VRBench: A Benchmark for Multi-Step Reasoning in Long Narrative Videos

Jiashuo Yu[1*] Yue Wu[1*] Meng Chu[1*] Zhifei Ren[1*] Zizheng Huang[2,3,1*] Pei Chu[1*]
Ruijie Zhang[1] Yinan He[1] Qirui Li[1] Songze Li[1] Zhenxiang Li[1] Zhongying Tu[1]
Conghui He[1] Yu Qiao[1] Yali Wang[4, 1✉] Yi Wang[1, 3✉] Limin Wang[2, 1✉]
[1]Shanghai Artificial Intelligence Laboratory [2]Nanjing University [3]Shanghai Innovation Institute
[4]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
VRBench.github.io

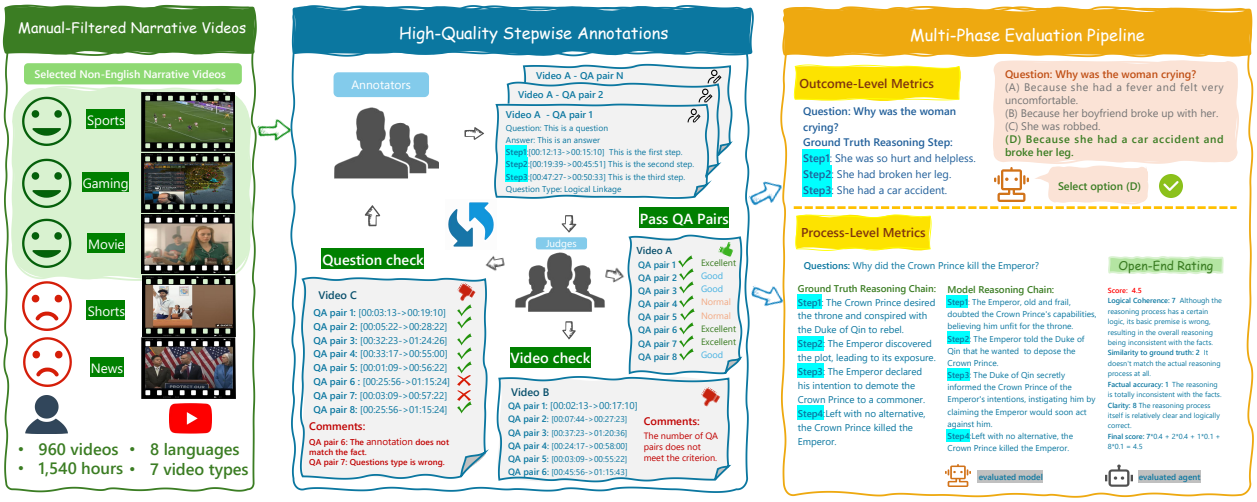

Figure 1. **Overview of VRBench.** We present VRBench, a long narrative video benchmark for multi-step reasoning. VRBench includes 960 **manual-filtered narrative videos**, covering 8 languages and 7 video categories that are suitable for reasoning about temporal relations. We also provide **high-quality stepwise annotations** for reasoning, which are labeled and reviewed by human experts. Each video incorporates 8-10 complex question-answer pairs, a multi-step reasoning chain, and fine-grained timestamps. To fully evaluate the capability of models in multi-step reasoning, we propose a **multi-phase evaluation pipeline** that assesses model results both from the process and outcome levels. Our VRBench is the first video reasoning benchmark that supports both multi-step annotation and evaluation.

Abstract

*We present **VRBench**, the first long narrative video benchmark crafted for evaluating large models' multi-step reasoning capabilities, addressing limitations in existing evaluations that overlook temporal reasoning and procedural validity. It comprises 960 long videos (with an average duration of 1.6 hours), along with 8,243 human-labeled multi-step question-answering pairs and 25,106 reasoning steps with timestamps. These videos are curated via a multi-stage filtering process including expert inter-rater reviewing to prioritize plot coherence. We develop a human-AI collaborative framework that generates coherent reasoning chains, each requiring multiple temporally grounded steps, spanning seven types (e.g., event attribution, implicit inference). VRBench designs a multi-phase evaluation pipeline that assesses models at both the outcome and process levels. Apart from the MCQs for the final results, we propose a progress-level LLM-guided scoring metric to evaluate the quality of the reasoning chain from multiple dimensions comprehensively. Through extensive evaluations of 12 LLMs and 19 VLMs on VRBench, we undertake a thorough analysis and provide valuable insights that advance the field of multi-step reasoning.*

1. Introduction

The rapid evolution of vision language models (VLMs) has heightened the need for benchmarks that rigorously eval-

*equal contributions. ✉corresponding authors.

uate complex reasoning capabilities. While existing standards like GSM8K [8] focus on domain-specific knowledge in mathematics and science, they neglect a critical reasoning dimension: contextual analysis in visual narrative content. Real-world applications increasingly demand temporal reasoning across interconnected elements: tracking character dynamics in films, interpreting gameplay strategies, or understanding cause-and-effect chains in documentaries. This capability gap persists because current video benchmarks [16, 53, 70, 77] primarily assess single-step perception rather than sustained reasoning processes.

We find three fundamental limitations in current evaluation paradigms: (1) an overemphasis on domain expertise rather than plot-driven reasoning, (2) the lack of temporally-grounded reasoning chains in annotations, and (3) Outcome-focused metrics that ignore procedural validity. To address these challenges, we present VRBench (as in Figure 1), the first benchmark specifically designed for multi-step reasoning in long-form narrative videos. Its construction involves the collection of narrative videos, a human-in-the-loop reasoning process method for annotation, and a multi-phase evaluation pipeline considering both procedure and outcome.

In data construction, VRBench aggregates 960 meticulously selected videos (1.6h average duration) across seven narrative categories (e.g., movies, sports, travelogues), sourced through a multi-stage filtering process. Our curation pipeline combines automated retrieval with expert validation (inter-rater reliability ρ=0.82), prioritizing plot coherence over domain-specific knowledge. This contrasts with existing benchmarks like MMVU [92] that emphasize disciplinary competence. To conduct multi-step reasoning chain annotation, we develop a human-AI collaborative approach, generating 8-10 QA pairs per video with explicit temporal grounding. Each question requires no less than 2 reasoning steps annotated with precise timestamps (Figure 3), validated through iterative expert review (95% inter-annotator agreement). The taxonomy spans seven reasoning types from event prediction to implicit inference, ensuring comprehensive coverage of narrative analysis skills.

When evaluating models on VRBench, we propose a multi-phase evaluation pipeline that combines outcome verification with process analysis. Beyond conventional multiple-choice accuracy that evaluates the outcome-level performance, we further introduce the process-level metric LLM-guided scoring to evaluate the overall reasoning process quality. This dual approach reveals critical insights – for instance, while GPT-4o achieves 81.25% outcome accuracy, its process rating scores drop to 56.11%, exposing reasoning fragility. Evaluations of 31 state-of-the-art large models reveal proprietary VLMs with long-context support outperform text-only LLMs by 13.82% absolute, emphasizing the importance of dense visual grounding. System-2 optimization strategies yield disproportionate gains in process metrics, and despite parameter parity, open-source VLMs lag behind proprietary counterparts by 12.30%, suggesting architectural limitations beyond scale. Our ablation studies show strong alignment between human and LLM evaluations and quantify the impact of test-time scaling—doubling context windows improves QwQ-32B's [65] accuracy by 12.43%.

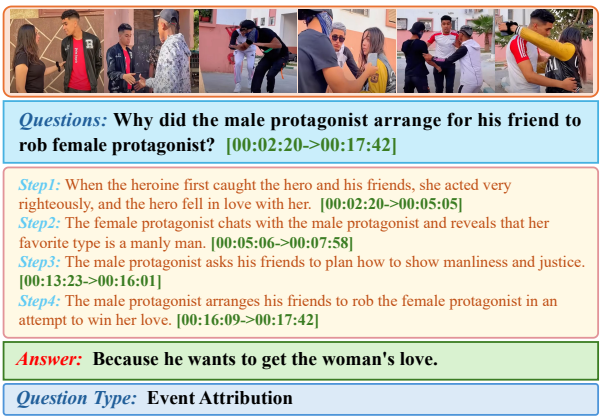

Figure 2. **An example of annotation in VRBench**. For each question, VRBench provides the question-answer pair, multi-step reasoning chain, question type, and the start-to-end timestamps of the entire question as well as each reasoning step.

VRBench sets a new standard for evaluating narrative reasoning, offering insights distinct from knowledge-centric benchmarks. By separating domain expertise from contextual analysis, our work aids in developing models capable of sustained, human-like understanding in real-world video content. The entire suite, including annotation protocols and evaluation tools, has been fully open-sourced to advance reasoning research.

2. Related Work

Reasoning Large Models. As the capabilities of LLMs continue to expand, frontier models have demonstrated significant potential in addressing high-order reasoning tasks. OpenAI pioneered this advancement with the release of o1 [48], an LLM capable of sophisticated reasoning tasks. Subsequently, a series of proprietary [12, 49, 59] and open-source [3, 18, 65] LLMs designed for advanced reasoning have emerged. Furthermore, researchers have shifted their focus towards developing VLMs possessing similar higher-order reasoning capabilities. Previously, some VLMs [2, 7, 30, 32, 37, 71, 73, 88, 91] to a certain extent exhibited good reasoning proficiency due to their proficiency in tackling long-sequence context. More recently, a series of VLMs [14, 66, 69, 81] trained on data incorporating higher-order Chains-of-Thought in RL-based algorithms [33, 52, 55] have been developed, enabling them

Dataset	#Size	#QA pairs	#Dur.(s)	QA types	Data Source	Anno Step	Eval Step	Eval Target	Anno Type	Clue	Multilingual
MMLU [21]	14,079	14,079	/	MCQ	Multi-Disc	Single	Single	LLM	M	✖	✖
MMLU-Pro [72]	12,032	12,032	/	MCQ	Multi-Disc	Single	Single	LLM	A+M	✖	✖
LiveCodeBench [26]	511	511	/	MCQ	Code	Multi	Single	LLM	M	✖	✖
SciEval [57]	15,901	15,901	/	MCQ+Open	Science	Single	Single	LLM	M	✖	✖
GSM8k [8]	8,792	8,792	/	MCQ+Open	Math	Multi	Multi	LLM	M	✖	✖
C-Eval [23]	12,342	12,342	/	MCQ	Multi-Disc	Single	Single	LLM	M	✖	✖
ScienceQA [54]	10,332	21,208	/	MCQ	Science	Multi	Single	VLM	M	✖	✖
VisScience [27]	3,000	3,000	/	MCQ+Open	Science	Single	Single	VLM	A+M	✖	✖
MMMU [85]	11,500	11,500	/	MCQ+Open	Multi-Disc	Single	Single	VLM	M	✖	✖
MMMU-Pro [86]	3,460	3,460	/	MCQ	Multi-Disc	Single	Single	VLM	A+M	✖	✖
MathVista [42]	6,141	6,141	/	MCQ+Open	Math	Multi	Single	VLM	M	✖	✖
MathVision [67]	3,040	3,040	/	MCQ+Open	Math	Multi	Single	VLM	M	✖	✖
CharXiv [74]	2,323	11,615	/	Open	Multi-Disc	Single	Single	VLM	M	✖	✖
OlympicArena [25]	7,571	11,163	/	MCQ+Open	Multi-Disc	Multi	Multi	VLM	M	✖	✖
MVBench [31]	3,641	4,000	16.0	MCQ	Open-Domain	Single	Single	VLM	A+M	✖	✖
EgoSchema [43]	5,063	5,063	180.0	MCQ	Egocentric	Single	Single	VLM	A	✖	✖
LongVideoBench [77]	3,763	6,678	473.0	MCQ	Open-Domain	Single	Single	VLM	M	✖	✖
LVBench [70]	103	1,549	4,101	MCQ	Open-Domain	Single	Single	VLM	M	✖	✖
CGBench [4]	1,219	12,129	1624.4	MCQ+Open	Open-Domain	Multi	Multi	VLM	M	✔	✖
MLVU [93]	1,730	3,102	930	MCQ	Narrative	Single	Single	VLM	M	✖	✖
VideoMME [16]	900	2,700	1017.9	MCQ	Open-Domain	Single	Single	VLM	M	✖	✔
Video-MMMU [22]	300	900	506.2	MCQ	Multi-Disc	Single	Single	VLM	M	✖	✖
MMWorld [20]	1,910	6,627	108.0	MCQ	Multi-Disc	Single	Single	VLM	A+M	✖	✖
MMVU [92]	1,529	3,000	51.4	MCQ+Open	Multi-Disc	Multi	Single	VLM	M	✖	✖
VRBench (Ours)	960	8,243	5,796.0	MCQ+Open	Narrative	Multi	Multi	LLM, VLM	M	✔	✔

Table 1. Comparison between VRBench and existing benchmarks. #Size is the number of text/images/videos, #Dur. means the average video duration, A and M indicate automatic and manual annotation type, respectively, and multi-disc denotes multi-disciplinary data source. Clue means clue-grounded annotation. Multilingual requires the number of data source languages to be greater than 2.

to solve multidisciplinary, knowledge-intensive problems through a combination of fast and slow thinking processes.

Reasoning Benchmarks. With the rapid advancement of LLMs and VLMs in reasoning capability, numerous reasoning benchmarks have been developed in text and image-text modalities to facilitate comprehensive evaluation. For the textual domain, several knowledge-driven benchmarks [5, 15, 21, 23, 57, 72] have been derived from exam and textbook data sources, and comprise expert-level questions that require multi-step reasoning. Meanwhile, some multimodal benchmarks also emerge to introduce multi-discipline tasks based on charts [44, 74, 78], plots [45], exam [10, 25, 36, 85, 89], or expert-level questions printed on the static images [27, 35, 54, 86]. It is noted that even in the image and text domains, benchmarks with process-level annotations [8, 87] are relatively scarce. For video understanding, evaluation also gradually shifts from short video clip perception [6, 9, 28, 29, 31, 39, 46, 50, 80] to long-form long video understanding [13, 16, 40, 43, 53, 56, 58, 70, 77, 90] and single-step reasoning. [4, 16, 19, 34, 76, 84]. More recently, some advances [13, 20, 92] propose several new video reasoning tasks in multi-disciplinary scenarios like healthcare, engineering, and science. Different from previous works, VRBench is the first narrative video-based benchmark that is purely used for multi-step reasoning evaluation. Table 1 further presents the detailed statistics of our dataset and distinguishes the difference between VRBench and existing text, image, and video reasoning benchmarks.

3. Benchmark

We present VRBench, a comprehensive multi-step reasoning benchmark consisting of a collection of long, multilingual, and narrative videos with corresponding question-answer pairs. Figure 1 gives the overall pipeline of the benchmark construction, and we illustrate it as follows.

3.1. Video Curation

We collect long narrative videos from YouTube, yielding an initial pool of over 10,000 public videos. Considering plausible reasoning in steps requires videos with rich and coherent plots, we source long-form footage using a cherry-picked tag set and a comprehensive criterion for narration.

Plot-related Tags for Queries. We employ a manually curated set of 7 semantic tags (e.g., Film & Animation, Sport, Travel & Events) to retrieve videos with strong narrative potential. This tag set is developed through iterative validation by domain experts, explicitly excluding non-narrative categories such as news broadcasts and lecture recordings. Our analysis revealed that these excluded categories exhibit limited visual-semantic dynamics (e.g., static

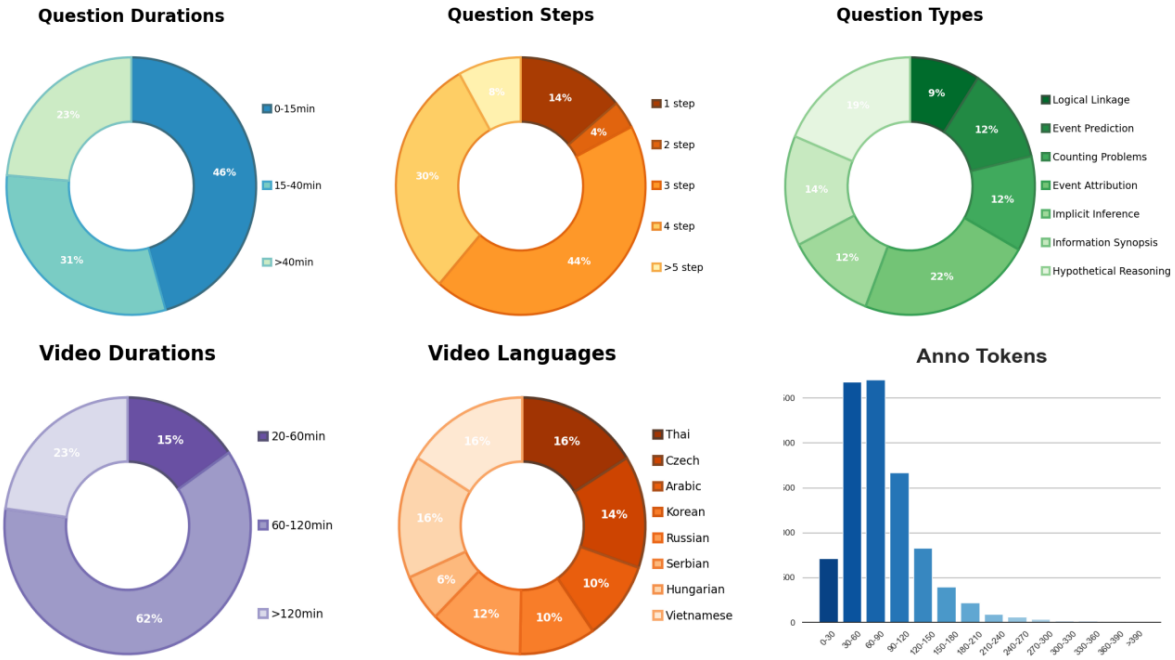

Figure 3. Statistics of VRBench. We provide the detailed distribution of videos and annotations of VRBench, including video languages and durations, steps, types, and temporal duration of questions, as well as the token numbers of answers and the reasoning process.

camera angles in talk formats) and minimal event progression, rendering them ineffective for benchmarking temporal reasoning capabilities.

Criterion for Narration. Our sourcing standards include video duration, its language source, and user ratings. The duration requirement (a minimum of 20 minutes and an average of 1.61 hours) ensures adequate temporal context for constructing reasoning chains. Our language diversity strategy intentionally excludes English and Chinese to counterbalance existing dataset biases, aligning with established findings that linguistic variety improves benchmark generalizability.

Upon the aforementioned two quantitative prerequisites, we organized a panel of 14 multilingual domain experts to evaluate candidate videos using a standardized 10-point scale based on plot coherence and richness. The scoring is guided by several instructions. Videos scoring below 7 are systematically excluded, resulting in a final curated set of 960 high-quality narratives. Detailed annotation protocols are documented in the supplementary materials.

3.2. Stepwise Reasoning Annotations

We label videos with reasoning steps via a two-stage human-in-the-loop framework. It first generates pseudo candidate QA pairs through automated pipelines, then employs expert-guided rewriting to curate high-quality annotations, ensuring rigorous adherence to benchmark specifications while maintaining multimodal reasoning fidelity. To further improve annotation quality, we also implement a comprehensive quality assurance.

Automatic Pipelines. We first employ AutoShot [94] to cut videos into several segments, and then use VideoChat2 [31] to caption them. For auditory content, we adopt whisper-large-v3 [51] to obtain speech transcripts, and DeepL [11] to translate them into English. The video captions and translated subtitles are then put into GPT-4o [47] to generate 6 pseudo QA pairs with a multi-step reasoning process. To ensure the quality of reasoning, we specify 7 multi-step reasoning types for narrative videos as:

- **Event prediction**: Forecast subsequent events in the video timeline.
- **Hypothetical reasoning**: Deduce plausible scenario outcomes from stated premises.
- **Event attribution**: Determine causal origins or underlying motivations of video events.
- **Implicit inference**: Extract unstated temporal, emotional, or relational context from visual cues.
- **Logical linkage**: Establish event-mediated connections between visual/narrative elements.
- **Information synopsis**: Condense critical information across multimodal inputs.
- **Counting problems**: Quantify state changes through arithmetic/combinatorial analysis.

Expert-guided Rewriting. We recruit and train 67 graduate students to generate 8-10 high-quality QA pairs per video, providing raw video footage, translated subtitles, and GPT-generated pre-annotations that offer contextual hints without meeting final benchmark standards. Each QA pair must satisfy four rigorous criteria: (1) Non-synopsis questions require $\geq$2 timestamped reasoning steps (start/end times documented); (2) Temporal distribution constraints ($\leq$4 questions from 0-15min, $\geq$3 from 15-40min, $\geq$ 1 from 40-120min); (3) Coverage of $\geq$5 predefined reasoning taxonomies (from 7 categories) ensuring diversity; (4) Mandatory multimodal grounding where solutions demand both visual analysis (excluding subtitle-only answers) and explicit reasoning (beyond basic perception).

Comprehensive Quality Assurance. To ensure annotations meet our stringent multimodal standards, we implement a rigorous verification protocol: 10 trained reviewers validate each annotation, with non-compliant entries returned to annotators for iterative revisions until fully compliant. Annotators and reviewers are exclusively recruited from top-tier universities to ensure academic rigor. We further enforce quality through dual safeguards: a systematic 5% random sampling audit across both annotation and review stages, coupled with full documentation of protocols, annotation guidelines, and quality assessment criteria in supplementary materials.

3.3. Multi-Phase Evaluation Pipeline

We benchmark VLMs and LLMs through a comprehensive multi-phase evaluation pipeline, which compares predictions against ground-truth annotations both at the process and outcome level. For the outcome-level evaluation stage, we adopt a multiple-choice question (MCQ) format, where false options are generated by DeepSeek-V3 [38] using carefully-designed prompts based on human-annotated answers. For the process-level stage, we propose the open-ended rating to fully evaluate the model's multi-step reasoning capability. Specifically, we adopt an LLM to evaluate the overall quality of the whole reasoning process through four 0-10 scores: logical coherence (40%), similarity to ground truth (40%, excluded for event prediction and hypothetical reasoning tasks), factual accuracy (10%), and clarity (10%). DeepSeek-V3 [38] serves as the judge due to its optimal balance between cost and human alignment (Section 4). For the event prediction and hypothetical reasoning tasks, we argue that there is no ground-truth reasoning steps since there might be multiple possible predictions, hence, we remove the similarity score and compute the final score with an 8:1:1 weight. All metrics are normalized to 0-100 scales, and we report the rankings by computing average scores across all metrics.

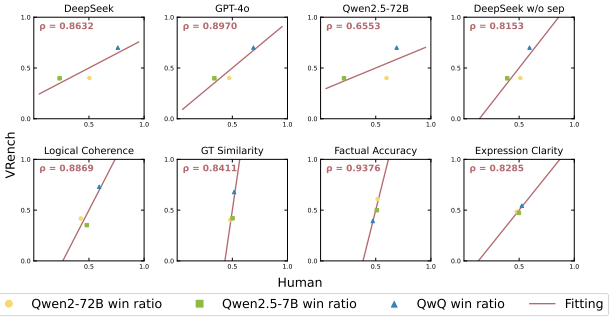

Figure 4. **Human preference alignment results.** For each plot, we show the win ratios of three different tested LLMs evaluated by human experts and VRBench. We then fit them with a straight line and quantify the correlation by calculating the Spearman correlation coefficient.

LLM Evaluation Support. For text-only LLMs, we convert videos to text inputs using Qwen2.5-72B-Instruct [83] to synthesize video captions and subtitles into fine-grained summaries. The process involves: 1) Dividing inputs into 5-minute chunks for duplicate removal and abstraction; 2) Merging chunk-level abstracts into coherent summaries with detailed audio-visual descriptions. These summaries enable LLM evaluations while addressing context window limits for VLMs.

4. Experiments

We test a number of both open-source and proprietary VLMs (e.g. Qwen2-VL [68], InternVL2.5 [7], InternVideo2.5 [73], GPT-4o [47], Gemini-2.0-Pro [61], etc.) and LLMs (DeepSeek-R1 [18], QwQ [65], OpenAI o1 [48], Claude-3.7-Sonnet [59], etc.) with a two-stage evaluation on VRBench. The configuration of each evaluated model is detailed in the supplementary material.

Evaluation Protocols. Our protocol involves two phases. Given a question, 1) models produce multi-step reasoning using Chain-of-Thought [75] prompts. Then 2) models choose final answers from multiple-choice options. For answer extraction, final answers are parsed from response endings to mitigate verbose option analysis, following [31]. While reasoning process evaluations use judge LLMs with structured prompts and example-guided rating extraction. All evaluation prompts and implementation details are included in the supplementary materials.

4.1. Main Results

Table 2 shows the CoT evaluation results of LLMs and VLMs on VRBench. Among LLMs, we found the proprietary model Gemini-2.0-Flash-Thinking [12] achieves the optimal performance while other reasoning models such as OpenAI o1-preview [48] and DeepSeek-R1 [18] also demonstrate strong results. It is noted that o1-preview [48]

Model	Overall	Results by Metric		Results by Taxonomy						
		MCQ-O	OE-P	EA	CP	HR	II	IS	EP	LL
LLMs										
Proprietary Models										
GPT-4o [47]	55.49	63.87	47.11	54.41	32.63	69.03	61.47	71.83	66.37	66.19
Claude-3.7-Sonnet [59]	58.07	62.91	53.23	55.87	35.11	69.83	63.29	72.61	67.91	67.89
o1-preview [48]	60.14	68.47	51.81	56.97	35.87	70.51	64.81	73.13	68.27	69.41
Gemini-2.0-Flash-Thinking [61]	60.97	67.53	54.41	58.61	38.11	72.09	67.57	74.19	70.43	71.13
Open-Source Models										
QwQ-32B-preview [64]	35.90	27.51	44.29	34.31	34.41	46.29	39.67	27.97	42.39	40.61
InternLM3-8B-Instruct [3]	47.81	50.31	45.31	45.47	34.87	60.33	51.77	55.47	56.97	56.81
Qwen2.5-7B-Instruct [83]	48.29	52.61	43.97	46.93	35.17	61.83	51.57	54.87	56.63	57.03
Llama3.3-70B-Instruct [41]	49.84	52.59	47.09	47.63	38.07	63.87	54.03	54.83	59.47	56.97
QwQ-32B [65]	52.52	56.01	49.03	49.53	36.03	62.01	54.03	52.03	58.01	59.03
Qwen2.5-72B-Instruct [83]	53.51	60.49	46.53	51.03	36.01	67.03	58.53	67.83	61.93	63.03
DeepSeek-V3 [38]	56.06	64.79	47.33	54.03	36.57	69.97	61.83	69.53	65.53	65.43
DeepSeek-R1 [18]	57.13	64.19	50.07	56.13	38.91	72.81	64.13	68.01	68.53	67.93
VLMs										
Proprietary Models										
Claude-3.7-Sonnet [59]	68.15	80.09	56.21	63.11	32.97	72.63	71.13	75.37	71.33	70.27
GPT-4o [47]	68.68	81.23	56.13	66.61	36.51	76.67	70.47	78.03	72.03	73.17
Gemini-2.0-Pro [61]	**74.61**	**83.29**	**65.93**	**71.09**	**65.21**	**81.01**	**75.73**	87.13	**77.93**	**75.89**
Open-Source Models										
DeepSeek-VL2 [79]	31.50	33.27	29.73	25.93	22.41	35.73	31.73	30.01	33.29	29.57
H2OVL Mississippi-2B [17]	47.15	52.33	41.97	35.17	32.37	51.71	41.41	60.03	49.97	41.91
Phi-3.5-Vision [1]	48.52	58.03	39.01	31.53	28.03	45.03	37.03	68.03	44.03	36.03
LongVA-7B [91]	50.14	67.81	32.47	27.61	25.07	38.69	33.77	76.91	38.27	32.07
InternVL2.5-8B [7]	50.41	69.31	31.51	26.63	26.31	37.99	31.21	85.97	37.37	30.69
MiMo-VL-7B-RL [60]	50.48	63.39	37.57	47.31	34.59	63.83	53.23	74.61	60.33	56.97
VideoChat-Flash-7B [32]	50.82	72.01	29.63	24.69	21.41	39.37	29.17	79.37	34.87	28.07
InternVideo2.5 [73]	51.94	75.63	28.25	23.81	24.09	34.51	27.07	84.57	33.61	26.81
LongVA-7B-DPO [91]	52.36	67.91	36.81	30.73	27.17	45.09	38.23	79.39	43.67	37.03
Qwen2-VL-7B [68]	54.08	72.01	36.15	30.49	26.61	46.71	33.69	85.27	43.31	35.29
Aria [30]	54.55	72.97	36.13	30.19	29.23	44.23	34.63	86.23	44.99	34.57
Qwen2.5-VL-7B [2]	56.52	69.61	43.43	37.07	33.17	54.27	42.27	83.91	50.17	42.87
Keye-VL-8B-Preview [63]	60.44	64.41	56.47	62.53	40.37	73.61	67.99	70.13	71.83	71.03
Qwen2.5-VL-72B [2]	61.71	66.85	56.57	51.87	46.13	67.13	54.13	**90.04**	63.67	60.77
Kimi-VL-A3B-Thinking-2506 [62]	61.82	61.67	61.97	64.53	47.91	71.37	69.57	74.47	71.23	71.47
InternVL2.5-78B [7]	62.31	76.61	48.01	43.77	38.67	58.21	46.13	87.53	54.43	47.57

Table 2. Evaluation results on VRBench across two evaluation metrics (outcome-level MCQ, process-level open-ended evaluation) and seven QA taxonomies (event attribution, counting problems, hypothetical reasoning, implicit inferences, information synopsis, event prediction, logical linkage). **Bold values** indicate the best results among all models, and underlined values are best results of LLMs.

achieves the best outcome-level MCQ accuracy of 68.47%. For the VLMs, Gemini-2.0-Pro [61] achieves 74.61% overall accuracy, surpassing all components by at least 5.93%. GPT-4o [47] and Claude-3.7-Sonnet [59] also show competitive results, which demonstrate 68.68% and 68.15% overall performance, respectively. For open-source models, InternVL2.5-78B [7], emerges as the best non-proprietary VLMs with a 62.31% overall accuracy.

We then delve into the specific results of each metric and taxonomy. For results in each evaluation stage, we find that most of LLMs and VLMs are capable of achieving high MCQ accuracy, yet compared to LLMs, VLMs struggle to demonstrate their reasoning steps and exhibit lower reasoning ratings. Through results across taxonomy, we found

that most of the large models are not proficient in counting problems since they require fine-grained visual perception, in contrast with other tasks. Since the perception of LLMs fully depends on the video summary and some VLMs only support a small number of frame inputs, they are unable to correctly perceive the original frames containing the target elements, resulting in the MCQ accuracy close to random guessing and low reasoning process ratings.

4.2. Ablations and Analysis

We then further discuss the observations from the evaluation results and highlight some insights on obtaining higher multi-step reasoning performance.

Does the evaluation of VRBench align with Human Preference? In contrast to MCQs that compute accuracy via a deterministic method, the correctness of process-level open-ended evaluation highly depends on the judgment of LLM's performance, which may include hallucinations that could influence the evaluation reliability. To this end, we select a subset of 30 videos with 300 questions to perform process-level human evaluation, aiming to quantify the correlation between human and LLM assessment results.

Specifically, the human annotators are asked to give process-level open-ended ratings following the same requirement for the judging LLM, i.e., four separated ratings are needed for each question to evaluate the logical coherence, similarity to ground-truth, factual accuracy, and expression clarity. We then compute the win ratio of pairwise comparison for each tested model following [24], where a model scores 1 if its rating is higher than its current opponent, 0 if lower, and 0.5 for a tie. Finally, we calculate the average win ratio of each model and the Spearman correlation coefficients (ρ) of the win ratios evaluated by LLMs and human annotators.

We first investigate the human preference alignment of several LLMs. We select GPT-4o [47], DeepSeek-V3 [38], Qwen2.5-7B [83], and Qwen2.5-72B [83] as evaluation models and give them the same rating prompt for a fair comparison. As shown in Figure 4, Qwen2.5 shows low human preference alignment compared with DeepSeek-V3 and GPT-4o, which both have correlation coefficients greater than 0.8. Though the human alignment of GPT-4o is slightly better than DeepSeek-V3, we observe the evaluation cost of GPT-4o is approximately 10 times that of DeepSeek, which imposes a significant burden on the benchmark users. Hence, we adopt DeepSeek-V3 as the judging model for its optimal trade-off between performance and cost.

We also probe the evaluation correctness of each open-ended sub-metric. Following the same procedure, we compute the win ratio and correlation coefficient of DeepSeek-V3 for each sub-metric, and the results are shown in Figure 4 (b-d). Results show the coefficients in all metrics are higher than 0.8, indicating a strong positive monotonic relationship between LLM and human evaluation results. Another alternative is to instruct the LLM to compute the final score based on the given weights of each sub-metric, which is denoted as *DeepSeek w/o sep* in Figure 4. Results show that adding this additional computing process diminishes the robustness of evaluation LLM, thus in practice, we ask the model to output all sub-metric rating separately and automatically compute the final score.

The role of System-2 thinking. o1-like LLMs, also denoted as system-2 models, have emerged with superior reasoning abilities in multi-disciplinary scenarios such as science and math. We further explore these models' performance on the narrative video benchmark. We both assess the role of open-source and proprietary o1-models, as well as the gap compared with their previous system-1 version. For proprietaries, we compare OpenAI o1-preview [48] with GPT-4o [47], and results show that the system-2 version o1 outperforms GPT-4o with 4.30% overall accuracy. For open-source model DeepSeek-V3 [38] and R1 [11], we observe that DeepSeek-R1 achieves higher scores both on multiple-choice questions and open-ended evaluations. It is noted that compared with Qwen2.5-72B-Instruct [82], QwQ-32B [65] gets higher reasoning ratings yet lower outcome-level MCQ results. This shows that though some system-2 models are more proficient in multi-step thinking, their ability to converge lengthy thinking processes into correct answers still needs improvement. Recently, some o1-like training strategies have also been utilized by some VLMs, such as LongVA-7B-DPO [91], a model that utilizes direct preference optimization [52] for long video understanding. It achieves an overall performance of 52.36% and surpasses its vanilla version, suggesting that training VLMs with different optimization methods is feasible and that developing system-2 VLM models is crucial for tackling multi-step reasoning questions in narrative videos.

Impact of Model Size. It is commonly believed that models with large scales are often accompanied by better perception and reasoning capabilities. To verify such claims, we conduct experiments on the models with different scales, such as Qwen2.5-7B [83] and Qwen2.5-72B, Qwen2.5-VL [2] with 7B and 72B parameters, and InternVL2.5-8B [7] and InternVL2.5-78B. As shown in Table 2, models with more parameters obtain higher overall accuracy (48.29% vs 53.51% for Qwen2.5, 56.52% vs 61.71% for Qwen2.5-VL, and 50.41% vs 62.31% for InternVL2.5), and the performance gap on process-level metrics is more apparent. This indicates that large-scale models are more likely to achieve advanced reasoning capabilities. However, we also noticed that models specifically trained on the reasoning corpus like QwQ-32B [65], can achieve comparable performance compared with its 72B Qwen2.5 counterpart (52.52% vs. 53.51% in overall accuracy) with a smaller

parameter amount. This suggests that training small-sized models with various preference optimization strategies and reasoning-centric corpora might improve their overall reasoning capability.

Long Context Helps. We observe a substantial performance gap between models with and without long context support. As shown in Table 2, models with long frame input tend to achieve higher overall accuracy. For example, Gemini-2.0-Pro [61] is capable of inference with 0.5 fps and supports large number frame inputs when answering questions of videos over 1 hour, and it achieves the optimal performance compared with all other fixed-length models. For the open-source models, Qwen2.5-VL-7B [2] and Kimi-VL-A3B-Thinking-2506 [62] with 256 frame input achieve 61.71% and 61.82% overall accuracy, which is 11.30% and 11.41% higher than the 8B InternVL-2.5 [7] model with 64 frames input. Since the average length of VRBench is 1.61 hours, the necessity of long context support is becoming more pronounced, so that models can perceive more plot elements that help the analysis of long video content.

LLMs vs. VLMs. Since tested LLMs are only provided with the automatically-generated video summary, yet VLMs are capable of directly perceiving the raw visual content, it is unfair to directly compare the numerical results of LLMs and VLMs. However, there are still some noteworthy phenomena that can be discussed. First, VLMs with coarse-grained visual input (e.g., 4 frames for H2OVL Mississippi-2B [17] and DeepSeek-VL2 [79]) perform significantly worse than several system-2 LLMs like DeepSeek-R1 [18] and Claude-3.7-Sonnet [59]. This indicates that randomly selecting a small number of frames is not enough to tackle narrative-based reasoning tasks that require long-range perception. Furthermore, VLMs with fine-grained visual perception, such as Gemini-2.0-Pro with a 0.5 fps perform better than top-tier LLMs on both process-level and outcome-level metrics, showing the inevitability of detailed visual content input for tackling questions in VRBench.

4.3. Test-Time Scaling Exploration

We here probe models' performance when scaling test-time compute cost. Since we tend to prompt the models with a Chain-of-Thought template when tackling complex reasoning tasks, it is observed that system-2 models are capable of dramatically expanding the inference budget to improve their performance. To this end, we set a series of token limitations on these models and require the models to think with different instructions. Specifically, we select QwQ-32B [65] as the experimental LLM and Qwen2-VL-7B [82] as the testing VLM on a subset with 300 videos and 2,403 questions. For each model, several parallel experiments are conducted with increasing maximum token limits from 256 to 2048. We instruct the models to output the reasoning process as much as possible for large token limitations, and let them think concisely when the token limit is low. Since we aim to investigate the quality of models' initial reasoning process and their capability to generate the right answers through thinking, we report the overall scores that both evaluate the outcome accuracy and process quality. Results are shown in Figure 5, where QwQ shows a remarkable rating boost from 48.91% to 61.34% when setting a large token limit. On the contrary, the small-sized system-1 model Qwen2-VL-7B performs even worse when using a large token number limitation and instructing the model to output more thinking process. The model tends to output more ambiguous outputs that lead to the wrong answer. This observation suggests that models with large parameter scales and system-2 capabilities benefit from the test-time scaling strategy. It also provides the insight that developing more test-time scaling approaches that guide the model to generate long reasoning traces could be a feasible and promising way to unlock more potential capabilities of system-2 models, thereby benefiting complex tasks that require multi-step reasoning.

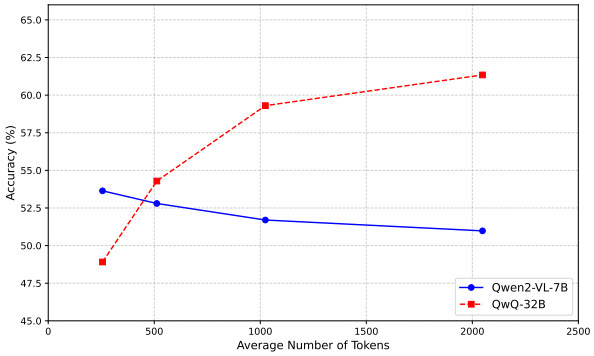

Figure 5. Test-Time Scaling Results. We report the average accuracy of outcome-level MCQ and process-level open-ended ratings.

5. Conclusion

This paper introduces VRBench, a comprehensive long-narrative video benchmark for evaluating multi-step reasoning. Through manually filtered narrative videos, high-quality stepwise annotations, and a multi-phase evaluation pipeline, VRBench distinguishes itself from other existing reasoning benchmarks and shows the robust capability to evaluate LLMs and VLMs both from the process and outcome perspectives. By assessing and analyzing 31 frontier large models, we thoroughly demonstrate the detailed performance of current reasoning models across various reasoning questions and provide valuable insights towards constructing more advanced multi-step reasoning models.

6. Acknowledgement.

This work is supported by the National Key R&D Program of China (No. 2022ZD0160101) and Jiangsu Frontier Technology Research and Development Program (No. BF2024076),

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 6

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6, 7, 8

[3] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 2, 6

[4] Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024. 3

[5] Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. *arXiv preprint arXiv:2305.12524*, 2023. 3

[6] Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*, pages 179–195. Springer, 2024. 3

[7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 2, 5, 6, 7, 8

[8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 2, 3

[9] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*, 2024. 3

[10] Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*, 2024. 3

[11] DeepL. Deepl translate: The world's most accurate translator. https://www.deepl.com/en/translator, 2025. 4, 7

[12] Google Deepmind. Gemini 2.0 flash thinking. https://deepmind.google/technologies/gemini/flash-thinking/, 2025. Accessed: 2025-01-21. 2, 5

[13] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2025. 3

[14] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*, 2024. 2

[15] Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*, 2024. 3

[16] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 2, 3

[17] Shaikat Galib, Shanshan Wang, Guanshuo Xu, Pascal Pfeiffer, Ryan Chesler, Mark Landry, and Sri Satish Ambati. H2ovl-mississippi vision language models technical report. *arXiv preprint arXiv:2410.13611*, 2024. 6, 8

[18] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 5, 6, 8

[19] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. *arXiv preprint arXiv:2411.14794*, 2024. 3

[20] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*, 2024. 3

[21] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 3

[22] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025. 3

[23] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010, 2023. 3

[24] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 7

[25] Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, et al. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *Advances in Neural Information Processing Systems*, 37: 19209–19253, 2025. 3

[26] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024. 3

[27] Zhihuan Jiang, Zhen Yang, Jinhao Chen, Zhengxiao Du, Weihan Wang, Bin Xu, Yuxiao Dong, and Jie Tang. Visscience: An extensive benchmark for evaluating k12 educational multi-modal scientific reasoning. *arXiv preprint arXiv:2409.13730*, 2024. 3

[28] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 3

[29] Bozheng Li, Yongliang Wu, Yi Lu, Jiashuo Yu, Licheng Tang, Jiawang Cao, Wenqing Zhu, Yuyang Sun, Jay Wu, and Wenbo Zhu. Veu-bench: Towards comprehensive understanding of video editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13671–13680, 2025. 3

[30] 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. *arXiv preprint arXiv:2410.05993*, 2024. 2, 6

[31] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 3, 4, 5

[32] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 2, 6

[33] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025. 2

[34] Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*, 2024. 3

[35] Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, et al. Mmsci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In *AI for Accelerated Materials Design-Vienna 2024*, 2024. 3

[36] Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. Scemqa: A scientific college entrance level multimodal question answering benchmark. *arXiv preprint arXiv:2402.05138*, 2024. 3

[37] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024. 2

[38] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 5, 6, 7

[39] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 3

[40] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Et bench: Towards open-ended event-level video-language understanding. *arXiv preprint arXiv:2409.18111*, 2024. 3

[41] Llama-3.3. Llama-3.3-70b-instruct. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct, 2025. 6

[42] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 3

[43] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 3

[44] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 3

[45] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2020. 3

[46] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023. 3

[47] OpenAI. Hello gpt4-o. https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2024-05-13. 4, 5, 6, 7

[48] OpenAI. Introducing openai o1. https://openai.com/o1/, 2024. 2, 5, 6, 7

[49] OpenAI. Openai o3-mini. https://openai.com/index/openai-o3-mini/, 2025. 2

[50] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023. 3

[51] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 4

[52] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 2, 7

[53] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024. 2, 3

[54] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022. 3

[55] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2

[56] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 3

[57] Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19053–19061, 2024. 3

[58] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 3

[59] Anthropic. Claude Team. Claude 3.7 sonnet. https://www.anthropic.com/claude/sonnet, 2025. 2, 5, 6, 8

[60] Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *https://arxiv.org/abs/2506.03569*, 2025. 6

[61] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 5, 6, 8

[62] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 6, 8

[63] Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl technical report. *arXiv preprint arXiv:2507.01949*, 2025. 6

[64] Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown. https://qwenlm.github.io/blog/qwq-32b-preview/, 2024. Accessed: 2024-11-28. 6

[65] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning. https://qwenlm.github.io/blog/qwq-32b/, 2025. Accessed: 2025-3-6. 2, 5, 6, 7, 8

[66] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025. 2

[67] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2025. 3

[68] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5, 6

[69] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024. 2

[70] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024. 2, 3

[71] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 2

[72] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 3

[73] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 2, 5, 6

[74] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2025. 3

[75] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 5

[76] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024. 3

[77] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2025. 2, 3

[78] Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. Chartinsights: Evaluating multimodal large language models for low-level chart question answering. *arXiv preprint arXiv:2405.07001*, 2024. 3

[79] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 6, 8

[80] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 3

[81] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 2

[82] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report, 2024. *URL https://arxiv. org/abs/2407.10671*, 2024. 7, 8

[83] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 5, 6, 7

[84] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. *Advances in Neural Information Processing Systems*, 37:57240–57261, 2025. 3

[85] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 3

[86] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 3

[87] Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. *arXiv preprint arXiv:2312.17080*, 2023. 3

[88] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 2

[89] Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, et al. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*, 2024. 3

[90] 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. *arXiv preprint arXiv:2312.04817*, 2023. 3

[91] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 2, 6, 7

[92] Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, et al. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*, 2025. 2, 3

[93] 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. *arXiv preprint arXiv:2406.04264*, 2024. 3

[94] Wentao Zhu, Yufang Huang, Xiufeng Xie, Wenxian Liu, Jincan Deng, Debing Zhang, Zhangyang Wang, and Ji Liu. Autoshot: A short video dataset and state-of-the-art shot boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2238–2247, 2023. 4