

Can MLLMs Reason in Multimodality?

EMMA: An Enhanced MultiModal ReAsoning Benchmark

Yunzhuo Hao^{1*}, Jiawei Gu^{2*}, Huichen Will Wang^{3*}, Linjie Li^{4*},
 Zhengyuan Yang⁴, Lijuan Wang⁴, Yu Cheng⁵

¹University of Electronic Science and Technology of China, ²Sun Yat-sen University,

³University of Washington, ⁴Microsoft, ⁵The Chinese University of Hong Kong

emma-benchmark.github.io

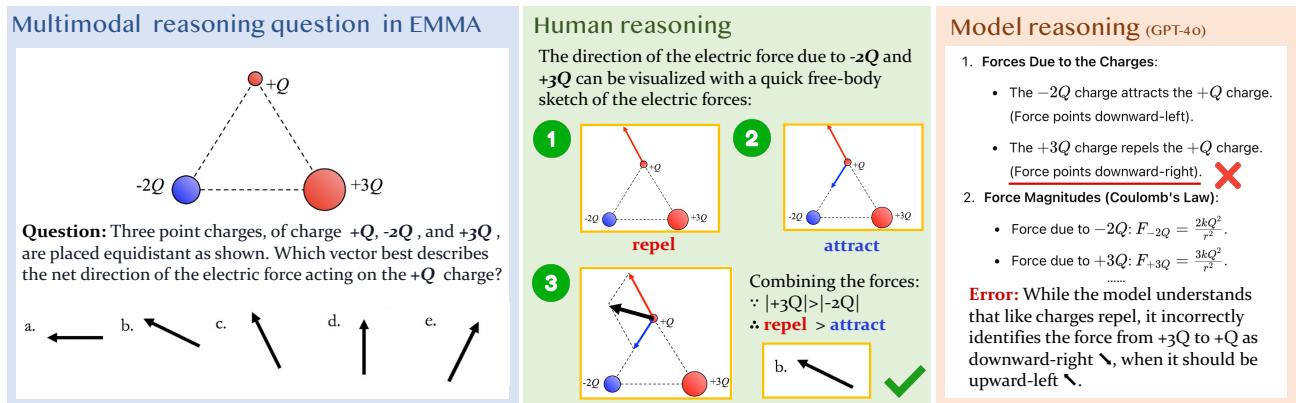


Figure 1. A sample multimodal reasoning question in EMMA. To solve it, humans engage in graphical reasoning (middle panel): guided by the principles of electric force, they draw force vectors with appropriate directions and visually compute their sum. While GPT-4o understands that like charges repel, it mistakes the direction of the repulsive force (right panel), highlighting its limitations in multimodal reasoning.

Abstract

The ability to organically reason over and with both text and images is a pillar of human intelligence, yet the ability of Multimodal Large Language Models (MLLMs) to perform such multimodal reasoning remains under-explored. Existing benchmarks often emphasize text-dominant reasoning or rely on shallow visual cues, failing to adequately assess integrated visual and textual reasoning. We introduce EMMA (Enhanced MultiModal reAsoning), a benchmark targeting organic multimodal reasoning across mathematics, physics, chemistry, and coding. EMMA tasks demand advanced cross-modal reasoning that cannot be addressed by reasoning independently in each modality, offering an enhanced test suite for MLLMs’ reasoning capabilities. Our evaluation of state-of-the-art MLLMs on EMMA reveals significant limitations in handling complex multimodal and multi-step reasoning tasks, even with advanced techniques like Chain-of-Thought prompting and test-time compute scaling underperforming. These findings underscore the need for improved multimodal architectures and training paradigms to close the gap between human and model reasoning in multimodality.

1. Introduction

Multimodal reasoning is fundamental to human problem-solving. For example, meteorologists interpret weather maps alongside numerical forecasts, and interior designers combine textual descriptions with mental imagery to optimize room layouts. Text-based reasoning allows us to analyze abstract concepts, while visual reasoning enables us to manipulate and draw insights from complex visual information. Combining these skills provides a robust framework for solving technical and creative challenges. In math, for example, formal symbolic reasoning and visual aids such as diagrams and spatial thinking work in tandem to explore ideas that neither modality could fully address alone [21].

Recent advancements in Large Language Models (LLMs) have significantly enhanced their reasoning abilities [18, 54, 56, 90], enabling strong performance on tasks such as formal logic reasoning [25], graduate-level academic question answering [59], and competitive programming [7, 14, 26]. Despite these successes, these models primarily focus on *text-only reasoning*, leaving an open question: can Multimodal LLMs (MLLMs) effectively reason across both language and visual inputs?

A major bottleneck in addressing this question is the lack

*Equal contribution.

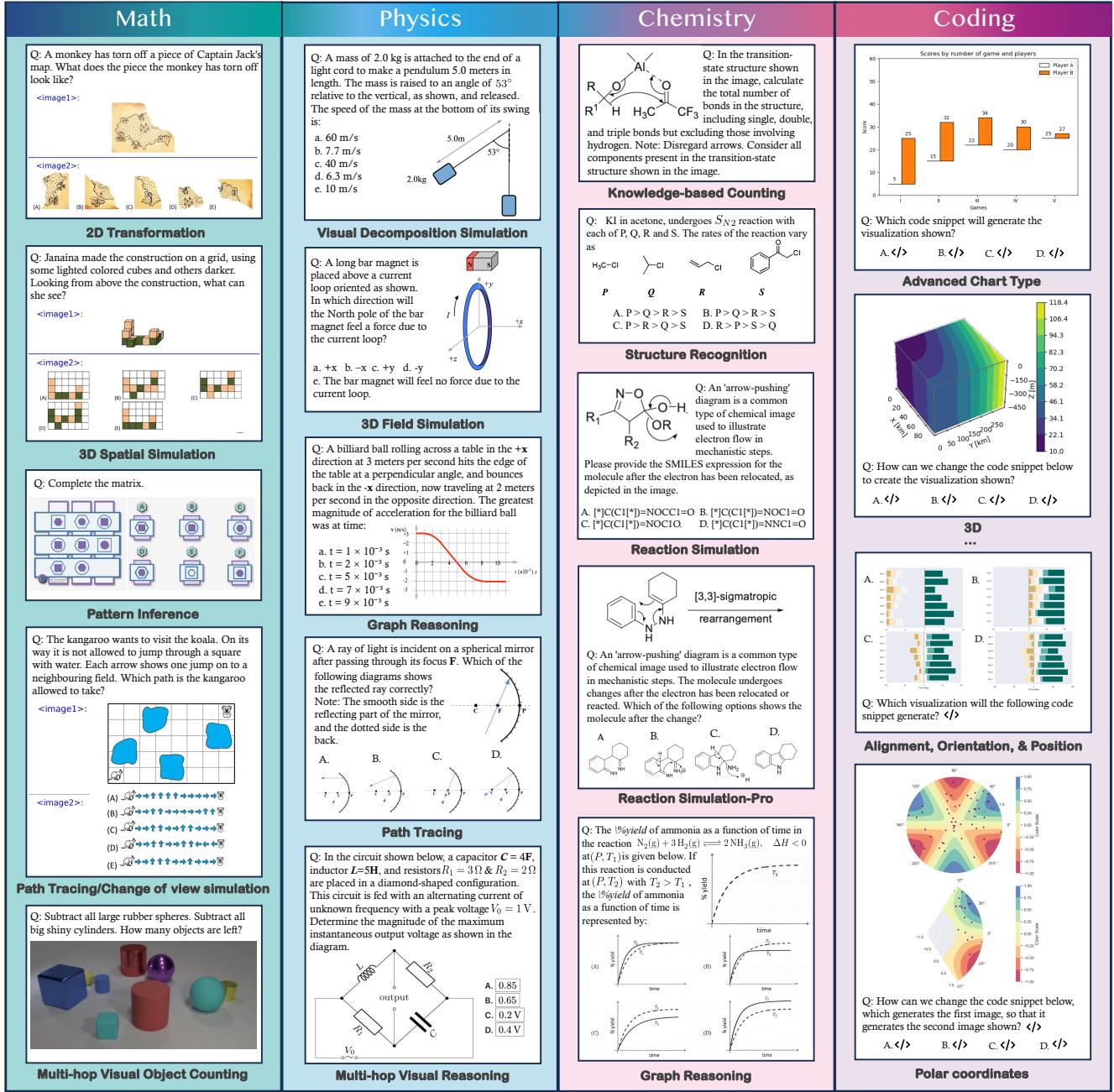


Figure 2. Overview of EMMA. EMMA covers four subjects: math, physics, chemistry, and coding. Questions are categorized based on specific skills. We provide answers to the sample questions in this figure on the GitHub page.

of appropriate benchmarks. Existing multimodal benchmarks largely test surface-level visual understanding [4, 6, 22, 31, 82] or textual knowledge recall with multimodal inputs [49, 83, 84]. While some benchmarks [47, 70] include math questions with images, studies [87] have shown that many of these tasks reduce to language-only reasoning, as the visual content is often fully described in text.

To address this gap, we introduce EMMA: an Enhanced MultiModal reAsoning benchmark, specifically designed to evaluate the ability to solve problems that require both

visual- and language-based problem-solving. EMMA features questions that *are difficult to solve by relying solely on text-based reasoning or a single visual pass*. Instead, solving these problems necessitates a back-and-forth process between interpreting visual inputs and applying multimodal reasoning steps, where visual aids are often integral or more efficient for arriving at the solution. For instance, Figure 1 illustrates a sample physics problem that asks for the direction of the net electric force. While GPT-4o understands that like charges repel, it mistakes the direction

of the repulsive force, highlighting its limitations in multimodal reasoning.

Unlike recent benchmarks [12, 58], which focus on spatial cognition, or visual puzzles that can be perfectly represented in text, EMMA introduces domain-specific challenges where reasoning can often be strengthened by visual aids. These include tasks like 3D spatial transformations, chemical structure recognition, multi-step physical simulations, and program output visualization (Figure 2). EMMA consists of 992 multimodal reasoning questions gathered from existing benchmarks through a rigorous filtering pipeline, and 1,796 newly constructed questions created manually in collaboration with domain experts. Our evaluation of nine state-of-the-art (SoTA) MLLMs on EMMA reveals three key findings:

- **MLLMs struggle with multimodal reasoning:** All models perform suboptimally on EMMA, regardless of the usage of Chain-of-Thought (CoT) prompting [73]. On the balanced subset of EMMA, the best-performing model, o1, scores only 45.75%, which is 8.5% higher than the best non-reasoning MLLM, Qwen2-VL [71], but still trails human experts by 32%. These results suggest a limitation of current MLLMs to perform in-depth multimodal reasoning.
- **Test-time compute scaling methods with textual CoTs are insufficient:** We explore test-time compute scaling of SoTA MLLMs with different methods (e.g., majority voting, best-of-N, and tournament) up to 16 times, yet they still fail to address the challenges posed by multimodal reasoning questions in EMMA. Our analysis reveals that simply increasing the number of candidate responses with textual CoTs does little to compensate for the models’ inability to produce valid visual reasoning steps, particularly for tasks requiring fine-grained spatial understanding or multi-step reasoning. In addition, current MLLMs and specialized reward models struggle with complex multimodal reasoning themselves, which can make their reward signals unreliable and limit the utility of test-time compute scaling.
- **Visual reasoning remains a fundamental bottleneck:** Through thorough error analysis, we find that SoTA MLLMs frequently struggle with tasks requiring precise spatial simulations, multi-hop visual reasoning, and integration of visual and textual information. These shortcomings are particularly pronounced in problems where visual aids offer a simpler or more natural path to the solution. Further, textual CoT negatively impacts model performance on visual-reasoning-heavy tasks, highlighting the need for new paradigms to improve visual reasoning.

These insights suggest that the performance gap between text-based and multimodal reasoning arises from MLLMs’ limited ability to perform fine-grained visual reasoning. EMMA highlights the need for new architectures and train-

ing paradigms that can better integrate and reason over diverse modalities, enabling models to leverage both visual and linguistic information more effectively.

2. Related Work

Multimodal Large Language Models Recent years have witnessed rapid progress in MLLM development. Building upon early techniques in vision-language modeling [9, 39, 46, 57, 66, 81, 86], modern MLLMs [3, 35, 37, 43, 45, 67, 69, 77] leverage the success of LLMs and achieve impressive performance in many multimodal tasks. In addition, various visual instruction tuning techniques [43, 44, 92] and the increasing availability of open-source and model-generated training data have further contributed to the robustness and zero-shot generalization ability of MLLMs.

LLM and MLLM Reasoning While early LLMs were widely regarded as mere next-token predictors with limited reasoning ability [29, 30], recent advancements in LLM research [18, 54, 56, 90] have begun to challenge this view. State-of-the-art models now achieve strong performance on tasks such as formal logic reasoning [25], graduate-level academic question answering [59], and competitive programming [14]. These advancements in text-based reasoning have spurred growing interest in multimodal reasoning, exemplified by visual CoT models such as Visual Chain-of-Thought [60] and Multimodal Chain-of-Thought [89], as well as visual CoT prompting techniques like Image-of-Thought [91]. Although visual CoT prompting techniques have shown promise, their focus is primarily on enhancing perception through methods like cropping images to simulate attention. Hence, these approaches offer limited support for tasks that demand more advanced visual reasoning skills, such as visual manipulation or imagination.

Multimodal Reasoning Benchmarks Most reasoning benchmarks to date are purely text-based (e.g., [13, 27, 32, 64, 65]). The growing demand for measuring multimodal reasoning has driven the development of multimodal reasoning benchmarks across diverse domains (e.g., [8, 11, 38, 40, 47, 70, 79, 80]). In particular, recent efforts have targeted spatial and relational reasoning [4, 58] and college-level reasoning that requires domain knowledge [83]. Nonetheless, a persistent shortcoming remains: many purported multimodal benchmarks contain redundancy between text and images, allowing models to shortcut through language reasoning alone. For example, Zhang et al. [87] observe that textual elements in math benchmarks often describe the visual content in detail, reducing the need for integrated reasoning across modalities. In light of this, MMMU-Pro [84] incorporates a filtering pipeline to better evaluate true multimodal reasoning. In this work, we further refine such approaches by curating a benchmark that

Statistic	Number
Total questions	2,788
- Multiple-choice questions	2,002 (72%)
- Free-form questions	786 (28%)
- Questions with answers	2,788 (100%)
- Questions newly added	1,796 (64%)
Image in the question	2,599 (93%)
Image in the option(s)	195 (7%)
Problems with multiple images	298 (10%)

Table 1. Key statistics of EMMA.

focuses explicitly on tasks requiring strong visual reasoning. Unlike existing benchmarks, our test suite emphasizes multimodal reasoning challenges that are difficult to solve with text-based reasoning and a single visual pass.

3. The EMMA Benchmark

3.1. Overview of EMMA

We introduce EMMA, an Enhanced MultiModal ReAsOning Benchmark. EMMA is composed of 2,788 problems, of which 1,796 are newly constructed, across four domains: math, physics, chemistry, and coding. The key statistics of EMMA are summarized in Table 1, and its composition is presented in Figure 3.

To provide fine-grained insights into how MLLMs might fail in multimodal reasoning, we assign labels to each problem in our benchmark. These labels are either created by domain experts or assigned by GPT-4o and subsequently verified by experts. As shown in Figure 2, questions in EMMA assess a wide array of multimodal reasoning skills. For example, the pattern inference problem in math challenges models to identify and generalize visual patterns; the visual decomposition simulation problem in physics requires graphically decomposing forces to determine resultant effects; the reaction simulation problem in chemistry demands precise interpretation and simulation of electron movement; the 3D visualization problem in coding¹ evaluates spatial imagination by requiring models to associate function calls with their corresponding 3D representations.

In addition, to ease the burden of evaluation, all questions in EMMA are in either multiple-choice or open-ended formats with short, easily checkable ground truth answers, obviating the need for using MLLMs as judges [61, 75, 88].

3.2. Data Curation

As discussed in Section 2, most existing multimodal reasoning benchmarks likely contain many problems that primarily measure text-based reasoning. To address this, we employ a two-step approach to constructing EMMA (Fig-

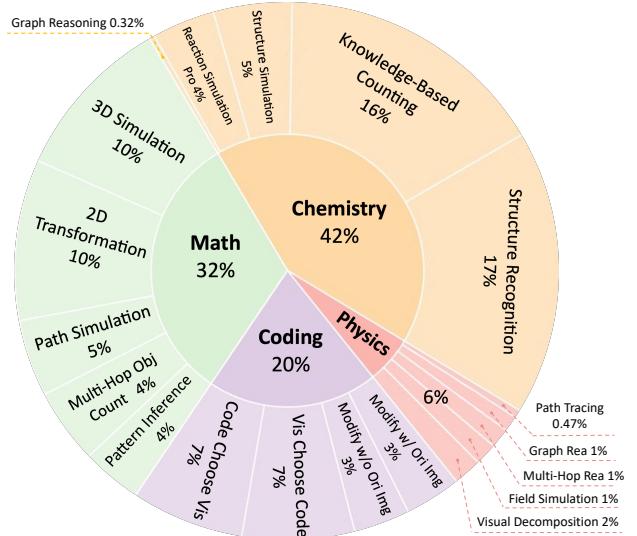


Figure 3. Composition of EMMA. EMMA comprises 2,788 questions across four subjects: math, physics, chemistry, and coding. Within each subject, we further provide fine-grained labels for each question based on the specific skills it measures.

ure 4). First, we source problems from existing multimodal reasoning benchmarks and apply rigorous filtering to exclude those solvable through text-based reasoning and a single visual pass. Next, we categorize the remaining problems for each subject into fine-grained multimodal reasoning skill taxonomies and manually collect more samples aligned with these taxonomies to expand our dataset.

Filtering Mechanisms To filter for questions that require multimodal reasoning, Yue et al. [84] provide only the text from multimodal reasoning questions to LLMs and discard questions that can be correctly answered this way. Nonetheless, some of the remaining questions may still not *truly* measure visual reasoning, as a single pass of visual perception and language understanding may suffice to answer them. We extend [84] one step further (illustrated in Figure 4): we first caption the images in multimodal reasoning questions using GPT-4o and then pass both the original text and our generated captions to MLLMs, filtering out questions that can be answered under this condition. Specifically, for each candidate question, we query Llama-3-70B-Instruct [19], GPT-4o, and Qwen2-72B-Instruct [77] ten times; if any model answers a question correctly at least five times, we discard it. This more stringent filtering ensures that the remaining questions require models to engage deeply with visual information. We introduce the data collection process for each project in detail below.

Math We first apply the filtering pipeline to Math-Vision [70] and MathVista [47], and then manually inspect the remaining set and craft a taxonomy consisting of five categories with a strong focus on multimodal reasoning, including 3D Simulation, 2D Transformation, Path Tracing,

¹Different from the other subjects in EMMA, coding questions can be assigned more than one category since our visualizations tend to employ multiple advanced techniques.

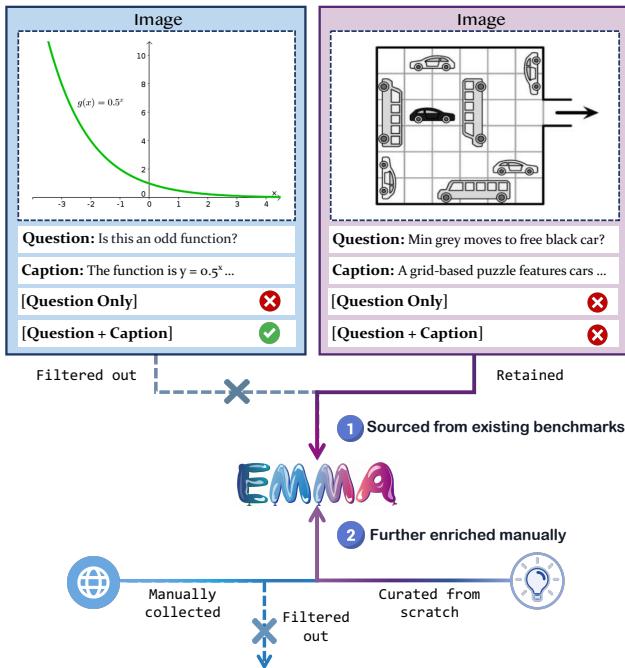


Figure 4. Data curation process for EMMA. We first apply a filtering pipeline to existing benchmarks, discarding problems that MLLMs can solve given only the text and image captions. Next, we categorize the remaining problems based on the skills they assess and manually collect or create additional problems to expand our dataset within these taxonomies.

Multi-hop Object Counting, and Pattern Inference. Next, we use GPT-4o to categorize all questions based on this taxonomy, followed by a manual verification. In addition, we supplement our benchmark with additional pattern inference questions from RAVEN [85], which inherently require multi-hop visual reasoning. This process results in a total of 892 math questions. This process ultimately results in

Physics We apply the filtering pipeline to multimodal physics problems in OlympiadBench [24], EXAMS-V [15], and MMMU [83], which yields only 80 problems. In addition, we manually collect more problems online from Learn AP Physics [55] and Khan Academy [2] and filter them, resulting in 76 more new problems. Through manual labeling, we verify that these problems span a wide range of topics, including 3D Field Simulation, Graph Reasoning, and Path Tracing. We note that despite our best efforts, multimodal physics problems meeting our criteria are difficult to source and construct.

Chemistry After filtering the chemistry portion of EXAMS-V [15] and MMMU [83], we are left with only 20 problems. These questions mostly involve reasoning about molecular formulas, which inspires us to manually construct a novel test suite on organic chemistry. Based on molecular images in SMiCRM [34], we analyze their chemical properties with RDKit [1], a computational chemistry toolkit, and develop 746 novel questions over

chemical structure recognition, bond counting, and structure simulation. In addition, we draw from the collection of chemical reactions in [36] and collaborate with PhD students in chemistry to annotate reaction outcomes, contributing another 210 questions on reaction simulation.

Coding In addition to STEM questions, we design four types of coding tasks to capture various real-world use cases of MLLMs for visualization creation. For instance, to evaluate the ability to reproduce a visualization, we construct “Vis Choose Code” questions where models select the code that generates a target chart. Since previous data visualization benchmarks all rely on using MLLMs as judges, we do not source from existing benchmarks, but manually construct all coding questions from scratch through a three-stage process. First, we identify “seed visualizations” that employ advanced visualization techniques from CharXiv [72], the matplotlib example gallery [68] (following Wu et al. [75]), and our prior experience. Next, we generate four variations for each seed visualization to form a “set” by introducing design variations (e.g., changes in spine configuration, line style, and axis scaling) either manually or through prompting MLLMs, with post-hoc manual verification. We also provide these design variations as labels for each problem. Finally, we construct four types of questions using these visualization sets. In addition to “Vis Choose Code”, we have “Code Choose Vis” questions that require models to identify the chart produced by a given code snippet. “Modify” questions, inspired by visualization debugging scenarios, provide a target chart, a code snippet, and the chart generated by the snippet, requiring models to select the modification needed to transform the code into the target chart. Similarly, “Modify without Original Image” presents the same task but excludes the image generated by the initial code snippet. In the end, our curation results in 564 multiple-choice coding questions in total.

3.3. Comparison with Existing Benchmarks

Our enhanced data filtering pipeline ensures that EMMA focuses on questions requiring in-depth multimodal reasoning, i.e., those that cannot be solved solely using text-based reasoning or a single visual pass. While MMMU-Pro [84] removes questions solvable through their text portion alone, it may still retain problems for which visual reasoning is inessential. In contrast, EMMA applies a stricter filtering criterion, discarding questions solvable with text and image captions. For instance, the left example in Figure 4 (adapted from MathVista) asks whether a depicted function is even or odd. Although unsolvable without the image, the problem can be shortcut by extracting the function’s text expression embedded in the image. In this case, the role of vision is more aligned with visual perception than with visual reasoning. By eliminating such problems, which MMMU-Pro’s approach would retain, EMMA better evaluates the

	CoT	EMMA					EMMA-mini				
		Math (892)	Phys. (156)	Chem. (1,176)	Coding (564)	Overall (2,788)	Math (100)	Phys. (100)	Chem. (100)	Coding (100)	Overall (400)
Random choice	—	14.01	25.64	16.50	25.71	18.08	13.00	23.00	27.00	28.00	22.75
Human Expert	—	—	—	—	—	—	75.00	64.50	86.00	85.50	77.75
Claude 3.5 Sonnet	✗	25.34	33.97	40.90	38.65	35.08	23.00	34.00	44.00	35.00	34.00
Gemini 2.0 Flash	✗	23.88	38.46	36.31	42.02	33.61	20.00	40.00	36.00	41.00	34.25
GPT-4o	✗	27.24	38.46	31.89	40.07	32.42	30.00	38.00	33.00	40.00	35.25
Qwen2-VL-72B-Instruct	✗	33.07	42.31	32.06	34.57	33.46	38.00	40.00	34.00	37.00	37.25
LLaVA-Onevision-72B	✗	27.69	35.90	25.26	28.72	27.33	25.00	32.00	24.00	28.00	27.25
InternVL2-Llama3-76B	✗	25.11	22.44	24.06	27.84	25.07	31.00	22.00	21.00	28.00	25.50
InternVL2.5-78B	✗	31.39	38.46	35.20	31.91	33.50	30.00	40.00	38.00	33.00	35.25
Claude 3.5 Sonnet	✓	29.37	41.03	41.07	40.60	37.23 (↑ 2.15)	30.00	38.00	41.00	39.00	37.00 (↑ 3.00)
Gemini 2.0 Flash	✓	25.90	38.46	24.66	40.96	29.12 (↓ 4.48)	24.00	41.00	36.00	44.00	36.25 (↑ 2.00)
GPT-4o	✓	25.56	43.59	33.67	39.01	32.71 (↑ 0.29)	27.00	44.00	35.00	38.00	36.00 (↑ 0.75)
Qwen2-VL-72B-Instruct	✓	27.69	34.62	24.57	29.43	27.12 (↓ 6.35)	35.00	34.00	32.00	23.00	31.00 (↓ 6.25)
LLaVA-Onevision-72B	✓	22.42	15.38	22.70	30.67	23.82 (↓ 3.52)	23.00	26.00	23.00	29.00	25.25 (↓ 2.00)
InternVL2-Llama3-76B	✓	22.20	32.05	19.73	30.32	23.35 (↓ 1.72)	27.00	33.00	21.00	32.00	28.25 (↑ 2.75)
InternVL2.5-78B	✓	25.56	39.74	27.47	25.18	27.08 (↓ 6.42)	31.00	36.00	24.00	19.00	27.50 (↓ 7.75)
Gemini 2.0 Flash Thinking	—	31.61	56.41	37.93	43.44	38.06	35.00	57.00	41.00	41.00	43.50
o1	—	—	—	—	—	—	41.00	49.00	40.00	53.00	45.75

Table 2. Evaluation results of state-of-the-art MLLMs, which are outperformed by human experts with wide margins. The highest model performance in each column is highlighted in green, and the second-highest is highlighted in blue. Performance improvements from CoT are indicated with upward green arrows, while reductions are marked with downward red arrows.

multimodal reasoning capabilities of models.

We also contribute 1,796 novel multimodal reasoning problems across physics, chemistry, and coding. After filtering physics and chemistry problems from all relevant benchmarks to our knowledge (e.g., [15, 24, 83]), only 100 remain. We expand this to 1,332 in EMMA by manually sourcing additional data and hiring domain experts. For coding, EMMA is the first benchmark to systematically evaluate data visualization skills using a multiple-choice format, enabling a standardized assessment and obviating the need for MLLMs as judges. Moreover, through meticulous manual labeling or verification, we provide fine-grained labels for each question (Figure 2), categorizing them based on the specific skills they assess. These labels enable a detailed analysis of MLLM performance, as we demonstrate in Section 5.2.

4. Experiments

4.1. Evaluation Settings

Data Split To create a more balanced subset of EMMA, we randomly sample 400 questions (100 per subject) from the benchmark, hereafter referred to as EMMA-mini. Within each subject, we aim for equal representation across categories to the extent possible.

Human Performance To estimate expert-level performance on EMMA-mini, we hire two human experts per subject and report their average score. This score serves as a

baseline contextualizing model performance.

Models We evaluate nine state-of-the-art MLLMs under the zero-shot setting, including four open-source models (Qwen2-VL (72B) [71], LLaVA-Onevision (72B) [35], InternVL2 (76B) [69], and InternVL2.5 (78B) [10]) and five proprietary ones (GPT-4o [52], Claude 3.5 Sonnet [5], Gemini 2.0 Flash [17], Gemini 2.0 Flash Thinking [16], and o1 [54]). Due to rate limits, we report o1 performance on EMMA-mini only. All other models are evaluated on the entire benchmark.

Prompting Strategies For all models except o1 and Gemini 2.0 Flash Thinking, we test two prompting strategies: (1) *Direct* prompting, which instructs models to output the answers without reasoning steps; and (2) *Chain-of-Thought (CoT)* prompting [73], where we prompt models to “think step-by-step” and output responses in a structured format.

4.2. Main Results

Table 2 compares the performance of different MLLMs and prompting strategies.

Are MLLMs Multimodal Reasoners? Table 2 demonstrates that all models perform suboptimally across the subjects in EMMA. On EMMA-mini, the best-performing model, o1, achieves an accuracy of 45.75%, trailing human experts by 32%. At the lower end, LLaVA-OneVision-72B scores only 25.25%, barely surpassing random choice by 2.5%. Drilling down into subjects, the best models show

Model	Method	Reward Model	N=1	N=2	N=4	N=8	N=16
GPT-4o	Majority Voting	—	—	37.25	36.25	38.25	
	BoN	GPT-4o (Self)	36.00	35.50	35.75	36.75	—
	BoN	Gemini Flash Thinking		40.75	36.25	36.5	—
	Tournament	Gemini Flash Thinking		40.75	39.25	41.25	35.25
	Pass@N	—		45.00	53.25	65.75	74.00
Gemini 2.0 Flash	Majority Voting	—	—	37.75	39.25	39.75	
	BoN	Gemini Flash (Self)	36.25	38.25	36.50	36.00	—
	BoN	Gemini Flash Thinking		36.75	37.00	40.25	—
	Tournament	Gemini Flash Thinking		36.75	37.25	40.75	38.75
	Pass@N	—		45.25	56.25	64.50	75.00
Gemini 2.0 Flash Thinking	Majority Voting	—	—	48.00	49.00	50.75	
	Tournament	Gemini Flash Thinking (Self)	43.50	45.50	47.25	47.25	48.00
	Pass@N	—		53.75	64.50	71.50	81.50
o1	—	—	45.75	—	—	—	—

Table 3. Results of different test-time scaling strategies on EMMA-mini. We also include Pass@N accuracies as upper bounds to scaling performance. While test-time scaling tends to improve model accuracy, they do not help models achieve near-human-level performance. Overall, Gemini 2.0 Flash Thinking is the best reward model. It also benefits the most from test-time scaling, gaining 7.25% in accuracy with majority voting at N=16 over N=1.

the smallest gap with human performance on physics, with Gemini 2.0 Flash Thinking scoring 7.5% lower than human experts. This smaller gap may reflect the inherent difficulty of physics problems, leading human experts to achieve a score of 64.5%. For other subjects, however, **the best-performing models lag significantly behind human experts, with gaps of 34%, 42%, and 32.5% in math, chemistry, and coding, respectively.** These results underscore the limitations of current MLLMs in addressing complex multimodal reasoning tasks.

On the full EMMA benchmark, closed-source models generally outperform open-source ones, particularly with CoT prompting. Across all subjects, Qwen2-VL-72B-Instruct is the only open-source model to place in the top two for any subject. Gemini 2.0 Flash Thinking scores best overall, ranking among the top two models in three out of four subjects, further highlighting the advantages of optimizing models for reasoning by training them to generate thought processes over traditional MLLM paradigms (note that o1 is not evaluated on the full set due to rate limits). In particular, Gemini 2.0 Flash Thinking performs exceptionally well in physics, leading by nearly 13% over the second-best model, GPT-4o. In contrast, Claude 3.5 Sonnet excels in chemistry, surpassing the next best model, Gemini 2.0 Flash Thinking, by over 2%. These results suggest interesting comparative strengths between models, which may result from differences in training data.

Does CoT help? We observe divergent tendencies in the effectiveness of CoT prompting across both closed- and open-source models. We exclude o1 and Gemini 2.0 Flash Thinking from this analysis, as these models inherently generate CoT as part of their responses. Under direct prompting, accuracies achieved by the best open-source models are well within 2% of Claude 3.5 Sonnet (the best closed-source model). However, the gap widens significantly under CoT

prompting, with the best open-source model underperforming by almost 10%. Comparing each model’s performance with and without CoT on EMMA and EMMA-mini, **CoT prompting generally improves performance for closed-source models, while reduces performance for open-source models.** Notably, Qwen2-VL-72B-Instruct and InternVL2.5-78B, the top two open-source models overall under direct prompting, suffer decreases of over 6% in accuracies on both EMMA and EMMA-mini. While some tasks might not benefit significantly from textual CoT, we hypothesize that this divergence arises because open-source models fail to fully leverage the potential of language to assist in multimodal reasoning tasks where language could be helpful. We elaborate on this hypothesis in detail in Section 5.2.

4.3. Results with Test-Time Compute Scaling

In this section, we test three test-time compute scaling methods [54, 62, 78] on EMMA-mini: majority voting, Best-of-N selection, and Tournament-Style selection. Both Best-of-N and Tournament-Style selection require a reward model to select the best response among multiple candidates. We use CoT prompting to generate the candidate responses, so that the reward model has enough context to score the responses. For each test-time scaling method, we experiment with N = 1, 2, 4, 8, and 16, as long as the context length of the reward model allows.

- **Majority Voting:** Majority voting selects the most frequent response among batches of N candidate responses. When there is a tie, we randomly choose one among the most frequent answers.
- **Best-of-N:** Best-of-N selection [13, 41] selects the highest-scoring response according to a reward model. We explore two configurations: using the base model itself or a stronger reasoning model (*e.g.* Gemini 2.0 Flash

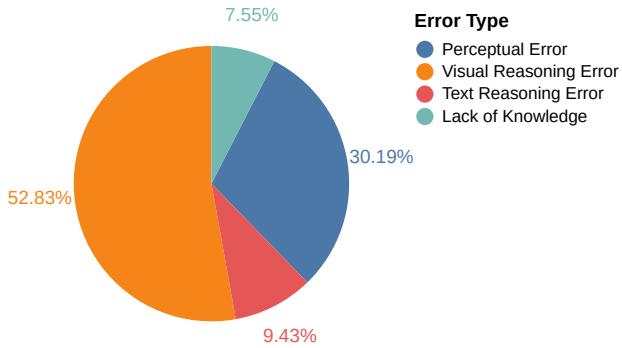


Figure 5. Distribution of error types made by o1 on the math and coding portions of EMMA-mini. The majority of errors arise in visual reasoning.

Thinking) as the reward model.

- **Tournament-Style Selection:** In Tournament-Style Selection [53, 63], responses are paired in matches, and the winners progress through successive rounds until a final selection is made.² We use the best-performing reward model identified in the Best-of-N experiments, which is Gemini 2.0 Flash Thinking.

Does test-time compute scaling help? Tables 3 presents results of test-time compute scaling methods on top of GPT-4o, Gemini 2.0 Flash, and Gemini 2.0 Flash Thinking.

Overall, test-time compute scaling improves model performance, but it fails to close the gap to human expert performance. The highest accuracy improvements are 5.25% for GPT-4o, 4.5% for Gemini 2.0 Flash, and 7.5% for Gemini 2.0 Flash Thinking. Notably, without test-time scaling ($N=1$), Gemini 2.0 Flash Thinking’s accuracy is 2.25% lower than that of o1, but it overtakes o1 by 5% with majority voting at $N=16$. Nonetheless, its best performance still lags human performance by 27%.

We also observe distinct patterns in test-time compute scaling performance across different models. While scaling beyond $N=8$ for GPT-4o and Gemini 2.0 Flash leads to performance degradation, Gemini 2.0 Flash Thinking continues to benefit incrementally from additional test-time compute, at least up to $N=16$. In fact, stronger base models also achieve higher Pass@N accuracy: Gemini 2.0 Flash Thinking’s Pass@N consistently surpasses those of the other two models by around 7%, suggesting that a stronger base reasoner is more likely to cover the correct response when given multiple attempts. In sum, these results suggest that using a stronger model as the base model raises the upper bound for test-time scaling.

Comparing scaling strategies for each model, we find that GPT-4o and Gemini 2.0 Flash achieve their greatest improvements when Gemini 2.0 Flash Thinking is used as the reward model. Additionally, tournament-style selection consistently outperforms Best-of-N (BoN) selection. These

X
Physics: 3D Field Simulation

Question: A long bar magnet is placed above a current loop oriented as shown. In which direction will the North pole of the bar magnet feel a force due to the current loop?
a. +x b. -x c. +y d. -y e. The bar magnet will feel no force due to the current loop.

o1: A current loop behaves like a small bar magnet whose magnetic-dipole moment points according to the Right-Hand Rule.
From the figure, the current goes around the loop in such a way that if you curl your right-hand fingers in the direction of the current, **your thumb points up (the +y direction). Consequently, the loop acts like a magnet whose “north side is on top (at +y).** Hence the force on the north pole is c. +y.

Error Analysis:
The error occurred because of **limited visual spatial imagination**, leading to the wrong conclusion about the direction of the thumb when applying the Right-Hand Rule. Additionally, the 3D spatial nature of the **magnetic field** is not **properly simulated**, leading to a faulty reasoning of the “north” pole’s position.

Human Reasoning (Ground Truth):
The current loop will produce a magnetic field B according to the Right-Hand Rule as shown below.

Then, as shown on the left, B points out from the middle of the current loop I , directed in the $+x$ direction at the location where the bar magnet is located. This magnetic field, pointing to the right, will have the effect of repelling the North pole of the bar magnet, forcing the bar magnet to the right, in the $+x$ direction. The correct answer is a.

Figure 6. A representative example of visual reasoning error. While o1 recognizes that solving this problem involves the Right-Hand Rule, it misapplies the rule due to limited visual spatial simulation skills, resulting in an incorrect conclusion about the direction of the magnetic field.

results suggest that employing a stronger model as the reward model enables less capable models to achieve better results, particularly when the reward model can make fine-grained decisions involving a couple candidate responses each time. This is intuitive, as evaluating responses also requires reasoning.

On the other hand, we find that self-reward modeling tends to be perform suboptimally. Even using Gemini 2.0 Flash Thinking for self-reward modeling yields performance consistently below that of majority voting. We conjecture that self-reward modeling may be less effective because the model’s evaluation criteria may be disrupted by its own generation patterns, making it less sensitive to differences in the reasoning of the generated responses compared to an independent reward model.

5. Error Analysis

5.1. Error Distribution

We present an analysis of the errors made by o1 on the math and coding portions of EMMA-mini. In total, o1 incorrectly answers 59 math questions and 47 coding questions. Figure 5 categorizes these errors into four types. **Perceptual errors**, such as misinterpreting visual information, account

²Tournament-style selection with $N=2$ is equivalent to Best-of-2.

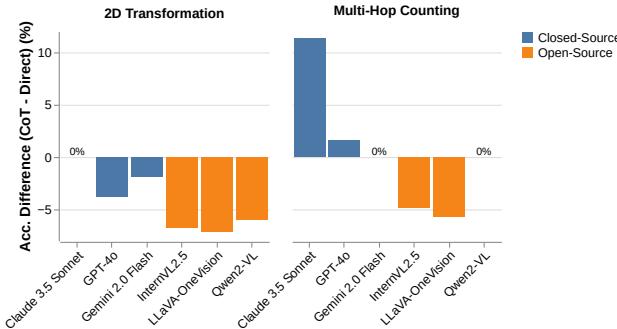


Figure 7. Performance differences with and without CoT for two math tasks: 2D Transformation and Multi-Hop Object Counting. The observed discrepancies in CoT’s effectiveness suggest that its impact might depend on the extent of visual reasoning required to solve the task.

for 30.19% of all errors. **Lack of knowledge errors**, including mistakes related to API usage, contribute 7.55%. **Visual reasoning errors**, such as failures to simulate 3D processes, constitute the largest category at 52.83%. Finally, **textual reasoning errors**, including calculation mistakes or logical missteps, represent 9.43%. The predominance of visual reasoning errors underscores the limitations of current models in addressing complex visual reasoning tasks. Figure 6 illustrates a representative case: while o1 correctly identifies that the problem calls for the application of the Right-Hand Rule, it fails to simulate where the thumb would point to when the right-hand fingers are curled in the direction of the current. We provide more error cases in the Appendix.

5.2. The Effects of Textual CoT

Results in Table 2 reveals a notable discrepancy: while CoT prompting improves reasoning in closed-source MLLMs, it tends to reduce performance in open-source models. Although the cause remains unclear without direct access to training data or methods, we analyze error rate per skill category to propose a conjecture. Figure 7 compares accuracy differences between CoT and Direct prompting for six models on two math tasks: 2D Transformation and Multi-Hop Object Counting. On 2D Transformation, all but one model shows reduced performance with CoT; on Multi-Hop Object Counting, while CoT generally helps closed-source models, it makes all but one open-source model perform worse. In fact, we notice that CoT prompting introduces more hallucinations for open-source models.

2D Transformation questions primarily test for visual simulation and spatial imagination, which are difficult to verbalize. The question in Figure 8, for example, calls for spatial imagination beyond the power of language. In contrast, Multi-Hop Object Counting can leverage language to describe the relative positions of objects. Hence, we conjecture that visual-centric tasks, such as 2D Transformation, are poorly suited for textual CoT. In contrast, tasks that ben-

Math: 2D Transformation

Question: Rebecca folds a square piece of paper twice. Then she cuts off one corner as you can see in the diagram. Then she unfolds the paper. What could the paper look like now?

GPT-4o with CoT : To solve this problem, we need to carefully analyze the folding and cutting process and determine the resulting pattern when the paper is unfolded. ✓ Step 1: Understand the folding ✓ Step 2: Understand the cutting process

- Rebecca cuts off one corner of the folded paper. ✓
 - Since the paper is folded into a smaller square, this ‘corner’ is actually a corner of the folded square, which corresponds to multiple corners of the original square. X

Final Answer:\boxed{E} X

Error Analysis: After folding into 1/4, the “corners” of the folded paper are not the original corners. The solution fails to properly simulate the unfolding process.

Direct GPT-4o : B. ✓

Human Reasoning (Ground Truth):

Figure 8. An example where textual CoT is unhelpful for solving the problem with current models. While GPT-4o correctly resolves the problem without CoT, it answers incorrectly with CoT. The thought process demonstrates a superficial association with spatial manipulations and relations rather than genuine visual reasoning.

efit from language-based reasoning, such as Multi-Hop Object Counting, allow models to achieve greater performance gains with textual CoT, as evidenced by the closed-source models.

6. Conclusion

We contribute EMMA, an Enhanced MultiModal reAsoning benchmark. EMMA features multimodal questions requiring advanced cross-modal reasoning, which cannot be solved by independently reasoning within each modality. Evaluation of nine state-of-the-art MLLMs reveals a substantial performance gap compared to human experts on EMMA, with techniques such as Chain-of-Thought prompting and test-time compute scaling offering only marginal gains. EMMA highlights the need for new architectures and training paradigms that can better integrate and reason over diverse modalities. Like any benchmark, EMMA has its limitations, which can be improved in future works. For example, future iterations could enrich the currently underrepresented physics section or expand the chemistry section to incorporate a broader range of chemistry topics. Nonetheless, EMMA sets a new standard for assessing MLLMs on multimodal reasoning.

References

- [1] Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* 2024 plot2codem, 8(31.10):5281, 2013. 5
- [2] Khan Academy. Khan academy. <https://www.khanacademy.org/science/>. 5
- [3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3
- [4] Syeda Nahida Akter, Sangwu Lee, Yingshan Chang, Yonatan Bisk, and Eric Nyberg. Visreas: Complex visual reasoning with unanswerable questions. *arXiv preprint arXiv:2403.10534*, 2024. 2, 3
- [5] Anthropic. Claude 3.5 sonnet. <https://www.anthropic.com/news/clause-3-5-sonnet>. 6
- [6] 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. 2
- [7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1
- [8] Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 3
- [9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3
- [10] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6
- [11] Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. *arXiv preprint arXiv:2412.12932*, 2024. 3
- [12] François Fleuret. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 3
- [13] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 3, 7
- [14] codeforces. Codeforces. <https://codeforces.com/>. 1, 3
- [15] Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*, 2024. 5, 6, 16, 17
- [16] Google Deepmind. Gemini 2.0 flash thinking mode. <https://ai.google.dev/gemini-api/docs/thinking-mode>. 6
- [17] Google Deepmind. Introducing gemini 2.0: our new ai model for the agentic era. <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/>. 6
- [18] Team DeepSeek. Deepseek-r1-lite-preview is now live: unleashing supercharged reasoning power! <https://api-docs.deepseek.com/news/news1120>, 2024. 1, 3
- [19] 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. 4
- [20] Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuvan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. Isobench: Benchmarking multimodal foundation models on isomorphic representations. *arXiv preprint arXiv:2404.01266*, 2024. 17
- [21] Marcus Giaquinto. The epistemology of visual thinking in mathematics, Feb 2020. 1
- [22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2
- [23] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023. 17
- [24] 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. 5, 6
- [25] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 1, 3
- [26] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 1
- [27] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. 3
- [28] Linmei Hu, Duokang Wang, Yiming Pan, Jifan Yu, Yingxia Shao, Chong Feng, and Liqiang Nie. Novachart: A large-scale dataset towards chart understanding and generation of multimodal large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3917–3925, 2024. 17
- [29] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022. 3
- [30] 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
- [31] 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. 2
- [32] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh conference on neural information processing systems*, 2023. 3
- [33] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 18
- [34] Ching Ting Leung, Yufan Chen, and Hanyu Gao. Smicrm: A benchmark dataset of mechanistic molecular images. *arXiv preprint arXiv:2407.18338*, 2024. 5, 16
- [35] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3, 6
- [36] JJ Li. Name reactions. a collection of detailed mechanisms and synthetic applications 4th edition, 2009. 5, 16
- [37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3
- [38] Kaixin Li, Yuchen Tian, Qisheng Hu, Ziyang Luo, and Jing Ma. Mmcode: Evaluating multi-modal code large language models with visually rich programming problems. *arXiv preprint arXiv:2404.09486*, 2024. 3
- [39] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXXI 16*, pages 121–137. Springer, 2020. 3
- [40] Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, Hyeyoung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, et al. Mmsci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In *AI for Accelerated Materials Design-Vienna 2024*, 2024. 3
- [41] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. 7
- [42] Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhui Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table trans-
- lation. *arXiv preprint arXiv:2212.10505*, 2022. 17
- [43] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 3
- [44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3
- [45] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 3
- [46] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3
- [47] 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. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 3, 4, 14, 15, 17, 18
- [48] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 16, 17
- [49] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 2
- [50] 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. 17
- [51] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2020. 17
- [52] OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>. 6
- [53] OpenAI. Introducing chatgpt pro. <https://openai.com/index/introducing-chatgpt-pro/>. 8
- [54] OpenAI. Learning to reason with llms. <https://openai.com/index/learning-to-reason-with-llms/>. 1, 3, 6, 7
- [55] Learn AP Physics. Learn ap physics. <https://www.learnappphysics.com>. 5
- [56] Team Qwen. Qwq: Reflect deeply on the boundaries of the unknown. <https://qwenlm.github.io/blog/qwq-32b-preview/>, 2024. 1, 3
- [57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages

- 8748–8763. PMLR, 2021. 3
- [58] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Krahenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? *arXiv preprint arXiv:2410.06468*, 2024. 3
- [59] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023. 1, 3
- [60] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuban Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024. 3
- [61] Chufan Shi, Cheng Yang, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, et al. Chartmimic: Evaluating lmm’s cross-modal reasoning capability via chart-to-code generation. *arXiv preprint arXiv:2406.09961*, 2024. 4, 17
- [62] 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. 7
- [63] Seonil Son, Ju-Min Oh, Heegon Jin, Cheolhun Jang, Jeong-beom Jeong, and Kuntae Kim. Varco arena: A tournament approach to reference-free benchmarking large language models. *arXiv preprint arXiv:2411.01281*, 2024. 8
- [64] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022. 3
- [65] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022. 3
- [66] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3
- [67] G Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024). URL <https://goo.gle/GeminiV1-5>. 3
- [68] Matplotlib Team. Examples. <https://matplotlib.org/stable/gallery/index.html>. 5
- [69] OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. <https://internvl.github.io/blog/2024-07-02-InternVL-2.0/>. 3, 6
- [70] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024. 2, 3, 4, 17, 18
- [71] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 6
- [72] 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. *arXiv preprint arXiv:2406.18521*, 2024. 5, 16
- [73] 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
- [74] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chauvin, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 18
- [75] Chengyue Wu, Yixiao Ge, Qiushan Guo, Jiahao Wang, Zhixuan Liang, Zeyu Lu, Ying Shan, and Ping Luo. Plot2code: A comprehensive benchmark for evaluating multi-modal large language models in code generation from scientific plots. *arXiv preprint arXiv:2405.07990*, 2024. 4, 5, 16, 17
- [76] Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, et al. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*, 2024. 17
- [77] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 3, 4
- [78] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024. 7, 18
- [79] John Yang, Carlos E. Jimenez, Alex L. Zhang, Kilian Lieret, Joyce Yang, Xindi Wu, Ori Press, Niklas Muenninghoff, Gabriel Synnaeve, Karthik R. Narasimhan, Diyi Yang, Sida I. Wang, and Ofir Press. Swe-bench multimodal: Do ai systems generalize to visual software domains?, 2024. 3
- [80] 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
- [81] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 3
- [82] 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
- [83] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024. 2, 3, 5, 6, 16, 17, 18
- [84] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhua Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 2, 3, 4, 5, 17
- [85] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327, 2019. 5, 15
- [86] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 3
- [87] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024. 2, 3
- [88] Zhehao Zhang, Weicheng Ma, and Soroush Vosoughi. Is gpt-4v (ision) all you need for automating academic data visualization? exploring vision-language models’ capability in reproducing academic charts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8271–8288, 2024. 4, 17
- [89] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 3
- [90] Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. Marco-ol: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*, 2024. 1, 3
- [91] Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*, 2024. 3
- [92] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3