

VideoAutoArena: An Automated Arena for Evaluating Large Multimodal Models in Video Analysis through User Simulation

Ziyang Luo^{1,2}, Haoning Wu⁴, Dongxu Li⁵, Jing Ma², Mohan Kankanhalli³, Junnan Li¹

¹Salesforce AI Research, ²Hong Kong Baptist University, ³National University of Singapore

⁴Nanyang Technological University, ⁵The Australian National University

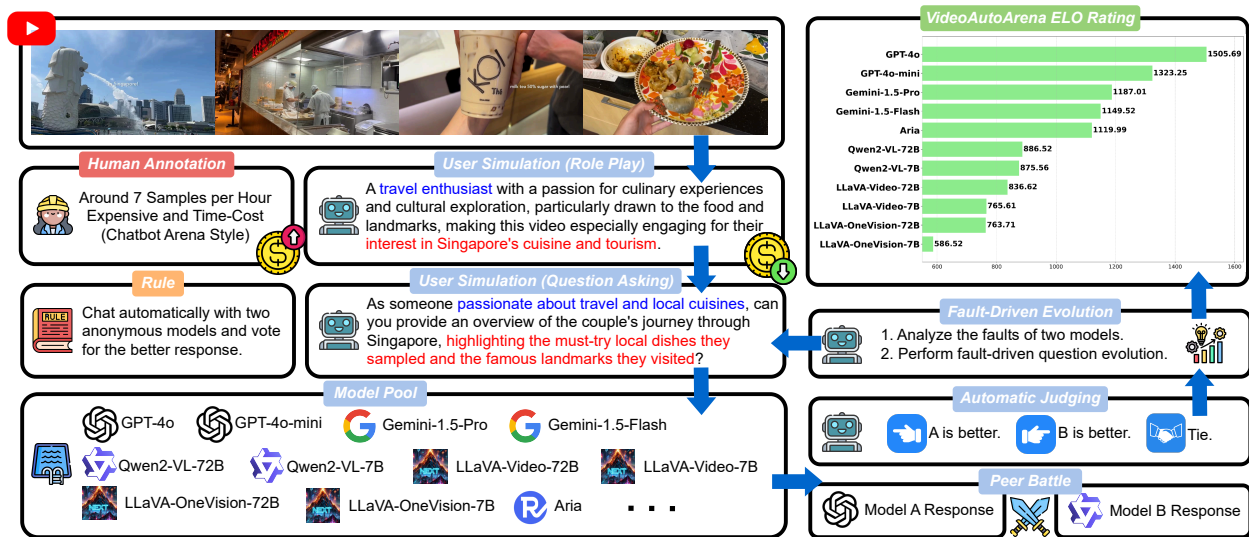


Figure 1. An overview of our **VideoAutoArena**, where we leverage LMMs for user simulation to automatically evaluate LMMs in video analysis, offering an efficient alternative to costly and time-consuming human annotations, distinct from platforms like LMSYS Chatbot Arena [14] and WildVision Arena [45]. In this figure, we showcase 4 sampled frames from a Singapore travel vlog video.

Abstract

Large multimodal models (LMMs) with advanced video analysis capabilities have recently garnered significant attention. However, most evaluations rely on traditional methods like multiple-choice question answering in benchmarks such as VideoMME and LongVideoBench, which are prone to lack the depth needed to capture the complex demands of real-world users. To address this limitation—and due to the prohibitive cost and slow pace of human annotation for video tasks—we introduce **VideoAutoArena**, an arena-style benchmark inspired by LMSYS Chatbot Arena’s framework, designed to automatically assess LMMs’ video analysis abilities. VideoAutoArena utilizes user simulation to generate open-ended, adaptive questions that rigorously assess model performance in video understanding. The benchmark features an automated, scalable evaluation

framework, incorporating a modified ELO Rating System for fair and continuous comparisons across multiple LMMs. To validate our automated judging system, we construct a “gold standard” using a carefully curated subset of human annotations, demonstrating that our arena strongly aligns with human judgment while maintaining scalability. Additionally, we introduce a fault-driven evolution strategy, progressively increasing question complexity to push models toward handling more challenging video analysis scenarios. Experimental results demonstrate that VideoAutoArena effectively differentiates among state-of-the-art LMMs, providing insights into model strengths and areas for improvement. To further streamline our evaluation, we introduce **VideoAutoBench** as an auxiliary benchmark, where human annotators label winners in a subset of VideoAutoArena battles. We use GPT-4o as a judge to compare responses against these human-validated answers. Together, VideoAutoArena and VideoAutoBench offer a cost-effective, and

◇ Project Page: <https://videoautoarena.github.io/>

1. Introduction

Recently, large multimodal models (LMMs) with advanced video understanding capabilities, such as GPT-4o [26], Gemini-1.5-Pro [2, 51], Aria [31], Qwen2-VL [57], and LLaVa-Video [30, 73], have garnered significant attention within the multimodal community. These models represent a major shift in the capabilities of artificial intelligence by extending beyond traditional image-based LMMs [1, 16, 32, 40] to encompass complex video inputs. Unlike their predecessors, which primarily focused on static visual data, these models are designed to handle dynamic, time-variant data, making them well-suited for processing lengthy and intricate video sequences. By leveraging language-based instructions from users, these LMMs can efficiently conduct a range of video analysis tasks, from understanding fine-grained scene details [31] to comprehending overarching narrative structures [26].

To better evaluate the video analysis capabilities of these models, recent works have introduced several widely used benchmarks, including MVBench [35], VideoMME [20], and LongVideoBench [59]. These benchmarks typically share a common feature: they pre-define core video analysis skills, such as object recognition in a single frame and action reasoning across a sequence of frames. They adopt an ability-centric approach, often using multiple-choice questions to assess performance. While these benchmarks have significantly contributed to the development of LMMs, they place limited emphasis on the types of questions real users might ask when seeking assistance with video analysis. In contrast, practical video analysis scenarios are far more complex and diverse in their requirements [14].

To address this gap, we can draw inspiration from the evaluation methods used for large language models (LLMs) [5, 17, 54, 55]. Traditional language benchmarks, such as MMLU [24], IFEval [76], HumanEval [10], and GSM8k [15], also suffer from limited alignment with real user interactions [14, 72, 74]. To mitigate this issue, platforms like LMSYS Chatbot Arena [14] have been introduced, providing an open, crowdsourced platform for evaluating LLMs based on human preferences. Chatbot Arena employs a pairwise comparison approach, collecting feedback from a diverse set of users, ensuring that evaluation questions are generated by real users. This approach overcomes many of the challenges posed by previous benchmarks and has become one of the most widely adopted methods for evaluating LLMs.

One straightforward solution is to adapt this idea directly to LMMs for video analysis. Indeed, recent work, such as WildVision Arena [45], has attempted to do so. However, an

analysis of the WildVision Video Arena leaderboard reveals certain challenges with this approach. The video arena has received only 256 votes across 11 models, with each model participating in an average of just 23 battles since its release approximately 6 months ago. This limited number of battles can likely be attributed to the increased complexity involved in formulating questions for video-based tasks. Unlike language and image data, where questions can be quickly generated or verified within seconds, videos are typically longer and contain richer, more complex contexts. Annotators must spend significantly more time watching and understanding the video content before formulating high-quality questions. In our preliminary attempts, we hired annotators to complete this task, and the time cost proved significant, with **a maximum output of only 7 samples annotated per hour**. This time investment hinders the scalability of generating high-quality questions for LMMs, a crucial factor for ensuring the effectiveness of arena-style evaluations.

To address the limitations of existing video analysis benchmarks, we propose **VideoAutoArena**, a fully automated, arena-style evaluation method for LMMs. Unlike human-driven platforms, VideoAutoArena leverages LMM agents for user simulation and preference selection, eliminating the need for costly human annotators and enabling scalable, efficient evaluations. The framework also integrates fault-driven hard prompt evolution, which generates progressively challenging questions based on model performance, ensuring more rigorous testing. By simulating real-world user behavior, VideoAutoArena bridges the gap between ability-centric evaluations and practical application demands. Our human preference experiments show that **84.20%** of the time, questions in VideoAutoArena better mimic real-world user question styles compared to VideoMME and LongVideoBench. Additionally, **87.29%** of the time, our automatic judging aligns with human preference selection.

Experiments on 11 well-known proprietary and open-source LMMs reveal that open-source models still lag behind the SOTA closed-source model GPT-4o in video analysis, with a significant performance gap (-385.7). This gap is notably larger than those observed in traditional multiple-choice question-answering benchmarks. The disparity becomes even more pronounced as video length increases or the difficulty of the questions rises. Furthermore, when focusing on user-background relevance and helpfulness, the performance gap widens further. These findings highlight how our benchmark offers a user-centric perspective, providing valuable insights for the development of LMMs.

To complement VideoAutoArena, we also introduce **VideoAutoBench**, a streamlined benchmark designed for faster, more accessible evaluation of LMMs in video anal-

The numbers are recorded on Nov. 14, 2024 at <https://huggingface.co/spaces/WildVision/vision-arena>.

Benchmark	Venue	Long Video Included	User-Centric	Scalable	Open-Ended	Automated
MVBench [35]	CVPR 24	✗	✗	✗	✗	✓
MLVU [77]	Arxiv 24	✓	✗	✗	✗	✓
LVBench [58]	Arxiv 24	✓	✗	✗	✗	✓
VideoMME [20]	Arxiv 24	✓	✗	✗	✗	✓
LongVideoBench [59]	NeurIPS 24	✓	✗	✗	✗	✓
WildVision Video Arena [45]	NeurIPS 24	?	✓	✗	✓	✗
VideoAutoArena (Ours)	CVPR 25	✓	✓	✓	✓	✓

Table 1. Comparison of recent popular benchmarks for video analysis. WildVision video data are not yet publicly available.

ysis. VideoAutoBench leverages a curated subset of battles from VideoAutoArena, where human annotators have selected the winning model responses. Using GPT-4o for automatic judging, VideoAutoBench compares model answers against these human-selected and rejected responses, providing an efficient, cost-effective assessment method. Our results show that the rankings from VideoAutoBench align closely with those from VideoAutoArena, with a significant gap between SOTA closed-source and open-source LMMs, underscoring the benchmark’s challenge.

2. Related Work

LMMs with advanced video understanding capabilities have garnered significant research attention. For the closed-source models, GPT-4o [26] and Google’s Gemini-1.5 [2, 51] demonstrate SOTA video analysis performance. Meanwhile, the open-source community has made notable strides [7, 9, 12, 13, 19, 21, 25, 27, 29, 34, 38, 39, 41, 42, 44, 46, 48, 52, 63–67, 69, 71]. Notably, the LLaVa series [40] has been updated to the LLaVa-Video [73] and LLaVa-OneVision models [30], along with the release of all training data. The Qwen-VL model [3] has also been upgraded to the Qwen2-VL version [57], and the first open-source multimodal mixture-of-experts (MoE) model, Aria [31], has recently been introduced. These contributions have significantly narrowed the gap between closed-source and open-source models in video understanding. To accelerate the development of LMMs in video analysis, the establishment of more comprehensive benchmarks is essential.

In the early phase, researchers primarily relied on benchmarks featuring short videos [11, 22, 23, 50], such as MSVD-QA [62], MSRVT-QA [62], NExT-QA [60], and MVBench [35]. However, these benchmarks have limitations due to their short video durations, averaging less than 50 seconds. This brevity restricts their ability to comprehensively evaluate the temporal understanding capabilities of LMMs, thereby hindering further advancements in LMMs development. To address these limitations, benchmarks like ActivityNet-QA [68] and EgoSchema [49] have extended video durations to approximately 180 seconds on average. More recently, research has introduced even more comprehensive benchmarks [43, 58, 70].

For instance, MovieChat-1K [53] assesses LMMs using movie videos with an average duration of 500 seconds, while LongVideoBench [59] focuses on long-context interleaved evaluation with an average duration of 473 seconds. Additionally, MLVU [77] presents a LMMs’ ability-centric benchmark, featuring substantial extensions of video lengths. Furthermore, VideoMME [20] introduces a highly comprehensive benchmark that includes short, medium, and long videos, further enhancing the evaluation of LMMs’ temporal understanding abilities. As discussed in the introduction, most current benchmarks are limited by multiple-choice questions that diverge from real user interaction. To address this, we introduce **VideoAutoArena**, which evaluates LMMs through open-ended, simulated human questions. Table 1 highlights the differences between our benchmark and other recent video-based benchmarks.

3. VideoAutoArena

3.1. Overview

As illustrated in Figure 1, the **VideoAutoArena** pipeline consists of four core components: user simulation, peer battles, automatic judging, and fault-driven evolution. Initially, an agent reviews a video to identify user personas likely to be interested in the content. Adopting one of these personas, the agent formulates a relevant question about the video. Two randomly selected models then engage in peer battles to respond to the question. A judging agent determines which model provides the better response, while an analysis agent evaluates the responses, performs fault analysis, and generates progressively challenging questions to further assess the models’ capabilities. To demonstrate the effectiveness of our VideoAutoArena, we use 2,881 videos from LongVideoBench, with an average duration of 479 seconds. These videos are categorized into four duration ranges—(8s, 15s], (15s, 60s], (180s, 600s], (900s, 3600s]—and 10 categories: Movie, Life Vlogs, Geography, History, News Programs, Art, STEM, Computer Science, Cooking Recipes, and Travel Guides. The data distribution are provided in Figure 2. Our benchmark is not restricted to specific videos, and new videos can be easily incorporated into the evaluation pipeline.

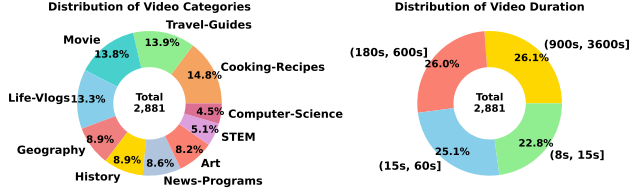


Figure 2. Video statistics by category and duration.

3.2. User Simulation

Ideally, real user queries would offer the direct evaluation of LMM performance in real life video analysis. However, given the complex contextual demands of video content, **human annotation is expensive and time-cost**. To address this, **VideoAutoArena** adopts user simulation with role-play, with SOTA LMMs acting as agents to generate realistic, user-centric questions and preferences, enabling a more practical evaluation. In language-based role-play [6, 8, 28, 33], there is typically no need to consider external context, such as videos, allowing role-play to freely generate diverse questions. In video analysis, however, questions are constrained by the video content. To address this, we introduce a novel role-play method called **video content-constrained persona generation**. Given a video, agents first identify the types of users likely to be interested in it, defining three user types: (1) users with backgrounds highly relevant to the video; (2) users with moderately relevant backgrounds; and (3) users with unrelated backgrounds who encounter the video by chance. This relevance-based categorization aims to emulate real users with varied backgrounds seeking assistance from LMMs for video analysis. Once user types are established, agents adopt these personas to generate persona-specific questions for detailed video analysis. This user-centric process sets our evaluation method apart from previous ability-centric benchmarks. The prompts used are provided in the Appendix A.1.

3.3. Peer Battles and Fault-Driven Evolution

Once the role-play agent generates a question for a specific video, we initiate peer battles between two anonymous models, following the Chatbot Arena style. Two models are randomly chosen from the model pool, presented with the same video and question, and asked to generate responses. Our goal is to fairly compare the quality of these responses to determine which LMM provides a better answer.

Similar to the concept of Hard Prompts in Chatbot Arena [36], incorporating more challenging questions can further push the boundaries of evaluating the abilities of current LMMs in video analysis. Thus, we aim to create a harder question set for evaluation purposes. Unlike Chatbot Arena, which can directly source hard prompts from millions of real user queries, our approach is limited to the

prompts generated by a user simulator. To derive harder questions, we draw inspiration from the famous instruction synthesis method, Evol-Instruct [47, 61], which evolves existing questions into more complex ones using predefined heuristic strategies. However, because Evol-Instruct generates questions based on a similar prompt structure for each evolution, it encounters limitations in question diversity. To address this, we introduce a **fault-driven question evolution** strategy that iteratively increases question complexity. Rather than relying on isolated prompts, each new evolution generates questions based on the results of the previous model battle. This approach creates a more adaptive and progressively challenging environment for the models, pushing them to respond to increasingly complex questions.

In this framework, a response analysis agent initially reviews responses from two competing models, identifying specific faults and performance weaknesses. Based on this assessment, role-play agents then generate tailored questions aimed at probing these weaknesses, making the question synthesis process progressively more fault-driven. After a new question is generated, a complexity evaluation agent assesses its difficulty. If the new question receives a higher overall complexity score than the previous one, it is retained for subsequent model battles. This iterative approach establishes a rigorous testing environment, challenging models with increasingly complex and contextually nuanced tasks, thus providing a deeper evaluation of each model’s video understanding capabilities. The prompts used are provided in the Appendix A.2.

3.4. Judging and Ranking

A key aspect of arena-style evaluation is determining the winner in each model comparison. In Chatbot Arena, human annotators directly express their preferences, but in **VideoAutoArena**, human annotation is costly and difficult to scale due to the time-intensive nature of video analysis. To address this, we aim to automate the judgment process. Drawing inspiration from automated judging benchmarks like Arena-Auto-Hard [36] and MT-Bench [75], we first define our judging standards as follows:

1. **Instruction Following**: The response should closely adhere to the user’s instructions, ensuring it directly addresses the specified task.
2. **Accuracy**: The response must utilize information from the video accurately, avoiding fabrication or misquotation. It should maintain factual correctness, avoid hallucinations, and demonstrate contextual coherence with precise terminology and knowledge.
3. **Relevance**: The response should consider the user’s background information and needs, providing a comprehensive, detailed answer that addresses the question directly without straying off-topic. Responses should be thorough, offering multiple perspectives where relevant.

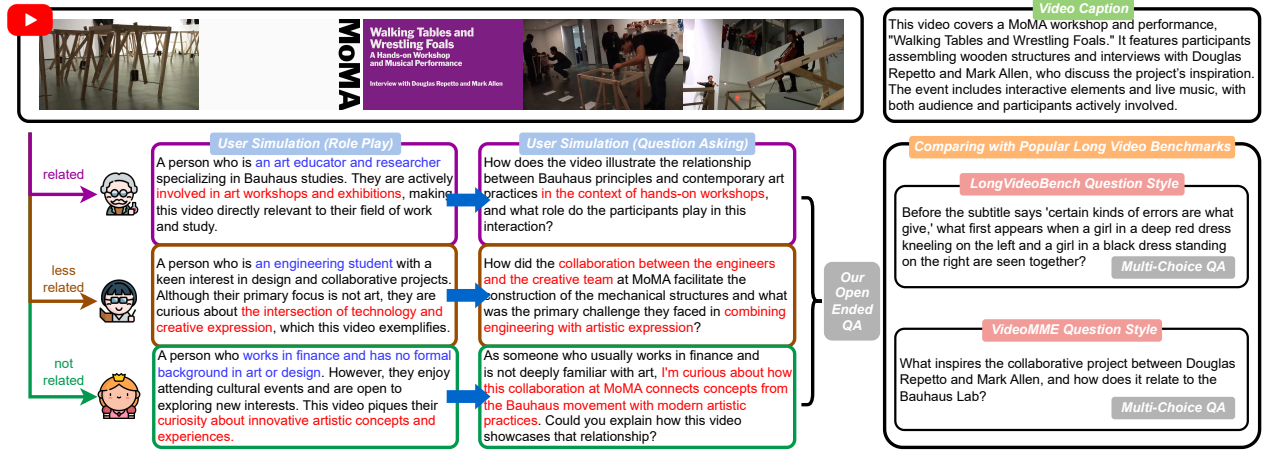


Figure 3. Examples of synthesized personas with three levels of relevance and corresponding synthesized questions. We also compare the style of our questions with those in popular long-video benchmarks, including LongVideoBench and VideoMME.

4. **Helpfulness:** The response should provide valuable information to aid the user in understanding or solving their issue, avoiding irrelevant or vague content.

Based on these standards, we adopt the LMM-as-a-Judge paradigm [36, 45, 75] to automatically determine the better response between two models. The prompts used are provided in the Appendix A.3.

After obtaining the automatic judging results, we utilize the ELO Rating System [18] to establish a dynamic evaluation platform that ranks LMMs through statistical modeling based on direct pairwise comparisons. Here, we briefly explain the Online ELO Rating and statistical estimation methods. The Online ELO Rating system calculates the probability that model i will win against model j using their respective ratings, R_i and R_j , where $i, j \in N$. For each comparison, we define a binary outcome Y_{ij} , where $Y_{ij} = 1$ if model i wins and $Y_{ij} = 0$ otherwise. The probability is computed as follows:

$$P(Y_{ij} = 1) = \frac{1}{1 + 10^{(R_j - R_i)/\alpha}},$$

where $\alpha = 400$ is the scaling factor in the ELO computation. After each comparison, player ratings are updated by:

$$R'_i = R_i + K \times (S(i, j) - P(Y_{ij} = 1)),$$

where $S(i, j)$ represents the actual outcome (1 for a win, 0.5 for a tie, and 0 for a loss). Higher-rated players gain fewer points if they win and lose more if defeated, while lower-rated players experience the reverse. Since ELO updates are sensitive to comparison order, we employ the Bradley-Terry model [4] for stable statistical estimation, as in Chatbot Arena.

The Bradley-Terry model refines ELO ratings through a logistic regression model using maximum likelihood esti-

mation. For N models with pairwise comparisons, where W_{ij} is the count of times model i wins over j , the log-likelihood function for all comparisons is:

$$\mathcal{R} = \sum_{i, j \in N, i \neq j} W_{ij} Y_{ij} P(Y_{ij} = 1),$$

where $\mathcal{R} = \{R_1, \dots, R_N\}$ represents each model's rating. Since this approach doesn't accommodate ties directly, we split all tie votes, counting half as wins for model i ($Y_{ij} = 1$) and the other half as wins for model j ($Y_{ij} = 0$). This ensures balanced ranking and fair statistical estimation across all model comparisons.

3.5. Experiments

Setup. To demonstrate the effectiveness of VideoAutoArena, we evaluate 11 SOTA LMMs, including GPT-4o/mini, Gemini-1.5-Pro/Flash, Aria, Qwen2-VL-72B/7B, LLaVa-Video-72B/7B, and LLaVa-OneVision-72B/7B, all of which have shown strong performance on the VideoMME. For response generation, each video was uniformly sampled to provide 64 frames as input. Since most of these LMMs do not support audio, the audio track was converted to subtitles and combined with the question as input. For automatic judging, each video was sampled to provide 128 frames and combined with subtitles. Additional experimental details are provided in the Appendix B.

User Simulation and Diversity. In our experiments, we generated an average of three personas per video across 2.9k videos, resulting in about 8.6k unique personas. As shown in Figure 3, we showcase examples of three personas with varying levels of relevance to a given video. Each persona includes motivations or reasons for their interest in

Claude-Series LMMs were excluded from our experiments due to limited support for long video input (maximum 20 frames).

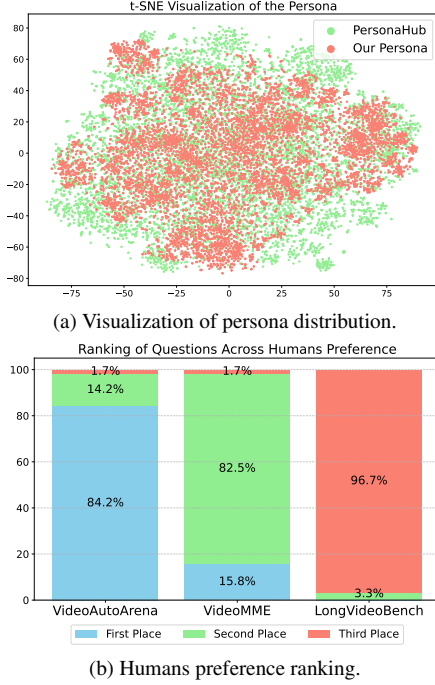


Figure 4. Our user simulation offers diverse personas and more effectively mirrors real-world users’ question styles.

the video, enabling more persona-specific question generation. Moreover, in Figure 4a, we compare the distribution of our synthesized personas with those from PersonaHub [6], which includes diverse personas automatically curated from web. To map the distribution, we used one of the SOTA sentence embedding models, `gte-large-en-v1.5` [37], to encode each persona description into vectors, then applied t-SNE [56] for dimensionality reduction. The results show that our synthesized personas achieve diversity comparable to that in PersonaHub. Notably, our personas are constrained by video content, while PersonaHub personas are synthesized without such constraints, highlighting that our personas effectively simulate a range of realistic user backgrounds. Additional examples of synthesized personas are provided in Appendix C.1.

After generating personas, our role-play agent adopts each persona to ask relevant questions about the video content, seeking help from AI assistants. In our experiments, we synthesized one question per unique personas for 2.9k videos. To assess how well our questions mimic real-world user queries, we randomly selected 120 questions and compared them with question styles from LongVideoBench and VideoMME. Since our videos are from LongVideoBench, we used the same set for consistency. However, since VideoMME uses different videos, we applied a style transfer method, adapting VideoMME questions to synthesize similar-style questions based on our benchmark’s videos. This allowed a fair comparison across the three benchmarks

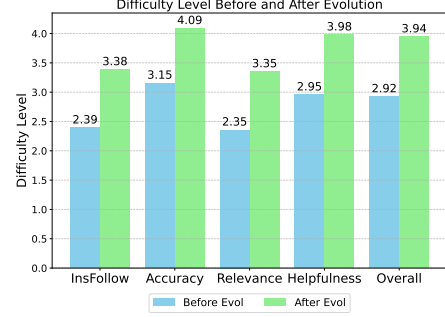


Figure 5. Our fault-driven evolution strategy generates increasingly challenging questions for video analysis.

on identical video samples. To evaluate the naturalness of each benchmark’s question style, we conducted a blind ranking task, asking three skilled annotators to rank which questions best mimic real-world user queries, with Rank 1 as the best. Annotation guidelines are detailed in the Appendix D.1. As shown in Figure 4b, for 84.2% of the comparisons, VideoAutoArena’s questions ranked first in mimicking real-world user question styles. This result highlights how our benchmark brings a unique perspective to evaluating LMMs for video analysis. Figure 3 further provides example questions to illustrate the style of our questions in mimicking real-world users. In Appendix C.2, we also include more synthesized question examples.

Question Evolution. To analyze the hard prompts generated by our fault-driven evolution, we perform an automated complexity assessment of the synthesized questions. Using the SOTA LMM, GPT-4o, we evaluate the difficulty level of these questions on a 1–5 scale before and after evolution. We further break down the difficulty into 5 categories—instruction following, accuracy, relevance, helpfulness and overall—based on our automatic judging standards. As shown in Figure 5, the evolved questions consistently achieve higher difficulty scores across all categories, demonstrating the effectiveness of our evolution strategy. The analysis prompts are included in Appendix A.4.

Automatic Judging. To establish a “gold standard” for evaluating the accuracy of automatic judging, framed as a 3-choice task (A, B, or Tie), we created a benchmark guided by our judging criteria. Since no public human annotations are available for this, we engaged annotators to carefully evaluate a subset of battles. Annotation guidelines are detailed in the Appendix D.2. Given the labor-intensive nature of this work—yielding about 7 annotations per hour—we randomly selected around 300 battles across various video lengths and models for annotation. Figure 6 presents the accuracy comparison among SOTA LMMs and voting methods for automated judging. GPT-4o demonstrated the highest alignment with human preferences, achieving an 87.29% agreement. Notably, employing a voting approach with multiple LMMs (Top2 to Top4) did not result in better

Models	Size	ELO	Win Rates	(8 s, 15 s]	(15 s, 60 s]	(180 s, 600 s]	(900 s, 3600 s]
<i>Proprietary Models</i>							
GPT-4o	-	1505.69	89.19	1447.86	1449.59	1575.34	1552.23
GPT-4o-mini	-	1323.25	76.90	1293.27	1343.28	1327.75	1349.29
Gemini-1.5-Pro	-	1187.01	65.11	1247.65	1171.82	1263.58	1291.64
Gemini-1.5-Flash	-	1149.52	62.07	1081.58	1131.27	1140.07	1260.36
<i>Open-Source Models</i>							
Aria	8×3.5B	1119.99	59.54	1147.45	1273.77	1110.67	1111.40
Qwen2-VL	72B	886.52	35.61	985.46	928.23	829.65	826.56
Qwen2-VL	7B	875.56	34.90	969.28	859.33	850.30	829.21
LLaVA-Video	72B	836.62	30.25	796.90	850.12	827.88	782.55
LLaVA-Video	7B	765.61	23.52	672.35	736.14	759.15	721.78
LLaVA-OneVision	72B	763.71	23.11	731.50	710.64	759.29	741.80
LLaVA-OneVision	7B	586.52	9.86	626.70	545.82	556.31	533.18

Table 2. Our **VideoAutoArena** Leaderboard. We show the overall ELO ratings and win rates within four different video lengths.

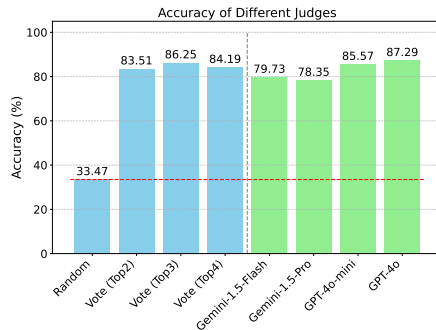


Figure 6. Evaluate the accuracy of various judging methods using human annotations as the gold standard. In the Vote (Top N) method, the top N models are used to cast votes.

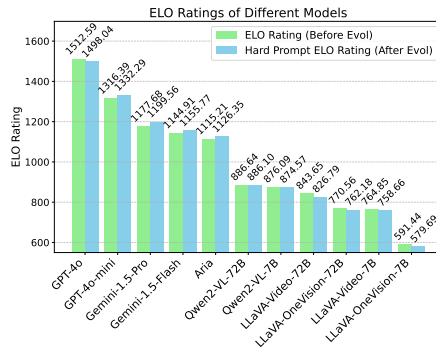


Figure 7. ELO ratings for models competing on questions before and after applying fault-aware evolution.

agreement. As a result, we selected GPT-4o as the primary judge for determining the winning responses due to its strong alignment with human judgments.

Leaderboard. As shown in Table 2, we present the ELO ratings and win rates across 4 video length categories for 11 SOTA LMMs in video analysis. Our evaluation involves a total of 12,479 head-to-head battles, with each model participating in roughly 1,600 battles, and each model pair competing approximately 150 times—significantly surpassing the scale of WildVision Video Arena and demonstrating the scalability of our method. In terms of ELO ratings,

the leading open-source LMM, Aria, lags behind the top proprietary LMM, GPT-4o, by a notable margin (-385.7), underscoring the benchmark’s strong discriminative power. Unlike in VideoMME, where the score gap among the top six models is less than 10%. For shorter videos (under 60 seconds), the gap between the top proprietary and open-source LMMs narrows, with Aria even surpassing the well-known Gemini-1.5-Pro on videos between 15 and 60 seconds. However, as video length increases, performance gaps widen considerably; for instance, Qwen2-VL sees a drop of over 100 ELO points, while Gemini-1.5-Flash gains around 180 ELO points when comparing results on short versus long videos. Notably, Aria demonstrates stable and strong performance across varying video lengths.

In addition, Figure 7 presents the ELO ratings for models competing on questions before and after applying fault-aware evolution. When challenged with more difficult prompts, all open-source LMMs except Aria exhibit lower scores, whereas all proprietary LMMs, apart from GPT-4o, show improved scores. Figure 8 further details the performance of various LMMs across different judging standards. Notably, the gap between proprietary and open-source LMMs is most evident in the helpfulness and relevance tracks, compared to the instruction-following and accuracy tracks. This gap suggests that many open-source LMMs are primarily optimized for traditional, ability-focused benchmarks like VideoMME, which overlook user-centered aspects such as contextual relevance and information helpfulness. By focusing on a more comprehensive evaluation, our VideoAutoArena bridges this gap, providing deeper insights into the limitations and future development potential of LMMs. In the Appendix C.3, we also include examples of different LMMs’ responses as case studies to better understand the weaknesses of these models.

Battle Example. As shown in Figure 9, we present an example of a battle between Aria and LLaVa-Video-72B. Both models follow the question and provide accurate infor-

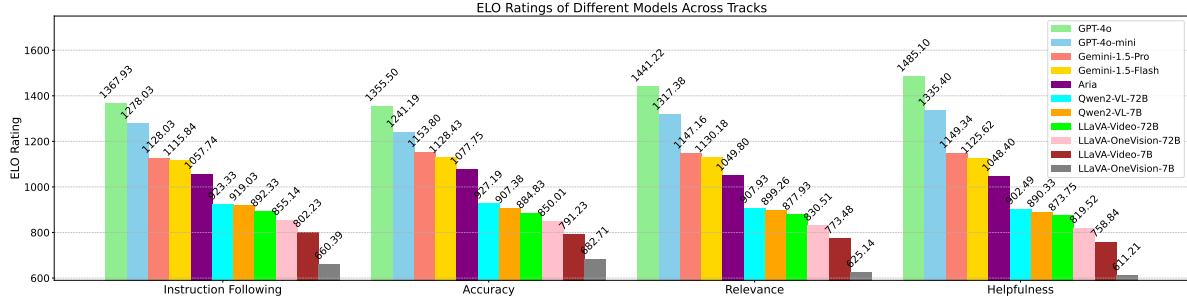


Figure 8. We evaluate the performance of various models based on four different judging standards.

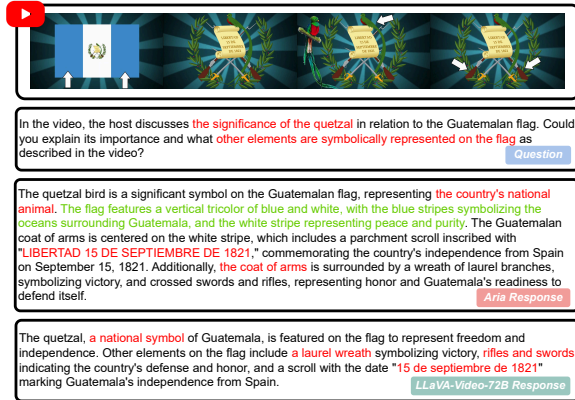


Figure 9. Example of a battle between Aria and LLaVa-Video-72B. Red highlights key content, while green highlights important details mentioned only by Aria.

mation about the Guatemalan flag. However, only Aria includes additional details, such as the flag’s vertical tricolor, while LLaVa-Video’s response lacks sufficient depth on these aspects. Consequently, Aria’s response is more helpful for users seeking to understand the flag’s symbolism. The automated judge also selects Aria as the winner. Traditional multiple-choice QA benchmarks, like VideoMME, cannot reveal such open-ended limitations in LMMs. More examples are included in the Appendix C.3.

4. VideoAutoBench

While VideoAutoArena offers a novel approach to evaluating LMMs in video analysis, it is less immediately user-friendly than multiple-choice benchmarks like VideoMME and LongVideoBench, which provide straightforward scores based on model responses. In contrast, VideoAutoArena requires the target model to engage in comparative battles with other models to generate results. To streamline this evaluation process, we introduce **VideoAutoBench**, which combines the user-centric assessment strengths of VideoAutoArena with the simplicity and speed of traditional benchmarks. In our automated judging experiments, we included human annotators to label winners for a subset of battles, using these questions and non-

Models	vs. Sel.	vs. Rej.	Avg.
GPT-4o	70.98	94.12	82.55
GPT-4o-mini	49.80	92.16	70.98
Gemini-1.5-Pro	28.24	82.74	55.49
Gemini-1.5-Flash	27.25	81.96	54.61
Aria	19.80	76.86	48.33
Qwen2-VL-72B	13.92	64.71	39.32
Qwen2-VL-7B	11.96	60.00	35.98
LLaVA-Video-72B	7.45	56.08	31.77
LLaVA-OneVision-72B	4.12	52.16	28.14
LLaVA-Video-7B	5.29	46.67	25.98
LLaVA-OneVision-7B	3.53	30.98	17.26

Table 3. LMMs compete against human selected or rejected answers in our **VideoAutoBench**.

tied responses as reference answers. In VideoAutoBench, we employ GPT-4o as the judging model to evaluate each model’s responses against the human-selected or rejected answers, with GPT-4o voting based on the same standards used in VideoAutoArena. When competing with human-selected answers, a model earns 1 point for a win, 0.5 for a tie, and 0 for a loss, with the final score being the average across all battles. When competing with human-rejected answers, only a win earns 1 point. Table 3 presents the performance of different LMMs, showing the similar rank as in VideoAutoArena.

5. Conclusion

We introduce VideoAutoArena, an automated arena-style benchmark that addresses the limitations of traditional multiple-choice video QA benchmarks. Using user simulation, peer battles, automated judging, and fault-driven evolution, VideoAutoArena enables a scalable, user-centric evaluation for complex video analysis tasks. Alongside, we present VideoAutoBench, a streamlined evaluation comparing model responses to human-labeled answers. Experiments on 11 SOTA LMMs reveal a notable performance gap between closed and open-source models, particularly for long videos, challenging questions, and scenarios involving user background relevance and response helpfulness.

Acknowledgment

This work is partially supported by National Natural Science Foundation of China Young Scientists Fund(No. 62206233) and Hong Kong RGC ECS (No. 22200722).

References

- [1] Gpt-4v(ision) system card. 2023. [2](#)
- [2] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023. [2](#), [3](#)
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023. [3](#)
- [4] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952. [5](#)
- [5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. [2](#)
- [6] Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *CoRR*, abs/2406.20094, 2024. [4](#), [6](#)
- [7] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, and Limin Wang. Videollm: Modeling video sequence with large language models. *CoRR*, abs/2305.13292, 2023. [3](#)
- [8] Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Shiwen Ni, and Min Yang. Agentcourt: Simulating court with adversarial evolvable lawyer agents. *CoRR*, abs/2408.08089, 2024. [4](#)
- [9] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. *CoRR*, abs/2406.04325, 2024. [3](#)
- [10] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021. [2](#)
- [11] 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. *CoRR*, abs/2311.14906, 2023. [3](#)
- [12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238, 2023. [3](#)
- [13] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476, 2024. [3](#)
- [14] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [1](#), [2](#)
- [15] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. [2](#)
- [16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In

Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. 2

- [17] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. 2
- [18] A.E. Elo. *The USCF Rating System: Its Development, Theory, and Applications*. United States Chess Federation, 1966. 5
- [19] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *CoRR*, abs/2408.14023, 2024. 3
- [20] Chaoyou Fu, Yuhao Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiaowu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *CoRR*, abs/2405.21075, 2024. 2, 3
- [21] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiaowu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. VITA: towards open-source interactive omni multimodal LLM. *CoRR*, abs/2408.05211, 2024. 3
- [22] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haefliger, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thureau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5843–5851. IEEE Computer Society, 2017. 3
- [23] Madeleine Grunze-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA: A benchmark for compositional spatio-temporal reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11287–11297. Computer Vision Foundation / IEEE, 2021. 3
- [24] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2
- [25] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower LLM to grasp video moments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14271–14280. IEEE, 2024. 3
- [26] OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alexandre Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codisoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, B. Ghorbani, Ben Leimberger, Ben Rossen, Benjamin D. Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Gertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Ming Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Chris Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mély, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Hai-Biao Bao, Haitang Hu,

Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Pondé de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian D. Kivlichan, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, İbrahim Cihangir Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub W. Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Ryan Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quiñero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Joshua Gross, Josh Kaplan, Josh Snyder, Josh Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Ouyang Long, Louis Feuvrier, Lu Zhang, Lukasz Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljube, Ma teusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Ali Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nickolas Stathas, Nick Turley, Nikolas A. Tezak, Niko Felix, Nithanth Kudige, Nitish Shirish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Phil Tillet, Prafulla Dhariwal, Qim ing Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Raphael Gontijo Lopes, Raul Puri, Reah Miyara, Reimar H. Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Ramilevich Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam S. Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Her-

mani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal A. Patwardhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiye Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card. 2024. 2, 3

- [27] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univ: Unified visual representation empowers large language models with image and video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13700–13710. IEEE, 2024. 3
- [28] Chuyi Kong, Ziyang Luo, Hongzhan Lin, Zhiyuan Fan, Yaxin Fan, Yuxi Sun, and Jing Ma. From general to specific: Utilizing general hallucination to automatically measure the role relationship fidelity for specific role-play agents. 2024. 4
- [29] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkan Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *CoRR*, abs/2305.03726, 2023. 3
- [30] 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. *CoRR*, abs/2408.03326, 2024. 2, 3
- [31] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. 2024. 2, 3
- [32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 19730–19742. PMLR, 2023. 2
- [33] Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunchwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents. *CoRR*, abs/2405.02957, 2024. 4
- [34] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355, 2023. 3
- [35] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*

- 2024, Seattle, WA, USA, June 16-22, 2024, pages 22195–22206. IEEE, 2024. 2, 3
- [36] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline, 2024. 4, 5
- [37] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. 6
- [38] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *CoRR*, abs/2311.10122, 2023. 3
- [39] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: on pre-training for visual language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26679–26689. IEEE, 2024. 3
- [40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 2, 3
- [41] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoli Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *CoRR*, abs/2408.15542, 2024. 3
- [42] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. ST-LLM: large language models are effective temporal learners. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVII*, pages 1–18. Springer, 2024. 3
- [43] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Shihuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 8731–8772. Association for Computational Linguistics, 2024. 3
- [44] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx MLLM: on-demand spatial-temporal understanding at arbitrary resolution. *CoRR*, abs/2409.12961, 2024. 3
- [45] Yujie Lu, Dongfu Jiang, Wenhui Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *CoRR*, abs/2406.11069, 2024. 1, 2, 3, 5
- [46] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *CoRR*, abs/2306.07207, 2023. 3
- [47] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 4
- [48] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *CoRR*, abs/2306.09093, 2023. 3
- [49] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 3
- [50] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doversch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchet, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 3
- [51] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024. 2, 3
- [52] Michael S. Ryoo, Honglu Zhou, Shrikant B. Kendre, Can Qin, Le Xue, Manli Shu, Silvio Savarese, Ran Xu, Caiming Xiong, and Juan Carlos Niebles. xgen-mm-vid (blip-3-video): You only need 32 tokens to represent a video even in vlms. 2024. 3
- [53] Enxin Song, Wenhao Chai, Guan hong 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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*,

- Seattle, WA, USA, June 16-22, 2024, pages 18221–18232. IEEE, 2024. 3
- [54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. 2
- [55] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruian Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. 2
- [56] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 6
- [57] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao 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. *CoRR*, abs/2409.12191, 2024. 2, 3
- [58] Weihai Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark. *CoRR*, abs/2406.08035, 2024. 3
- [59] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *CoRR*, abs/2407.15754, 2024. 2, 3
- [60] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9777–9786. Computer Vision Foundation / IEEE, 2021. 3
- [61] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 4
- [62] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 3
- [63] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See-Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *CoRR*, abs/2404.16994, 2024. 3
- [64] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. *CoRR*, abs/2408.10188, 2024.
- [65] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A GPT-4V level MLLM on your phone. *CoRR*, abs/2408.01800, 2024.
- [66] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multimodal large language models. *CoRR*, abs/2408.04840, 2024.
- [67] En Yu, Liang Zhao, Yana Wei, Jinrong Yang, Dongming Wu, Lingyu Kong, Haoran Wei, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Wenbing Tao. Merlin: Empowering multimodal llms with foresight minds. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part IV*, pages 425–443. Springer, 2024. 3
- [68] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9127–9134. AAAI Press, 2019. 3
- [69] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 543–553. Association for Computational Linguistics, 2023. 3
- [70] Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao. Movqa: A benchmark of versatile question-answering for long-form movie understanding. *CoRR*, abs/2312.04817, 2023. 3
- [71] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkan Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *CoRR*, abs/2406.16852, 2024. 3

- [72] Shudan Zhang, Hanlin Zhao, Xiao Liu, Qinkai Zheng, Zehan Qi, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Natural-CodeBench: Examining coding performance mismatch on HumanEval and natural user queries. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7907–7928, Bangkok, Thailand, 2024. Association for Computational Linguistics. [2](#)
- [73] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. 2024. [2](#), [3](#)
- [74] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [2](#)
- [75] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. [4](#), [5](#)
- [76] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911, 2023. [2](#)
- [77] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: A comprehensive benchmark for multi-task long video understanding. *CoRR*, abs/2406.04264, 2024. [3](#)