

MINERVA: Evaluating Complex Video Reasoning

Arsha Nagrani¹, Sachit Menon^{2*}, Ahmet Iscen¹, Shyamal Buch¹, Ramin Mehran¹, Nilpa Jha¹,
 Anja Hauth¹, Yukun Zhu¹, Carl Vondrick², Mikhail Sirotenko¹, Cordelia Schmid¹, Tobias Weyand¹

¹ Google DeepMind, ² Columbia University

Abstract

Multimodal LLMs are turning their focus to video benchmarks, however most video benchmarks only provide outcome supervision, with no intermediate or interpretable reasoning steps. This makes it challenging to assess if models are truly able to combine perceptual and temporal information to reason about videos, or simply get the correct answer by chance or by exploiting linguistic biases. To remedy this, we provide a new video reasoning dataset called MINERVA for modern multimodal models. Each question in the dataset comes with 5 answer choices, as well as detailed, hand-crafted reasoning traces. Our dataset is multimodal, diverse in terms of video domain and length, and consists of complex multi-step questions. Extensive benchmarking shows that our dataset provides a challenge for frontier open-source and proprietary models. We perform fine-grained error analysis to identify common failure modes across various models, and create a taxonomy of reasoning errors. We use this to explore both human and LLM-as-a-judge methods for scoring video reasoning traces, and find that failure modes are primarily related to temporal localization, followed by visual perception errors, as opposed to logical or completeness errors. The dataset, along with questions, answer candidates and reasoning traces is publicly available under <https://github.com/google-deepmind/neptune?tab=readme-ov-file#minerva>.

1. Introduction

Video understanding has long been a holy grail for computer vision. Video benchmarks have evolved from testing action recognition on short clips [17, 26, 50] to more recent benchmarks testing advanced capabilities over longer time horizons (LVBench [45], VideoMME [15], Neptune [34], and others). The development of the latter has been supercharged by the rise of multimodal large language models [1, 6, 39], which perform complex video tasks on videos up to an hour long [45]), and are steadily approaching human performance

on these recent benchmarks [15, 34].

While these datasets aim to measure complex video understanding capabilities, they do not provide further insight into *how* or *why* models succeed or fail beyond the correctness of their final answers. One way to inspect model failures is by exposure to step-by-step rationales generated by the model. Attempts to improve these rationales have led to a paradigm shift towards models (particularly in the text domain), that spend more time “thinking” [3] i.e. producing intermediate thoughts towards a final answer, often via additional compute at inference time [24, 42]. ‘Thinking’ or reasoning in the video domain, however, looks very different to reasoning in the text or coding domain. In the text domain, models can directly quote sections or edited versions of the input as thoughts – these intermediates are easier to assess by humans, or by another language model. This is not the case for the video domain. The high dimensionality and multi-modality of video means that solving a video task often requires temporal localization, perceptually recognizing key objects, events and actions (from multiple modalities such as speech or frames) and applying logical reasoning to tie these intermediates all together. Complex queries in video question answering (videoQA) hence naturally require multi-step processes to solve, where each step may require a different skill and/or access to a different modality input. We henceforth refer to this multi-step process as a ‘reasoning trace’ for videoQA. This includes the set of the steps required to solve the question, including perception and localization.

Yet, despite the fundamentally multi-step nature of this capability, existing video benchmarks only evaluate final answers: they only check the *outcome* and not the *reasoning*. It is not clear, however, if a model arrives at a correct answer due to a successful execution of key steps, pure chance, linguistic bias, or the process of elimination of answer choices. Conversely, if it fails to produce the correct answer, was it completely unable to solve the problem, or did it get close but ultimately make a mistake that changed the final answer? Using final answers alone makes it difficult to design datasets that will remain challenging but still provide signal; all the more pertinent as new models improve at breakneck speed.

*Work done during an internship at Google DeepMind

Question: How many times does the ball change hands on the play that results in Pampanga having 52 points on the scoreboard?

- (a) 5 (b) **3** (c) 4 (d) 2 (e) 1

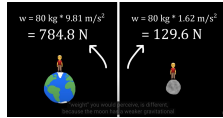


I watched the play that increased Pampanga's score in the game to 52, which was played from 07:07 - 07:19, and counted the number of times the **ball changed hands** from the start of the play to the score. I counted the first possession change at 07:09, when the player with the number 8 jersey **passes the ball** to the player with the number 19 jersey. I watched the player with the number 19 jersey **hand off the ball** to the player in the number 15 jersey at 07:12, marking the second possession change. I then watched the third possession change take place from 07:13 - 07:14 when number 15 **passed the ball** back to number 19, who then **scored** at 07:16 to increase Pampanga's score to 52. Thus, three total possession changes occurred during the play before the basket.

Skills: Counting, Reading, Temporal Reasoning, Event Localization

Question: Using the magnitude of the gravitational force of the moon given in the video and a value of $d=5$, what would the answer be to the equation given at 03:16?

- (a) 8.25 (b) 8.0 (c) **8.1** (d) 8.5 (e) 7.9



I heard the narrator say that weight relates to the force of gravity at 02:20. Then, I observed an example of the same mass on the earth versus the moon at 02:26. I read that the weight was mass times the force of gravity. I read that the force of gravity in the example for the moon was 1.62 m/s^2 at 02:26. I moved to 03:16 to find the equation. I heard the equation was " $W = Fd$ " and heard that the variables were "Work" "Force" and "Distance". Therefore, I used the gravitational force of the moon for F and the value of 5 given from the question for D . I calculated $(1.62 \text{ m/s}^2)(5 \text{ m}) = 8.1 \text{ J}$.

Skills: Numerical Reasoning, Reading, Listening, Temporal Reasoning

Question: Approximately how long would it take the woman to knead the dough if she was working twice as fast?

- (a) 30 seconds (b) **45 seconds** (c) 2 minutes(d) 3 minutes (d) 20 seconds



I watched the video and kept track of how long it takes the woman to **knead the dough** in the video. She **begins mixing** the dough at 01:11 with a spoon to incorporate the water. Then, at 01:21, she begins **mixing and kneading** by hand. She continues **kneading** the dough until 02:40. Therefore, it takes her 1 minute and 29 seconds to knead the dough in the video, or 89 seconds. If she was working twice as fast, it would have taken her approximately 45 seconds.

Skills: Counting, Numerical Reasoning, Counterfactual

Question: Why does the gravedigger pull a gun on the man in the khaki suit?

- (a) To save his parents (b) **To save the strangers** (c) To save his children (d) To save himself (e) To save his enemies



I watched the entire video and never saw the gravedigger **interact** with the two young men or acknowledge them, and therefore determined he did not know them. I watched the video until the 02:58 mark and saw that the man is digging a grave. At 03:02 there's a clear shot of him noticing the two men being held at gunpoint. I continued to 03:38 where it becomes clear the man in the khaki suit plans to kill the young men by covering them in driveway sealant. At 03:51 the gravedigger pulls a gun on the man in the khaki suit and commands him to stop what he's doing to the two young men. Therefore, it became clear that the gravedigger pulls a gun on the man in the khaki suit to save the lives of strangers.

Skills: Goal Reasoning, 'Event Occurrence, Temporal Reasoning

Figure 1. Examples from MINERVA: We introduce MINERVA, a complex video question-answering dataset. Unlike existing video datasets, the answer to each question is accompanied by a detailed *reasoning trace*, which outlines the steps required to come to the answer. Videos cover multiple domains such as (clockwise) – sports, cooking, short films and science lectures. Reasoning traces are detailed, including *timestamps* (highlighted in green) and *key actions* (highlighted in pink). We show a single frame from each video.

In light of this, we release MINERVA¹ – a benchmark for video reasoning which consists of question, answers and reasoning traces. MINERVA is challenging and complex – every question requires multiple steps to solve, and even the very strong Gemini 2.5 Pro Thinking achieves only 66.2% accuracy, while humans are able to achieve 92.5%. The dataset is hence well positioned for hill-climbing and future assessment of multimodal models. Every question in MINERVA requires complex reasoning using two or more skills (for example numerical reasoning, temporal reasoning, spatial navigation). Videos also span multiple domains (short films, sports, instructional videos etc), with various video lengths (from 2 minutes to *over 1.5 hours*), making the dataset diverse. For each question we also provide the hand-crafted, detailed reasoning trace, with the steps that are required to come to the correct answer. Unlike datasets that provide auxiliary information in a single format such as timestamps (LITA [21], CG-Bench [10] + others), the reasoning trace is an unconstrained block of text - allowing flexibility. Each reasoning trace does include timestamps where necessary to refer to relevant sections of the video, but also describes key actions, objects, as well as outlines logical reasoning steps (see Fig. 1 for examples). For each question, we also release a number of less challenging sub-questions, with each sub-question typically representing a single step in the reasoning chain.

¹Multimodal InterpretABLE Reasoning Video Annotations. Minerva is also the Roman goddess of thinking and wisdom.

We benchmark both open-source and proprietary models on our dataset, and manually compare some of their reasoning traces to the ground truth reasoning annotations to identify failure modes. We use these failure modes to build a taxonomy of errors (in other words, a rubric) for video reasoning. This rubric is specific to the video domain, and highlights the following broad categories of errors - (1) Perceptual Correctness, (2) Temporal Localization, (3) Logical Reasoning and (4) Completeness. We use this rubric to assess how both humans and LLMs grade reasoning outputs along these criteria when given the ground truth reasoning, illuminating where automatic graders correlate to humans and where they may not, highlighting directions for potential improvement for reference-based metrics. In addition, we also find in some cases that LLM performance on video reasoning improves simply by being prompted with the rubric, in rough analogy to 'Let's think step by step' [27] in text reasoning. To summarize, we make the following contributions: (i) We introduce MINERVA, a challenging video reasoning benchmark for LMMs consisting of 1,515 hand-crafted questions and 4,065 less challenging sub-questions. For each question, we provide 5 answer choices, as well as detailed, manually-annotated reasoning traces; (ii) We evaluate multiple frontier models on MINERVA, both open-source and proprietary, and perform ablations to show low textual bias and the importance of visual information; (iii) we perform fine-grained error analysis to identify common failure modes across various models, and create a taxonomy of reasoning

errors. We find that failure modes are primarily related to temporal localization and visual perception. Preliminary analysis of LLM-based metrics show areas for future work.

2. Related Work

2.1. Recent VideoQA Datasets

There exist many valuable video datasets. Temporal-Bench [7] and PerceptionTest [36] include a variety of tasks, such as video QA, captioning, and grounding, but contain shorter videos (most videos under 2.5 minutes). Similarly, CinePile [38] and VideoVista [29] focus on short-form content (average of 160s). VideoVista [29] is notable for its broad coverage of 19 understanding and 8 reasoning tasks. These benchmarks, along with InfiniBench [5] (maximum 52-minute videos), and Neptune [34], rely on semi-automatic pipelines using LLMs for annotation. In contrast, our dataset is entirely manually annotated. TVBench [11] specifically addresses the challenge of temporally difficult answer candidates. E.T. bench [32] focuses on multi-event and time-sensitive understanding, using human-annotated videos with a maximum length of 795 seconds. Video-MME [15] bridges the gap between short and long-form content, and datasets such as LVBench [45] (average length 68 minutes) and HourVideo [8] (egocentric videos, 20 to 120 minutes) extend to even longer videos. In contrast to these, our work provides not only the final outputs but also human-annotated reasoning traces, allowing us to assess the model’s reasoning process in addition to its accuracy.

2.2. Reasoning Traces for Video

As discussed in the previous section, few datasets provide auxiliary dataset beyond final answers. ScienceQA [40] consists of science questions about images, and provides the accompanying lectures, which when used for training results in increased performance with fewer data points. Some recent works have aimed to develop automatic labeling pipelines to instead generate noisy auxiliary information at scale. Visual CoT [41] provides bounding box pseudolabels generated by off-the-shelf models for image VQA, while VideoEspresso [18] does something similar for video, constructing a pipeline connecting different frozen models together to label bounding boxes for sparse key frames. VideoCoT [47] is perhaps the closest to our work, aiming to generate text-based chain of thought for videos; however, we note key differences here. The primary goal of VideoCoT is to present a semiautomatic pipeline using LLMs and VLMs to scale up for training purposes, while we present a fully manually annotated, high quality dataset for evaluation purposes. The videos in VideoCoT are all taken from established Kinetics700 [26], making them shorter and limited in domain. Finally, their automated rationales tend to contain substantial information about the video that does not relate to the

particular query, rather than providing specific reasoning for the given question.

2.3. Reasoning Analysis

With the growth of reasoning models, substantial work has been conducted into analyzing reasoning in text-only settings and remains an open problem [33]. Previous works demonstrate that correctness of final answers alone often does not indicate correct reasoning, commonly due to hallucination of reasoning steps or answers [16, 19, 37, 44]. These works fall into (i) reference-based, which use additional human annotation as reference, and (ii) reference-free (evaluation without such additional information). Unsurprisingly, multiple works find that reference-based evaluation proves consistently more reliable [13], and that language models are poor evaluators of reasoning in isolation [22, 25, 44]. In this work, we primarily focus on contributing a high-quality dataset with reference annotations for video reasoning, which may spur the development of and provide a comparison for further research into both reference-based and, by comparison, reference-free metrics for video reasoning.

Earlier work in reference-based evaluation (before LLMs) explored rule based methods such as n -grams for summarization with ROUGE scores [30] or for machine translation with BLEU [35]. ROSCOE [16] observed the need for evaluation of reasoning for early LLMs, inspired by chain-of-thought prompting [48], using various customized metrics such as word vector similarity. As LLMs advanced, they began to be used for evaluation themselves, such as via LLM-as-a-Judge Zheng et al. [52]. MLLM as a judge [9] and LLaVA-Critic [49] show these capabilities in multimodal settings, but this has thus far been largely explored for the image-text domain. Unlike these works, we explore reference-based LLM-as-a-judge strategies for analysis of video reasoning traces. In both text and multimodal contexts, well-designed, itemized evaluation criteria for rating different model responses give substantially more consistent, well-aligned LLM scores [31, 52] – we follow along these lines of work by providing a rubric for scoring video reasoning, and provide a preliminary analysis of its effectiveness.

3. MINERVA

We introduce MINERVA, a videoQA dataset consisting of 1,515 challenging questions for videos of varying length. Each question has 5 answer options and a detailed reasoning trace accompanying the final answer (see Fig. 1 for examples). The key features of our dataset are: (i) **multistep**, **complex** questions. Each question requires multiple steps and multiple skills to solve, (ii) **multimodality**, with both frames and ASR needed for some questions, (iii) **high quality** – the entire dataset is hand crafted by experienced annotators, and most importantly (iv) **intermediate reasoning** – detailed, hand-crafted reasoning traces. These reasoning

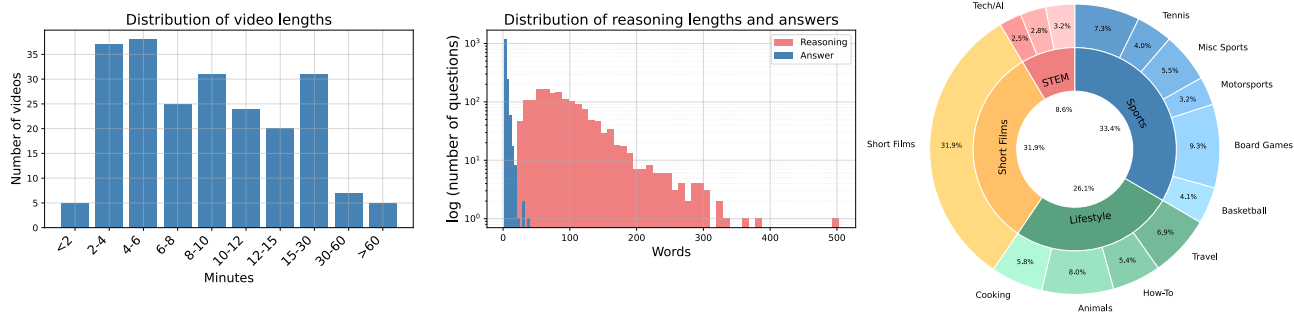


Figure 2. **Dataset statistics.** Video lengths (left), lengths of answers and reasoning (middle), and domains (right). Videos cover a wide range of lengths, with some longer than 100 minutes. Every question comes with a reasoning trace which is long and detailed, mean number of words is 92 (middle). Domains are hand-selected to include videos that lend themselves well to complex reasoning questions.

traces allow us to perform an analysis of where models are making errors, beyond just their final answer outputs. We do this via **reference-based** analysis of model outputs, which can be performed solely in the text domain (and is therefore much cheaper than reference-free assessments). We showcase the value of this annotation via both human and LLM-based judgments of model outputs (in Sec. 5), based on a taxonomy of errors and detailed rubric.

Statistics: Statistics are provided in Fig. 2. The dataset covers a wide range of video lengths, with the longest video being 100 minutes while the shortest video is less than 2 minutes (mean of 12 minutes). There are multiple questions per video, with the distribution following a min/mean/max of 1/6.8/10 questions per video (223 videos in total). Reasoning traces are long and detailed, with the mean number of words in a reasoning trace being 92 (Fig. 2). 99.6% of reasoning traces contain timestamps, with the mean number of timestamps per reasoning trace being 4. Each question also consists of a number of sub-questions. Each sub-question represents a single step in the intermediate reasoning chain, and comes with a ground truth open-ended answer, a list of skills required to solve the sub-question, and a list of dependencies. The dependencies are pointers to previous sub-questions that need to be solved before the sub-question in question, giving us a reasoning *tree*. The average number of sub-questions per question in the dataset is 2.7.

3.1. Dataset Construction

As multimodal models continue to improve, it becomes increasingly non-trivial to come up with challenging questions. Questions should be complex, requiring multiple steps to solve, and not solvable with only the speech or external knowledge (full criteria provided to raters can be found in Appendix 8.1.1). Our dataset construction pipeline consists of the following steps:

1. Video Selection: We begin by selecting video domains

from YouTube that lend themselves well to questions fulfilling the desiderata above.

2. Manual Annotation: Raters propose questions, answers and reasoning traces.

3. Quality Review: Questions are reviewed by other raters.

4. Adversarial Filtering: We attempt to mitigate textual bias using consensus from multiple frontier text-only models.

3.1.1. Video Selection

Video selection is non-trivial, as many videos online contain simple storylines, few shots, or are dominated by talking heads. To avoid this, we identify a (non-exhaustive) set of domains that lend themselves well to reasoning. These are described below (and shown in Fig. 2).

Short Films: We search for videos tagged as “short film”, which are complex, multi-event videos that typically have a story line. Short films tend to contain relationship and event arcs over varying time horizons. Note that this set does not include mainstream movies, so there is less chance of training data contamination or external knowledge bias.

Sports and Board Games: This set includes sports such as tennis, basketball, motorsports, hockey, cricket, and board games such as chess, scrabble and risk. These videos lend themselves well to questions that require reasoning against the rule set of the sport or game, as well as recognition of fine-grained actions and identification of player or piece positions on screen.

Educational: Inspired by reasoning works in the maths and textual domain, we include some STEM lectures. However this domain is only 8% of the dataset (Fig. 2, right), as it is challenging to find lectures where the speech does not dominate (and visual reasoning is important).

Lifestyle: Our final domain category consists of VLOG-style lifestyle videos, including cooking, general how-to videos, as well as travel vlogs and pet videos. These are typically multi-event videos where events may follow a logical

Table 1. **Frame and ASR ablations on MINERVA.**

Method	# of Frames	ASR	MCQ-Acc. %
GPT-4o [1]	64	✓	39.74
GPT-4o [1]	250 (max)	✗	44.82
GPT-4o [1]	250 (max)	✓	45.54
Gemini 2.0 Flash [39]	64	✓	47.85
Gemini 2.0 Flash [39]	256	✗	49.70
Gemini 2.0 Flash [39]	256	✓	53.47
Qwen-2.5VL [6]	64	✓	33.00
Qwen-2.5VL [6]	768 (max)	✗	34.39
Qwen-2.5VL [6]	768 (max)	✓	35.05

causal flow (*e.g.* steps in a recipe). Travel and vlogging style videos also add egocentric viewpoints, and lend themselves well to questions requiring spatial reasoning.

3.1.2. Annotation and Quality Review

Once videos are identified, raters then propose complex questions, answers, decoys, reasoning traces and label question types. The raters are instructed to attempt to propose questions such that each question requires at least two of the following skills: Temporal Reasoning, Counting, Cause and Effect, Goal Reasoning, Situational Awareness, Event Occurrence, State Changes, Reading (OCR), Listening (identifying a detail in the audio track), Spatial Perception, Numerical Reasoning (all math operations other than counting), Object Recognition, Counterfactual Reasoning (‘what if’, but with an objective outcome). Each annotation is verified by at least one other rater (details in Appendix 8.1.4). The annotations are also periodically reviewed by the papers’ authors, who provide detailed feedback to the raters.

3.1.3. Filtering for Quality and Biases

Finally, we address quality and potential for biases. We begin by filtering examples where the reasoning traces are of low quality or too short (<20 words). While raters are explicitly instructed to avoid proposing questions that can be solved from text alone, we find (very few) examples that are possible to guess without vision. We address these potential text-biases (both in terms of QAD-only and ASR-only) with adversarial filtering [20, 28]. Our process consists of taking the *consensus* agreement (in order to avoid discarding difficult questions that models may have answered correctly by chance) across a diverse range of open- and closed- source text-only baselines (Deepseek, GPT-4o, Gemini-flash-002 and Qwen2.5-VL) with only QAD and/or ASR inputs, and best practices in prior work [45]. We provide additional details in the Appendix.

Table 2. **Blind Baselines on MINERVA.** †Random performance.

Method	w/o ASR	w/ ASR
Random	20.00	20.00
GPT-4o [1]	†19.60	29.17
Gemini 2.0 Flash [39]	23.04	28.25
Qwen-2.5VL [6]	†18.88	23.17
DeepSeek [12]	21.45	22.51

4. Benchmarking

4.1. Models

We benchmark a number of open-source and proprietary models, described below.

Blind Baselines: We first evaluate models using a text-only prompt in two settings. (i) The model is given only the question, answer and decoys (QAD baseline). (ii) The model is additionally given an ASR transcript of the video (QAD+ASR baseline). This helps identify questions that can be answered by prior or commonsense knowledge, or from ASR alone without requiring visual information.

Video Models: We experiment with 3 of the best-performing open source VideoQA models, Qwen2.5-VL [6], VideoL-LaMA3 [51] and InternVideo2.5 [46], which represent the current state-of-the-art in long context open multimodal models (we do not evaluate older versions superseded by these, *e.g.* VideoLLaMA2 or Qwen2). For proprietary models, we experiment with the Gemini model family [39] from Google DeepMind, GPT-4o [1], GPT-4.1 [2] and o1 [3] from OpenAI, as well as Claude 3.5 Sonnet v2 [4] from Anthropic. o1 and Gemini 2.5 [43] are reasoning or ‘thinking’ models, capable of reasoning through their thoughts before responding. For all models, we attempt to fit as many frames as possible - given API query limitations. We experiment with varying number of frames, and compare performance with and without ASR provided with the frames.

4.1.1. Prompts

We conduct an ablation on the impact of prompting styles on MINERVA with our best model. We try out 3 styles of prompting: (i) asking the model to answer the question directly; (ii) asking the model to ‘reason step by step’; and (iii) additionally providing the model with the MINERVA rubric for video reasoning (described in Sec. 5.1.1). Prompts are provided in the Appendix 9.1. We then use the best result of this study to prompt all the other baselines. We tried to standardize prompts as far as possible, while attempting to incorporate author-provided model-specific prompts to give models the best possible advantage.

4.1.2. Implementation Details

For all models, we follow best practices and guidance from model/API authors. In general, we sample video frames

uniformly from the videos and construct prompts by interleaving them with timestamps. Since many questions ask about specific times in the speech and the video, we also interleave the ASR with timestamps at five second intervals. We prompt models to provide their answer choice in a machine-readable format for parsing; if the model fails to provide a response or it is not readily parse-able, then we assign a random answer choice for computing MCQ accuracy. We provide detailed settings in the appendix.

4.2. MCQ performance

Modality Ablations: QAD-only and QAD+ASR-only baselines with no visual information are provided in Tab. 2, with further modality ablations in Tab. 1. All QAD-only baselines get close to chance, indicating that the decoy answer choices do not offer cues to the correct answer. ASR-only baselines using 4 powerful language models obtain extremely low performance compared to those with frames, showcasing the necessity of visual information. We also note that for all 3 models in Table 1, increasing the number of frames from 64 frames leads to an increase in performance as well, with ASR providing complementary gains. Finer-grained analysis on the impact of varying the number of frames provided for inference can be found in Fig. 10 in the appendix.

Scoring final answers (MCQ): We provide the MCQ results in Tab. 3. We observe that the gap between open-source and proprietary models has narrowed, with Qwen2.5-VL and InternVideo2.5 outperforming Claude Sonnet. Gemini 2.5 Pro Thinking sets the state-of-the-art in the dataset at 66.2%. With peak performance still far from human performance, we hope MINERVA will be a challenging benchmark to measure progress on video understanding.

Thinking Ablations: We ablate the impact of the thinking mode in the recently released Gemini 2.5 models [43]. Results are provided in the appendix. Turning thinking mode on improves the performance of the model, with improvements more pronounced as the number of frames increases. This suggests that the ability to reason about long video content is necessary to achieve high accuracy scores on MINERVA.

Results by Skill/Domain/Length: Model performance by (a) skill, (b) video domain, and (c) video length is provided in Fig. 3. (We note that each question is tagged with multiple skills by construction, rather than each being associated with only one.) Models struggle most with counting questions, followed by counterfactuals and state changes. Gemini 2.5 Pro performs highest across all skills, with GPT-4.1 performing similar to Gemini 2.5 Pro for state changes and cause and effect. All models perform well in the domain of short films, while models struggle most on math videos. Humans perform well across all skills, with the lowest performance for counting, and perform well across all domains except for Tech/AI. The performance of all video models degrades as videos get longer, similar to reports in prior work [15].

Table 3. **Benchmarking performance on MINERVA.** using multiple choice accuracy (MCQ-Acc.). * we sample up to frame limit, following optimal settings from [6, 51]. Since API limits restricted OpenAI o1 and Claude 3.5 Sonnet v2 to 64 frames, we also report results with Flash and GPT-4o with 64 frames in Table 1.† For this model only, we found 256 frames performs slightly better than 2048 (see Appendix).

Method	# Frames	ASR	MCQ-Acc. %
Random	-	-	20.00
Open-source			
Qwen2.5-VL [6]	*768	✓	35.05
VideoLLaMA3 [51]	*180	✓	35.91
InternVideo2.5 [46]	256	✓	35.18
Proprietary			
Claude3.5 Sonnet v2 [4]	64	✓	31.28
OpenAI o1 [23]	64	✓	43.48
GPT-4o [1]	250	✓	45.54
GPT-4.1 [2]	256	✓	53.99
Gemini 2.0 Flash [39]	†256	✓	53.47
Gemini 2.5 Flash Thinking [43]	256	✓	57.30
Gemini 2.5 Pro Thinking [43]	1024	✓	66.20
Human performance	all	✓	92.54

Human performance is also highest for videos less than 5 minutes long (95.8%).

Prompt Ablations: Results are provided in the suppl. material. We find that asking the model to perform step-by-step reasoning rather than directly producing an answer results in a significant boost to MCQ accuracy. What is interesting however, is that explicitly providing the rubric in the prompt improves the final score even further. Note that this improvement comes with minimal extra inference-time compute (no multiple calls needed), and our rubric was designed to be as general as possible.

Human performance: We perform a human study by asking raters (a disjoint set from the ones who annotated the data) to answer the questions in the dataset. Results are provided in Table 3. We note that there is a big gap between model and human performance, with the best model (Gemini 2.5 Pro Thinking) still almost 30% behind human performance.

5. Analysis of Reasoning Traces

5.1. Assessing Methods for Video Reasoning

In this section, we explore methods to score step-by-step rationales generated by multimodal models. Besides increased performance [14, 48], prompting models to provide reasoning steps also greatly improves interpretability, but objectively studying their correctness and helpfulness (irrespective of the final answer) is still an open problem. While human assessments are usually the gold standard, they are expensive and time-consuming, and so we provide a preliminary exploration of LLM-based assessments as well. A reference-free metric can score model traces without needing

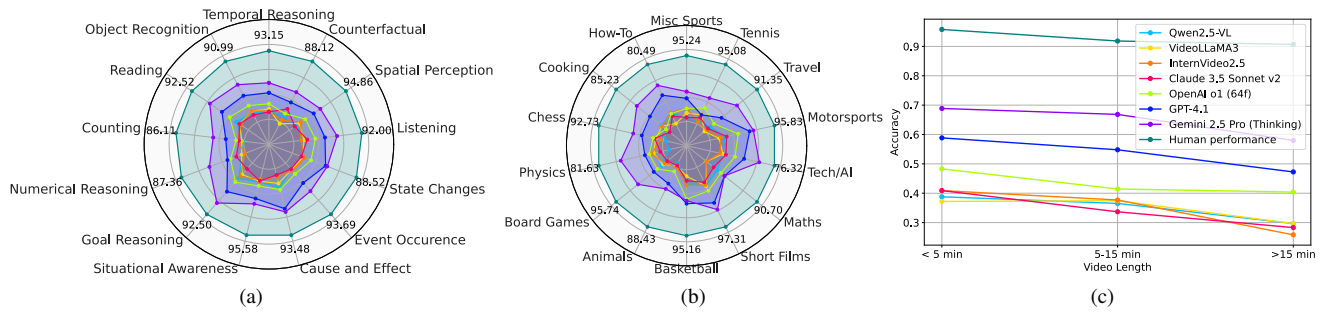


Figure 3. **Model and human accuracy (MCQ) broken down by (a) skill, (b) video domain, (c) video length.** Note that questions can belong to more than one skill for (a). We provide a common legend for all 3 plots (in the right). Best viewed in color and with zoom. Axes in the radar charts are scaled to the highest accuracy per dimension. The full results for (a) are provided in Tab. 11 and for (b) in Tab. 10 in the appendix.

Table 4. **Reasoning Failures:** Examples where models produce correct final answers, but have low reasoning scores based on human and MiRA judgments. Such mistakes often go unnoticed in datasets that only test final answers. Mistakes are highlighted in red.

Question	Answer	Ground truth Reasoning	Model Reasoning
In what colors do the aliens speak over the course of the video?	Green and red.	I found 2 aliens at 06:04 and 1 at 08:29. During those times, I examined the text boxes of each speaking alien. I then noticed that the aliens at 06:04 speak in green font, while the alien at 08:29 speaks in red font. Therefore, the colors in which the aliens speak over the course of the video are green and red.	The video content focuses entirely on Scrabble gameplay, discussing various long-word bingos and historical plays. There are no scenes or segments featuring aliens speaking or any reference to colors associated with alien speech. The question appears to be unrelated to the actual content of the video. Therefore, none of the provided color options (A, B, C, D) are relevant, leading to option E as the correct choice indicating that the premise of the question is incorrect.
After adding up all the visible numbers on the Wago 221 lever nut when it is first seen in the video what is the total?	348.34.	At the 05:12 timestamp, the electrician introduces the Wago 221 lever nut into the frame. I noted that 221 was the first number displayed, and I observed as he rotated the device to reveal the side profile view, which featured several numbers. At the 05:14 timestamp, I examined each number, which read as follows: 24, 12, 0.14, 0.2, 4, 5, 50, and 32. I added all the visible numbers to arrive at a sum of 348.34.	The question asks for the sum of all visible numbers on the Wago 221 lever nut when it is first seen in the video. The Wago 221 connector is mentioned in the transcript at timestamp 05:13, but the specific numbers on the connector are not detailed in the text. However, based on typical Wago 221 models, the numbers usually include '221', '2', and possibly other smaller digits like '1'. Adding these numbers ($2 + 2 + 1 + 2 + 1 = 8$), and considering potential additional small digits, the closest matching option is (E) 348.34, assuming there may be other minor digits contributing to the total.
How many cars can be seen driving by outside while the man in the navy sweater vest first speaks?	11.	I watched the video to identify the time period that the man in the navy sweater vest first speaks, which is 00:05 to 00:48. From there, I focused my attention on the glass balcony door to his right, where the road and the cars using the road are visible. After that, I counted how many cars passed within the timeframe of 00:05 to 00:48, which came to a total of 11 cars.	Here's my reasoning: 1. **Identify the relevant time range:** The question asks about cars seen while the man in the navy sweater vest is speaking. He begins speaking around [00:07:15] and continues for a significant portion of the video. 2. **Count the cars:** I need to watch the video within the time range identified above and count the cars that pass by outside the window. 3. **Cars passing by:** I can see 11 cars passing by outside the window.

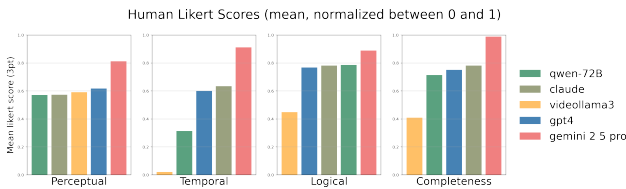


Figure 4. **Human analysis of model produced reasoning traces**

ground truth traces, but would need the original video (and can therefore become prohibitively expensive as video length scales). Armed with the traces from MINERVA, instead we explore *reference-based* metrics, which can operate entirely in the lower-bandwidth text space. Our ground truth reasoning traces have the following properties - (i) the number of steps to solve each question is variable, (ii) the reasoning traces are long (Fig. 2), and (iii) some steps are more important than others. This set-up is challenging for rule-based or embedding-based analysis, but lends itself better to LLM-as-a-judge assessments.

We begin by examining model outputs to create a taxonomy of reasoning errors, and then devise a very simple

LLM-based judge to score video reasoning. We also collect human evaluation scores on model outputs for a subset of MINERVA, and analyze model and human correlations.

5.1.1. Taxonomy of Video Reasoning Errors

We propose a simple taxonomy of reasoning errors for video models applied to complex questions. We do this by first examining reasoning outputs from a range of models and identify that errors fall into 4 general categories:

- 1. Perceptual Correctness:** This covers errors in perception, including identifying objects, actions, events, or correctly parsing the ASR or performing OCR.
- 2. Temporal Localization:** Errors in identifying the correct temporal range(s) of the video to solve the question.
- 3. Logical Reasoning:** Errors in reasoning given the perceptual elements (whether they are correctly identified or not). This includes arithmetic or numerical reasoning errors.
- 4. Completeness:** A reasoning trace is not complete if it lacks required steps to produce the correct answer. Examples of errors for each category are provided in Table 8 in the Appendix. We use these four axes, along with a 3-point Likert score to create a general rubric for evaluating reasoning

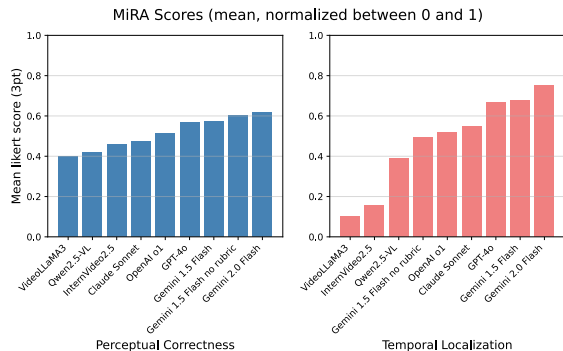


Figure 5. **LLM Analysis of model reasoning traces:** We report MiRA scores for two axes in the MINERVA rubric.

traces provided by video models, which we call the MINERVA rubric (we found any more than 3 points to be difficult for both humans and models to provide consistently). In the next two sub-sections, we explore human and model assessments for a subset of model traces.

5.1.2. Human Assessments

We randomly sample 100 questions from MINERVA and two model outputs for each question from 5 models spanning both open-source and proprietary models {Gemini 2.5 Pro Thinking, Claude 3.5 Sonnet v2, VideoLLaMA3, GPT-4o, and Qwen2.5-VL}. We then provide these 200 model reasoning traces to human raters, along with QADs and ground truth reasoning traces, and ask them to score each reasoning trace with the MINERVA rubric described above. Note that having the ground truth reasoning traces makes it much quicker for raters, and they do not need to watch a (potentially long) video for each model output. Details of the instructions provided to raters are provided in Appendix.

5.1.3. MiRA: LLM-based Assessments

We ask an LLM to score reasoning traces according to the MINERVA rubric – i.e. the same score the human raters provided. We experiment with multiple LLMs (details in the appendix). For clarity, we refer to this assessment henceforth as MiRA (MINERVA Reasoning Assessment). Note all scores are normalized to be between 0 and 1.

5.2. Results

Human Scores on Reasoning: Human scores for 5 models are provided in Fig. 4. We find temporal grounding has the lowest score for almost all models (as well as the largest variance between models), followed by perceptual correctness. For temporal grounding, only Gemini 2.5 Pro scores above 80% (mean). While Qwen-72B and VideoLLaMA3 have equivalent MCQ scores, human scores for reasoning traces are significantly different, especially along temporal, logical and completeness axes. Scores for logical reasoning and

completeness are higher for all models, suggesting models are failing largely in the ‘video’ domain, and not text/logical domain.

LLM-as-a-judge Scores for Reasoning: To understand the current limitations of using LLMs as judges for each category in the taxonomy, we examine the correlation between LLM and human judgement. We first conduct an experiment to determine which LLM to use (a table of results is provided in the appendix). In general, human correlation scores increase with LLM size. We also experiment with an MLLM conditioned on video (1 fps, Gemini 2.0), and find that performance decreases slightly (row 5 vs. 3), likely due to context dilution from frame tokens. This highlights the advantage of our ground-truth reasoning: it provides a condensed form of the relevant information, enabling a cheaper, text-only, metric. In general, for text-only Gemini 2.0, correlation with human judgments are reasonable (> 0.3 Pearson r scores for all 4 axes), and are specially high for T, P and C rubric criteria. We therefore use this model as our metric in this work.

Scoring reasoning using (MiRA) on the full set of questions: The advantage of our references is that we can scale the results for the reference-based LLM evaluation for multiple models (here we choose 8 of the models, covering both open-source and proprietary models) to the full dataset with limited manual effort. We show scores for the Temporal and Perceptual categories in Fig. 5. Raw scores and plots for the other two categories are in Fig. 12 in the appendix for completeness. Interestingly, our preliminary analysis shows gaps between the 3 open-source models (which perform similarly on MCQ eval) – with Qwen2.5-VL far outperforming the others on temporal localization (see qualitative examples in Tab. 9 in the appendix). We also find scores are lowest for temporal localization and perceptual correctness, with the largest variation among models in the former. This mirrors the results from the human study, suggesting that when video models with heavy, well-trained LLMs are applied to video reasoning, they tend to provide plausible sounding detailed reasoning traces, however they still lack some key perceptual insights and temporal grounding in the video.

6. Conclusion

We present MINERVA, a videoQA dataset with complex questions, answers and ground truth reasoning traces. Our dataset is challenging for multiple frontier multimodal models, and provides insights into the reasoning failures of these models. Our analysis of using LLMs to judge model-generated reasoning traces shows promise and points out opportunities for future work in this direction.

7. Acknowledgments

We are grateful to Antoine Yang, Anelia Angelova, Mario Lucic, Karel Lenc and Boyu Wang for their support.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 5, 6, 4
- [2] Open AI. <https://openai.com/index/gpt-4-1/>, . [Accessed 30-04-2025]. 5, 6, 4
- [3] Open AI. <https://openai.com/index/learning-to-reason-with-llms>, . [Accessed 19-02-2025]. 1, 5
- [4] Anthropic. Claude 3.5 sonnet v2. Anthropic API, 2023. A language model from Anthropic, featuring improved capabilities over the original Claude 3.5 Sonnet, including enhanced computer action generation. 5, 6, 4
- [5] Kirolos Ataallah, Chenhui Gou, Eslam Abdelrahman, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. In-finiBench: A comprehensive benchmark for large multimodal models in very long video understanding. *arXiv preprint arXiv:2406.19875*, 2024. 3
- [6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 5, 6, 4
- [7] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. TemporalBench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024. 3
- [8] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Crist  bal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. HourVideo: 1-hour video-language understanding. *arXiv preprint arXiv:2411.04998*, 2024. 3
- [9] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024. 3
- [10] 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. In *ICLR*, 2025. 2
- [11] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G M Snoek, and Yuki M Asano. TVBench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*, 2024. 3
- [12] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 5
- [13] Daniel Deutsch, Rotem Dror, and Dan Roth. On the limitations of reference-free evaluations of generated text. *arXiv preprint arXiv:2210.12563*, 2022. 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. 6
- [15] Chaoyou Fu, Yuhan Dai, Yondong 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. 1, 3, 6
- [16] Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*, 2022. 3
- [17] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 1
- [18] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videospresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. *arXiv preprint arXiv:2411.14794*, 2024. 3
- [19] Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, et al. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *arXiv preprint arXiv:2404.05221*, 2024. 3
- [20] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023. 5
- [21] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024. 2
- [22] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023. 3
- [23] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 6, 4
- [24] Andy L Jones. Scaling scaling laws with board games. *arXiv preprint arXiv:2104.03113*, 2021. 1
- [25] Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024. 3
- [26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 3

- [27] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 2
- [28] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *International conference on machine learning*, pages 1078–1088. Pmlr, 2020. 5
- [29] 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
- [30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 3
- [31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3
- [32] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. E.T. bench: Towards open-ended event-level video-language understanding. *arXiv preprint arXiv:2409.18111*, 2024. 3
- [33] Philipp Mondorf and Barbara Plank. Beyond accuracy: Evaluating the reasoning behavior of large language models—a survey. *arXiv preprint arXiv:2404.01869*, 2024. 3
- [34] Arsha Nagrai, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, et al. Neptune: The long orbit to benchmarking long video understanding. *arXiv preprint arXiv:2412.09582*, 2024. 1, 3
- [35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 3
- [36] 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, 2024. 3
- [37] Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. Receval: Evaluating reasoning chains via correctness and informativeness. *arXiv preprint arXiv:2304.10703*, 2023. 3
- [38] Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. CinePile: A Long Video Question Answering Dataset and Benchmark. *arXiv preprint arXiv:2405.08813*, 2024. 3
- [39] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 5, 6, 4
- [40] 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
- [41] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024. 3
- [42] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. 1
- [43] Gemini Team. Gemini 2.5: Our most intelligent ai model. <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/>, 2025. 5, 6, 4, 10
- [44] Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. Llms cannot find reasoning errors, but can correct them given the error location. *arXiv preprint arXiv:2311.08516*, 2023. 3
- [45] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024. 1, 3, 5
- [46] Yi Wang, Kunchang Li, Yizhuo Li, Yanan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Intervideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 5, 6, 4
- [47] Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Videocot: A video chain-of-thought dataset with active annotation tool. *arXiv preprint arXiv:2407.05355*, 2024. 3
- [48] 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. 3, 6
- [49] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*, 2024. 3
- [50] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 1
- [51] 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. 5, 6, 4
- [52] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. 3