MAVERIX: Multimodal Audio-Visual Evaluation and Recognition IndeX

Liuyue Xie[1*], Avik Kuthiala[1*], George Z. Wei[1*], Ce Zheng[1], Ananya Bal[1], Mosam Dabhi[1], Liting Wen[1], Taru Rustagi[1], Ethan Lai[1], Sushil Khyalia[1], Rohan Choudhury[1], Morteza Ziyadi[2], Xu Zhang[2], Hao Yang[2], László A. Jeni[1]

[1]Carnegie Mellon University
[2]Amazon

Abstract

We introduce MAVERIX (Multimodal Audio-Visual Evaluation and Recognition IndeX), a unified benchmark to probe video understanding in multimodal LLMs, encompassing video, audio, and text inputs with human performance baselines. Although recent advancements in audiovisual models 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 questions that closely mimic the multimodal understanding experiences available to humans during 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, with the website publicly available below.

Project Page — https://maverix-benchmark.github.io

Introduction

Human cognition seamlessly integrates visual and auditory information to reason, infer, and interact within dynamic environments. Replicating this ability in Multimodal Large Language Models (MLLMs) 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; James W. A. Strachan 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. 2025), simple recognition, or questions that can be solved through unimodal cues, such as transcripts (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).

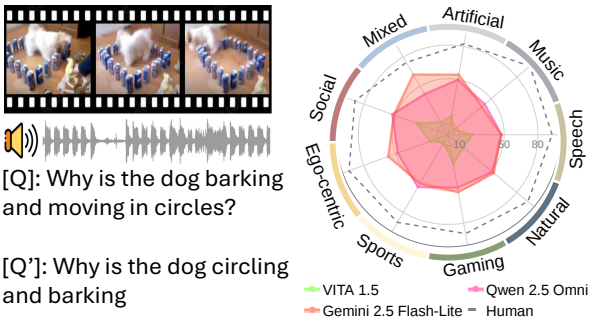

Figure 1: An illustration of our proposed benchmark, which includes highly audiovisual correlated questions and paraphrased questions, can be used to evaluate the model's underlying comprehension abilities and their gaps to humans.

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 insight into the underlying interpretations (Li et al. 2025; Hong et al. 2025). Many existing video-language benchmarks reduce to visible-object recognition or dialog parsing, bypassing the need to synthesize audiovisual dependencies (Patraucean et al. 2023; Kesen 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, senti-

*Equal contribution

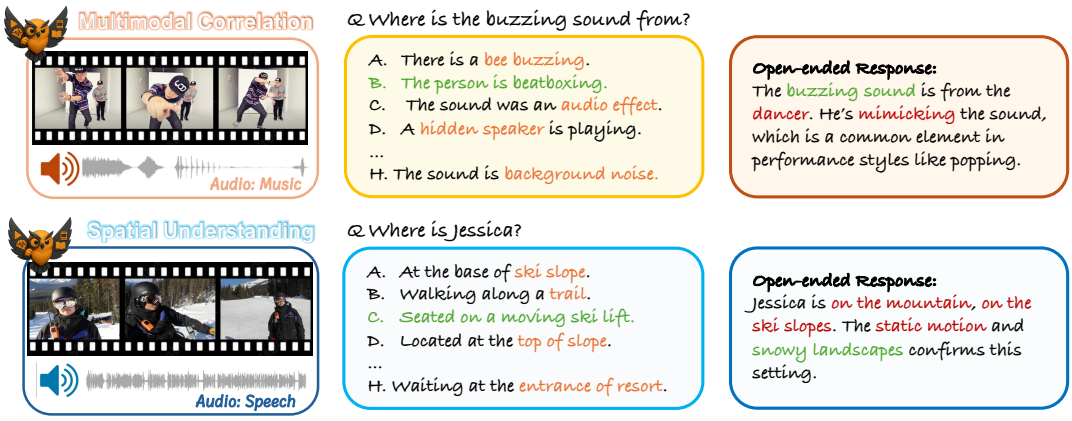

Figure 2: Example Agentic Categories and corresponding QAs in the MAVERIX benchmark.

ment analysis, temporal recall, situational awareness, context understanding, social interaction understanding, and emotional synthesis, covering 700 videos and 2,556 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 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 FL, 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 (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 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 Motivations and Principles

While previous video-understanding benchmarks curate multiple-choice question-answer pairs over different topics, some still suffer from 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 Shortcuts. 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 subtopic 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

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

covering a wide range of taxonomies, where the videos have strong audiovisual correlations. MSR-VTT 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 egocentric, long-duration videos to probe models' understanding of daily interactions for agentic scenarios. The videos are selected by human annotators according to the principles described above, ensuring a thorough distribution between 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 (Amazon Web Services 2005) with 219 unique participants for gauging common consensus, and are

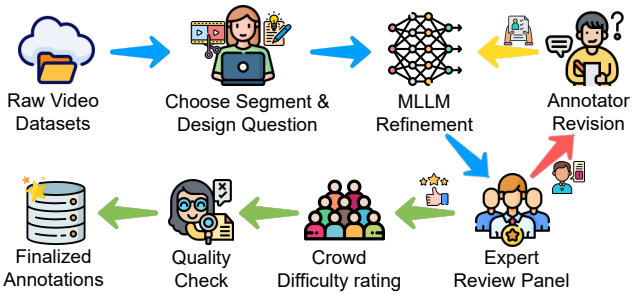

Figure 3: The framework to construct annotation sets with hybrid annotator and MLLM-as-judge quality assurance.

determined based on the subtlety of cross-modal cues, the depth of understanding required, and the ease of locating relevant information in 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 2,556 questions offers one ground-truth answer and seven carefully crafted distractors. Annotators design distractors that remain semantically

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	
Agentic Abilities									
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
QA Lengths									
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
Video–Audio Statistics									
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).

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 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 an 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 all evaluation dimensions to ensure a thorough evaluation. To mitigate positional bias, the 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 addressing modality interdependence and human 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 behavior 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$26–55%, 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, models that trained on smaller datasets, Ola and VITA 1.5, 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 do-

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	+9.5
Qwen-2.5-VL		7B	ViT-LLM	SFT	–	40.3	46.9	46.9	55.3	–	55.3	+8.4
InternVL2		8B	ViT-LLM	SFT	–	24.1	26.3	26.3	33.1	–	33.1	+6.8
LLaVA-OneVision		7B	SigLIP-LLM	SFT	–	44.5	46.8	46.8	55.6	–	55.6	+8.8
DeepSeekVL2-small		2.8B	Hybrid Enc.-MoE	SFT	–	34.3	33.2	34.3	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 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); Δ*Multi* = *Best-Multi* − *Best-Uni*.

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

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

	Taxonomy (AV)			
Models	**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

	Audio Category (AV)				
Models	**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

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

mains.

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 agentic 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% on easy items and +5.3% on hard items. Qwen 2.5 Omni shows a different pattern with substantial benefit on hard items as well. These mixed results suggest that current systems use 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 example of respectable unimodal performance. Humans also benefit from aligned cues and strong priors, yet the jump from

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

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.

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), followed by domain-specific and knowledge-intensive settings (Saikh et al. 2022; Lu et al. 2023). 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; 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; 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; **?**). 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 an open question (James W. A. Strachan 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.

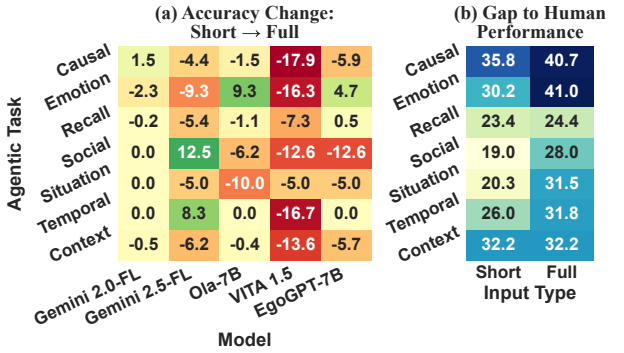

Figure 4: Impact of Video Length on Agentic Category Performances. (a) Accuracy change (%) from short to full-length videos across models and questions. (b) Accuracy gap (%) to human performance for short and full inputs.

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

Acknowledgments

This work was financially supported by Amazon.

References

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.

Amazon Web Services, I. 2005. Amazon Mechanical Turk. https://www.mturk.com/. Crowdsourcing platform service.

Amirizaniani, M.; Martin, E.; Sivachenko, M.; Mashhadi, 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. https://www.anthropic.com/news/claude-3-5-sonnet. 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.; Giallanza, 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.

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

James W. A. Strachan, G. B. O. P. E. S. S. G. K. S. A. R. S. P. G. M. M. S. A. G. C. B., Dalila Albergo. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7): 1285–1295. Publisher Copyright: © The Author(s) 2024.

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, 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, 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. 2025. 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)*.

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 multimodal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, 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.

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. https://docs.x.ai/docs/models/grok-4-0709. 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*.