to proprietary models.

Documentation:

- We provide a comprehensive datasheet explaining the benchmark datasets purpose and intended usage.

License:

- MAVERIX will be made publicly available under the MIT License. Do note that Ego4D videos are publicly available under the Ego4D License.

Versioning and Updates:

- We will maintain MAVERIX, with all updates and new versions announced publicly.

B. Rationale and limitations

Broader Impact. Robust audiovisual understanding is critical for the next generation of agentic systems that interact with humans and environments in real-time. Benchmarks like MAVERIX can help guide the development of models capable of supporting complex tasks such as collaborative decision-making, social interaction understanding, embodied navigation, and context-aware assistance. By evaluating models on realistic, multimodal scenarios, MAVERIX encourages progress toward AI systems that can reason beyond static or unimodal inputs, making them better suited for applications like personal assistants, robotics, and human-computer interaction.

However, improved performance on MAVERIX may not directly translate to safe or ethical deployment. Many categories in MAVERIX, such as social sentiment and situational awareness, involve subjective interpretation and context-sensitive judgment. Without careful handling, these capabilities risk reinforcing biases, misunderstanding cultural cues, or misinterpreting social dynamics, especially when applied across diverse real-world contexts. We encourage users of MAVERIX to consider these risks and to accompany performance improvements with a thorough evaluation of fairness, robustness, and societal impact.

Limitations. While MAVERIX advances the evaluation of multimodal reasoning, it has several limitations. First, the benchmark focuses on short- to medium-length video segments, which may not fully capture the long-horizon dependencies found in extended real-world interactions. Second, although MAVERIX emphasizes modality interdependence, complete isolation of modalities is challenging, and some tasks may still be partially solvable through strong unimodal priors. Third, certain categories, such as social interaction, show lower human agreement due to the inherent subjectivity of emotional and interpersonal interpretation, which may limit the upper bound of achievable accuracy. Finally, while MAVERIX supports open-ended evaluation using GPT-4o as an automatic judge, automated scoring of generative responses remains imperfect and may introduce noise or bias in assessments.

Despite these limitations, MAVERIX offers a step toward a more comprehensive and realistic evaluation of multimodal LLMs, with the goal of driving future work in building models capable of deeper, contextually grounded, and socially aware reasoning.

C. Additional experiments

This section presents extended experimental results to further analyze model performance across various dimensions of the benchmark. In addition to model evaluation, we provide human performance baselines segmented by difficulty levels, input modalities, and agentic task categories for direct comparison.

C.1 Standardized Prompt Design

To ensure fairness in evaluation, we adopt a standardized prompt format across all models. This eliminates potential biases introduced by prompt engineering and ensures that differences in performance are attributed to model capabilities rather than variations in input phrasing. For tasks requiring multiple modalities, the prompts incorporate the modalities in the inputs. In multiple-choice settings, the prompt is presented as follows:

```
[images] + [audio] (if available) This
video's subtitles are listed below:
OR No subtitles available. [Subtitles]
Select the best answer to the following
multiple-choice question based on the
video and the subtitles. Respond with
only the letter (A, B, C, D, E, F, G, or
H) of the correct option. [Question] The
best answer is:
```

For open-ended questions, the following prompt is used:

```
[images] + [audio] (if available) This
video's subtitles are listed below:
OR No subtitles available. [Subtitles]
Select the best answer to the following
open-ended question based on the video
and the subtitles. [Question] The best
answer is:
```

C.2 GPT-4o as Judge Criteria

For evaluating the open-ended answer qualities, we divide the evaluations in 4 different aspects: correctness, level of details, context understanding, and temporal consistency. We prompt the judge model as follows for correctness:

```
1  [
2    {
3      "role": "system",
4      "content": "You are an
→ intelligent chatbot designed
→ for evaluating the factual
→ accuracy of generative outputs
→ for video-based question-answer
→ pairs. Your task is to compare
→ the predicted answer with the
→ correct answer and judge
→ factual consistency.\n
→ nINSTRUCTIONS:\n- Focus on
→ factual consistency; avoid
→ misinterpretations or
→ misinformation.\n- Ensure the
→ prediction aligns with the
→ video's content.\n- Treat
→ synonyms or paraphrases as
→ valid.\n- Output a single score
→ reflecting factual accuracy."
5    },
6    {
7      "role": "user",
8      "content": "Please evaluate the
→ following video-based question-
→ answer pair:\n\nQuestion:
→ QUESTION \nCorrect Answer:
→ REFERENCE_ANSWER \nPredicted
→ Answer: MODEL_ANSWER\nReturn
→ only a Python dict string like
→ {'score': 4}, where 'score' is
→ an INTEGER from 0 (lowest) to 5
→ (highest). Do not include any
→ other text."
9    }
10 ]
```

The template prompt for level of detail orientation:

```
1  [
2    {
3      "role": "system",
4      "content": "You are an
→ intelligent chatbot designed
→ for evaluating the detail
→ orientation of generative
→ outputs for video-based
→ question-answer pairs. Your
→ task is to compare the
→ predicted answer with the
→ correct answer and judge its
→ level of detail, considering
→ both completeness and
→ specificity.\n\nINSTRUCTIONS:
→ - Check whether the prediction
→ covers all major points; do not
```

```

    ↵ omit key aspects.\n- Prefer
    ↵ specific, grounded details over
    ↵ generic statements; tie
    ↵ information to concrete
    ↵ elements of the video.\n- Treat
    ↵ synonyms or paraphrases as
    ↵ valid.\n- Output a single score
    ↵ reflecting overall detail
    ↵ orientation (completeness +
    ↵ specificity)."
5  },
6  {
7    "role": "user",
8    "content": "Please evaluate the
    ↵ following video-based question-
    ↵ answer pair:\n\nQuestion:
    ↵ QUESTION\nCorrect Answer:
    ↵ REFERENCE_ANSWER\nPredicted
    ↵ Answer: MODEL_ANSWER\n\nReturn
    ↵ only a Python dict string like
    ↵ {'score': 4}, where 'score' is
    ↵ an INTEGER from 0 (lowest
    ↵ alignment) to 5 (highest
    ↵ alignment). Do not include any
    ↵ other text."
9  }
10 ]

```

The template prompt for level of context understanding:

```

1  [
2  {
3    "role": "system",
4    "content": "You are an
    ↵ intelligent chatbot designed
    ↵ for evaluating the contextual
    ↵ understanding of generative
    ↵ outputs for video-based
    ↵ question-answer pairs. Your
    ↵ task is to compare the
    ↵ predicted answer with the
    ↵ correct answer and judge
    ↵ whether the response aligns
    ↵ with the overall context of the
    ↵ video.\n\nINSTRUCTIONS:\n-
    ↵ Check that the prediction
    ↵ aligns with the video's context
    ↵ ; it should not introduce out-
    ↵ of-context or contradictory
    ↵ information.\n- Ensure the
    ↵ prediction captures the main
    ↵ themes and sentiments of the
    ↵ video.\n- Treat synonyms or
    ↵ paraphrases as valid.\n- Output
    ↵ a single score reflecting
    ↵ overall contextual
    ↵ understanding."
5  },
6  {
7    "role": "user",
8    "content": "Please evaluate the
    ↵ following video-based question-
    ↵ answer pair:\n\nQuestion:
    ↵ QUESTION\nCorrect Answer:
    ↵ REFERENCE_ANSWER\nPredicted

```

```

    ↵ Answer: MODEL_ANSWER\n\nReturn
    ↵ only a Python dict string like
    ↵ {'score': 4}, where 'score' is
    ↵ an INTEGER from 0 (lowest
    ↵ alignment) to 5 (highest
    ↵ alignment). Do not include any
    ↵ other text."
9  }
10 ]

```

The template prompt for level of temporal consistency:

```

1  [
2  {
3    "role": "system",
4    "content": "You are an
    ↵ intelligent chatbot designed
    ↵ for evaluating the temporal
    ↵ understanding of generative
    ↵ outputs for video-based
    ↵ question-answer pairs. Your
    ↵ task is to compare the
    ↵ predicted answer with the
    ↵ correct answer and judge
    ↵ whether the response preserves
    ↵ the temporal sequence of events
    ↵ .\n\nINSTRUCTIONS:\n-
    ↵ Verify
    ↵ temporal consistency; the
    ↵ prediction should reflect the
    ↵ order and timing of events as
    ↵ presented in the video.\n-
    ↵ Accept synonyms or paraphrases
    ↵ only if the temporal order is
    ↵ maintained.\n- Output a single
    ↵ score reflecting overall
    ↵ temporal accuracy."
5  },
6  {
7    "role": "user",
8    "content": "Please evaluate the
    ↵ following video-based question-
    ↵ answer pair:\n\nQuestion:
    ↵ QUESTION\nCorrect Answer:
    ↵ REFERENCE_ANSWER\nPredicted
    ↵ Answer: MODEL_ANSWER\n\nReturn
    ↵ only a Python dict string like
    ↵ {'score': 4}, where 'score' is
    ↵ an INTEGER from 0 (lowest
    ↵ temporal consistency) to 5 (
    ↵ highest). Do not include any
    ↵ other text."
9  }
10 ]

```

C.3 Human Performance by Categories

We report performance across distinct agentic task categories in Tab. 7 to showcase the strengths and weaknesses of the human study participants within the defined contexts. We note that these statistics may be biased due to the composition of the participant pool, and further details on the recruitment process can be found in the Questionnaire Survey section.

Modality	Task	Worst (%)	Best (%)	Average (%)
Visual	Social	66.71	94.72	79.31
	Sports	28.91	93.84	74.34
	Information Querying	77.84	95.67	89.83
	Sentiment	28.91	92.43	68.96
	Shopping	51.62	94.11	83.37
	Gaming	54.20	96.34	84.18
	Overall	28.91	96.34	80
Audio	Social	22.10	53.04	32.05
	Sports	26.52	56.83	44.51
	Information Querying	30.60	48.63	44.12
	Sentiment	28.42	53.04	46.89
	Shopping	30.60	79.56	58.48
	Gaming	33.15	33.15	33.15
	Overall	22.10	79.56	43.2
Both (A+V)	Social	78	95.81	92.86
	Sports	50.15	97.20	94.71
	Information Querying	85.80	94.91	94.77
	Sentiment	96.28	96.70	96.87
	Shopping	70.20	95.17	93.60
	Gaming	66.85	93.42	83.22
	Overall	50.15	97.20	92.67

Table 7: Worst, Best, and Average human performance (%) across tasks and modalities.

C.4 Quantitative GPT Evaluations

To further assess model performance beyond multiple-choice accuracy, we conduct a quantitative evaluation of the open-ended question responses using GPT-based scoring. Specifically, as discussed in the paper, we prompt GPT-4o to act as an automated grader, evaluating the correctness and relevance of the model’s free-form answers based on predefined guidelines consistent with the task objectives in Tab. 8 and Tab. 9.

This automatic assessment provides an additional perspective on the model’s generative capabilities, capturing nuances that are not reflected in standard accuracy metrics. The results offer insight into how well the model can produce contextually appropriate and informative responses when not constrained by fixed answer options.

We report the GPT-assigned scores across the full benchmark as well as per agentic task category, enabling a more granular understanding of the model’s strengths and weaknesses in open-ended scenarios. These findings complement the multiple-choice evaluations and expose areas where the model demonstrates strong language generation and areas where further improvement is needed. However, we note that these scores may be biased due to limitations in GPT-4o’s own reasoning and evaluation capabilities (Li et al. 2025a), which can introduce alignment artifacts or systematic preferences.

C.5 Quantitative Evaluations by Categories

To provide a comprehensive understanding of model performance across the diverse challenge dimensions of the benchmark, we report quantitative results for all probed models across each defined split. This includes evaluations broken down by agentic task categories, difficulty levels, audio characteristics, topics, and input modalities, as shown in

Tab. 10.

These category-wise results allow us to examine how different models handle specific functional scenarios, content types, and levels of complexity. Through these experiments, we hope to reveal both general patterns and areas of strength or weakness. By comparing performance across these splits, we observe how well models generalize to varying real-world situations and identify which aspects of the benchmark remain most challenging.

Together, these fine-grained evaluations provide a detailed view of model behavior across the full spectrum of tasks and shed light on future work on multimodal understanding and generalization.

C.6 Common Error Modes

To better understand the limitations of the state of the art models, we analyze common error patterns observed across the benchmark. We find that model failures often arise from modality-specific weaknesses, such as misinterpretation of complex visual cues, inability to process nuanced audio signals, or challenges in aligning multimodal information with nuanced context. For illustration, we highlight representative failure cases from OpenAI 01 and Gemini 1.5 Pro, chosen because they are among the strongest available models; however, the same error patterns recur across the models we evaluated. Common failure modes are showcased in Fig. 5, 6, 7, 8, 9, 10.

D. Questionnaire Survey

To establish human performance baselines for comparison with the probed models, we conducted a questionnaire study approved by our institutional IRB. We recruited participants with proficiency in at least one language present in the dataset and prior familiarity with video content to ensure they could reasonably complete the tasks. Participants answered questions presented in the same format as those used for model evaluation.

The survey included a total of 261 questions curated from the dataset, covering diverse subcategories within each defined split. For each participant, the questionnaire consisted of either one multiple-choice question (MCQ) with eight answer options, or an open-ended question from a different video. To minimize content overlap and reduce potential bias, no video was reused between questions within a single questionnaire. This approach ensured broad coverage of the dataset’s taxonomy while avoiding contamination across tasks. We recruited 382 participants through Amazon Mechanical Turk service with approval rate >96% to answer the questions with A, S, V, A+V, S+V as inputs with media from MAVERIX and MAVERIX-Long. Both MCQ responses and open-ended responses are recorded.

For evaluation, MCQ responses were scored against the ground truth to calculate accuracy, while open-ended answers were assessed using the same GPT-4o evaluation pipeline applied to model outputs, ensuring consistency across human and model scoring.

We tested human performance across three different conditions: audio only, visual only, and audio plus video. Each

Model	Temporal Consistency			Correctness		
	Video	Audio/Sub	V + A/Sub	Video	Audio/Sub	V + A/Sub
Human	1.85	1.69	2.83	1.93	1.46	2.65
<i>Open-source MLLMs</i>						
EgoGPT-7B	1.3955	1.2523	1.5634	1.2477	1.1843	1.4930
Ola-7B	0.9178	1.2500	1.3638	0.9038	1.2171	1.3005
VITA 1.5	0.6643	1.2160	0.7042	0.6561	1.0587	0.7007
Qwen-2-VL	1.4554	1.3638	1.6772	1.3498	1.2254	1.6092
Qwen-2.5-VL	1.6819	1.2793	1.9730	1.4014	1.1279	1.7477
Qwen-2.5-Omni	1.6819	1.2793	1.9730	1.4014	1.1279	1.7477
InternVL2	0.9742	1.3169	1.1925	0.9237	0.9143	1.1315
LLaVA-OneVision	1.5962	1.3685	1.8052	1.4002	1.2899	1.5763
DeepSeekVL2-small	0.9507	1.3157	1.3286	0.9953	1.1866	1.3650
<i>Proprietary MLLMs</i>						
Gemini 2.0-FL	1.6526	1.5880	2.1784	1.4296	1.4390	1.9014
Gemini 2.5-FL	1.7864	1.6244	2.1678	1.5751	1.4507	1.9167
Claude Sonnet 3.5	2.0798	1.9155	2.6009	1.6538	1.5857	2.2300
GPT-4o	1.5434	1.8462	2.4390	1.4167	1.5634	2.2207
GPT-4o-mini	1.9859	1.9272	2.3850	1.5739	1.6185	2.0117
Grok 4	2.2782	2.1854	2.8028	1.9507	1.7066	2.4354
NOVA-Lite	0.7453	0.9167	1.0129	0.9812	1.1068	1.2066
NOVA-Pro	0.9484	0.9930	1.3333	1.1901	1.1221	1.5387

Table 8: GPT-4o evaluation of Temporal Consistency and Correctness scores (higher is better). Columns show performance for *Video*, *Audio/Subtitles*, and their combination.

condition was toggle-enabled within the survey toolkit, allowing participants to be assigned to a specific modality. The survey interfaces for the conditions are shown in Fig. 12 for audio only, Fig. 13 for visual only, Fig. 11 for both modalities. To prevent cross-condition contamination and ensure focused evaluation, each participant completed the entire study under only one selected modality.

This human study provides us with valuable reference points for interpreting model results and understanding how humans perform under comparable multimodal constraints.

E. Dataset Distribution Continued

In this section, we provide a detailed breakdown of the taxonomy distributions within each defined agentic task category. For each task type, we analyze the composition of its associated taxonomies to illustrate the diversity and balance of the dataset across different functional scenarios. The corresponding distributions are visualized in Fig. 14 for Gaming, Fig. 15 for Egocentric Agent, Fig. 16 for Information Querying, Fig. 17 for Sentiments, Fig. 18 for Shopping, Fig. 19 for Social, and Fig. 20 for Sports. These visualizations offer insight into the internal structure of each agentic task and highlight the variety of situations represented within the dataset.

Model	Context			Detail Orientation		
	Video	Audio/Sub	V+A/Sub	Video	Audio/Sub	V+A/Sub
Human	2.13	1.93	3.11	1.84	2.02	2.96
<i>Open-source MLLMs</i>						
EgoGPT-7B	1.6420	1.5599	1.8732	1.2042	1.1244	1.3509
Ola-7B	1.3146	1.6491	1.7254	1.0047	1.2805	1.3345
VITA 1.5	0.9906	1.4918	1.0610	0.7465	1.2735	0.7958
Qwen-2-VL	1.7570	1.6549	2.0505	1.4343	1.4178	1.7430
Qwen-2.5-VL	1.9824	1.6197	2.2829	1.7336	1.3885	2.0669
InternVL2	1.3462	1.3521	1.5164	1.1303	1.0505	1.2523
LLaVA-OneVision	1.8662	1.7136	2.0129	1.5129	1.4284	1.7031
DeepSeekVL2-small	1.3920	1.6984	1.7418	1.0340	1.5998	1.3580
<i>Proprietary MLLMs</i>						
Gemini 2.0-FL	1.8204	1.8439	2.1784	1.5880	1.5293	2.0117
Gemini 2.5-FL	1.9930	1.9120	2.3216	1.7653	1.6678	2.0692
Claude Sonnet 3.5	2.1725	2.0493	2.6960	2.0129	1.9096	2.5575
GPT-4o	1.7570	2.0634	2.6103	1.4178	1.7946	2.3169
GPT-4o-mini	2.1444	2.1397	2.5716	1.7277	1.8662	2.2336
Grok 4	2.5141	2.2488	2.9636	2.4660	2.1185	2.9566
NOVA-Lite	1.4178	1.5516	1.6843	1.1749	1.3451	1.3955
NOVA-Pro	1.7312	1.5387	2.0610	1.5282	1.3439	1.8204

Table 9: GPT-4o evaluation of Context score and Detail Orientation score (higher is better). Columns show performance for *Video only*, *Audio/Subtitles only*, and *Video+Audio/Subtitles*.

	GPT-4o	Gemini 2.5 FL	Qwen 2.5 Omni	Nova Lite	Nova Pro	Ola-7B
Difficulty	Easy	66.7	59.8	52.1	55.0	62.5
	Medium	67.4	54.9	48.8	52.4	54.9
	Hard	49.7	43.0	45.2	38.4	43.0
Agentic Tasks	Social	62.6	55.6	47.5	48.5	50.5
	Sentiment	56.8	54.3	58.0	42.0	55.6
	Egocentric Agent	59.3	39.0	42.4	44.1	54.2
	Information Querying	68.3	60.0	45.3	52.1	54.7
	Sports	60.5	47.1	49.6	50.4	58.0
	Gaming	61.2	49.5	41.7	39.8	47.6
	Shopping	68.3	61.9	64.3	69.0	67.5
Audio Type	Natural Sound	52.9	51.5	52.9	35.3	45.6
	Speech	66.4	56.7	47.8	53.8	57.6
	Music	61.5	50.0	53.8	42.3	57.7
	Artificial Sound	64.0	52.0	48.0	48.0	50.0
	Mixed Sounds	63.0	53.1	51.5	51.5	56.1
Topics	Humanities and Society	71.4	64.7	56.4	54.1	63.9
	Geography and Travel	56.2	50.0	37.5	56.2	31.2
	Technology and Gaming	65.6	55.0	42.0	43.5	48.9
	Science and Knowledge	60.7	50.0	50.0	46.4	35.7
	Movies, TV and Animations	53.7	37.3	34.3	43.3	52.2
	Arts and Performance	51.7	44.8	51.7	44.8	44.8
	Pets and Animals	57.8	66.7	55.6	40.0	53.3
	Business and Commerce	69.8	66.3	68.6	73.3	69.8
	Life and Practical Skills	72.1	60.6	51.9	60.6	61.5
	Sports and Adventure	61.1	49.2	48.4	46.0	58.7
	Social Trends and Reactions	53.7	48.1	44.4	46.3	40.7
	Vehicles and Transportation	66.7	33.3	55.6	44.4	77.8
	Low Quality and Extended Content	58.3	29.2	25.0	41.7	50.0
	Overall	64.0	54.7	49.5	50.9	55.8
	Overall	64.0	54.7	49.5	50.9	53.1
Multimodal Abilities	Visual Reasoning	58.3	54.4	45.6	50.0	51.8
	Emotional Inference	53.5	65.1	51.2	39.5	44.2
	Situational Reasoning	70.0	65.0	35.0	55.0	65.0
	Causal Reasoning	67.2	49.3	44.8	49.3	55.2
	Factual Recall	67.3	55.4	52.9	53.8	58.6
	Spatial Reasoning	55.2	41.4	48.3	34.5	55.2
	Relationship	68.8	50.0	50.0	56.2	62.5
	Temporal Distance	75.0	50.0	50.0	41.7	50.0

Table 10: Performance comparison across different models and categories with *Video + Audio/Subtitle* modalities.



Factual Recall: Geography and Travel

Question: The video shows a cityscape accompanied by recognizable instrumental music. What could be the connection between the visuals and the audio?

Options:

- (A) The music is a coincidence, playing in the background while the video simply showcases the sights of the city.
- (B) The music was composed to accompany a fictional city tour, making the video feel like a scene from a movie.
- (C) The music is played by someone on a bus traveling through the city, adding to the atmosphere.
- (D) The music is a traditional British composition that enhances the feeling of the city being part of British culture.
- (E) The music is the theme song of a famous British franchise, and the cityscape evokes the same nostalgic feeling.
- (F) The music is unrelated and was simply chosen because it has a whimsical tone that complements the fast-paced nature of the tour.
- (G) The city is supposed to represent London, and the music is a subtle reference to British history, tying the visuals and the audio together.
- (H) The music is an instrumental cover of a famous British song, chosen to tie the video to the country's popular culture.

Subtitles: None



o1: (G) The city is supposed to represent London, and the music is a subtle reference to British history, tying the visuals and the audio together.

Error Category: Perceptual Limitation

Error Reason: o1 is not able to get sufficient signal from the audio because the audio playing is not speech. As a result, o1 is not able to know that the music playing is from the Harry Potter franchise and reason why that connects with the visuals.

Ground Truth: (E) The music is the theme song of a famous British franchise, and the cityscape evokes the same nostalgic feeling.

Figure 5: Error analysis showing that *o1* fails to correctly answer the question when the audio cannot be transcribed into text-based subtitles, leading to an incorrect connection between the cityscape video and the instrumental music.



Factual Recall: Pets & Animals

Question 1: Who is Hika-Chan?

Question 2: Which one is Hika Chan?

Subtitles (Whisper-base-v3): みちゃん上手! おいちおいちおいち! ゆるちゃん! あ、上手! あ、手! おる! メイルもできるようになってきちゃった! はい、やりこわさ! 可愛いね、みるこわい! おすわりしたごらん! うん、上手だよ! おすわり! お手の練習だから! はい、おすわり! おすわりは? そんな前こないで! あ、手! ひかちゃんは? はい、ひかちゃんお手は? ひかちゃんお手! お手は? ごめん! お手は? なもんなによ! お手は? そこにやんちゃっぽいお手やね! お! ちょ、ちょ、すごい! お、ひかちゃんのお手は? お! そんな感じね! 今あげるね! みるちゃんも今順番だからね! おやつあげてからね! はい、ちご! はい、ひかちゃん! みるき! みるき! はい、お!

Translation: Mii-chan, good job! Yummy, yummy, yummy! Yuru-chan! Ah, well done! Ah, paw! Oh! You're getting better at mail too! Okay, let's try! So cute, Miru, so scary! Try sitting down! Yes, good job! Sit! This is practice for shaking hands! Okay, sit! Sit? Don't come too close! Ah, paw! Hika-chan? Okay, Hika-chan, shake! Hika-chan, shake! Shake? Sorry! Shake? What's that? Shake? That's a bit of a mischievous shake! Oh! Whoa, whoa, amazing! Oh, Hika-chan, shake? Oh! That's how it is! I'll give it now! Miru-chan, it's your turn now! After giving a treat! Okay, Chigo! Okay, Hika-chan! Milky! Milky! Okay, here you go!



o1 Answer 1: Based on the images and the conversation in the subtitles, Hika-chan is the gray-and-white cat shown in the video

o1 Answer 2: From the interactions shown in the latter part of the clip (where Hika Chan is specifically called to practice giving a paw), Hika Chan is the Shiba Inu wearing the rainbow colored collar (the slightly darker coat of the two Shibas). The one wearing the pink collar is a different dog.

Error Category: Inconsistent responses

Error Reason: Although the model correctly identified Hika-chan in the original question, it failed to maintain consistency when the wording was slightly changed, showing a lack of robustness to simple rephrasing.

Ground Truth: The Shiba Inu dog with the rainbow collar

Figure 6: Error analysis showing inconsistent responses from *o1*, where it failed to accurately identify Hika-chan, when the question was rephrased slightly.



Factual Recall: Arts & Performance

Question: What does the singer want to take home in this parody song?

Options:

- (A) Pie.
- (B) Christmas tree.
- (C) Stocking.
- (D) Gift.
- (E) Tree ornaments.
- (F) Mistletoe branch.
- (G) Santa Hat.
- (H) Lights.

Subtitles: No more lies through hate

Keep a guy with this until Christmas party

Can I take this girl?

Have some...



o1: (F) Mistletoe branch

Error Category: Temporal Synchronization Limitation

Error Reason: The model failed to correctly synchronize the words “can I take this” with the visual action of the woman picking up tree ornaments, resulting in a misalignment between the audio and the corresponding visual cues.

Ground Truth: (E) Tree ornaments.

Figure 7: Error analysis highlighting *o1*'s failure in correctly synchronizing the audio with visual cues, leading to misinterpretation of the woman picking up tree ornaments.



Emotional Inference: Humanities & Society

Question: Why is the man smiling?

Options:

- A. The man is smiling because the dog's tail wagging reminds him of a fan.
- B. The man is smiling because the barking sound of the dog resembles gunfire.
- C. The man smiles because he is amused by a funny memory of his dog.
- D. The man is smiling because the way the dog jumps looks like it's dancing.
- E. The man smiles after hearing a joke from his friend in the background.
- F. The man is smiling because the dog's howl matches the rhythm of a popular song.
- G. The man smiles when the dog playfully tugs at his shoelaces.
- H. The man is smiling because the dog's bark sounds like a crying baby.

Subtitles (Whisper-base-v3): I don't know what I'm talking about, but I don't know what I'm talking about



o1: C. The man smiles because he is amused by a funny memory of his dog

Gemini Pro 1.5: H. The man is smiling because the dog's bark sounds like a crying baby.

Error Category: Contextual Misinterpretation

Error Reason: Instead of recognizing the smile as a reaction to a specific, humorous interaction with the dog, it attributes it to a general amusing memory that may not align with the context.

Error Category: Visually induced audio hallucination

Error Reason: Gemini incorrectly assumes the dog is barking based on its open-mouth appearance, even though no barking sound is present. This leads to the mistaken belief that the dog's bark resembles a crying baby.

Ground Truth: B. The man is smiling because the barking sound of the dog resembles gunfire

Figure 8: Error analysis highlighting *o1*'s contextual misinterpretation, attributing the man's smile to a general amusing memory rather than the dog's barking resembling gunfire. Additionally, *Gemini Pro 1.5* exhibits visually induced audio hallucination, mistaking the dog's open-mouth appearance for barking and associating it with a crying baby.



Causal Reasoning: Sports & Adventure

Question: When the commentator spoke of an action done by LeBron, how many opposing defenders were around (within a 1 meter radius) him?

Options: A. 0.

- B. 1.
- C. 2.
- D. 3.
- E. 4.
- F. 5.
- G. 6.
- H. 7.

Subtitles (Whisper-base-v3): The look to be addressed is to give you to pick up that third foul on Nellow.

Chalmers into Douglas rebound by LeBron.

That is the fifth second quarter offensive rebound by Miami tonight.



o1: E. 4

Gemini Pro 1.5: C. 2

Error Category: Spatial Misjudgment + Temporal Misalignment

Error Reason: The models may have struggled to distinguish actual close defenders from those further away or referenced a moment before or after the described action, leading to an error.

Ground Truth: D. 3

Figure 9: Error analysis showing *o1* and *Gemini Pro 1.5* both struggling with spatial misjudgment and temporal misalignment, leading to errors in estimating the number of defenders due to difficulties in distinguishing actual close defenders and potential misalignment with the described action.



Situational Reasoning: Humanities & Society

Question 1: What caused the glass to break in the video?

Question 2: What unique skill or phenomenon is demonstrated by the man in the video?

Subtitles (Whisper-base-v3): Now that's proof that it's not a trick.

That was too easy, man.

NOOOOO!



Gemini Pro 1.5 Answer 1: The man in the video breaks the glass by singing a high note.

Gemini Pro 1.5 Answer 2: The man in the video demonstrates the unique skill of throat singing. Throat singing, also known as overtone singing, is a vocal technique that allows a singer to produce two or more distinct pitches simultaneously. This creates a unique sound that is often described as otherworldly or ethereal.

Error Category: Inconsistent responses

Error Reason: Given the clear visuals and a high-pitched audio signal, this question was relatively easy to answer. However, a longer rewording of the question still caused Gemini Pro 1.5 to fail.

Ground Truth: The man's voice resonating with the glass

Figure 10: Error analysis showing *Gemini Pro 1.5* providing inconsistent responses despite clear visuals and audio cues, failing to accurately identify the cause of the glass breaking due to rephrased question structure.

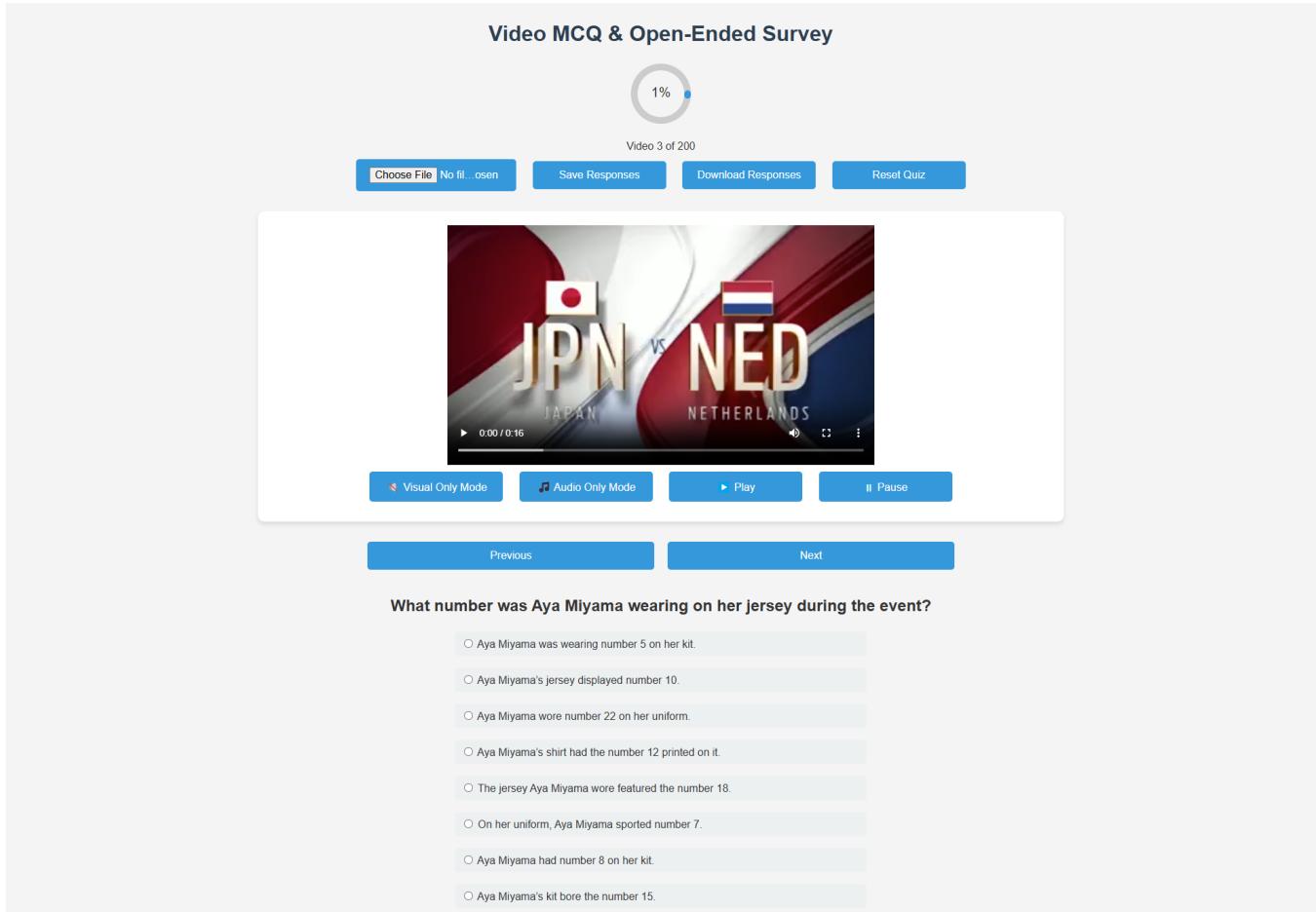


Figure 11: Screenshot of the questionnaire under the audio-plus-visual condition, showcasing the interface used to assess human performance across multiple-choice questions. Results offer a baseline for comparison with multimodal model performance across diverse tasks and difficulty levels.

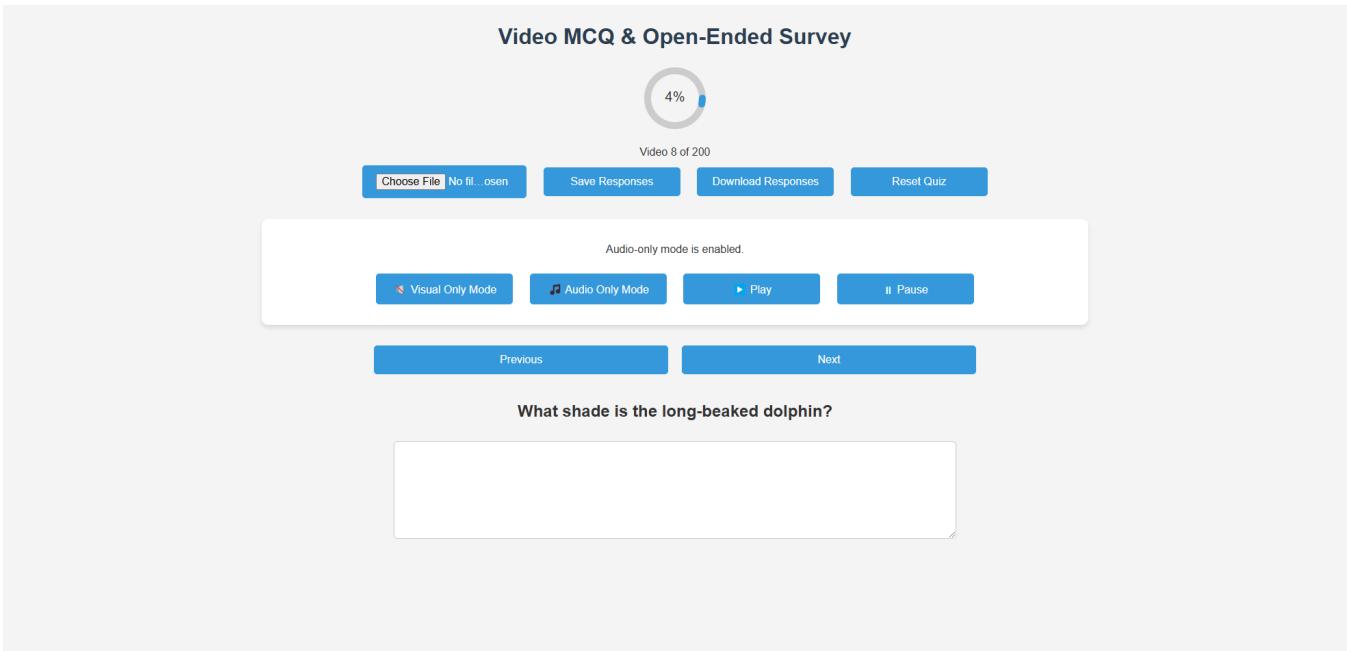


Figure 12: Questionnaire screenshot showing open-ended questions under the audio-only condition. Responses were evaluated using the GPT-4o grading pipeline, capturing the quality of language generation based solely on audio context.

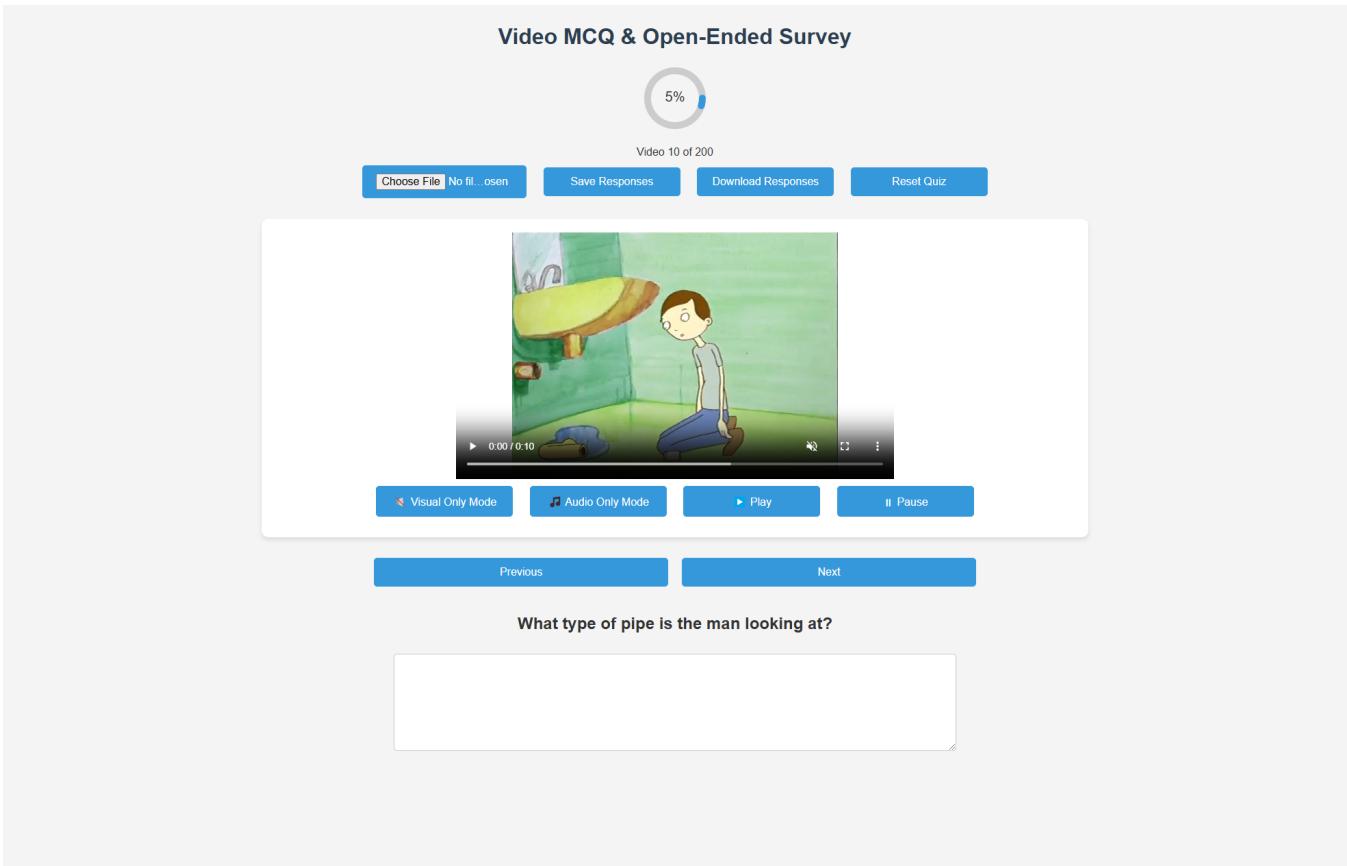


Figure 13: Questionnaire screenshot showing open-ended questions under the visual-only condition. GPT-4o evaluation scores assess human understanding based solely on visual information, without the aid of audio context.

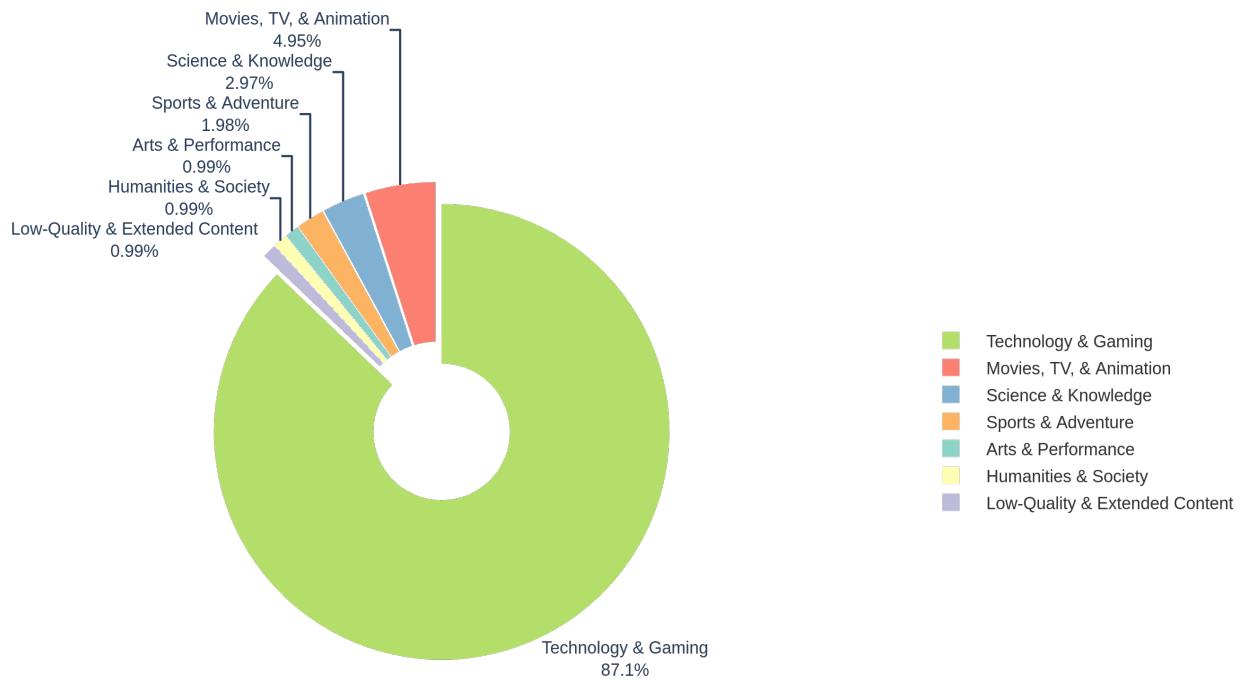


Figure 14: Taxonomy composition in Gaming.

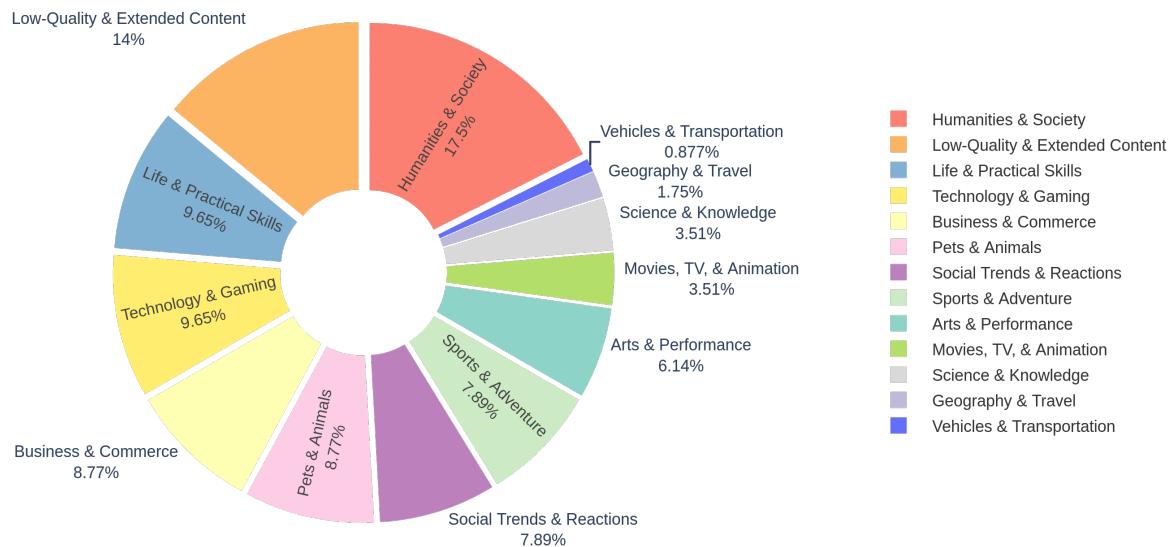


Figure 15: Taxonomy composition in Egocentric Agent.

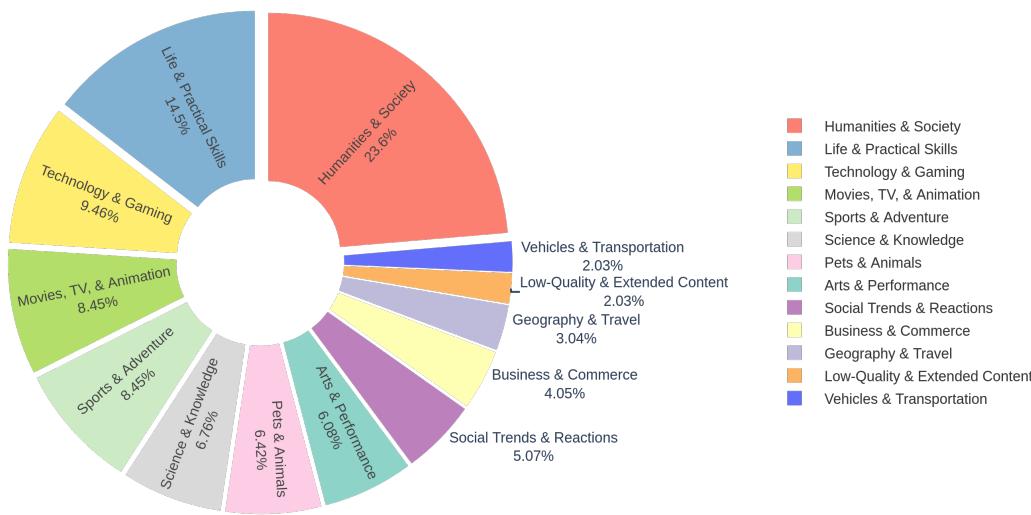


Figure 16: Taxonomy composition in Information Querying.

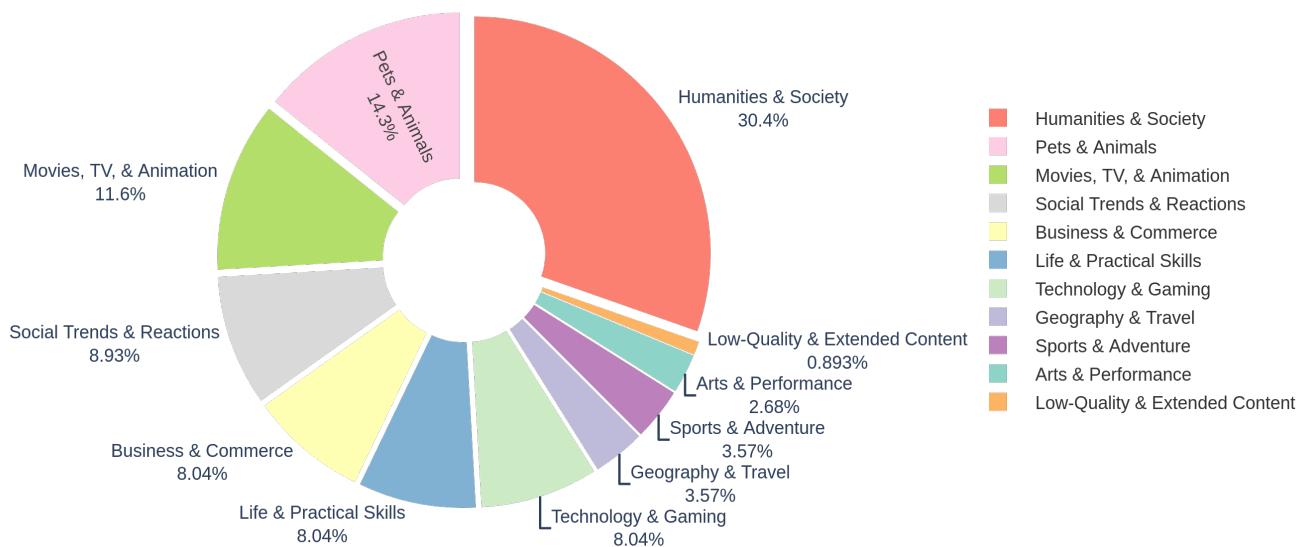


Figure 17: Taxonomy composition in Sentiments.

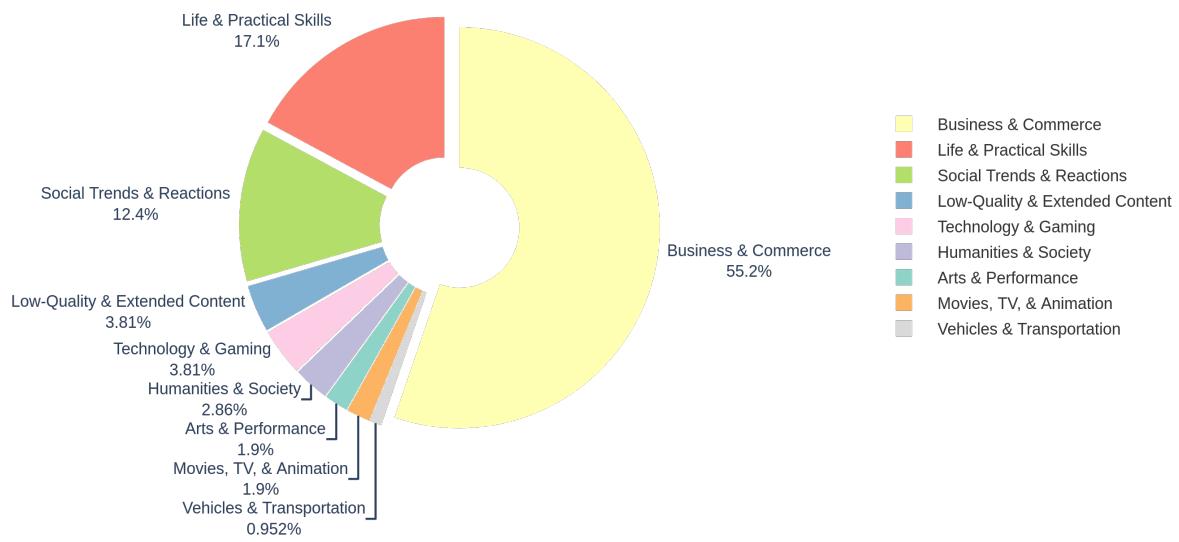


Figure 18: Taxonomy composition in Shopping.

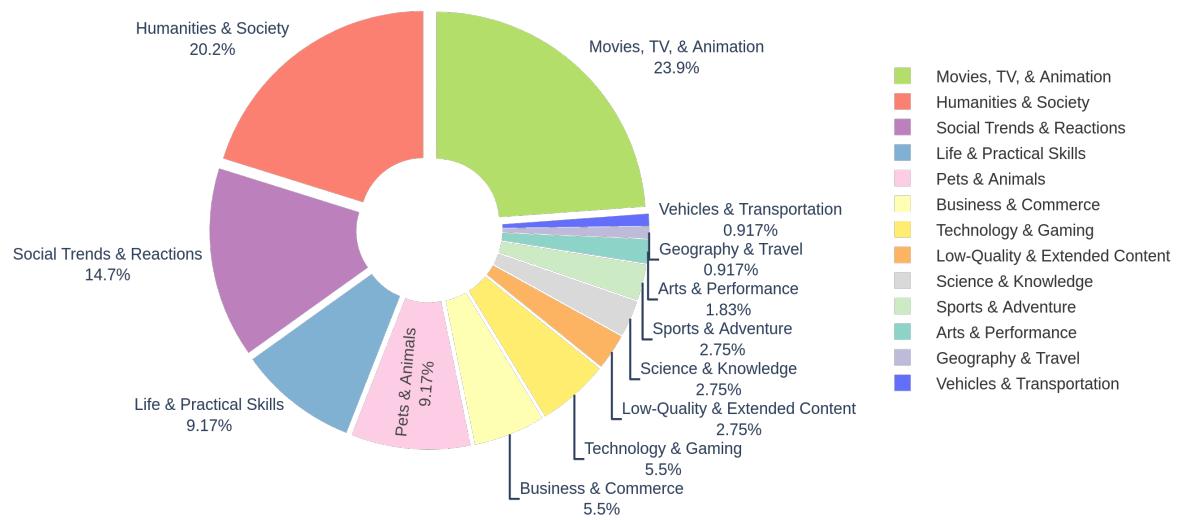


Figure 19: Taxonomy composition in Social.

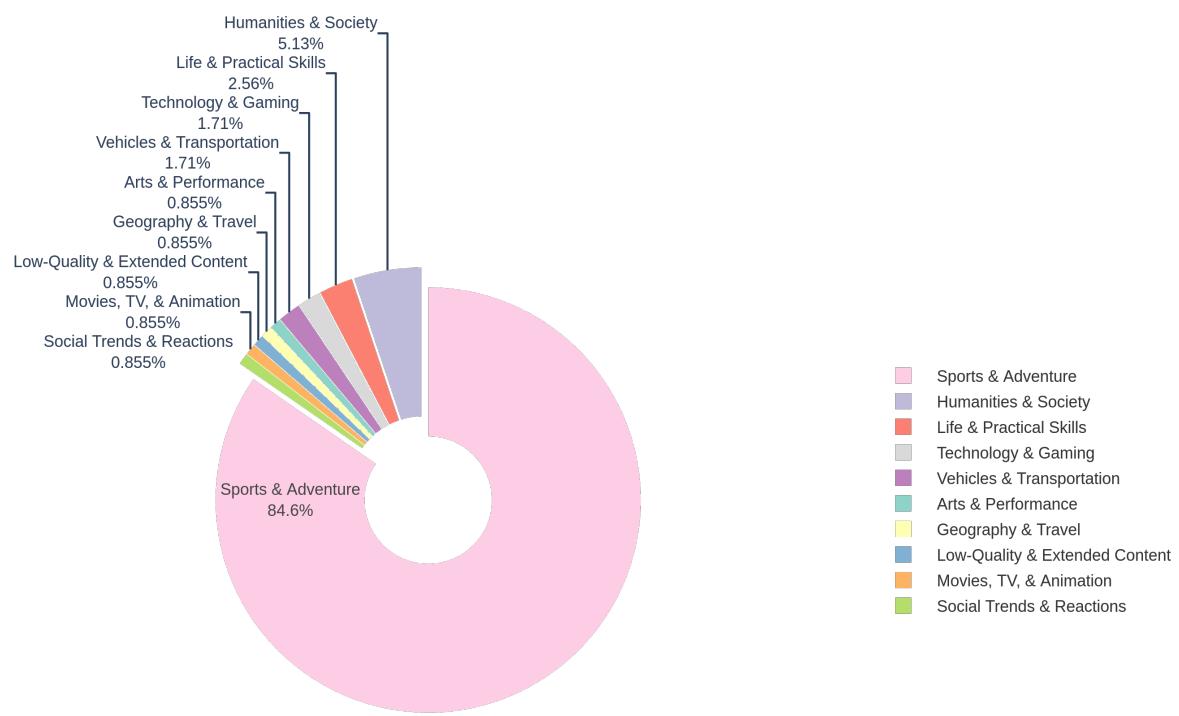


Figure 20: Taxonomy composition in Sports.