

MAVERIX: Multimodal Audio-Visual Evaluation Reasoning IndeX

Liuyue Xie^{*1} George Z. Wei^{*1} Avik Kuthiala^{*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

¹Carnegie Mellon University ²Amazon

^{*}Equal contribution

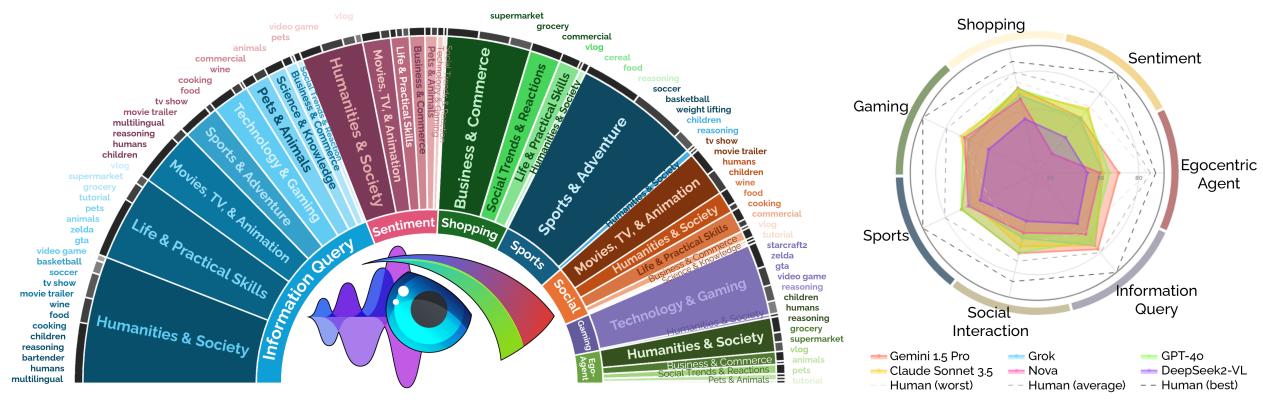


Figure 1. MAVERIX contains over 2,500 audio-visual tasks spanning diverse domains with tight modality interdependence (left) and remains challenging, as state-of-the-art multimodal frontier models are consistently outperformed by average human participants (right).

Abstract

Frontier models have either been language-only or have primarily focused on vision and language modalities. Although recent advancements in models with vision and audio understanding capabilities have shown substantial progress, the field lacks a standardized evaluation framework for thoroughly assessing their cross-modality perception performance. We introduce MAVERIX (Multimodal Audio-Visual Evaluation Reasoning IndeX), a novel benchmark with 700 videos and 2,556 questions explicitly designed to evaluate multimodal models through tasks that necessitate close integration of video and audio information. MAVERIX uniquely provides models with audiovisual tasks, 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. Experiments with state-of-the-art models, including Gemini 1.5 Pro and o1, show performance approaching human levels (around 70% accu-

racy), while human experts reach near-ceiling performance (95.1%). With standardized evaluation protocols, a rigorously annotated pipeline, and a public toolkit (maverix-benchmark.github.io)¹, MAVERIX establishes a challenging testbed for advancing audio-visual multimodal intelligence.

1. 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 [1, 4, 32].

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

¹Project page will be available soon.

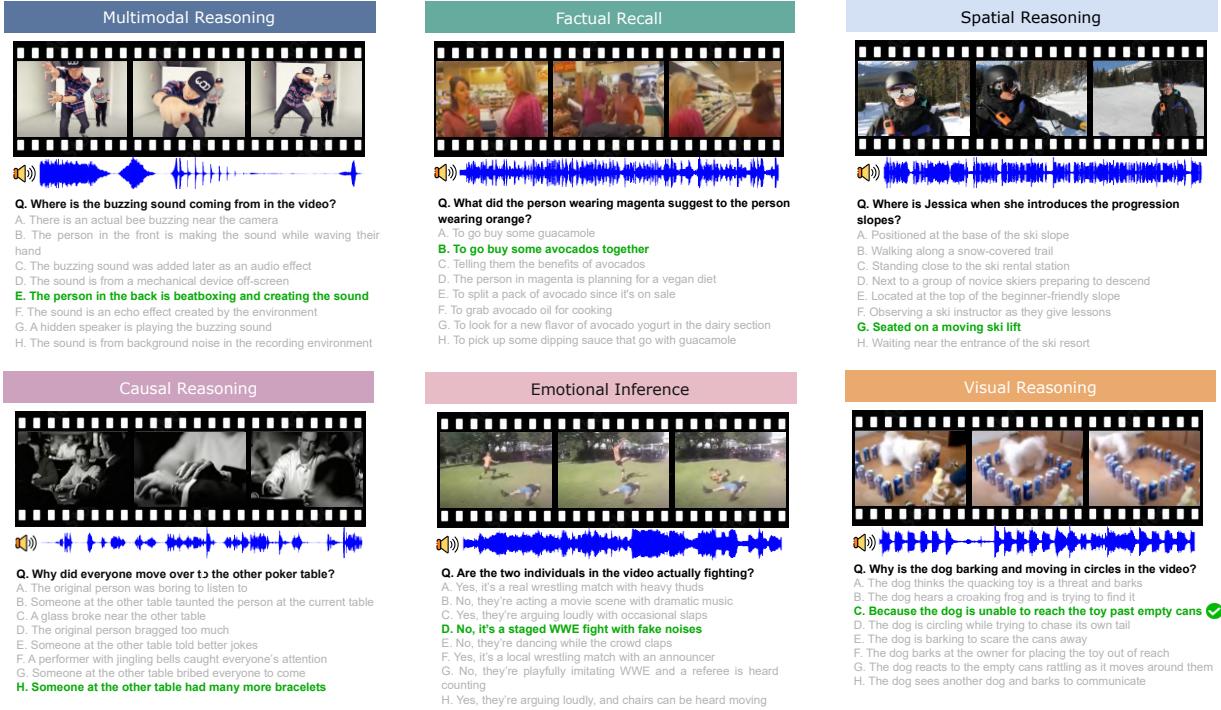


Figure 2. Example questions and answers in the MAVERIX benchmark.

images [3, 11], simple recognition, or tasks that can be solved through unimodal cues like transcripts [13, 44]. 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 [10, 36].

A core obstacle in designing effective multimodal benchmarks is ensuring that tasks genuinely require multimodal understanding rather than allowing models to exploit unimodal shortcuts. Many existing video-language benchmarks reduce to visible-object recognition or dialogue parsing, sidestepping the need to synthesize audio-visual dependencies [25, 29, 40].

To address this, we introduce **MAVERIX**, a benchmark designed to evaluate multimodal video-audio understanding through tasks that have tight modality interdependence. MAVERIX features four challenging tasks: situational awareness, contextual summarization, social sentiment analysis, and dynamic reasoning, covering 700 videos and over 2,500 carefully 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 tasks.

Evaluations of state-of-the-art models, including Gemini 1.5 Pro [49], GPT-4o [39], and o1 [38], reveal significant gaps. Gemini 1.5 Pro, even with direct audio-video inputs, achieves 71.9% accuracy on multi-

ple choice questions, marginally lower than human performance (80%). 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 [17, 37].

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.

2. Benchmark Design and Constructions

Designing benchmarks for audiovisual understanding requires tasks that demand genuine cross-modal reasoning, not just unimodal pattern matching. In real-world scenarios, one modality often dominates, such as visuals in a silent event or audio in a podcast, but effective understanding depends on integrating cues across modalities. Existing benchmarks typically focus on narrow tasks or merge disjointed datasets, limiting their ability to evaluate general-purpose multimodal LLMs (MLLMs). Moreover, many video-language benchmarks rely on multiple-choice questions (MCQs) for their simplicity, but small answer spaces of the common four-option MCQs inflate scores through random guessing and fail to assess free-form reasoning — a core

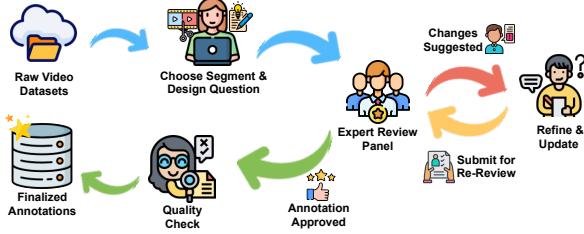


Figure 3. Annotation workflow outlining the stages of data selection, review, and refinement.

aspect of human-like comprehension.

To address these limitations, MAVERIX introduces a unified benchmark of diverse real-world tasks that systematically test the interdependence of modality while supporting a dual evaluation framework of *eight-option MCQs* and *open-ended questions*. The expanded eight-option format reduces guess bias (12.5% baseline), providing more reliable accuracy metrics, while open-ended tasks require models to generate nuanced, context-aware responses. For open-ended evaluation, we employ GPT-4o as an automated judge, scoring responses against human-curated references using a five-dimensional rubric: correctness of information, attention to detail, contextual understanding, temporal coherence, and consistency under paraphrasing. MAVERIX’s taxonomy and revealed performance gaps are shown in Fig. 1.

This hybrid design balances scalability and depth: MCQs provide standardized large-scale evaluation, while open-ended questions probe the models’ ability to synthesize and express complex audiovisual reasoning. Together, they create a robust assessment ecosystem tailored to intertwined multimodal understanding. We detail our pipeline used to generate MAVERIX’s 2,500+ high-quality questions and answers.

2.1. Dataset Generation Pipeline

Dataset Collection To ensure broad coverage of multimodal reasoning, we designed four domain hierarchies spanning agentic tasks, topics, video categories, and multimodal abilities, as shown in Fig. 2. Under agentic tasks, we defined seven key areas—social interactions, sentiment expression, information querying, sports analysis, gaming commentary, egocentric navigation, and multilingual contexts—each reflecting real-world scenarios that require complementary audiovisual cues. For example, gaming videos involve coordinating visual gameplay with voice chat strategies. In total, we annotated 155 fine-grained activity tags (e.g., “referee decision analysis” under sports) to capture diverse reasoning challenges (see Fig. 5).

We sourced videos from YouTube-8M [2], MSR-VTT [53], UR-FUNNY-V2 [20], Ego4D [19], and Au-

Multimodal Abilities	Agentic Task	Topic
Causal Reasoning	399	Humanities and Society 399
Emotional Inference	393	Tech and Gaming 393
Situational Reasoning	378	Sports and Adventure 378
Relationship	312	Practical Skill 312
Spatial Reasoning	258	Business and Commerce 258
		Geography and Travel 48
		Science and Knowledge 84
		Arts and Performance 87
		Pets and Animals 135
		Life and Practical Skills 312
		Social and Reactions 162
		Vehicles and Transportation 27

Difficulty
Easy
Medium
Hard

Figure 4. Overview of the dataset distribution. The corresponding number of samples in each category is shown.

dioSet [18], balancing short (< 1 min), medium (2–10 min), and long (10–65 min) formats. Each video includes subtitles, categorized audio (e.g., speech, environmental, artificial), and visual keyframes sampled at 1-second intervals. After deduplication and quality checks, MAVERIX comprises 700 videos with balanced coverage across domains and durations.

Multimodal Question-Answer Annotation To evaluate models’ ability to synthesize audiovisual information, we created 852 eight-option multiple-choice questions through a three-stage annotation protocol. A team of 11 researchers proficient in English (or another second language) with expertise in vision-language learning ensured high-quality, semantically precise questions.

The process involved (1) selecting videos and identifying clips requiring understanding multiple modalities, (2) drafting questions designed to be answerable only through joint audiovisual analysis, and (3) applying dual-validation checks to eliminate unimodal shortcuts by verifying that neither audio-only nor vision-only playback sufficed. The correct option is manually annotated, then paraphrased and modified into semantically different options with structurally similar alternative options. Surviving questions underwent peer review, with ambiguous cases resolved by consensus.

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.

Quality Assurance To ensure the reliability of MAVERIX’s videos and annotations, we apply a two-stage quality control protocol combining expert review and model-based validation. Each QA pair undergoes four checks by a separate 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

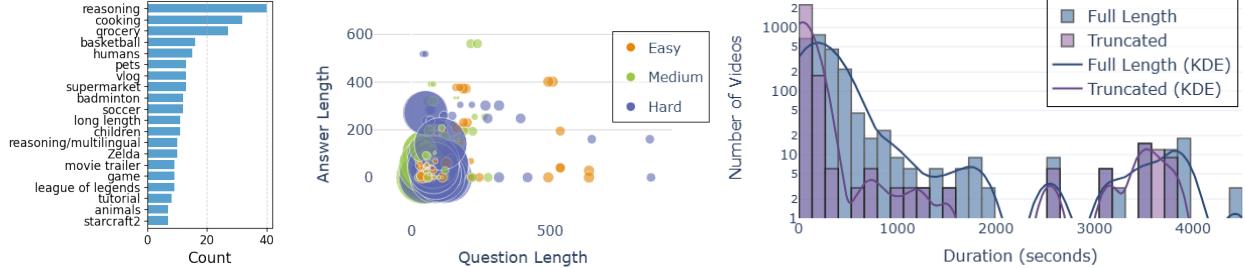


Figure 5. Dataset statistics. Left: distribution of video sources. Center: question and answer length across difficulty levels. Right: distribution of video durations for full-length and truncated clips.

identical options, and (4) modality interdependence, using cross-modal invalidation tests from Section 3.1 (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.

To guard against textual or visual biases, we test each question with GPT-4o for text-only and video keyframes + subtitles inputs. Questions answered correctly without full audiovisual reasoning are revised. For example, if GPT-4o answers “Why did the mechanic sigh?” from subtitles alone, we refine the question. Reviewers document model failures and work with annotators to iteratively resolve ambiguous cases. This combined human-model process ensures MAVERIX’s questions demand proper multimodal understanding.

2.2. MAVERIX Statistics

MAVERIX comprises a diverse set of videos spanning 155 real-world scenarios across seven agentic task domains, such as social interactions, sentiment analysis, and egocentric understanding. 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 reflecting its multimodal complexity as judged by human annotators. A detailed breakdown of these splits is provided in Fig. 4.

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 tasks, 75.7% medium-length videos (1–10 minutes) for sustained narrative reasoning, and 7.4% long-form content (10–65 minutes) for testing temporal coherence. The average video length is 5.7 minutes, reflecting real-world use cases like troubleshooting tutorials or sports highlights.

Each video is paired with 3–4 questions on average (2,556 total), including eight-option multiple-choice

questions (MCQs) and open-ended prompts. Questions are evenly rebalanced across tasks to ensure a balanced evaluation. To mitigate positional bias, answer labels are uniformly redistributed across options (e.g., correct answers appear equally in positions A–H). Fig. 4 details the dataset’s domain composition, and Fig. 5 shows the duration distribution and question-type breakdown.

3. Experiments

3.1. Evaluation Protocol

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

Evaluation is conducted under two settings: localized, where models access only the *timestamped video segment* relevant to each question, and global, where the full-length video is provided. The localized setting mitigates information leakage by limiting the context to the specific temporal window required for reasoning.

For MCQs, we report both task-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-based scoring pipeline, adapted from Video-ChatGPT [35], which assesses outputs on five dimensions using a 0–5 scale. Scoring is calibrated on 200 human-annotated examples to ensure consistency, with penalties applied for failures in audiovisual reasoning, such as overlooking modality contradictions.

Results are aggregated across tasks (Tab. 1), with separate analyses for easy, medium, and hard videos to diagnose reasoning limitations. This protocol ensures reproducibility while addressing modality interdependence, which is the core challenge MAVERIX is designed to evaluate.

3.2. Baselines

We evaluate MAVERIX on a diverse suite of 12 MLLMs, encompassing both commercial and open-source architectures, to assess their ability to rea-

son over intertwined audiovisual modalities. Commercial models include Gemini 1.5 Pro [49], GPT-4o [39], o1 [38], Grok2-Vision [52], Claude Sonnet 3.5 [5], Nova-Lite [23], and Nova-Pro [23], while open source representatives feature InternVL2 [14], Qwen2.5-VL [6], LLaVA-OneVision [27], DeepSeek-VL2-Tiny [51], and DeepSeek-VL2-Small [51]. Gemini 1.5 Pro is uniquely equipped with native audiovisual processing, enabling direct ingestion of raw video-audio streams. For these 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-base-v3 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 Gemini 1.5 Pro processes 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 visual cues like the startled expression of a character with an auditory context like an off-screen crash to match human-like reasoning.

3.3. Results

We select a wide range of open-sourced and proprietary MLLMs on the MAVERIX benchmark. We report the quantitative results grouped by difficulties in Tab. 1 for their performances with and without subtitles and by different video lengths. The detailed quantitative GPT consistency scores can be found in Tab. D.2, and visualized in Fig. 6. We then showcase the fine-grained quantitative evaluations of Gemini 1.5 Pro in different multimodal input settings and grouped by the splits in Tab. 2. Refer to Section D for additional experimental results. Below, we provide our key observations.

Open-sourced Models vs. Proprietary Models: Overall, the proprietary models achieved significantly higher results than their open-source counterparts in both multiple-choice scores and open-ended consistency scores evaluated by GPT-4o. Open-source models typically have shorter sequence lengths; for example, DeepSeek-VL2 is restricted to 4096 for both its *Tiny* and *Small* versions, limiting its ability to reason from videos. Thereby, the sampling strategy and resolutions of the input frames and the temporal horizon of the frames would have a significant impact. To study this, we evaluated both the truncated videos and the

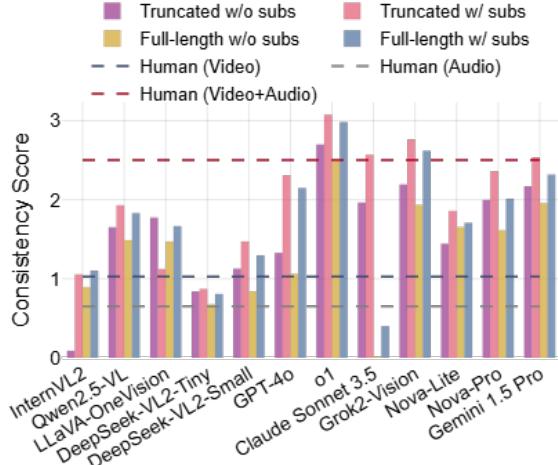


Figure 6. Consistency of model responses on open-ended questions, measuring how reliably models provide coherent and relevant answers across different inputs.

full-length videos. Qwen2.5-VL achieves an accuracy of 52.6% on full videos with subtitles, demonstrating its stronger capabilities to reason from longer videos. However, there still exists an aggregated performance gap of 30.9% in truncated videos and 19.1% on the full-length videos between open-source models and proprietary models. Among open-source models, including subtitles in truncated videos increased performance by 4.0%, while proprietary models showed an increase of 7.2% with subtitles. This suggests a potentially more balanced reliance on text and visual modalities in proprietary models. This trend dissipates in full-length videos, where the differences become 6.8% and 6.5%. Interestingly, we observe inconsistencies between MCQ accuracy and open-ended performance. While models that excel in MCQs generally perform well in open-ended responses, certain models like Nova-Pro and DeepSeek-VL2 show marginal gaps.

Comparing Multimodal Gains Between MLLMs and Humans: While current MLLMs have achieved remarkable performance, it remains unclear whether they rely on modalities in a manner similar to humans or if they can navigate questions using only unimodal information. To investigate this, we compare MLLMs with human performance in Tab. 1. Humans show a substantial increase in performance of $\sim 25.0\%$ with the addition of audio compared to $\sim 55.3\%$ from visual information alone. MLLMs also benefit from additional modalities, though to a lesser extent. Models with stronger reasoning capabilities can still make educated guesses and achieve moderately high scores using only a single modality, but their performance generally plateaus roughly within an incremental range from their unimodal performance. Beyond this threshold, mod-

Models	Params	Easy (%)		Medium (%)		Hard (%)		Overall (%)	
		w/o sub	w/ sub	w/o sub	w/ sub	w/o sub	w/ sub	w/o sub	w/ sub
<i>Random Guess</i>									
Human (Video)	-	12.4		12.7		12.3		12.5	
Human (Audio)	-	60.2		48.8		59.4		55.3	
Human (Video+Audio)	-	34.9		27.4		40.8		31.1	
		83.6		76.2		84.4		80.3	
<i>Open-source MLLMs (Truncated Videos)</i>									
InternVL2 [14]	2B	31.1	35.2 (+4.1)	27.7	35.1 (+7.4)	17.8	24.3 (+6.5)	27.5	33.4 (+5.9)
Qwen2.5-VL [6]	7B	30.8	34.9 (+4.1)	25.9	31.9 (+6.0)	19.3	20.7 (+1.4)	26.8	31.3 (+4.5)
LLaVA-OneVision [27]	7B	26.6	28.9 (+2.3)	26.5	28.2 (+1.7)	29.9	29.9 (+0.0)	27.1	28.8 (+1.7)
DeepSeek-VL2-Tiny [51]	3.37B	16.8	15.1 (-1.7)	14.5	20.6 (+6.1)	20.5	17.3 (-3.2)	16.4	17.9 (+1.5)
DeepSeek-VL2-Small [51]	16.1B	39.9	46.4 (+6.5)	38.8	46.4 (+7.6)	25.9	29.6 (+3.7)	37.2	43.8 (+6.6)
<i>Closed-source MLLMs (Truncated Videos)</i>									
GPT-4o [39]	-	58.9	69.8 (+10.9)	60.7	67.0 (+6.3)	41.5	49.6 (+8.1)	56.9	65.4 (+8.5)
o1 [38]	-	74.0	79.3 (+5.3)	70.7	76.0 (+5.3)	55.6	67.4 (+11.8)	69.6	75.9 (+6.3)
Claude Sonnet 3.5 [5]	-	27.4	68.9 (+41.5)	39.6	64.6 (+25.0)	27.4	44.4 (+17.0)	40.3	63.1 (+22.8)
Grok2-Vision [52]	-	58.3	65.1 (+6.8)	55.1	61.2 (+6.1)	45.9	47.4 (+1.5)	54.9	60.6 (+5.7)
Nova-Lite [23]	-	51.8	59.8 (+8.0)	49.3	55.9 (+6.6)	37.8	43.0 (+5.2)	48.5	55.4 (+6.9)
Nova-Pro [23]	-	53.0	51.5 (-1.5)	51.7	47.5 (-4.2)	35.6	34.1 (-1.5)	49.6	46.9 (-2.7)
Gemini 1.5 Pro [49]	-	64.5	69.5 (+5.0)	65.4	69.4 (+4.0)	55.6	49.6 (-6.0)	63.5	66.3 (+2.8)
<i>Open-source MLLMs (Full-Length Videos)</i>									
InternVL2 [14]	2B	23.0	26.3 (+3.3)	18.9	24.4 (+5.5)	14.2	11.0 (-3.2)	19.7	23.0 (+3.3)
Qwen2.5-VL [6]	7B	47.4	55.9 (+8.5)	45.3	51.7 (+6.4)	37.9	47.2 (+9.3)	44.8	52.6 (+7.8)
LLaVA-OneVision [27]	7B	34.5	40.1 (+5.6)	34.3	39.5 (+5.2)	30.7	35.4 (+4.7)	33.8	39.1 (+5.3)
DeepSeek-VL2-Tiny [51]	3.37B	17.1	26.0 (+8.9)	15.4	21.5 (+6.1)	21.3	19.7 (-1.6)	17.0	23.0 (+6.0)
DeepSeek-VL2-Small [51]	16.1B	28.3	44.1 (+15.8)	29.9	41.6 (+11.7)	23.6	26.0 (+2.4)	28.3	40.0 (+11.7)
<i>Closed-source MLLMs (Full-Length Videos)</i>									
GPT-4o [39]	-	57.6	64.8 (+7.2)	58.4	62.2 (+3.8)	43.3	46.5 (+3.2)	55.6	60.6 (+5.0)
o1 [38]	-	72.4	73.0 (+0.6)	68.3	72.4 (+4.1)	52.0	65.4 (+13.4)	67.2	71.5 (+4.3)
Claude Sonnet 3.5 [5]	-	7.9	17.4 (+9.5)	8.1	14.8 (+6.7)	7.9	17.3 (+9.4)	8.0	16.3 (+8.3)
Grok2-Vision [52]	-	55.6	59.2 (+3.6)	55.8	55.2 (-0.6)	45.7	53.5 (+7.8)	54.1	56.5 (+2.4)
Nova-Lite [23]	-	47.4	59.5 (+12.1)	45.3	55.9 (+10.6)	37.8	43.0 (+5.2)	44.9	55.3 (+10.4)
Nova-Pro [23]	-	45.2	58.2 (+13.0)	48.4	58.1 (+9.7)	28.9	41.7 (+12.8)	44.1	55.5 (+11.4)
Gemini 1.5 Pro [49]	-	65.1	69.4 (+4.3)	64.0	68.0 (+4.0)	54.3	55.1 (+0.8)	62.8	66.5 (+3.7)

Table 1. Performance comparison of different MLLMs across various settings with and without subtitles.

els begin to diverge: some remain limited by unimodal priors, while others demonstrate the ability to integrate multimodal information effectively.

Effect of Task Difficulty: Beyond overall multimodal gains, we also examined performance across difficulty levels. When comparing performance across difficulty levels, we observe that some models benefit more from multimodal inputs on *easy* questions but tend to rely more heavily on vision for the harder ones. For example, Claude 3.5 Sonnet shows a gain of +41.5% on *easy* videos and +17.0% on *hard* videos, while Nova-lite achieves increases of +8.0% and +5.2%, respectively. GPT models remain relatively consistent across difficulties, with comparable multimodal gains across *easy*, *medium*, and *hard* questions. In contrast, Gemini shows a noticeable drop in multimodal benefit when moving from *easy* to *hard* questions.

Are Audio and Text Equally Informative? Refer to Tab. 2 for detailed results. We evaluated Gemini 1.5 Pro, which was selected for its support for audio, text, and video modalities. Across all categories, Gemini 1.5 Pro consistently achieves higher scores with audio-only

input compared to text-only input, suggesting that the model may extract more informative or discriminative cues from audio signals than from textual descriptions in this setting. Furthermore, when comparing multimodal configurations, video+audio yields the highest overall performance, surpassing video+subtitle.

Although video + audio generally provide the most significant benefit, certain categories, including sentiment understanding, situational reasoning, causal reasoning, and egocentric agent tasks, experience particularly large performance gains under this configuration. These tasks likely benefit from the richer contextual and temporal cues present in audio, such as tone, ambient sounds, and speaker intent, which are difficult to fully capture through visual or textual modalities alone.

Misaligned Category Trends Between MLLMs and Humans: The category-level analysis in Fig. 1 reveals notable differences in how MLLMs and humans perform across tasks. The quantitative comparisons are detailed in Tab. D.6. Humans tend to excel in dynamic contextual categories such as egocentric reasoning, sports analysis, and gaming strategy, which require interpreting

		Video only		Audio only		Text only		Video + Sub		Video + Audio	
		Truncated	Full	Truncated	Full	Truncated	Full	Truncated	Full	Truncated	Full
<i>Level</i>	Easy	64.5	65.1	56.8	57.6	45.6	51.6	69.5	69.4	76.0	74.0
	Medium	65.4	64.0	57.8	52.3	47.0	46.8	69.4	68.0	71.5	70.3
	Hard	55.6	54.3	46.7	49.6	34.8	41.7	49.6	55.1	63.0	62.2
<i>Affective Tasks</i>	Social	55.6	58.3	47.5	49.0	41.4	54.2	62.6	59.4	61.6	67.7
	Sentiment	69.1	70.4	53.1	51.9	34.6	42.0	66.7	64.2	76.5	74.1
	Egocentric Agent	64.4	61.7	49.2	53.2	33.9	31.9	57.6	55.3	78.0	68.1
	Information Querying	66.8	65.9	58.9	55.6	48.3	50.0	69.4	68.3	75.1	72.2
	Sports	49.6	50.5	54.6	47.6	45.4	45.6	64.7	66.0	63.9	62.1
	Gaming	63.1	61.2	53.4	49.4	41.7	43.5	58.3	64.7	63.1	62.4
	Shopping	72.2	67.6	62.7	65.8	51.6	54.1	74.6	76.6	82.5	81.1
<i>Audio Type</i>	Natural Sound	57.4	64.6	47.1	50.8	29.4	32.3	54.4	61.5	64.7	61.5
	Speech	63.5	61.4	58.7	55.2	51.6	55.6	69.3	70.3	76.2	73.1
	Music	69.2	72.0	53.8	56.0	7.7	24.0	69.2	56.0	76.9	76.0
	Artificial Sound	60.0	74.0	36.0	40.0	20.0	26.0	56.0	58.0	56.0	68.0
	Mixed Sounds	65.3	61.5	56.5	55.5	44.7	45.4	66.0	63.8	69.1	67.9
<i>Topics</i>	Humanities and Society	70.7	71.3	63.9	65.9	48.9	51.9	77.4	79.1	82.7	80.6
	Geography and Travel	68.8	68.8	56.2	56.2	62.5	68.8	50.0	68.8	75.0	75.0
	Technology and Gaming	64.9	64.0	52.7	49.1	42.0	41.2	61.1	62.3	65.6	64.9
	Science and Knowledge	57.1	63.0	39.3	51.9	39.3	37.0	53.6	63.0	67.9	70.4
	Movies, TV and Animations	61.2	61.2	35.8	37.3	29.9	44.8	55.2	59.7	59.7	65.7
	Arts and Performance	58.6	60.7	48.3	42.9	34.5	39.3	58.6	57.1	62.1	71.4
	Pets and Animals	64.4	66.7	55.6	48.9	40.0	44.4	64.4	60.0	71.1	73.3
	Business and Commerce	73.3	68.6	68.6	71.4	55.8	57.1	79.1	77.1	86.0	87.1
	Life and Practical Skills	72.1	64.7	69.2	60.0	54.8	58.8	77.9	64.7	84.6	77.9
	Sports and Adventure	53.2	56.9	52.4	48.6	43.7	45.9	65.1	69.7	63.5	63.3
	Social Trends and Reactions	51.9	50.0	40.7	46.2	38.9	48.1	57.4	57.7	59.3	55.8
	Vehicles and Transportation	55.6	55.6	55.6	44.4	22.2	11.1	55.6	44.4	66.7	55.6
	Low Quality and Extended Content	41.7	41.7	54.2	54.2	29.2	37.5	37.5	50.0	66.7	62.5
<i>Multimodal Abilities</i>	Visual Reasoning	63.2	65.7	54.8	57.4	43.4	46.6	64.5	64.2	72.8	71.1
	Emotional Inference	67.4	68.3	55.8	51.2	37.2	34.1	62.8	58.5	72.1	73.2
	Situational Reasoning	70.0	77.8	55.0	50.0	50.0	44.4	70.0	66.7	80.0	77.8
	Causal Reasoning	65.7	60.3	47.8	43.1	37.3	48.3	67.2	65.5	70.1	60.3
	Factual Recall	64.3	61.0	57.9	55.1	46.2	48.6	68.2	69.0	72.5	71.7
	Spatial Reasoning	37.9	45.8	51.7	33.3	41.4	41.7	58.6	45.8	55.2	50.0
	Relationship	62.5	62.5	43.8	62.5	50.0	75.0	62.5	75.0	62.5	75.0
	Temporal Distance	66.7	81.8	58.3	54.5	58.3	72.7	58.3	81.8	83.3	81.8
Overall		63.5	62.8	55.6	53.9	44.5	47.9	66.3	66.5	71.9	70.5

Table 2. Performance comparison of Gemini 1.5 Pro across different modalities, difficulties, and categories.

spatial cues, tracking actions over time, and understanding complex interactions. However, their performance is less consistent in domains like shopping and social interaction. Shopping tasks may prioritize factual or procedural details that are less engaging or familiar to humans. For social interaction tasks, inherent subjectivity, such as interpreting cultural contexts, intentions, or unspoken tensions, can lead to diverse and sometimes conflicting human judgments, reducing overall precision.

MLLMs show a different trend. They perform comparatively well on shopping-related tasks, which often involve concrete retrieval of explicit audiovisual details like product descriptions, prices, or transactional cues. These structured object-centric queries align well with the strengths of large models trained on web-scale corpora rich in product and factual information. In contrast, MLLMs consistently struggle in categories that require nuanced temporal reasoning, social understanding, and dynamic context integration, such as predicting team strategies in games or recognizing subtle emotional cues in conversation.

These trends suggest that current MLLMs are more adept at tasks grounded in static, factual associations but remain limited in handling complex, situational reasoning that requires modeling interactions, intent, and evolving context. Closing this gap will likely require advances in temporal modeling, sentiment cue understanding, and the ability to generalize beyond explicit audiovisual patterns.

4. Related Work

MLLM Benchmarks: The development of MLLM benchmarks has progressed from static image-based question answering (QA) to dynamic video understanding. Early benchmarks for evaluating vision-language models (VLMs) focused primarily on images, emphasizing visual grounding and object recognition. Earlier datasets [3, 11, 41, 59] provided human-annotated captions and QA pairs to assess these core capabilities. Building on this foundation, subsequent benchmarks [34, 34, 42, 44] introduced more domain-specific and knowledge-intensive tasks. A-OKVQA [43] chal-

lenges abilities to read and interpret texts within images. More recently, MMMU [54] and MMMU-Pro [55] have expanded these efforts with diverse tasks designed to probe a model’s ability to simultaneously ‘see’ and ‘read’ from prompted images.

Despite their advances, image-based benchmarks lack the temporal and acoustic context necessary for richer multimodal understanding. This limitation motivated the shift toward video-based evaluation. Fine-grained temporal tasks assessing a model’s ability to perceive motion and temporal sequences have been included in video benchmarks [16, 22, 25, 29, 30, 35, 40, 45]. However, these datasets and recent extensions [12, 21, 36, 37, 46] tend to feature short clips with constrained domains and limited semantic diversity. They primarily probe local temporal, visual semantics, and factual information, offering limited insight into how models handle the complexities of daily activities, situational awareness, or social reasoning. Most recent works [9, 10, 25, 29, 30, 40] take a step towards everyday contexts by incorporating human actions, but their scopes are restricted to human-centric activities and do not handle visual-audio cues. Video-MME [17] is a comprehensive video benchmark with short, medium, and long videos. However, it only employed the MCQ format, overlooking the quality and richness of free-form expressions generated by MLLMs from open-ended questions.

MAVERIX extends prior benchmarks by explicitly elevating audio as a primary modality, with its questions requiring complex cross-modal reasoning, such as resolving conflicts between dialogue and visual cues. Unlike existing datasets, it systematically evaluates a model’s ability to integrate complete multimodal signals through both descriptive open-ended answers as well as multiple-choice questions, comparisons shown in Tab. 3.

Video Understanding Models: Models based on contrastive pre-training over image-text pairs have demonstrated significant efficacy in learning transferable representations across various visual tasks. These methods have been adapted to specialized domains, including multi-disciplinary scientific understanding and multimodal reasoning. Building upon advancements in pre-training, subsequent work [6, 32, 48, 50, 56, 57] has sought to enhance the multimodal reasoning capabilities of complex spatio-temporal video data and improve their alignments with human preferences through post-training reinforcement learning from human feedback (RLHF). Other efforts [7, 15, 31, 47, 51] integrating a mixture of experts (MoE) architectures [33], which conditionally activates subsets of specialized ‘expert’ networks during processing, have shown improved scalability and performance without a proportional increase in computational costs. However, whether these develop-

Benchmarks	#Videos	Len.	#Questions	Subtitle	Audio	Open	Multi-Split
MSRVTT-QA	2,990	15.2	72,821	✗	✗	✗	✗
MSVD-QA	504	9.2	50,505	✗	✗	✗	✗
TGIF-QA	9,575	3.6	165,165	✗	✗	✗	✗
ActivityNet-QA	5,800	14.9	58,000	✗	✗	✗	✗
TVQA	2,179	15.2	152,545	✓	✗	✗	✗
How2QA	1,517	11.5	71,812	✓	✓	✗	✗
STAR	914	11.9	59,871	✗	✗	✗	✗
NExT-QA	4,269	12.1	52,631	✗	✗	✗	✗
MVBench	3,641	16.0	20,000	✗	✗	✗	✗
Video-Bench	5,917	56.0	17,000	✗	✗	✗	✗
EgoSchema	6,365	18.0	36,054	✗	✗	✗	✗
AutoEval-Video	327	32.1	450	✗	✗	✗	✗
TempCompass	410	11.4	1,540	✗	✗	✗	✗
Video-MME	900	1071.9	2,700	✓	✓	✗	✗
HourVideo	500	2742	12,976	✓	✗	✗	✗
MAVERIX-T	700	105.8	2,556	✓	✓	✓	✓
MAVERIX	700	344.9	2,556	✓	✓	✓	✓

Table 3. Comparison of video QA benchmarks with various attributes.

ments enable models to reason in ways comparable to or surpass humans remains an open question [1, 4, 8, 24], as the assessment of human-like multimodal reasoning, particularly in real-world visual-audio contexts, remains underexplored.

This gap is revealed by our evaluations on MAVERIX, where even the strongest models like Gemini 1.5 Pro and o1 fall approximately 20% below human performance. Persistent limitations in models’ abilities to integrate visual and auditory information over time and reason about subtle cues native to social context have been exposed. These results urge a need for both improved evaluation frameworks and model architectures that can more effectively capture and prioritize multimodal information reflective of human perception.

5. Conclusion

Agentic tasks such as assisting collaborative work and navigating dynamic environments require strong audiovisual reasoning capabilities, and yet these abilities have not been fully assessed in the recent MLLMs. To address this, we introduce **MAVERIX**, a benchmark designed to evaluate models in complex real-world audiovisual scenarios. It contains 700 videos and 2,533 human carefully crafted questions providing not only a challenging benchmark but also a comprehensive evaluation suite. We conduct a thorough benchmarking of mainstream closed-source and open-source MLLMs on the proposed benchmark. Our evaluations reveal that cross-modal video-audio understanding is essential for solving real-world perception problems. Additionally, they show that even state-of-the-art MLLMs still exhibit a performance gap compared to humans, particularly in socially grounded or dynamic tasks. We hope MAVERIX will serve as a foundation for advancing research in multimodal reasoning and guiding future work towards more context-aware and socially intelligent systems.

References

- [1] Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024. Publisher Copyright: © The Author(s) 2024. 1, 8
- [2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. In *ArXiv*, 2016. 3
- [3] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 2, 7
- [4] Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. 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*, page 34–44, New York, NY, USA, 2024. Association for Computing Machinery. 1, 8
- [5] Anthropic. Introducing claude 3.5 sonnet \ anthropic, 2024. [Online; accessed 2025-03-06]. 5, 6
- [6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5, 6, 8
- [7] Ruisi Cai, Saurav Muralidharan, Greg Heinrich, Hongxu Yin, Zhangyang Wang, Jan Kautz, and Pavlo Molchanov. Flextron: Many-in-one flexible large language model, 2024. 8
- [8] Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M. Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor W. Webb. Understanding the limits of vision language models through the lens of the binding problem, 2024. 8
- [9] Rajatsubhra Chakraborty, Arkaprava Sinha, Dominick Reilly, Manish Kumar Govind, Pu Wang, Francois Bremond, and Srijan Das. Llavidal: Benchmarking large language vision models for daily activities of living. *arXiv preprint arXiv:2406.09390*, 2024. 8
- [10] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. In *Advances in Neural Information Processing Systems*, 2024. 2, 8
- [11] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 7
- [12] Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*, pages 179–195. Springer, 2024. 8
- [13] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2
- [14] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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*, pages 24185–24198, 2024. 5, 6
- [15] DeepSeek-AI and et al Aixin Liu. Deepseek-v3 technical report, 2025. 8
- [16] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2025. 8
- [17] Chaoyou Fu, Yuhuan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 2, 8
- [18] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017. 3
- [19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, and et al. Antonino Furnari. Ego4d: Around the world in 3, 000 hours of egocentric video. *CoRR*, abs/2110.07058, 2021. 3
- [20] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 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)*, pages 2046–2056, Hong Kong, China, 2019. Association for Computational Linguistics. 3
- [21] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*, 2024. 8
- [22] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie un-

- derstanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 8
- [23] Amazon Artificial General Intelligence. The amazon nova family of models: Technical report and model card. *Amazon Technical Reports*, 2024. 5, 6
- [24] Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Dee Guo, Sreenivas Gollapudi, and Ahmed Qureshi. Remi: A dataset for reasoning with multiple images, 2024. 8
- [25] Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, et al. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. *arXiv preprint arXiv:2311.07022*, 2023. 2, 8
- [26] Bo Li*, Peiyuan Zhang*, Kaichen Zhang*, Fanyi Pu*, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimodal models, 2024. 1
- [27] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 5, 6
- [28] Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference leakage: A contamination problem in llm-as-a-judge, 2025. 5
- [29] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 2, 8
- [30] Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024. 8
- [31] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-l lava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024. 8
- [32] Jessy Lin, Yuqing Du, Olivia Watkins, Danijar Hafner, Pieter Abbeel, Dan Klein, and Anca Dragan. Learning to model the world with language. 2023. 1, 8
- [33] Jiacheng Liu, Peng Tang, Wenfeng Wang, Yuhang Ren, Xiaofeng Hou, Pheng-Ann Heng, Minyi Guo, and Chao Li. A survey on inference optimization techniques for mixture of experts models, 2025. 8
- [34] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun-yuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 7
- [35] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 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)*, 2024. 4, 8
- [36] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 2, 8
- [37] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023. 2, 8
- [38] OpenAI. Openai o1 system card, 2024. [Online; accessed 2025-03-06]. 2, 5, 6
- [39] OpenAI, :, and et al Aaron Hurst. Gpt-4o system card, 2024. 2, 5, 6
- [40] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023. 2, 8
- [41] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 7
- [42] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022. 7
- [43] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. Aokvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 7
- [44] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 2, 7
- [45] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding, 2024. 8
- [46] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 8

- [47] Xingwu Sun, Yanfeng Chen, and et al Yiqing Huang. Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent, 2024. 8
- [48] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. 2023. 8
- [49] Gemini Team and et al Petko Georgiev. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 2, 5, 6
- [50] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhi-hao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 8
- [51] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. 5, 6, 8
- [52] xAI. Bringing grok to everyone, 2024. [Online; accessed 2025-03-06]. 5, 6
- [53] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 3
- [54] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi 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*, pages 9556–9567, 2024. 8
- [55] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 8
- [56] Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 8
- [57] Gongrui Zhang. Vl-rlhf: A rlhf infrastructure for vision-language model. <https://github.com/TideDra/VL-RLHF>, 2024. 8
- [58] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu.
- Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. 1
- [59] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016. 7

MAVERIX: Multimodal Audio-Visual Evaluation Reasoning IndeX

Supplementary Material

MAVERIX Supplementary Material

- Section A: MAVERIX Release v1.0
- Section B: Rationale and Limitations
- Section C: Qualitative Comparisons by Modalities
- Section D: Additional Experiments
 - Section D.1: Standardized Prompt Design
 - Section D.2: Performance by Difficulties
 - Section D.4: Quantitative GPT Evaluations
 - Section D.5: Performance by Categories
 - Section D.6: Common Error Modes
- Section E: Details on the Human Survey
- Section F: Additional Data Distribution Analysis

A. MAVERIX Release v1.0

We are releasing **MAVERIX v1.0**, our proposed benchmark dataset for multimodal video-language understanding. 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 provide all 700 video clips used in our benchmark, 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.

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 without 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/**
 - video_utils.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.
 - MAVERIXXX_dev: Individual setting files specifying the test settings for different modes. The currently supported modes are: audio_only, video_only, text_only, video_text, and video_text_audio.
 - **Remark:** Except for the open-sourced models listed in the paper or available through LMMs-Eval[26, 58], all other experiments require access to proprietary models.

Documentation:

- We provide a comprehensive datasheet explaining the benchmark dataset's 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 care-

ful 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. Qualitative Comparisons by Modalities

We conduct a comprehensive analysis of the performance of Gemini 1.5 Pro, providing a breakdown across each custom-defined split and input modality to better understand its strengths and limitations throughout the benchmark. Specifically, we evaluate the model under different input configurations like Audio Only, Video Only, and Audio+Video, as well as across diverse content-based splits covering topics, agentic task types, audio characteristics, and difficulty levels. The corresponding results are shown in Fig. C.1 for modality comparison, Fig. C.2 for topic-wise performance, Fig. C.3 for agentic task types, Fig. C.4 for audio type variation, and Fig. C.5 for difficulty levels. This detailed analysis highlights how well Gemini 1.5 Pro generalizes across different challenge dimensions and input conditions.

D. Additional experiments

This section presents extended experimental results to further analyze model performance across various dimensions of the benchmark. In addition to model eval-

uation, we provide human performance baselines segmented by difficulty levels, input modalities, and agentic task categories for direct comparison.

D.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. Respond with
only the letter (A, B, C, or D) 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:
```

D.2. Model Performance by Difficulties

We evaluate model performance across different difficulty levels to assess robustness and generalization from simpler to more complex scenarios. The results are presented in Figure D.1, which illustrates how the model's accuracy declines as task complexity increases and showcases its relative performance compared to human participants.

D.3. Human Performance by Categories

We report performance across distinct agentic task categories 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 Sec. E.

Performance by Modality - Multimodal Abilities

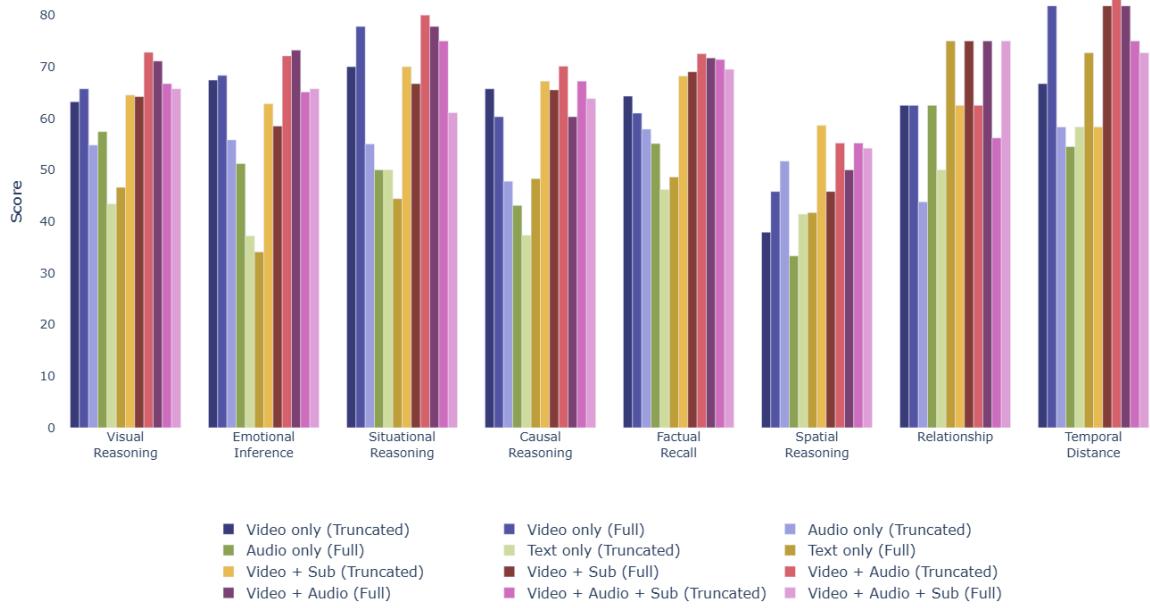


Figure C.1. Evaluation of multimodal abilities by Categories.

Performance by Modality - Topics

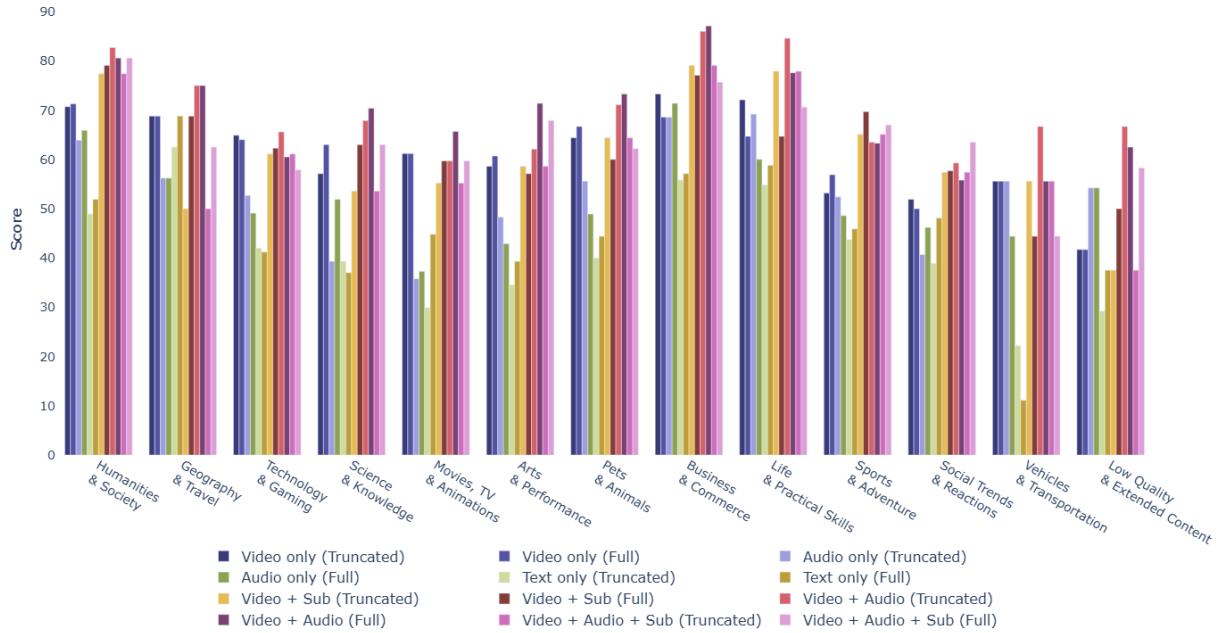


Figure C.2. Evaluation of multimodal abilities by Topics.

D.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 Sec. 2.2, we prompt GPT-4o to act as an automated grader, evaluating

Performance by Modality - Agentic Tasks

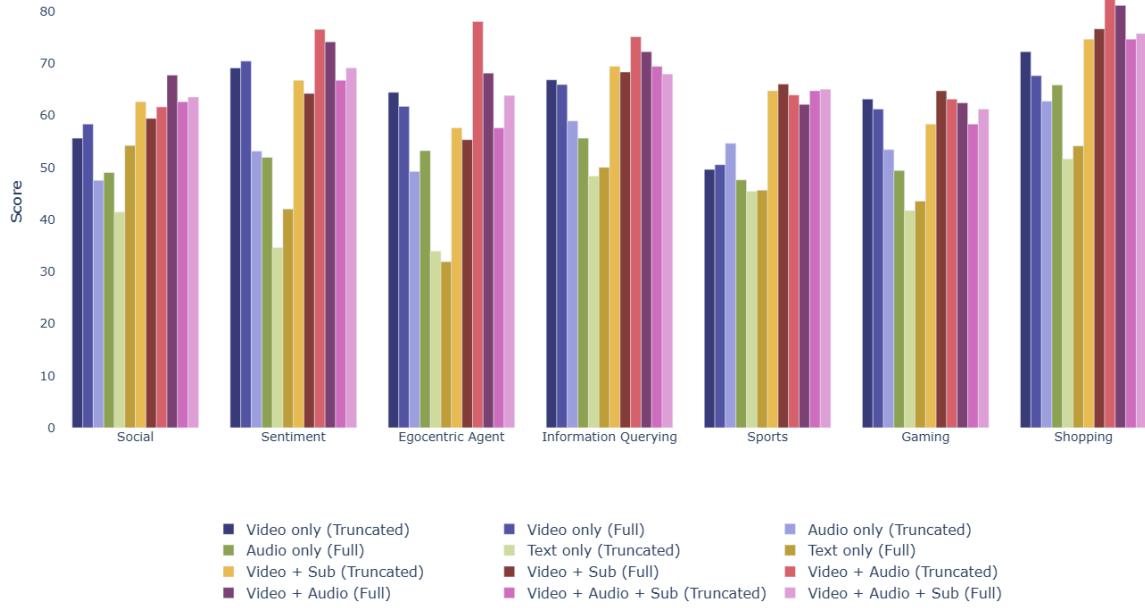


Figure C.3. Evaluation of multimodal abilities by Agentic Tasks.

Performance by Modality - Audio Type

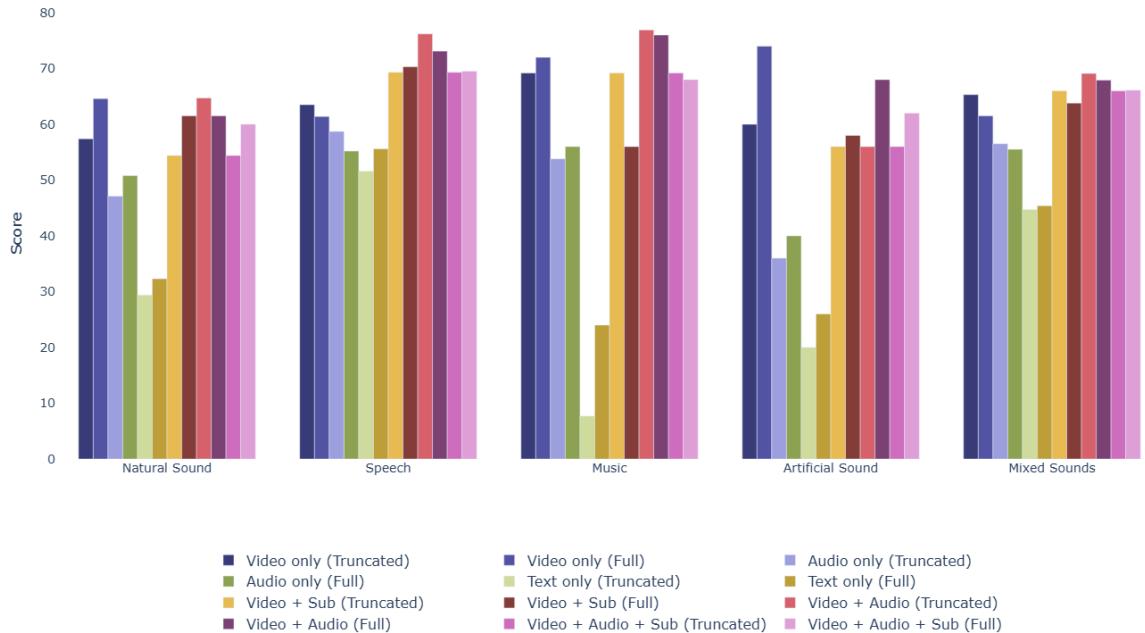


Figure C.4. Evaluation of multimodal abilities by Audio Types.

the correctness and relevance of the model's free-form answers based on predefined guidelines consistent with the task objectives.

This automatic assessment provides an additional perspective on the model's generative capabilities, capturing nuances that are not reflected in standard accu-

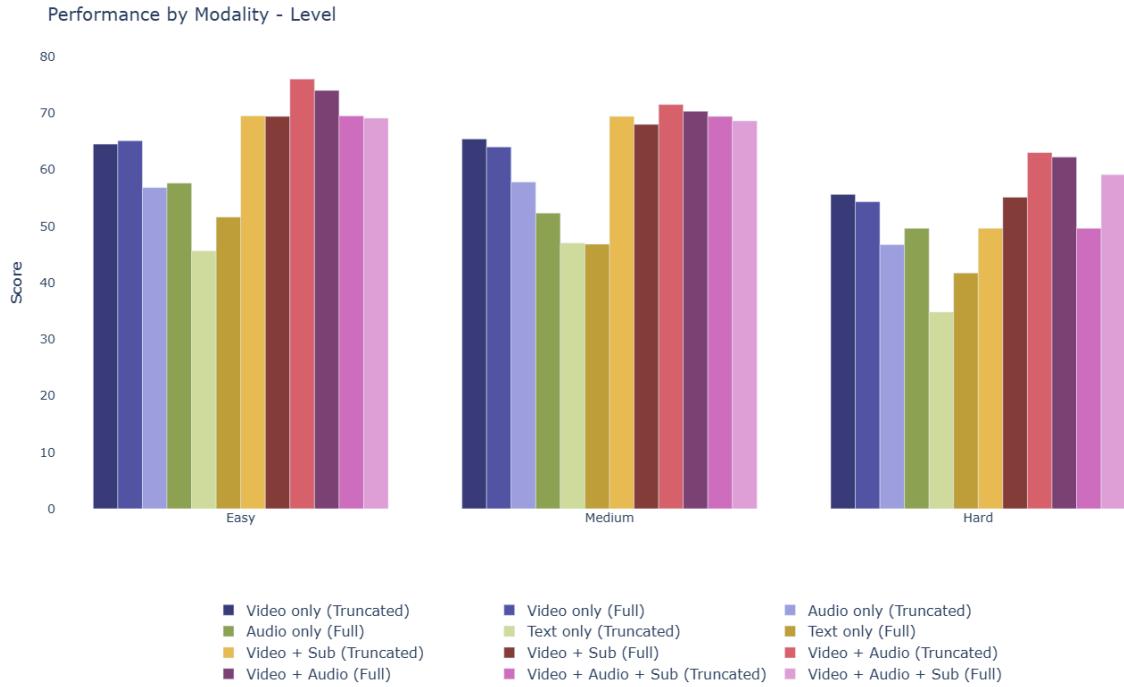


Figure C.5. Evaluation of multimodal abilities by the level of difficulties.

racy 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 [28], which can introduce alignment artifacts or systematic preferences.

D.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.

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.

D.6. Common Error Modes

To better understand the limitations of the evaluated 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. Common failure modes are showcased in Fig. D.2, D.3, D.4,

E. 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

MLM Performance By Difficulty

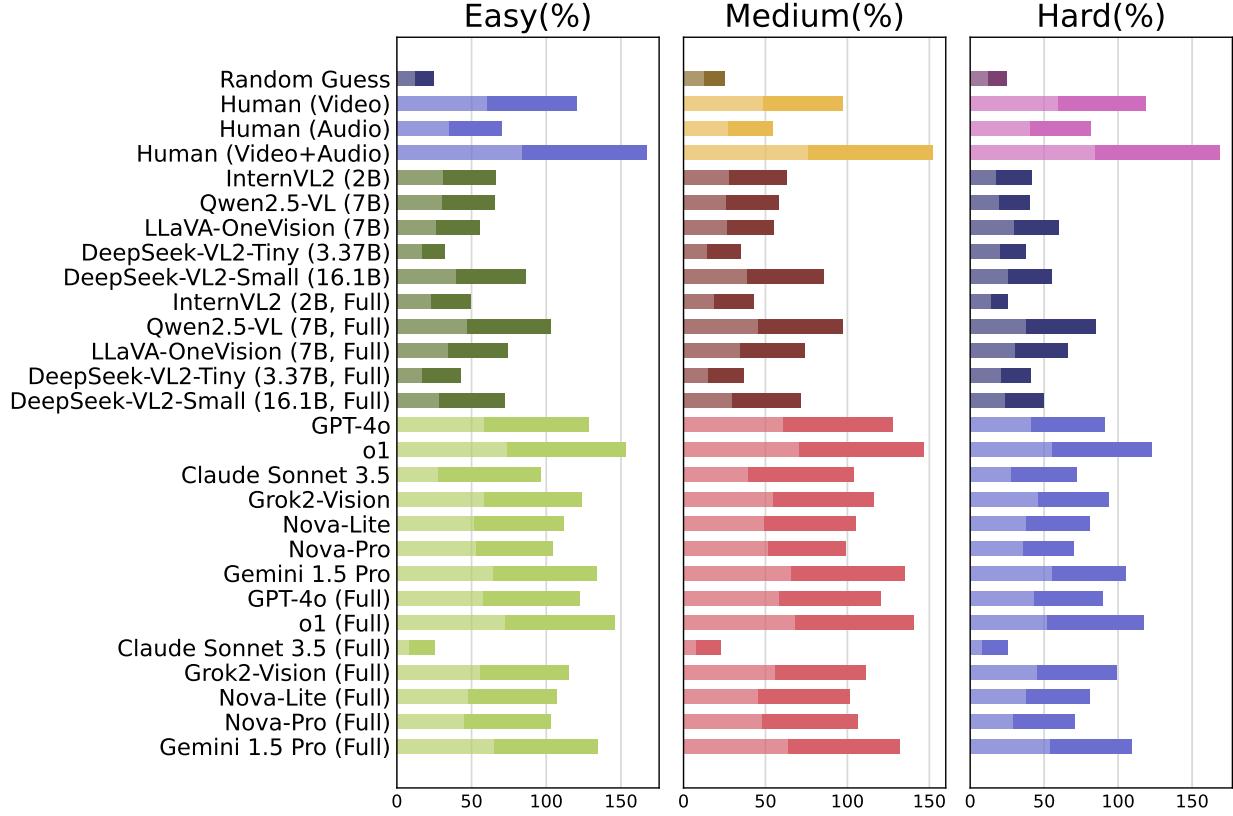


Figure D.1. Probed model performance across different difficulty levels. This analysis illustrates how the models’ accuracy changes with increasing task complexity and their relative positions against humans.

the same format as those used for model evaluation.

The survey included a total of 200 questions curated from the dataset, covering diverse subcategories within each defined split. For each participant, the questionnaire consisted of two parts: one multiple-choice question (MCQ) with eight answer options, followed by one 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.

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 condition was toggle-enabled within the survey toolkit, allowing participants to be assigned to a specific modality. The survey interfaces for the con-

ditions are shown in Fig. E.2 for audio only, Fig. E.3 for visual only, Fig. E.1 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.

F. 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. F.1 for Gaming, Fig. F.2 for Egocentric Agent, Fig. F.3 for Information Querying, Fig. F.4 for Sentiments, Fig. F.5 for Shopping, Fig. F.6 for Social, and Fig. F.7 for Sports. These visualizations offer in-

Modality	Task	Worst (%)	Best (%)	Average (%)
Visual	Social	46.15	66.67	54.87
	Sports	20.00	100.00	51.43
	Information Querying	53.85	71.43	62.15
	Sentiment	20.00	73.33	47.71
	Shopping	35.71	75.00	57.68
	Gaming	37.50	100.00	58.24
	Overall	35.20	81.07	55.68
Audio	Social	16.67	40.00	24.17
	Sports	20.00	42.86	33.57
	Information Querying	23.08	36.67	33.27
	Sentiment	21.43	40.00	35.36
	Shopping	23.08	60.00	44.10
	Gaming	25.00	25.00	25.00
	Overall	21.88	40.42	32.58
Both	Social	66.67	100.00	79.37
	Sports	42.86	100.00	80.95
	Information Querying	73.33	89.66	81.00
	Sentiment	86.67	93.33	89.17
	Shopping	60.00	100.00	80.00
	Gaming	57.14	87.50	71.13
	Overall	64.78	95.08	80.94

Table D.1. Worst, Best, and Average human performance (%) across tasks and modalities.

sight into the internal structure of each agentic task and highlight the variety of situations represented within the dataset.

Models	Overall	
	w/o subs	w/ subs
<i>Open source MLLMs w/ Truncated Videos</i>		
InternVL2	0.896	1.055
Qwen2.5-VL	1.651	1.929
LLaVA-OneVision	1.774	1.124
DeepSeek-VL2-Tiny	0.839	0.870
DeepSeek-VL2-Small	1.128	1.471
<i>Closed-source MLLMs w/ Truncated Videos</i>		
GPT-4o	1.327	2.308
o1	2.697	3.075
Claude 3.5 Sonnet	1.964	2.568
Grok2-Vision	2.193	2.765
Nova-Lite	1.444	1.859
Nova-Pro	1.998	2.364
Gemini 1.5 Pro	2.170	2.538
<i>Open source MLLMs w/ Full-Length Videos</i>		
InternVL2	0.895	1.102
Qwen2.5-VL	1.489	1.830
LLaVA-OneVision	1.471	1.666
DeepSeek-VL2-Tiny	0.675	0.808
DeepSeek-VL2-Small	0.839	1.295
<i>Closed-source MLLMs w/ Full-Length Videos</i>		
GPT-4o	1.061	2.148
o1	2.484	2.983
Claude 3.5 Sonnet	0.017	0.401
Grok2-Vision	1.935	2.622
Nova-Lite	1.653	1.708
Nova-Pro	1.615	2.012
Gemini 1.5 Pro	1.956	2.318

Table D.2. GPT-4o evaluation of consistency score.

Models	Overall	
	w/o subs	w/ subs
<i>Open source MLLMs w/ Truncated Videos</i>		
InternVL2	0.129	1.007
Qwen2.5-VL	1.404	1.702
LLaVA-OneVision	1.168	1.005
DeepSeek-VL2-Tiny	0.562	0.614
DeepSeek-VL2-Small	1.087	1.851
<i>Closed-source MLLMs w/ Truncated Videos</i>		
GPT-4o	1.277	2.103
o1	2.484	2.803
Claude 3.5 Sonnet	2.122	2.283
Grok2-Vision	1.901	2.434
Nova-Lite	1.287	1.602
Nova-Pro	1.518	1.741
Gemini 1.5 Pro	1.970	2.299
<i>Open source MLLMs w/ Full-Length Videos</i>		
InternVL2	1.333	0.975
Qwen2.5-VL	1.275	1.660
LLaVA-OneVision	1.277	1.572
DeepSeek-VL2-Tiny	0.687	1.168
DeepSeek-VL2-Small	0.873	1.268
<i>Closed-source MLLMs w/ Full-Length Videos</i>		
GPT-4o	1.277	2.103
o1	2.484	2.803
Claude 3.5 Sonnet	0.319	0.415
Grok2-Vision	1.693	2.189
Nova-Lite	1.196	1.481
Nova-Pro	1.384	1.678
Gemini 1.5 Pro	1.808	2.136

Table D.3. GPT-4o evaluation of correctness score.

Models	Overall	
	w/o subs	w/ subs
<i>Open source MLLMs w/ Truncated Videos</i>		
InternVL2	0.586	1.120
Qwen2.5-VL	1.746	2.033
LLaVA-OneVision	0.953	1.090
DeepSeek-VL2-Tiny	0.653	0.68
DeepSeek-VL2-Small	1.162	1.493
<i>Closed-source MLLMs w/ Truncated Videos</i>		
GPT-4o	1.284	2.220
o1	3.012	3.236
Claude 3.5 Sonnet	1.984	2.613
Grok2-Vision	2.454	2.968
Nova-Lite	1.339	1.701
Nova-Pro	1.812	1.841
Gemini 1.5 Pro	2.374	2.716
<i>Open source MLLMs w/ Full-Length Videos</i>		
InternVL2	1.107	1.159
Qwen2.5-VL	1.559	1.996
LLaVA-OneVision	1.40	1.679
DeepSeek-VL2-Tiny	0.839	0.953
DeepSeek-VL2-Small	0.970	1.382
<i>Closed-source MLLMs w/ Full-Length Videos</i>		
GPT-4o	1.032	1.967
o1	2.880	3.159
Claude 3.5 Sonnet	0.146	0.494
Grok2-Vision	2.237	2.792
Nova-Lite	1.273	1.562
Nova-Pro	1.567	1.832
Gemini 1.5 Pro	2.288	2.618

Table D.4. GPT-4o evaluation of detailed orientation score.

Models	Overall	
Human (Video)	1.26	
Human (Audio)	0.58	
Human (Video+Audio)	2.71	
<hr/>		
	w/o subs	w/ subs
<i>Open source MLLMs w/ Truncated Videos</i>		
InternVL2	0.129	1.393
Qwen2.5-VL	2.007	2.249
LLaVA-OneVision	1.168	1.311
DeepSeek-VL2-Tiny	0.839	0.870
DeepSeek-VL2-Small	1.505	1.851
<i>Closed-source MLLMs w/ Truncated Videos</i>		
GPT-4o	1.586	2.499
o1	3.032	3.291
Claude 3.5 Sonnet	2.122	2.734
Grok2-Vision	2.468	2.960
Nova-Lite	1.742	2.039
Nova-Pro	2.046	2.238
Gemini 1.5 Pro	2.435	2.801
<i>Open source MLLMs w/ Full-Length Videos</i>		
InternVL2	1.332	1.366
Qwen2.5-VL	1.818	2.187
LLaVA-OneVision	1.725	1.995
DeepSeek-VL2-Tiny	1.067	1.168
DeepSeek-VL2-Small	1.292	1.655
<i>Closed-source MLLMs w/ Full-Length Videos</i>		
GPT-4o	1.274	2.250
o1	2.912	3.214
Claude 3.5 Sonnet	0.319	0.533
Grok2-Vision	2.240	2.751
Nova-Lite	1.653	1.936
Nova-Pro	1.915	2.2348
Gemini 1.5 Pro	2.257	2.618

Table D.5. GPT-4o evaluation of context score.

		GPT-4o	o1	Gemini 1.5 Pro	Nova Pro	Grok2	Claude 3.5 Sonnet
<i>Difficulty</i>	<i>Easy</i>	69.8	79.3	69.5	51.5	65.1	68.9
	<i>Medium</i>	67.0	76.0	69.4	47.5	61.2	64.6
	<i>Hard</i>	49.6	67.4	49.6	34.1	47.4	44.4
<i>Agentic Tasks</i>	Social	64.6	69.7	62.6	49.5	60.6	66.7
	Sentiment	56.8	74.1	66.7	46.9	51.9	51.9
	Egocentric Agent	62.7	69.5	57.6	40.7	54.2	64.4
	Information Querying	67.5	75.5	69.4	47.2	62.6	64.9
	Sports	66.4	79.0	64.7	40.3	58.8	56.3
	Gaming	64.1	72.8	58.3	54.4	52.4	59.2
	Shopping	68.3	85.7	74.6	47.6	73.0	73.0
<i>Audio Type</i>	Natural Sound	57.4	66.2	54.4	35.3	50.0	48.5
	Speech	67.7	76.7	69.3	44.8	61.7	65.0
	Music	57.7	61.5	69.2	38.5	57.7	57.7
	Artificial Sound	70.0	78.0	56.0	58.0	54.0	60.0
	Mixed Sounds	63.4	78.2	66.0	52.3	63.0	64.9
<i>Topics</i>	Humanities and Society	71.4	78.9	77.4	51.1	66.9	72.2
	Geography and Travel	56.2	68.8	50.0	37.5	43.8	43.8
	Technology and Gaming	65.6	71.8	61.1	50.4	53.4	57.3
	Science and Knowledge	60.7	64.3	53.6	28.6	50.0	53.6
	Movies, TV and Animations	55.2	68.7	55.2	40.3	43.3	55.2
	Arts and Performance	51.7	65.5	58.6	41.4	58.6	75.9
	Pets and Animals	62.2	68.9	64.4	48.9	62.2	57.8
	Business and Commerce	72.1	86.0	79.1	60.5	76.7	75.6
	Life and Practical Skills	74.0	79.8	77.9	51.9	71.2	69.2
	Sports and Adventure	63.5	80.2	65.1	46.0	57.1	57.9
	Social Trends and Reactions	55.6	77.8	57.4	33.3	61.1	59.3
	Vehicles and Transportation	88.9	100.0	55.6	55.6	55.6	55.6
	Low Quality and Extended Content	54.2	58.3	37.5	16.7	50.0	54.2
<i>Multimodal Abilities</i>	Visual Reasoning	59.2	73.2	64.5	42.1	57.5	59.6
	Emotional Inference	60.5	69.8	62.8	46.5	53.5	48.8
	Situational Reasoning	65.0	80.0	70.0	40.0	75.0	75.0
	Causal Reasoning	65.7	73.1	67.2	50.7	58.2	68.7
	Factual Recall	68.9	79.2	68.2	49.2	63.2	65.0
	Spatial Reasoning	62.1	62.1	58.6	41.4	44.8	55.2
	Relationship	75.0	81.2	62.5	50.0	81.2	75.0
	Temporal Distance	66.7	66.7	58.3	58.3	50.0	66.7
Overall		65.4	75.9	66.3	46.9	60.6	63.1

Table D.6. Performance comparison across different models and categories on truncated videos.

		GPT-4o	o1	Gemini 1.5 Pro	Nova Pro	Grok2	Claude 3.5 Sonnet
<i>Difficulty</i>	<i>Easy</i>	64.8	73.0	69.4	58.2	59.2	17.4
	<i>Medium</i>	62.2	72.4	68.0	58.1	55.2	14.8
	<i>Hard</i>	46.5	65.4	55.1	41.7	53.5	17.3
<i>Agentic Tasks</i>	Social	66.7	65.6	59.4	63.5	62.5	9.4
	Sentiment	54.3	67.9	64.2	46.9	55.6	24.7
	Egocentric Agent	55.3	61.7	55.3	19.1	34.0	14.9
	Information Querying	63.9	75.4	68.3	60.7	57.9	17.5
	Sports	55.3	69.9	66.0	55.3	59.2	14.6
	Gaming	60.0	71.8	64.7	48.2	48.2	22.4
	Shopping	60.4	75.7	76.6	64.0	62.2	10.8
<i>Audio Type</i>	Natural Sound	70.5	79.8	61.5	53.8	43.1	30.8
	Speech	56.2	81.2	70.3	57.3	60.7	14.4
	Music	55.3	70.2	56.0	48.0	56.0	16.0
	Artificial Sound	51.9	59.3	58.0	48.0	48.0	28.0
	Mixed Sounds	58.2	59.7	63.8	55.0	54.6	12.8
<i>Topics</i>	Humanities and Society	57.1	67.9	79.1	66.7	64.3	27.1
	Geography and Travel	60.0	68.9	68.8	56.2	56.2	18.8
	Technology and Gaming	61.4	72.9	62.3	50.9	50.0	17.5
	Science and Knowledge	69.4	75.3	63.0	55.6	51.9	11.1
	Movies, TV and Animations	55.0	68.8	59.7	52.2	46.3	11.9
	Arts and Performance	57.7	75.0	57.1	46.4	57.1	17.9
	Pets and Animals	44.4	88.9	60.0	57.8	57.8	17.8
	Business and Commerce	62.5	62.5	77.1	65.7	67.1	12.9
	Life and Practical Skills	57.8	68.6	64.7	57.6	62.4	8.2
	Sports and Adventure	43.9	65.9	69.7	52.3	55.0	16.5
	Social Trends and Reactions	72.2	77.8	57.7	51.9	51.9	15.4
	Vehicles and Transportation	53.4	63.8	44.4	44.4	33.3	0.0
	Low Quality and Extended Content	63.3	73.9	50.0	20.8	50.0	8.3
<i>Multimodal Abilities</i>	Visual Reasoning	62.5	66.7	64.2	51.5	54.4	14.7
	Emotional Inference	68.8	75.0	58.5	53.7	43.9	19.5
	Situational Reasoning	81.8	90.9	66.7	55.6	77.8	16.7
	Causal Reasoning	53.8	56.9	65.5	55.2	55.2	25.9
	Factual Recall	63.8	71.9	69.0	58.1	58.8	15.9
	Spatial Reasoning	52.0	60.0	45.8	41.7	41.7	16.7
	Relationship	66.0	82.0	75.0	68.8	68.8	0.0
	Temporal Distance	56.4	73.9	81.8	54.5	45.5	18.2
Overall		60.6	71.5	66.5	55.5	56.5	16.3

Table D.7. Performance comparison across different models and categories on full-length videos.



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 D.2. 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? ゴメン! 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 D.3. 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 D.4. 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 D.5. 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 D.6. 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 D.7. 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.

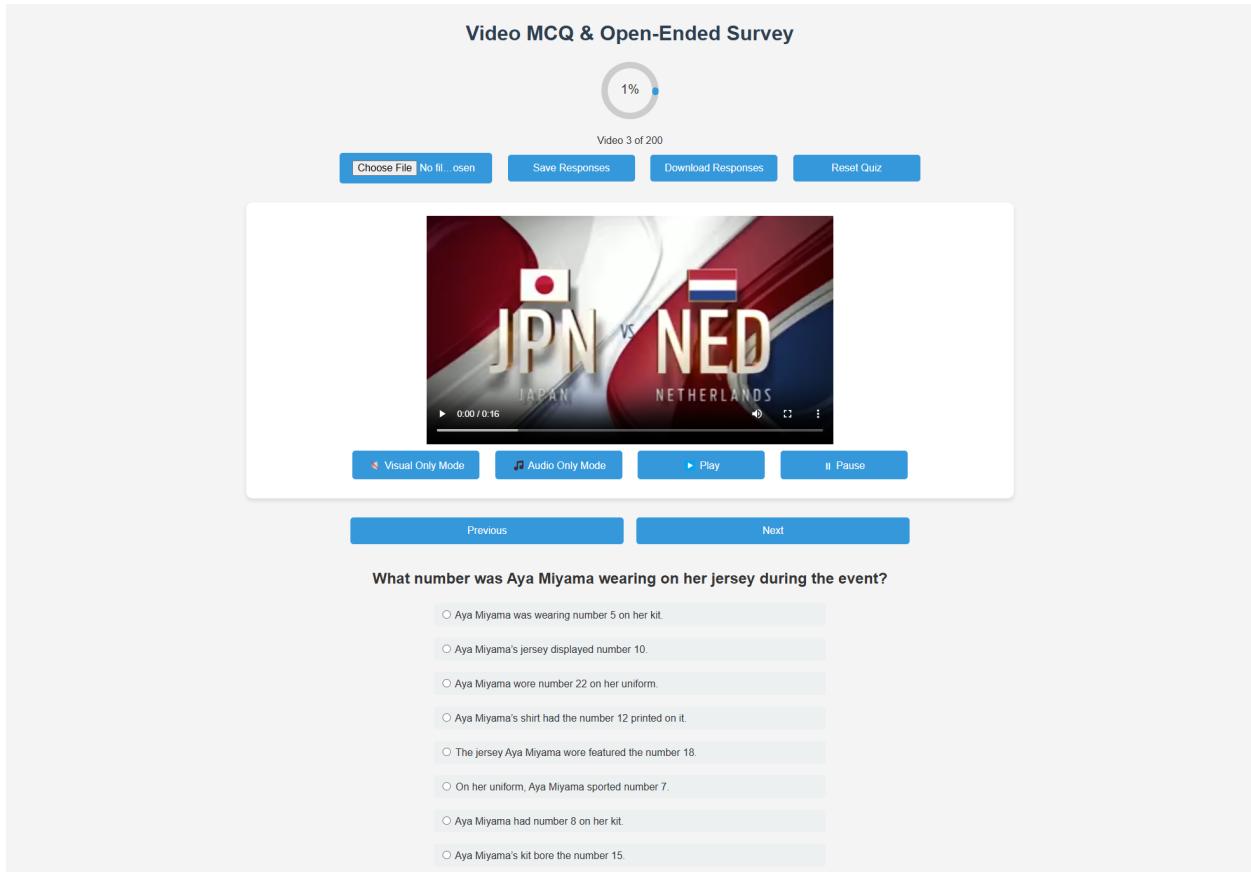


Figure E.1. Screenshot of the questionnaire under the audio-plus-visual condition, showcasing the interface used to assess human performance across multiple-choice questions. Results offer a baseline for comparison with multimodal model performance across diverse tasks and difficulty levels.

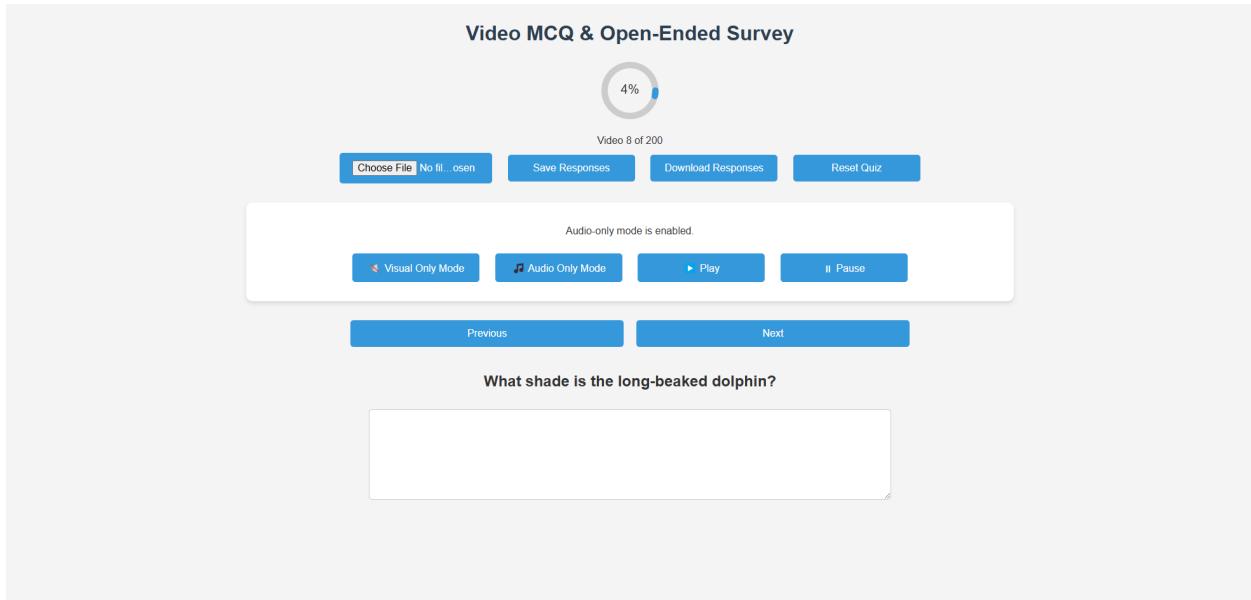


Figure E.2. Questionnaire screenshot showing open-ended questions under the audio-only condition. Responses were evaluated using the GPT-4o grading pipeline, capturing the quality of language generation based solely on audio context.

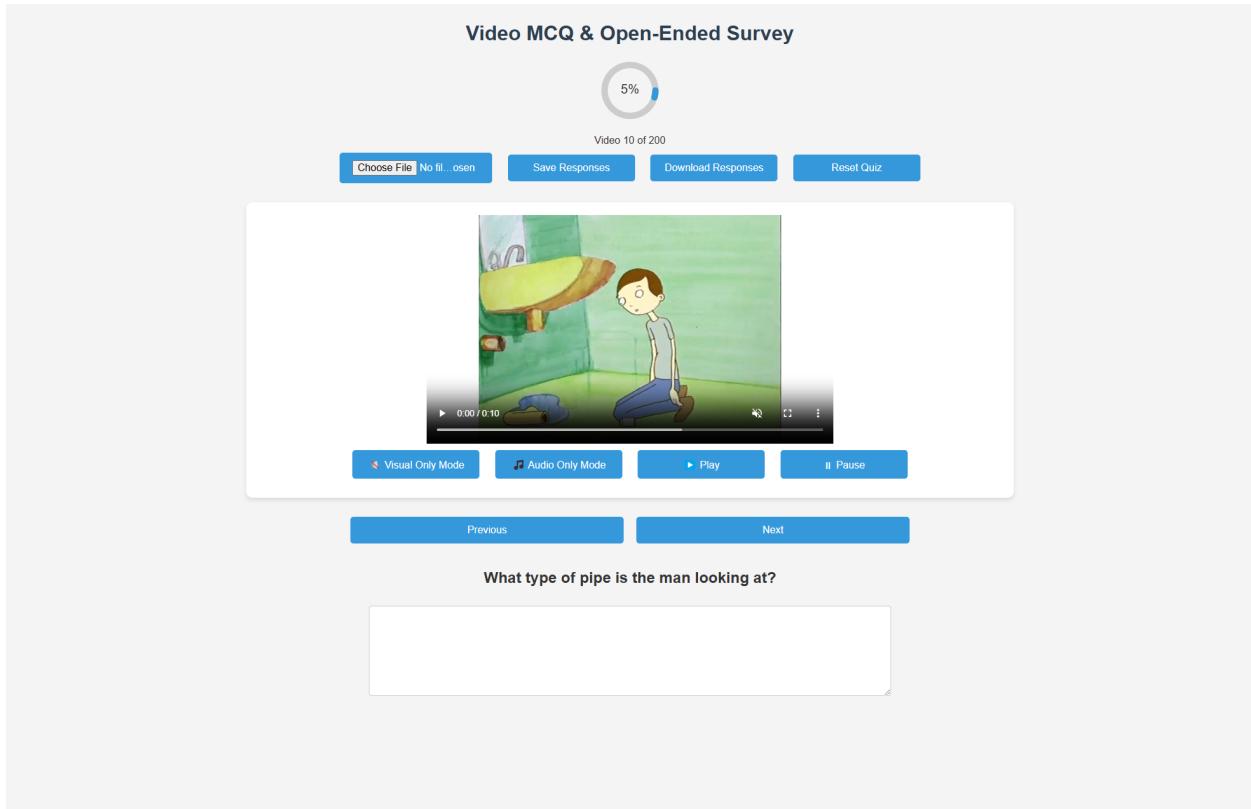


Figure E.3. Questionnaire screenshot showing open-ended questions under the visual-only condition. GPT-4o evaluation scores assess human understanding based solely on visual information, without the aid of audio context.

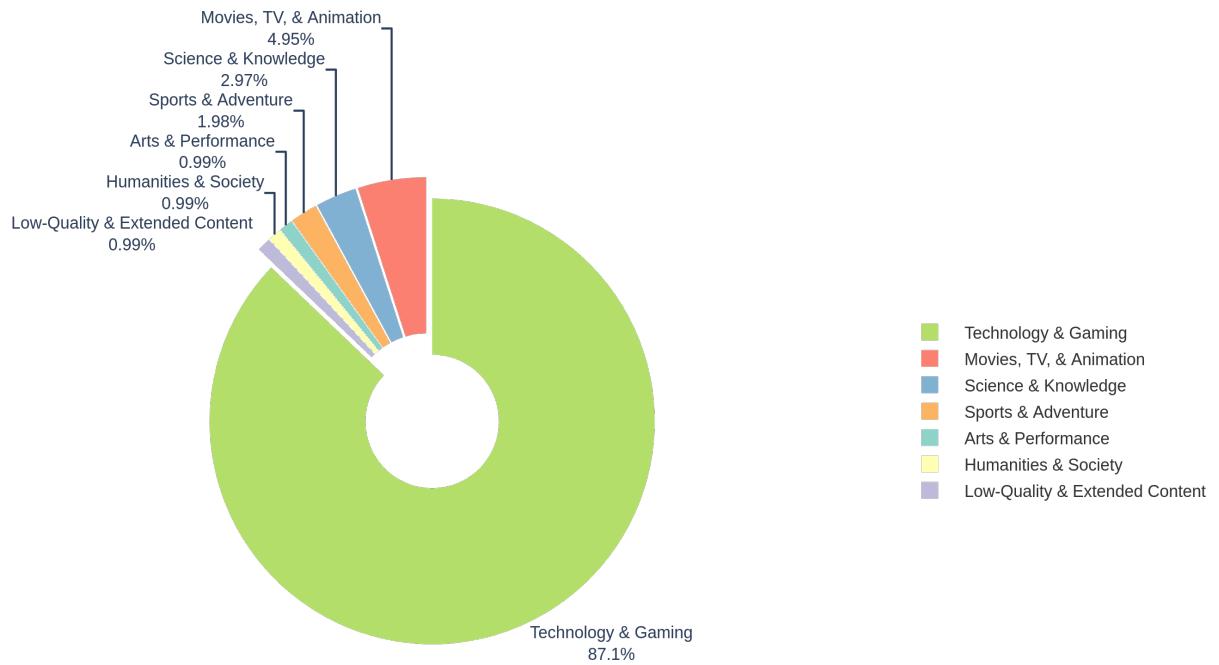


Figure F.1. Taxonomy composition in Gaming.

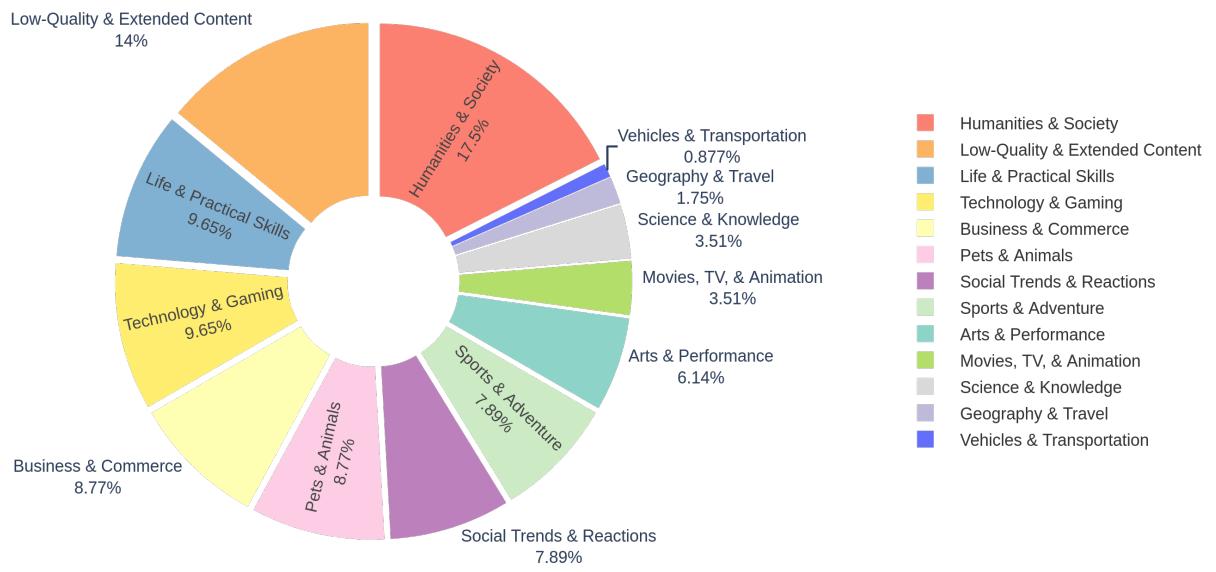


Figure F.2. Taxonomy composition in Egocentric Agent.

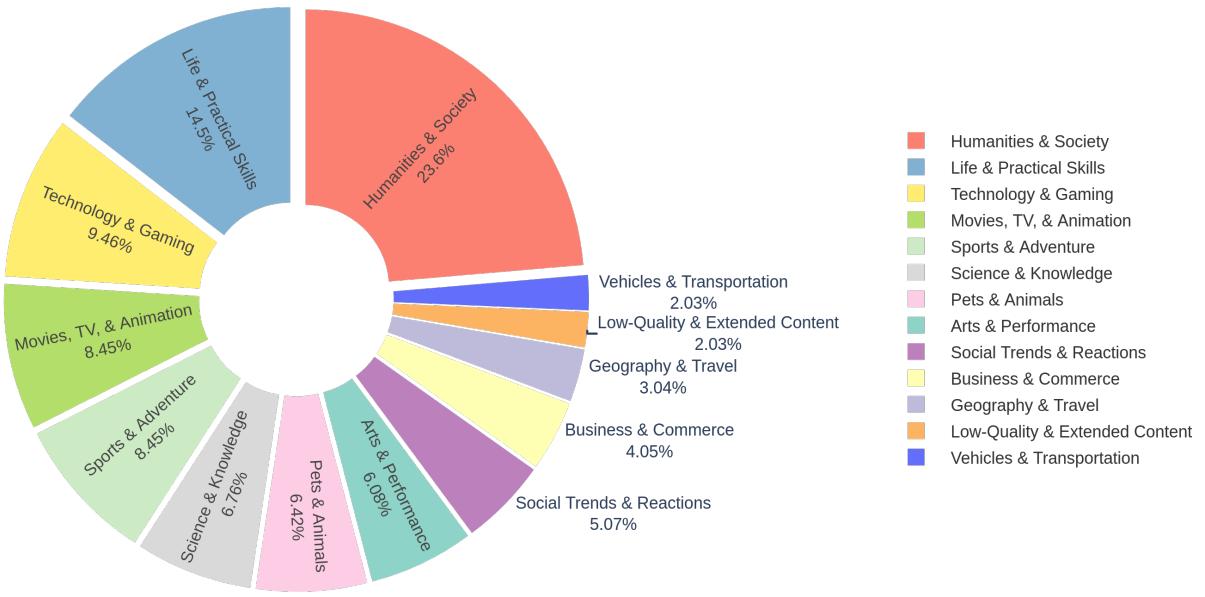


Figure F.3. Taxonomy composition in Information Querying.

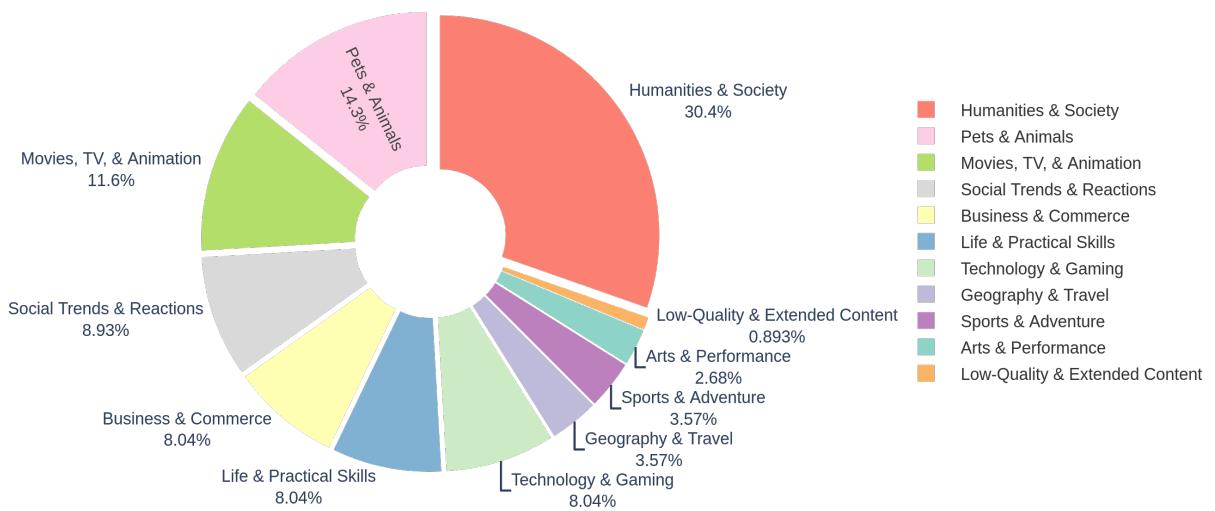


Figure F.4. Taxonomy composition in Sentiments.

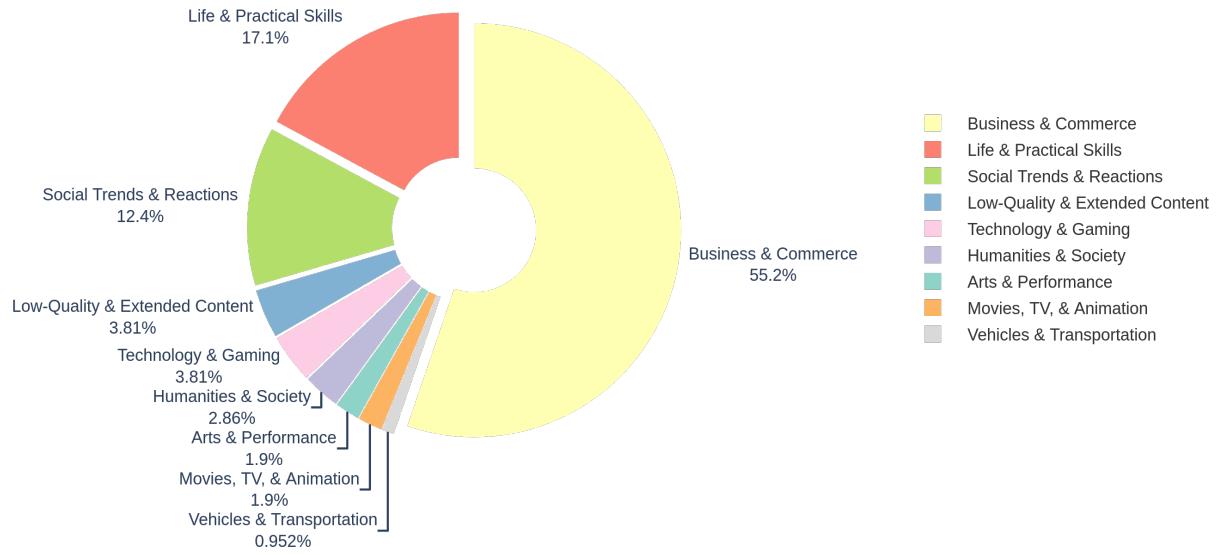


Figure F.5. Taxonomy composition in Shopping.

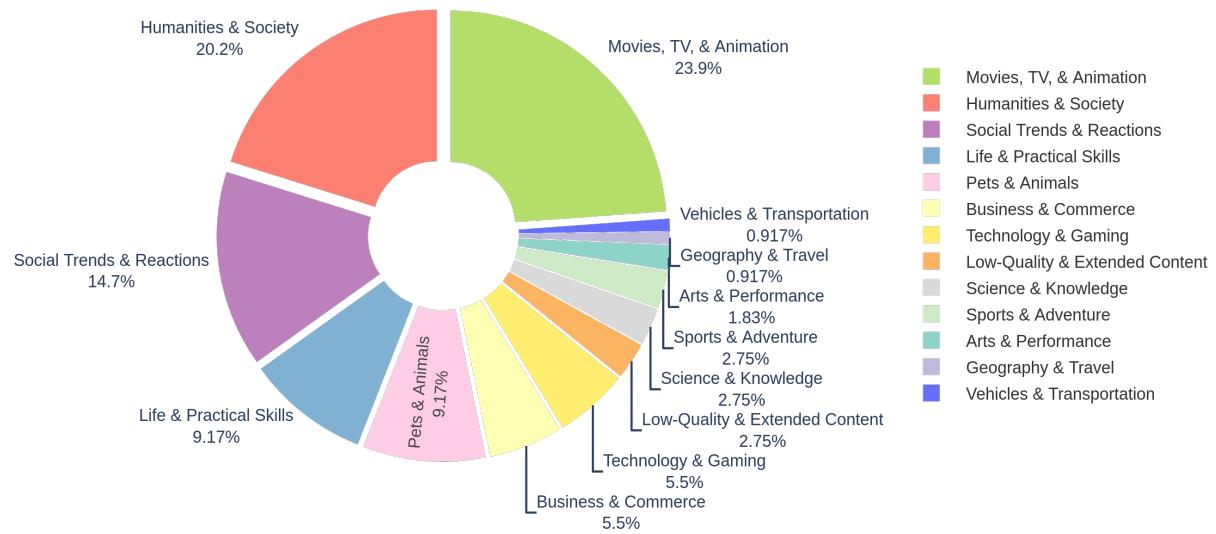


Figure F.6. Taxonomy composition in Social.

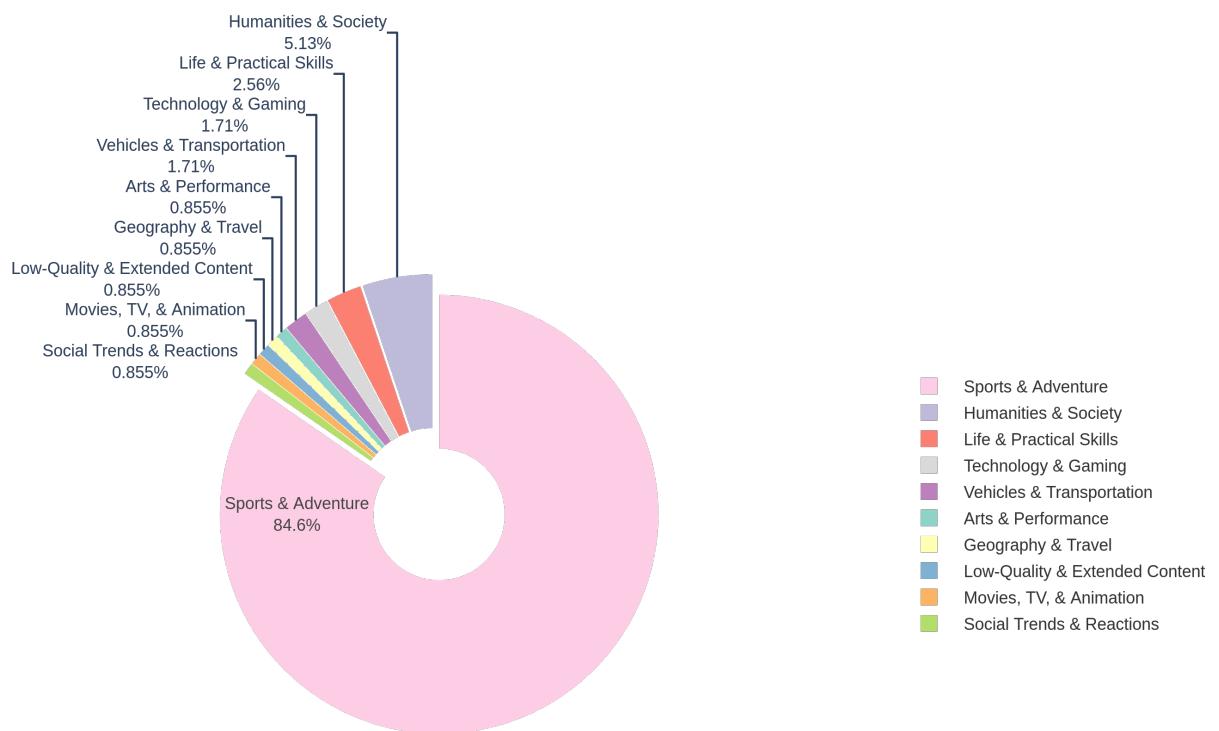


Figure F.7. Taxonomy composition in Sports.