

MMReason: An Open-Ended Multi-Modal Multi-Step Reasoning Benchmark for MLLMs Toward AGI

Huanjin Yao¹, Jiaxing Huang^{1✉}, Yawen Qiu², Michael K. Chen¹, Wenzheng Liu⁴, Wei Zhang⁵
 Wenjie Zeng², Xikun Zhang¹, Jingyi Zhang¹, YuXin Song³, Wenhao Wu³, Dacheng Tao¹

¹Nanyang Technological University ²Tsinghua University ³Baidu Inc.
⁴University of California ⁵University of Science and Technology of China

✉ Corresponding author

Abstract

Reasoning plays a crucial role in advancing Multimodal Large Language Models (MLLMs) toward Artificial General Intelligence. However, existing MLLM benchmarks often fall short in precisely and comprehensively evaluating long-chain reasoning abilities from three key aspects: (1) lack of difficulty and diversity, (2) susceptibility to guessability and memorization, (3) inadequate assessment of intermediate reasoning steps. To fill this gap, we introduce **MMReason**, a new benchmark designed to precisely and comprehensively evaluate MLLM long-chain reasoning capability with diverse, open-ended, challenging questions. First, we curate challenging questions requiring multi-step reasoning from various fields (i.e., 6 disciplines) and multiple difficulty levels (i.e., from pre-university to university, and from foundational to competition tiers). Second, these questions are reformulated into an open-ended format and filtered using a multi-model voting technique to eliminate shortcut cases related to guessing and memorization, ensuring robust reasoning evaluations. Third, we annotate the questions with detailed step-by-step solutions, and design a reference-based ternary scoring mechanism to reliably assess intermediate reasoning steps. With **MMReason**, we benchmark popular leading MLLMs and provide an in-depth analysis of their reasoning capabilities. We hope **MMReason** will serve as a valuable resource for advancing MLLM reasoning research.

1. Introduction

Recent breakthroughs in the reasoning capabilities of Large Language Models (LLMs) [11, 18, 23, 36, 44] are progressively paving the way toward AGI. Inspired by these advances, Multimodal Large Language Models (MLLMs) [35, 45, 46, 59] have shifted their focus toward

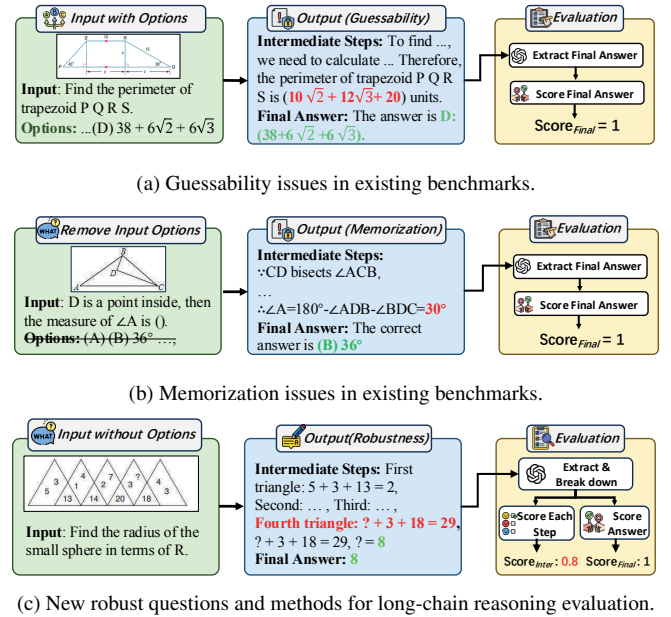


Figure 1. Existing benchmarks are susceptible to (a) guessability, where MLLMs arrive at correct answer by guessing despite flawed reasoning, and (b) memorization, where models recall leaked question types and answers rather than reasoning. We propose (c) a robust open-ended benchmark designed to evaluate long-chain reasoning by assessing both intermediate steps and final results.

long-chain reasoning, and aim to integrate textual and visual information to generate long-chain reasoning steps, ultimately solving complex problems across various domains. Leading-edge MLLMs, such as GPT-o1 [18], have exhibited exceptional understanding and reasoning abilities, even surpassing human performance. For example, on MathVista [25], GPT-o1 achieves a score of 73.9%, which significantly outperforms the human performance of 60.3%. Despite these advancements, a critical question arises: *how well do existing benchmarks truly evaluate the reasoning*

abilities of MLLMs?

While existing benchmarks [9, 10, 21, 24, 25, 51, 53] have laid the foundation for early-stage MLLM research and development, they often fall short in precisely and comprehensively evaluating MLLM long-chain reasoning capabilities. We identify three key limitations: (1) *Limited Difficulty and Diversity*. Most existing MLLM benchmarks consist primarily of simple textual questions paired with easily interpretable images from limited domains, posing minimal challenges for today’s advanced models. These homogeneous, low-complexity questions require little long-chain reasoning, making them insufficient for comprehensively assessing MLLMs’ long-chain reasoning capabilities. (2) *Susceptibility to Guessability and Memorization*. Most existing MLLM benchmarks are dominated by multiple-choice questions (MCQ), e.g., 94% of MMMU is MCQ. However, as illustrated in Figure 1, we observe that MCQ questions are susceptible to guessing and memorization issues. Specifically, MLLMs may arrive at the correct final answer either via random guessing or by recalling leaked and memorized benchmark data, even when their intermediate reasoning steps are incorrect. These shortcuts undermine the reliability of existing benchmarks in precisely measuring true long-chain reasoning capabilities. (3) *Lack of Intermediate Steps Evaluation*. Most existing MLLM benchmarks assess only the final answer, neglecting the evaluation of intermediate reasoning steps, especially for multi-disciplinary exam questions that require complex, multi-step reasoning and critical thinking. This oversight prevents a thorough assessment of MLLMs’ step-by-step reasoning processes.

To tackle these challenges, we propose MMReason, an open-ended, Multi-Modal, and Multi-Step Reasoning benchmark designed for precisely and comprehensively evaluating MLLM long-chain reasoning capabilities. MMReason is characterized by three key aspects:

1. **Collect challenging and diverse reasoning questions.**

We first reuse existing benchmarks [3, 5, 53, 54] by carefully selecting and collecting questions that requires long-chain reasoning from them. On the other hand, given the limited difficulty and diversity of these benchmarks, we further curate more challenging questions by sourcing new ones from the internet across various education grades and difficulty levels, covering pre-university school exams, Olympiad competitions, university exams, and university-level competitions. These questions span six widely studied disciplines (i.e., Mathematics, Business, Science, Engineering, Social Science, and Health), encompassing both foundational and competition-level topics. Solving them requires critical thinking and multi-step reasoning, making them well-suited for assessing MLLMs’ advanced long-chain reasoning capabilities.

2. **How can we mitigate imprecise reasoning evaluation caused by guessability and memorization?**

To reduce evaluation biases stemming from these issues, we employ two data-filtering strategies. First, since the collected data includes multiple-choice questions, we select those with unambiguous answers and manually reformulate them into open-ended questions to minimize guessability. Second, to address memorization issues, we design a multi-model voting technique to filter out leaked and memorized questions. This approach also enhances the multimodal relevance of the collected questions. In this way, if a MLLM can still derive the correct final answer after potential shortcuts (e.g., random guessing from choices and recalling leaked answers) have been eliminated, we believe its intermediate reasoning steps are much more likely to be valid. This, in turn, enables MMReason to provide a more precise assessment of MLLMs’ final answer reasoning abilities while also indirectly reflecting their intermediate reasoning processes.

3. **How can we evaluate intermediate reasoning steps precisely and reliably?**

We annotate questions with detailed step-by-step solutions and design a reference-based ternary scoring mechanism to reliably evaluate MLLMs’ intermediate reasoning steps. Specifically, we use GPT-4o to break down model responses into key steps, and prompt GPT-4o to score each step with the aid of annotated references, categorizing them as correct, unverifiable, or incorrect. With these references, this process mirrors a human grader evaluating a student’s problem-solving steps. Moreover, since some problems have multiple valid solutions, a grader may be unable to determine a step’s correctness. In such cases, assigning a neutral score is a more reasonable choice. Through this approach, MMReason provides a more precise and reliable evaluation of MLLMs’ intermediate reasoning ability.

With MMReason, we evaluate the long-chain reasoning abilities of popular and leading MLLMs from the perspectives of both intermediate reasoning steps and the reasoned final answer. Our contributions are summarized as follows:

(1) We curate new challenging and diverse questions across multiple grade levels and difficulty tiers to comprehensively assess MLLMs’ long-chain reasoning abilities.

(2) We investigate the issues of guessability and memorization in existing multimodal benchmarks. To address them, we introduce two data-filtering strategies to eliminate shortcuts (e.g., random guessing from choices and recalling leaked answers), aiming to precisely evaluate the multi-step reasoning capabilities of MLLMs.

(3) We annotate questions with detailed step-by-step solutions and design a reference-based ternary scoring mechanism to reliably evaluate MLLMs’ intermediate steps.

(4) We evaluate various leading MLLMs with MM-

Benchmark	Question Source (Grade & Difficulty)				Diverse Disciplines	Questions Filtering	Intermediate Steps Evaluation
	pre-university	university	Foundational	Competition			
MathVista [25]	✓	✗	✗	✗	✗	✗	✗
MMMU [53]	✗	✓	✓	✗	✓	✗	✗
MM-Math	✓	✗	✓	✗	✗	✗	✗
OlympiadBench [12]	✓	✗	✗	✓	✗	✗	✗
MME-CoT [19]	✓	✗	✓	✗	✗	✗	✓
MMReason (Ours)	✓	✓	✓	✓	✓	✓	✓

Table 1. Comparison of our MMReason with existing multimodal benchmarks

Reason, demonstrating it is a challenging benchmark for long-chain reasoning. Notably, the best model Gemini-2.5-pro [6] achieves 45.1% final answer reasoning accuracy and 79.2% intermediate steps reasoning score.

2. Related Work

2.1. Reasoning MLLMs

The rapid growth of MLLMs [2, 8, 17, 20, 31, 33, 38, 43, 47, 49, 55] has endowed them with extensive knowledge and robust multitasking capabilities, enabling their application across complex and diverse domains [13, 14, 56]. More recently, inspired by advancements in reasoning within LLMs [11, 18, 34, 36, 40], *e.g.*, Openai-o1’s [18] deliberative thinking time before responding and DeepSeek R1’s [11] large-scale reinforcement learning for enhanced reasoning, MLLMs have increasingly shifted their focus toward reasoning as well [4, 15, 22, 28, 35, 45, 46, 48, 57, 59]. Several works [30, 45, 59], *e.g.*, LLaVA-CoT [45], leverage chain-of-thought to enhance MLLMs reasoning capabilities. Others [42, 46] employ MCTS-based methods to strengthen their reasoning abilities, such as Mulberry [46] which introduces collective knowledge to MCTS to search reasoning paths, thereby enhancing reasoning and reflection capabilities. Moreover, some open-source projects [4, 39] apply reinforcement learning to multimodal field, following DeepSeek-R1 [11]. The advancement of reasoning MLLMs prompts us to develop a more precise and comprehensive multimodal benchmark for evaluating their intermediate and final reasoning capabilities.

2.2. Multimodal Benchmark

Previous benchmarks [1, 3, 5, 9, 10, 12, 16, 19, 21, 24, 26, 29, 41, 50–54, 58, 60] effectively evaluate the performance of earlier MLLMs from various perspectives, including perception, cognition, and hallucination, laying the groundwork for their development. However, as MLLMs rapidly advance in reasoning, existing benchmarks have inevitably become simplistic and homogeneous, with many dominated by multiple-choice and true/false questions, which limits their effectiveness in providing precise and comprehensive evaluations. Recently, MMMU [53] evaluates MLLMs using multi-disciplinary university exam questions but remains heavily reliant on multiple-choice ques-

tions. OlympiadBench [12] introduces a more challenging competition-level benchmark but is limited to mathematics and physics problems. MMStar [3] mitigates data leakage in existing benchmarks but also overlooks shortcuts in the questions. More recently, MME-CoT [19] benchmarks chain-of-thought reasoning in MLLMs across three dimensions: quality, robustness, and efficiency. In comparison, our proposed benchmark covers both foundational and competition-level problems, consisting of newly collected and reformulated questions, spanning various educational levels and 6 diverse disciplines. Besides, we observe that guessability and memorization issues exist in most current benchmarks, affecting the precision of evaluations. To address them, we first conduct an in-depth study in Section 3, then propose optimization methods by removing shortcuts (*i.e.*, options) in questions and applying multi-model voting-based filtering. Additionally, we employ a reference-based ternary scoring mechanism to reliably assess intermediate reasoning steps.

3. Analysis and Motivation

In this section, we delve into the issues of **guessability** and **memorization** in multimodal benchmarks, which hinder the effective evaluation of MLLM reasoning abilities. Additionally, we provide illustrations and experimental results to support our observations.

(1) **Guessability.** Current multimodal benchmarks predominantly feature multiple-choice and true/false questions due to their ease of collection and evaluation. However, questions containing shortcuts [(*i.e.*, candidate true answers)] have inherent flaws that limit their ability to effectively evaluate the reasoning capabilities of MLLMs. As illustrated in Figure 2 (Left), during the reasoning process, the model generates incorrect reasoning but still selects the correct answer. This suggests that MLLMs have a high likelihood of guessing the correct answer [among the several candidate answers] rather than reasoning, leading to imprecision in reasoning evaluation.

(2) **Memorization.** Different from previous studies [3] that analyze unintentional data leakage in current benchmarks, this paper conducts a more in-depth study by removing shortcuts to further illustrate the issue of memorization in evaluation. With the rapid development of MLLMs,

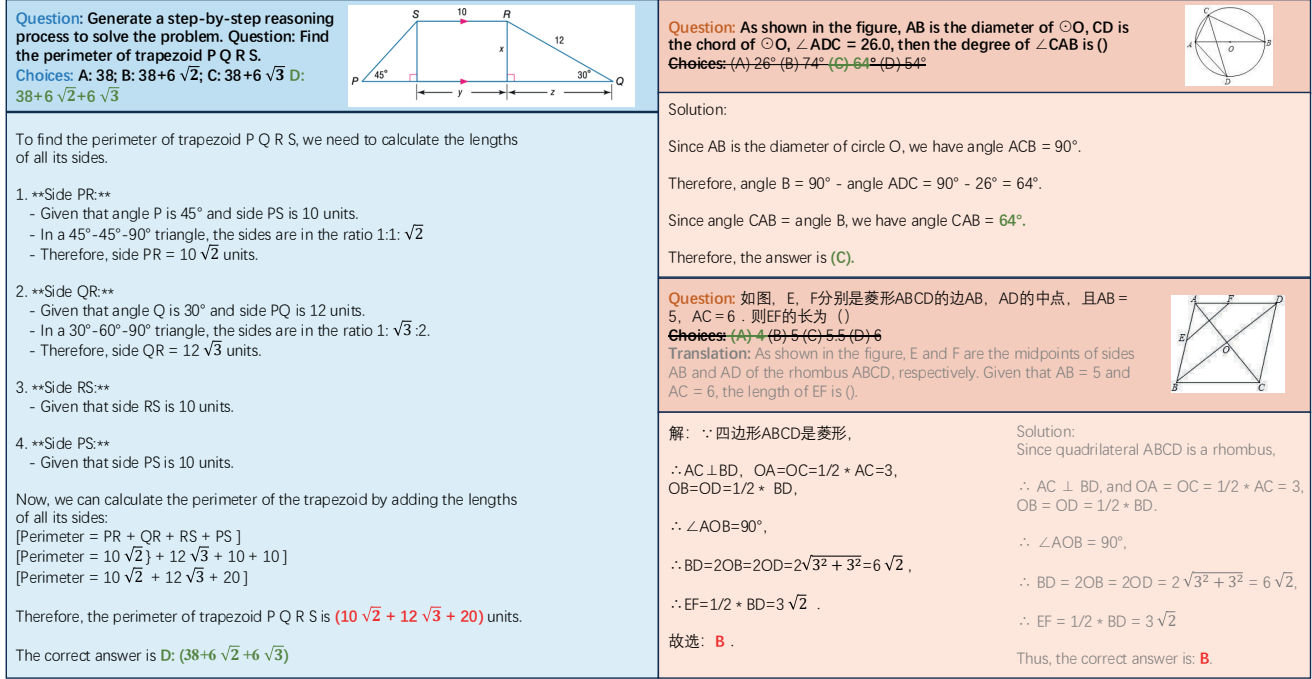


Figure 2. Examples of **Guessability (Left)** and **Memorization (Right)** issues. In the left figure, the model arrives at an incorrect numerical answer through multi-step reasoning but ultimately guesses the correct option by chance. In the right figure, we remove the choices from the original question before inputting it into MLLMs, yet the model still generates an answer choice, indicating the presence of memorization issues in existing benchmark.

the amount of data they are exposed to and their memorization ability continue to grow. Therefore, it is inevitable that models encounter similar questions and memorize their answers during training. As shown in Figure 2 (right), we remove the options from the input, but MLLMs still generate the correct option, indicating that they have already memorized the answers rather than learning to reason, leading to the correct answer.

To support our observations, we conduct experiments on existing benchmarks by comparing the original accuracy with the accuracy after removing these shortcuts, as shown in Table 2. Experimental results show that when options in existing benchmarks are removed, MLLMs fail to produce the correct answers, even for questions they originally answered correctly. Notably, the performance of Qwen2-VL-7B on MathVista dropped sharply from 58.2% to 37.8%. This indicates that MLLMs have not truly learned to reason, and in most cases, the intermediate responses generated by MLLMs are incorrect. Therefore, this motivates us to construct a more effective benchmark for precisely evaluating the multi-step reasoning capabilities of MLLMs.

4. MMReason

In this section, we provide a detailed introduction to the construction process and evaluation strategy of MMReason. Section 4.1 elaborates on our dataset curation process,

Benchmark	Input	GPT-4o	Qwen2-VL-7B
MathVista [25]	{Q, C, I}	63.8	58.2
	{Q, I}	40.8	37.8

Table 2. Performance Comparison of Removing Choices from Existing Benchmarks. By removing answer choices, instances of guessability and memorization are reduced as MLLMs are required to reason and conclude the answers by themselves instead of simply picking one, leading to a more precise evaluation of MLLMs’ long-chain reasoning capabilities.

while Section 4.2 introduces our multi-model voting filtering method for selecting unmemorized and visually relevant data. Finally, we outline the strategies for evaluating intermediate and final reasoning capability in Section 4.3.

4.1. Data Curation Process

To construct a new benchmark for precisely and comprehensively evaluating multi-step reasoning capability of MLLMs with diverse and challenging open-ended questions, we collect and reformulate questions from existing benchmarks and web, and annotate the questions with reference solutions. The details are as follows:

Existing Benchmark Collection. Existing multimodal benchmarks include a vast collection of multiple-choice questions spanning various domains. However, directly converting multiple-choice questions from existing bench-

marks into open-ended formats could lead to inaccuracies in evaluation, particularly due to multiple valid answers and variations in wording. To address this issue, we select several challenging benchmarks (*i.e.*, MMMU [53], MMMU-Pro [54], MMStar [3], M³CoT [5], MathVerse [58] and MathVision [37]) and identify questions with unique answers (*e.g.*, numerical calculations, chromosomal compositions, and chemical compounds), while discarding questions like “Which of the following statements is correct?”. We then reformulate them into an open-ended format.

New Data Collection. To further systematically collect new diverse and challenging open-ended questions, we first review mainstream subjects across different grade levels, prioritizing those that require multi-step reasoning and extensive vision-text interaction. Based on this, we identify six disciplines, *i.e.*, Math, Business, Science, Engineering, Social Sciences, and Health. We then source both foundational and competition-level questions for each discipline from the Internet, spanning pre-university to university levels. For foundational problems, we prioritize diversity, knowledge coverage, and critical thinking, while for competition-level questions, we focus on the necessity of visual information given their inherent difficulty. Moreover, we intentionally gather a wide range of open-ended questions types (*e.g.*, when, what, how) to enhance question diversity and assess MLLMs’ reasoning ability across different question formats. Finally, since most collected questions are in multiple-choice format and language-only, we extract multimodal questions and manually reformulate them into an open-ended format, ensuring each question has a unique answer.

Intermediate Steps Annotation To reliably evaluate intermediate reasoning steps generated by MLLMs, we introduce a reference-based ternary scoring mechanism with detailed strategies outlined in Section 4.3. To implement the reference mechanism, we annotate questions with multi-step solutions. Specifically, we carefully select a subset of questions from our newly collected data across different disciplines and manually annotate the intermediate steps with GPT assistance to ensure accuracy at each step.

In total, we collect 4300 instances from existing benchmarks, which, combined with the 756 newly collected instances, results in a total of 5056 instances.

4.2. Robustness Enhancement via Voting

To further mitigate memorization issues and enhance the visual relevance, we employ a multi-model voting mechanism with K powerful MLLMs to filter out potentially memorized or visually weakly relevant instances from our collected multimodal questions Q , as detailed in algorithm 1. Specifically, we leverage a group of powerful MLLMs, denoted as $\{\pi_1, \pi_2, \dots, \pi_K\}$, which perform filtering through voting iteratively for T rounds. During evaluation, we only

Algorithm 1: Multi-modal Voting filtering

Input: a group of MLLMs $\{\pi_1, \pi_2, \dots, \pi_K\}$; A set of multimodal questions $Q = \{T, V\}$ containing M questions.

```

1 for  $i = 1$  to  $T$  do
2   for  $k = 1$  to  $K$  do
3     Input only text  $T$  during evaluation:
4     Evaluation( $\pi_k, Q_T = \{T\}$ )
5   Traverse the question set  $Q$ :
6   for  $j = 1$  to  $M$  do
7     If question  $q_j$  is answered correctly more
      than 0 times, remove it:
8     if Count( $q_j$ ) > 0 then
9       Remove  $q_j$  from  $Q$ 
```

Output: Filtered Data Q

input the open-ended textual questions while excluding the image to test whether the MLLMs can answer correctly without visual input. If the model can still answer correctly without any shortcuts or visual cues, we consider the instance potentially memorized or visually weakly relevant. After each round of evaluation by all MLLMs, we remove any correctly answered questions and proceed with the updated question set Q' for the next round of filtering. After T iterations, we obtain 3275 robust and vision-related multimodal questions, forming MMReason.

Finally, MMReason is divided into two subsets: testmini and test. The testmini subset contains 1,000 examples and is designed to enable more efficient model development and validation, as long-chain reasoning is more time-consuming. The test set comprises the remaining 2,275 examples and is used for standard evaluation.

4.3. Evaluation Strategy

Intermediate step reasoning ability evaluation. A straightforward approach to evaluate intermediate steps is to utilize a highly capable MLLM (*e.g.*, GPT-4o) to extract key steps from reasoning responses and directly assign binary scores to them. However, directly using powerful MLLMs to score these segmented steps is unreliable, particularly for questions beyond model’s own knowledge. To address this, we adopt a reference-based ternary scoring mechanism (*i.e.*, correct, unverifiable, and incorrect), corresponding to scores of 0, 0.5, and 1, respectively, to ensure a more reliable assessment of intermediate reasoning steps. Here, ‘Unverifiable’ refers to steps where MLLMs are still unable to make an accurate judgment despite having access to the reference solutions. Specifically, we use GPT-4o to break down intermediate reasoning responses into N steps, denoted as $[s_1, s_2, \dots, s_N]$. Next, we provide GPT-4o with

Models	Size	Overall		Discipline-wise final reasoning accuracy					
		Final	Intermediate	Math	Business	Science	Engineering	Social	Health
Closed-Source Models									
Gemini-2.5-pro-preview-03-25 [6]	-	45.1	79.2	51.0	59.1	36.8	33.3	45.7	63.8
Claude-sonnet-4-0514-thinking [32]	-	33.2	61.7	42.1	47.0	21.1	20.7	37.1	53.2
GPT-4.1 [27]	-	27.5	60.9	22.9	47.7	22.6	18.2	34.3	55.3
GPT-4o-1120 [17]	-	21.1	53.9	14.0	45.5	16.5	10.6	31.4	51.1
Open-Source Models									
Deepseek-VL2 [43]	4.1B [†]	11.5	47.1	7.2	22.7	8.4	4.5	28.6	27.6
LLaVA-OneVision [20]	7B	11.0	33.4	10.9	19.7	5.3	4.0	22.8	29.8
Mulberry [46]	7B	11.8	47.1	9.6	21.9	8.0	6.1	21.4	27.7
Qwen-2.5-VL [38]	7B	17.3	39.6	15.4	29.5	11.9	10.6	25.7	40.4
R1-VL [57]	7B	17.9	49.5	19.2	28.9	11.4	8.6	30.1	34.7
MiniCPM-V-2.6 [49]	8B	9.6	26.6	8.2	15.1	6.1	4.0	24.3	23.4
InternVL-2.5-MPO [38]	8B	13.5	35.3	13.0	21.9	9.2	6.5	28.6	23.4
LLaMA-3.2-Vision [8]	11B	9.7	31.4	5.1	17.4	8.0	3.5	24.3	29.8
InternVL-2.5-MPO [38]	78B	23.1	50.2	25.7	40.9	17.2	13.1	25.7	48.9
Qwen-2.5-VL [2]	72B	23.4	47.7	23.3	48.5	16.9	8.6	28.6	44.7

Table 3. **Main Results** on testmini. We provide final-answer reasoning accuracy and intermediate-step reasoning scores for 14 mainstream MLLMs, as well as final answer reasoning accuracy across different disciplines. [†] denotes activated parameters of MoE model.

referenced multi-step solutions and evaluate each step using a ternary score. After scoring, we compute the final score by summing and averaging the scores across all steps:

$$S_{inter} = \frac{1}{N} \sum_n^N Score(s_n) \quad (1)$$

Final answer reasoning ability evaluation. To assess final answer reasoning ability of MLLMs, we depart from the common approach of using a prompt like “Answer the question using a single word or phrase.” which forces MLLMs to output only the final result. Instead, we encourage MLLMs to generate multi-step responses leading to the final answer, allowing us to evaluate their reasoning capabilities rather than memorizing. Since MMReason involves long-chain reasoning and unrestricted answer formats for each question, it is challenging to determine correctness through predefined rules. Therefore, following previous methods [25, 53, 58], we prompt GPT-4o to extract the final answer from response and then assess its correctness, obtaining the final answer reasoning accuracy S_{final} .

5. Experiments

5.1. Experiment Setup

Filtering Setting We implement $K = 4$ well-known MLLMs for multi-model voting filtering to mitigate memorization issues and enhance visual relevance. These models include GPT-4o [17], LLaMA3.2-Vision-11B [8], InternVL2.5-8B-MPO [38], and Qwen2.5-VL-7B [2]. These MLLMs are widely studied by both industry and academic, including in data distillation and post-training.

Models	Reformulated from				Newly collect (Ours)
	M ³ CoT	MMStar	MMM _U .Pro	MMM _U	
DeepSeek-vl2 [43]	16.9	22.7	11.6	12.4	8.3
InternVL2.5-78B-MPO [38]	29.1	28.6	20.1	18.6	14.2
Qwen2.5-VL-72B [2]	32.3	30.3	21.1	21.6	18.5
GPT-4o-1120 [17]	28.9	32.7	24.9	24.5	21.8

Table 4. **Final answer reasoning accuracy over different data sources.** We analyze the accuracy of different question sources in MMReason. Our newly collected data has the lowest accuracy, highlighting its difficulty for MLLMs.

Models	Text-only	Text & Visual	Multimodal Relevance Rate [†]
<i>Before filtering</i>			
LLaMA-3.2-11B	6.51	15.0	56.6
InternVL2.5-8B-MPO	7.23	17.9	59.6
Qwen2.5-VL-7B	9.36	23.8	60.7
GPT4o	13.4	30.8	56.5
<i>After filtering</i>			
LLaMA-3.2-Vision-11B	0.65	11.7	94.5
InternVL2.5-8B-MPO	0.69	12.4	94.5
Qwen2.5-VL-7B	0.73	18.3	96.1
GPT4o	0.78	25.7	97.0

Table 5. **Comparison before and after multi-model voting filtering.** ‘Text’ and ‘Visual’ indicate whether textual questions and visual information are provided, respectively.

Therefore, leveraging them for filtering in $T = 2$ rounds enhance the robustness for our benchmark.

Evaluation Setting. We benchmark the reasoning capabilities of 14 leading MLLMs on MMReason, including 4 closed-source and 10 open-source models. For closed-source MLLMs, we benchmark Gemini-2.5-pro-preview-03-25 [6], GPT-4o-1120 [17], Claude-sonnet-4-0514-thinking [32] and GPT-4.1 [27]. For open-source models, we evaluate Deepseek-VL2 [35], LLaVA-OneVision-

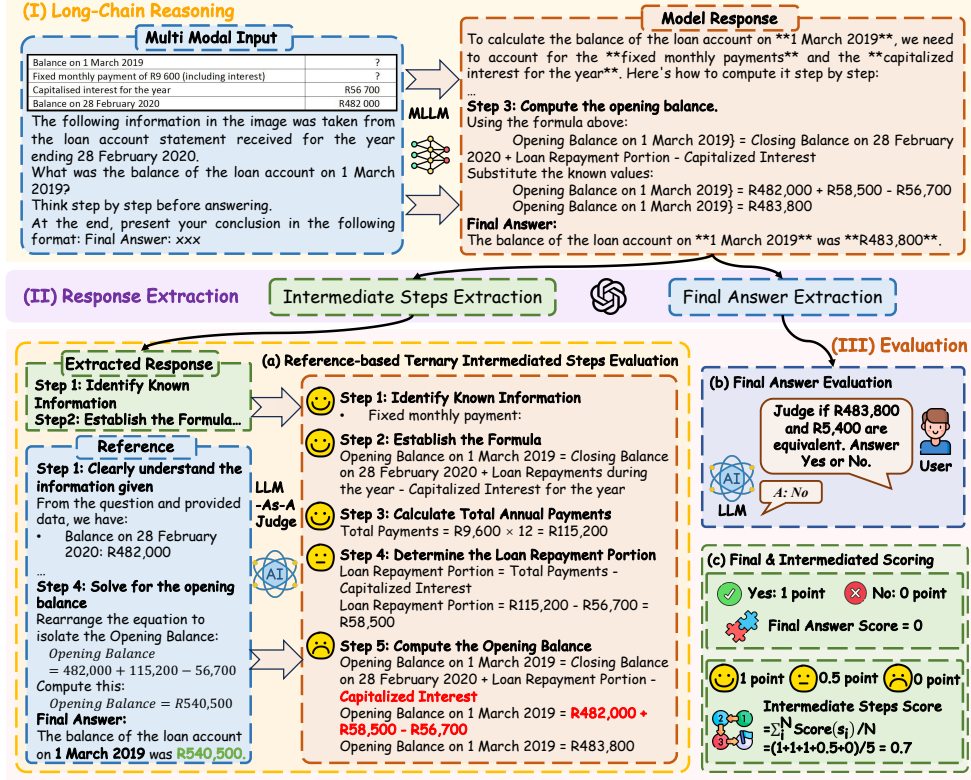


Figure 3. Qualitative analysis.

7B [20], Mulberry-7B [46], R1-VL-7B [57], Mini-CPM-V-2.6-8B [49], LLaMA-3.2-V-11B [8], InterVL-2.5-MPO-8B&78B [38], Qwen2.5-VL-7B&72B[2]. We use VLMEvalKit [7] to evaluate these models on MMReason.

5.2. Experimental Results

5.2.1. Main Results

In this section, we comprehensively compare the reasoning capabilities of various MLLMs in both intermediate steps and final answers on MMReason. Table 3 presents their overall final-answer reasoning accuracy and intermediate-step reasoning accuracy, with final-answer reasoning accuracy across different disciplines listed in right.

Final answer reasoning accuracy. MMReason benchmark presents a significant challenge to current MLLMs, requiring them to derive correct answers through multi-step reasoning without shortcuts (*i.e.*, answer choices). Table 3 shows that even the highest accuracy, obtained by Gemini-2.5-pro [6], reaches only 45.1% on MMReason testmini set. Notably, Qwen-2.5-VL-72B [2] significantly outperforms other open-source models, achieving an accuracy of 23.4%, which is comparable to proprietary models, *i.e.*, 21.1% on GPT-4o [17], 27.5% on GPT-4.1 [27]. Among smaller reasoning models, the o1-style Mulberry-7B [46] achieves 11.8% accuracy, while the MoE-based DeepSeek-VL2 reaches 11.5%. The accuracy of final answer reasoning reflects the challenging nature of our benchmark.

Intermediate steps reasoning scores. We design a reference-based ternary scoring mechanism to evaluate the intermediate steps generated by MLLMs, with the results shown in Table 3. Among them, Gemini-2.5-pro [6] achieves the highest score in intermediate step reasoning evaluation, with a score of 79.2%. Compared to accuracy of the final answer reasoning, closed-source models perform relatively better in intermediate step reasoning, showing a larger gap from open-source models. This demonstrates their stronger capability in long-chain reasoning.

Discipline-wise final reasoning accuracy. We investigate the accuracy of final answer reasoning across different disciplines. The results show that Gemini-2.5-pro Pro exhibits the strongest mathematical reasoning ability, achieving 51.0%. Moreover, engineering-related questions are generally more challenging. For example, LLaMA3.2-Vision achieves only 3.5% accuracy on these questions, as they require extensive domain knowledge and are most likely derived from university-level. The diversity of questions in MMReason enables us to evaluate the reasoning abilities of different models across various disciplines.

5.3. Analysis

Final answer reasoning accuracy from different sources.

In Table 4, we provide accuracy across different data sources, including questions collected and reformulated from existing benchmarks, as well as newly collected

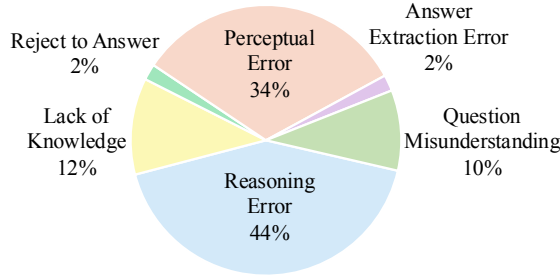


Figure 4. Error distribution over 50 incorrect responses from GPT-4o, with reasoning and perception errors being the most prevalent.

ones. The results are evaluated on three powerful reasoning models (*i.e.*, LLaMA-3.2-vision-11B [8], InternVL2.5-8B-MPO [38], Qwen2.5-VL-7B [2], and GPT-4o [17]), showing that our newly collected diverse data is the most challenging. This is attributed to our data collection strategy, which incorporates varying difficulty levels and educational stages across multiple disciplines.

Memorization. In Table 5, we present the accuracy before and after filtering, with and without visual input. We also use the multi-modal relevance rate to analyze the extent of memorization in different models. Specifically, we first evaluate the initially open-ended questions under two conditions: text-only input and combined text-visual input. The results in Table 5 indicate that before filtering, the collected dataset exhibits significant memorization effects or visual irrelevance, *i.e.*, many of these questions can be answered correctly without requiring images. For example, GPT-4o achieves an accuracy of 13.4% with text-only input. Therefore, we remove these memorized data to enhance the robustness of our benchmark. After T rounds of iterative filtering, we conduct another evaluation to ensure a higher multimodal relevance rate in MMReason. These results indicate that the retained questions in MMReason are largely free from memorization and require visual input to be solved, ensuring that our benchmark precisely evaluates the long-chain reasoning capabilities of MLLMs.

5.3.1. Qualitative analysis

Reasoning Analysis. Figure 3 represents the overall evaluation process of MMReason’s Qualitative Results, including both final answer reasoning evaluation and intermediate step evaluation. For the final answer evaluation, we remove the shortcut (*i.e.*, the options), preventing the model from guessing the answer and forcing it to derive the final answer step by step through reasoning. The results in the figure indicate that the model misunderstands “Capitalized interests”, leading to incorrect reasoning. This demonstrates that MMReason can precisely evaluate the model’s reasoning ability. For intermediate step evaluation, the presence of references allows for more accurate assessment of the intermediate reasoning steps. Additionally, the use of a ternary scoring system provides a more reliable intermediate score

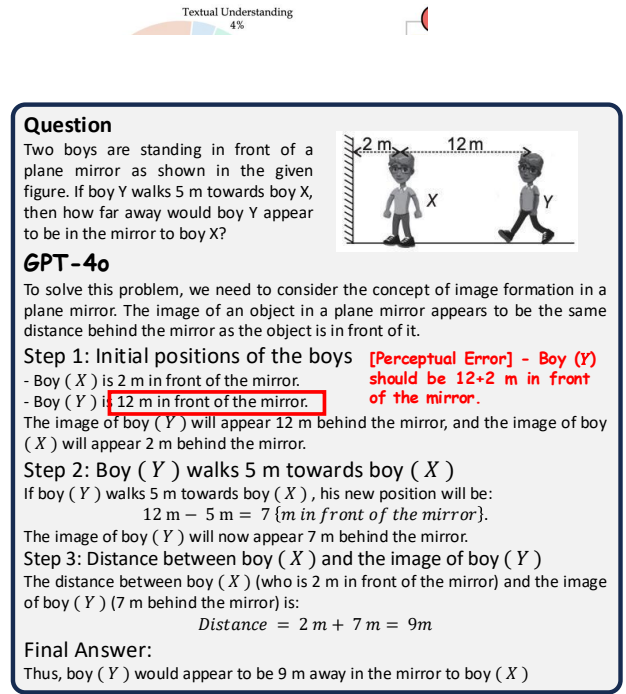


Figure 5. Qualitative results of incorrect reasoning.

for steps that the scoring model cannot confidently judge.

Error Analysis. In this section, we analyze the errors made by powerful MLLM GPT-4o in its multi-step reasoning responses. We hope this study will provide valuable insights to enhance the reasoning performance of MLLMs in the future. Specifically, we randomly select 50 cases of final reasoning errors and analyzed their causes. We identified six potential factors contributing to these errors, *i.e.*, Lack of Knowledge, Reject to Answer, Answer Extraction Error, Question Misunderstanding, Reasoning Error, and Perceptual Error. The error distribution is shown in Figure 4, with additional examples provided in the Appendix.

6. Conclusion

In this paper, we introduce MMReason, a new benchmark aiming to precisely and comprehensively evaluate long-chain reasoning capability of MLLMs with diverse and challenging open-ended questions that require multi-step-reasoning and span various education grades and difficulty levels. We evaluate the reasoning performance of multiple mainstream MLLMs on MMReason. Notably, Gemini-2.5-pro achieves the highest performance, yet its accuracy remains only 45.1%. Besides, our analysis of GPT-4o’s incorrect responses reveal that perception and reasoning error constitute a significant portion. These findings suggest that MLLMs still have considerable potential for improvement in reasoning. We hope MMReason offers valuable insights to advance the reasoning capabilities of MLLMs.

Acknowledgement. This project is supported by the National Research Foundation, Singapore, under its NRF Professorship Award No. NRF-P2024-001.

References

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 3
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-hai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 6, 7, 8, 13
- [3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 2, 3, 5, 12
- [4] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than \$3. <https://github.com/Deep-Agent/R1-V>, 2025. Accessed: 2025-02-02. 3
- [5] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M³cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proc. of ACL*, 2024. 2, 3, 5, 12
- [6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blisstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3, 6, 7, 13
- [7] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024. 7
- [8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3, 6, 7, 8, 13
- [9] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiwu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 2, 3
- [10] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 2, 3
- [11] 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. 1, 3
- [12] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024. 3
- [13] Jiaxing Huang, Jingyi Zhang, Kai Jiang, Han Qiu, and Shijian Lu. Visual instruction tuning towards general-purpose multimodal model: A survey. *arXiv preprint arXiv:2312.16602*, 2023. 3
- [14] Jiaxing Huang, Jingyi Zhang, Kai Jiang, and Shijian Lu. Open-vocabulary object detection via language hierarchy. *arXiv preprint arXiv:2410.20371*, 2024. 3
- [15] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 3
- [16] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 3
- [17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3, 6, 7, 8, 13
- [18] 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. 1, 3
- [19] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025. 3
- [20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3, 6, 7, 13
- [21] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 2, 3
- [22] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025. 3
- [23] 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. 1

- [24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 2, 3
- [25] 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. 1, 2, 3, 4, 6
- [26] 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
- [27] OpenAI. Introducing gpt-4.1 in the api, 2025. 6, 7, 13
- [28] Yi Peng, Peiyu Wang, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, et al. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*, 2025. 3
- [29] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024. 3
- [30] 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
- [31] Anthropic Team. Claude 3.7 sonnet, 2025. 3
- [32] Anthropic Team. Claude 4 sonnet, 2025. 6, 13
- [33] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 3
- [34] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025. 3
- [35] Qwen Team. Qvq: To see the world with wisdom, 2024. 1, 3, 6
- [36] Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, 2024. 1, 3
- [37] 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, 2024. 5, 12
- [38] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024. 3, 6, 7, 8, 13
- [39] Xiaodong Wang and Peixi Peng. Open-r1-video. <https://github.com/Wang-Xiaodong1899/Open-R1-Video>, 2025. 3
- [40] Yibo Wang, Li Shen, Huanjin Yao, Tiansheng Huang, Rui Liu, Naiqiang Tan, Jiaying Huang, Kai Zhang, and Dacheng Tao. R1-compress: Long chain-of-thought compression via chunk compression and search. *arXiv preprint arXiv:2505.16838*, 2025. 3
- [41] 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
- [42] Jinyang Wu, Mingkuan Feng, Shuai Zhang, Ruihan Jin, Feihu Che, Zengqi Wen, and Jianhua Tao. Boosting multimodal reasoning with mcts-automated structured thinking. *arXiv preprint arXiv:2502.02339*, 2025. 3
- [43] 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. 3, 6, 13
- [44] Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*, 2025. 1
- [45] 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. 1, 3
- [46] Huanjin Yao, Jiaying Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024. 1, 3, 6, 7, 13
- [47] Huanjin Yao, Wenhao Wu, Taojiannan Yang, YuXin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. Dense connector for mllms. *Advances in Neural Information Processing Systems*, 37:33108–33140, 2024. 3
- [48] Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, et al. R1-sharev1: Incentivizing reasoning capability of multimodal large language models via share-grpo. *arXiv preprint arXiv:2505.16673*, 2025. 3
- [49] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 3, 6, 7, 13
- [50] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024. 3
- [51] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang.

- Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. [2](#)
- [52] Jiakang Yuan, Tianshuo Peng, Yilei Jiang, Yiting Lu, Renrui Zhang, Kaituo Feng, Chaoyou Fu, Tao Chen, Lei Bai, Bo Zhang, et al. Mme-reasoning: A comprehensive benchmark for logical reasoning in mllms. *arXiv preprint arXiv:2505.21327*, 2025.
- [53] 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. [2](#), [3](#), [5](#), [6](#), [12](#)
- [54] 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. [2](#), [3](#), [5](#), [12](#)
- [55] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruvi Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024. [3](#)
- [56] Jingyi Zhang, Jiaying Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024. [3](#)
- [57] Jingyi Zhang, Jiaying Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025. [3](#), [6](#), [7](#)
- [58] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024. [3](#), [5](#), [6](#), [12](#)
- [59] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024. [1](#), [3](#)
- [60] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024. [3](#)