

BENCHMARKING MULTIMODAL MATHEMATICAL REASONING WITH EXPLICIT VISUAL DEPENDENCY

Anonymous authors

Paper under double-blind review

ABSTRACT

Recent advancements in Large Vision-Language Models (LVLMs) have significantly enhanced their ability to integrate visual and linguistic information, achieving near-human proficiency in tasks like object recognition, captioning, and visual question answering. However, current benchmarks typically focus on knowledge-centric evaluations that assess domain-specific expertise, often neglecting the core ability to reason about fundamental mathematical elements and visual concepts. We identify a gap in evaluating elementary-level math problems, which rely on explicit visual dependencies-requiring models to discern, integrate, and reason across multiple images while incorporating commonsense knowledge, all of which are crucial for broader AGI capabilities. To address this gap, we introduce VCBENCH, a comprehensive benchmark for multimodal mathematical reasoning with explicit visual dependencies. VCBENCH includes 1,720 problems across six cognitive domains, featuring 6,697 images (averaging 3.9 per question) to ensure multi-image reasoning. We evaluate 26 state-of-the-art LVLMs on VCBENCH, revealing substantial performance disparities, with even the top models unable to exceed 50% accuracy. Our findings highlight the ongoing challenges in visual-mathematical integration and suggest avenues for future LVLM advancements.

1 INTRODUCTION

Recent advancements in Large Vision-Language Models (LVLMs) Anthropic (2025); Deepmind (2025); OpenAI et al. (2024); Bai et al. (2023) have made significant strides in bridging the gap between visual understanding and language processing. These models have achieved remarkable performance across a range of tasks, demonstrating near-expert human-level proficiency in domains such as object recognition, caption generation, and visual question answering Lin et al. (2015); Agrawal et al. (2016). Among the various domains explored, LVLMs have shown particular promise in tasks that require both visual and linguistic reasoning, making them increasingly relevant for real-world applications.

While many visual mathematics benchmarks, such as MathVista Lu et al. (2023) and MathVision Wang et al. (2024a), focus on knowledge-centric evaluations that assess domain-specific mathematical or geometric expertise, they often fail to evaluate a model’s core ability to perceive and reason about fundamental mathematical elements and visual concepts. Moreover, these knowledge-centric evaluations are easily influenced by the pre-existing knowledge embedded in large language models, which may obscure true reasoning capabilities. To advance towards Artificial General Intelligence (AGI), a more holistic approach to multi-modal reasoning is needed-one that goes beyond task-specific benchmarks and better captures generalizable cognitive abilities.

In this context, we identify a gap in the evaluation of models on elementary-level math problems Cobbe et al. (2021a); Wei et al. (2023b). These problems, typically at the elementary school level, do not require complex mathematical or geometric reasoning but rely heavily on explicit visual dependencies-the ability to discern and integrate visual features across images and understand how different visual elements relate to one another to solve problems. This mirrors the cognitive development of children, who rely on similar skills to solve problems despite not yet possessing advanced reasoning abilities. Understanding and modeling this form of reasoning is crucial, as it represents a fundamental cognitive ability essential for advancing toward broader AGI capabilities.

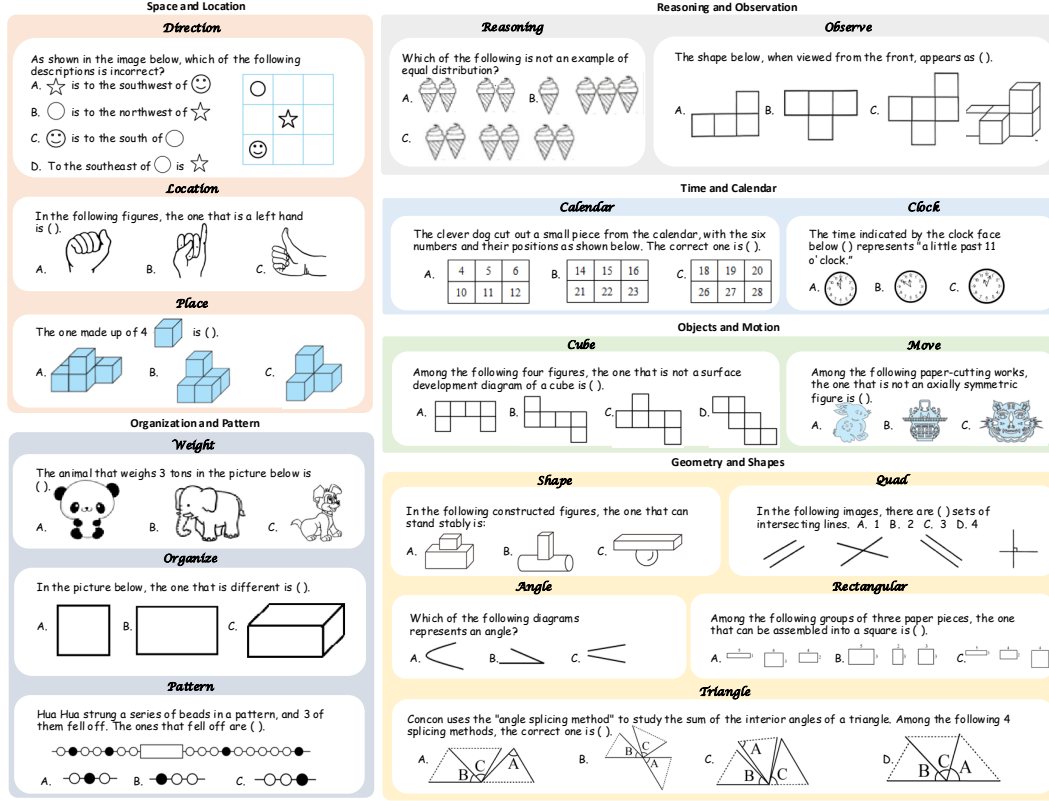


Figure 1: Representative examples from the VCBENCH, showcasing diverse question types and categories including Space and Location (Direction, Location and Place), Reasoning and Observation (Reasoning and Observe), Time and Calendar (Calendar and Clock), Objects and Motion (Cube and Move), Organization and Pattern (Weight, Organize and Pattern), and Geometry and Shapes (Shape, Quad, Angle, Rectangular and Triangle).

To address this gap, we introduce VCBENCH, a comprehensive benchmark designed to assess multimodal mathematical reasoning tasks with explicit visual dependencies. Specifically targeting elementary-level math problems (grades 1–6), VCBENCH focuses on tasks that require reasoning across multiple images to derive solutions. As shown in Figure 1, it covers six key cognitive domains: Time and Calendar, Spatial and Positional Awareness, Geometry and Shapes, Objects and Motion, Reasoning and Observation, and Organization and Patterns. It also evaluates five competencies: temporal reasoning, geometric reasoning, logical reasoning, spatial reasoning, and pattern recognition. These competencies span a broad spectrum, from basic temporal and spatial understanding to more advanced geometric and logical reasoning, providing a thorough evaluation of multimodal model performance. Comprising 1,720 QA pairs and 6,697 images (averaging 3.9 images per question), VCBENCH ensures models must reason across multiple visual inputs, rather than relying on single-image comprehension. With this holistic framework, VCBENCH serves as a valuable resource for advancing research in multimodal mathematical reasoning.

Table 1: Comprehensive Statistics of the VCBENCH Dataset, Including Detailed Break-down of Question-Image Pairs, Image Distribution, and Question Length Metrics.

Examples (Q&A pairs)	1,720
Images	6,697
Avg. images per question	3.9
Avg. question length	136.2
Max. # images in question	18
Min. # images in question	2

In our extensive experimental evaluation, we assessed 26 state-of-the-art LVLs across 17 distinct task categories within VCBENCH. Despite achieving near-perfect accuracy on normal human-level performance, the best-performing visual models were unable to exceed 50% accuracy. Many of these

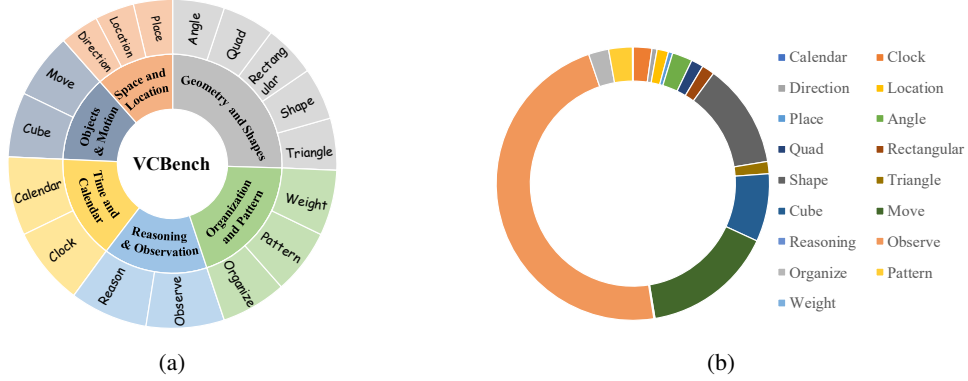


Figure 2: (a) Overview of the VCBENCH dataset structure, highlighting its six main categories and associated subcategories, designed to assess multimodal reasoning capabilities of LVLMs. (b) Distribution of question types in the VCBENCH, illustrating the relative frequency across different visual reasoning subcategories

state-of-the-art models exhibited a notable lack of pattern recognition in images, especially when it came to reasoning tasks that required integrating visual cues across multiple images. Interestingly, we observed that these same tasks could be easily answered by normal human. This highlights a significant gap in current benchmarks, which fail to adequately assess vision-centric mathematical reasoning abilities.

We make several key contributions with VCBENCH:

Unlike existing benchmarks that focus on knowledge-centric evaluations, we emphasize vision-centric assessments. VCBENCH targets problems that do not require specialized knowledge but rely on the common perceptual reasoning of mathematical images and concepts. This approach aligns with the way children learn-first mastering visual reasoning and later acquiring domain-specific knowledge.

VCBENCH is designed around multi-image tasks, with each question containing an average of 3.9 images. This requirement challenges models to explicitly integrate visual cues across multiple images and reason about how they interact, which better reflects real-world scenarios where information is often distributed across multiple visual inputs.

Our benchmark provides a holistic evaluation of various visual reasoning capabilities, such as temporal reasoning, spatial understanding, and pattern recognition. While these tasks may seem simple to children, they represent fundamental reasoning abilities that LVLMs often struggle with. Our experiments demonstrate that tasks considered easy for children-such as identifying time sequences or spatial relationships-prove challenging for state-of-the-art LVLMs, highlighting the gaps in current multimodal reasoning capabilities.

2 RELATED WORK

Large Vision-Language Models. Large Vision-Language Models (LVLMs) have significantly advanced the integration of vision and language, demonstrating strong performance in tasks such as image captioning, visual question answering (VQA), and complex multimodal reasoning Wang et al. (2024b); Wu et al. (2023). Recent developments, such as Gemini-2.0 Deepmind (2025), QVQ Team (2024), and Calude-3.7-Sonnet Anthropic (2025), showcase emergent abilities in cross-modal instruction-following and chain-of-thought reasoning.

Despite these advancements, quantitatively evaluating LVLMs, particularly in visual mathematical reasoning, remains challenging. Existing benchmarks like MathVista Lu et al. (2023), MathBench Liu et al. (2024), and Math-LLMs Liu et al. (2023) typically assess models within narrow domains, such as arithmetic word problems or geometry-based visual environments. Consequently, these benchmarks primarily measure foundational skills like geometric or spatial reasoning, limiting their

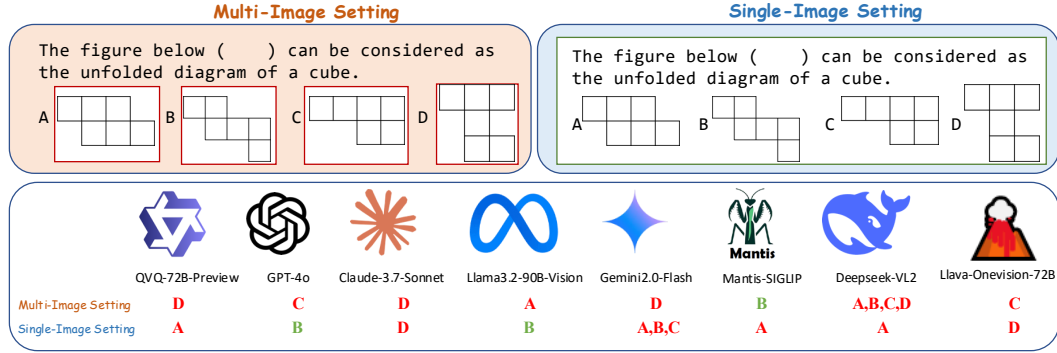


Figure 3: Comparative evaluation of various LVLMs under Multi-Image and Single-Image settings for the same question. The letters (A, B, C, D) indicate models’ predictions, with correct answers marked in green and incorrect answers in red.

capacity to comprehensively evaluate broader cognitive integration and reasoning abilities. To address this limitation, we introduce VCBENCH, a systematic evaluation framework designed to rigorously assess LVLMs performance across diverse multimodal mathematical reasoning tasks with explicit visual dependencies.

Visual Mathematical Reasoning. Mathematical reasoning is a core cognitive ability increasingly explored within the context of LVLMs research Hendrycks et al. (2021); Cobbe et al. (2021b). While earlier benchmarks such as GSM8K Cobbe et al. (2021b) and MATH Hendrycks et al. (2021) primarily focused on text-based mathematical problems, recent research has expanded toward visual mathematical reasoning, incorporating diagrams, charts, and geometry-based problem-solving Wang et al. (2024b); Yang et al. (2024).

Multimodal mathematical reasoning requires LVLMs to integrate visual perception and logical reasoning, presenting a greater challenge compared to purely textual problems. Recent benchmarks like MathVista Lu et al. (2023) and MathGLM-Vision Yang et al. (2024) have advanced evaluation efforts but still suffer from issues including ambiguous annotations, dependency on GPT-based scoring methods, and limited evaluation of generalizable cognitive abilities Yan et al. (2024).

To overcome these challenges, we propose VCBENCH, a comprehensive benchmark explicitly designed for multimodal mathematical reasoning with visual dependencies. VCBENCH encompasses 17 distinct subtasks, systematically assessing foundational cognitive skills such as temporal reasoning, logical reasoning, spatial reasoning, geometric reasoning, and pattern recognition. By standardizing task instructions and employing a multiple-choice evaluation format, VCBENCH provides objective, reproducible evaluations, offering deeper insights into the strengths and limitations of current LVLMs.

3 VCBENCH

3.1 BENCHMARK CONSTRUCTION

For VCBENCH, we employed a systematic approach to collect high-quality multimodal mathematical reasoning problems that explicitly require visual reasoning. We started by examining elementary school mathematics online question banks, manually filtering for problems that contained at least two images. In our manual review, we further excluded problems that contained non-English annotations not inferable from visual cues, were in multiple-choice format, had low-resolution or unclear visuals, relied on region-specific or cultural knowledge, or were ambiguous or not confidently understood by the reviewers. The benchmark prioritizes vision-centric evaluation through perceptual reasoning tasks that avoid specialized knowledge, while simultaneously challenging models to implicitly integrate and synthesize visual information across multiple images - a critical capability for real-world applications where understanding emerges from connecting disparate visual cues. During our selection process, we enforced strict criteria to ensure quality and consistency. We only retained problems with unique, unambiguous answers to facilitate objective evaluation. After the initial collection phase, we utilized large language models to translate all problems into English (the specific prompts used

are available in the Appendix), followed by rigorous human verification to maintain translation accuracy. The human verification process served as a filtering mechanism, where we eliminated problems containing non-English content in images, as well as those with unclear visual elements or ambiguous instructions. This meticulous curation process ensured that our benchmark evaluates genuine reasoning abilities rather than testing models on their capacity to handle poorly defined problems. Through this methodology, we assembled our final collection of problems that encompass various mathematical domains while maintaining consistent quality standards.

3.2 BENCHMARK STATISTICS

VCBENCH comprises a diverse collection of multimodal mathematical reasoning problems, carefully organized into **six** major categories to provide comprehensive coverage of different cognitive dimensions. As shown in Table 1, our benchmark contains 1,720 question-answer pairs featuring a total of 6,697 images. Each question is paired with, on average, 3.9 images; some problems are highly complex and include as many as 18 images, while the minimum per question is 2 images. To systematically evaluate the breadth of reasoning skills, we classified our problems into six major domains, each capturing distinct aspects of mathematical cognition. This domain-specific organization enables a granular assessment of model performance across a diverse set of cognitive abilities, ranging from visual perception and spatial understanding to arithmetic and logical reasoning. Such structured categorization not only facilitates targeted diagnostics of model strengths and weaknesses, but also mirrors the multifaceted nature of human mathematical problem-solving. Furthermore, we deliberately constrained the vocabulary used in VCBench to 2,312 unique words, minimizing confounding effects from linguistic complexity and ensuring that evaluation focuses squarely on reasoning capability. With an average question length of 136.2 characters, each problem remains concise, yet provides sufficient detail and context to support an accurate solution. The six domains are:

- **Time and Calendar:** Problems testing temporal reasoning across two subcategories (*Calendar* and *Clock*) that require understanding time intervals, and calendar-based calculations.
- **Space and Location:** Challenges focused on spatial reasoning (*Direction*, *Location*, and *Place*) that assess understanding of relative positions, directions, and spatial relationships.
- **Geometry and Shapes:** Problems spanning five subcategories (*Angle*, *Quad*, *Rectangular*, *Shape*, and *Triangle*) that test fundamental geometric comprehension from basic shape recognition to more complex property analysis.
- **Objects and Motion:** Tasks in two subcategories (*Cube* and *Move*) that evaluate the understanding of three-dimensional objects and motion transformations.
- **Reasoning and Observation:** Problems in two subcategories (*Reasoning* and *Observe*) designed to test logical reasoning and careful visual observation skills.
- **Organization and Pattern:** Challenges across three subcategories (*Organize*, *Pattern*, and *Weight*) that assess pattern recognition, sequencing, and organizational logic.

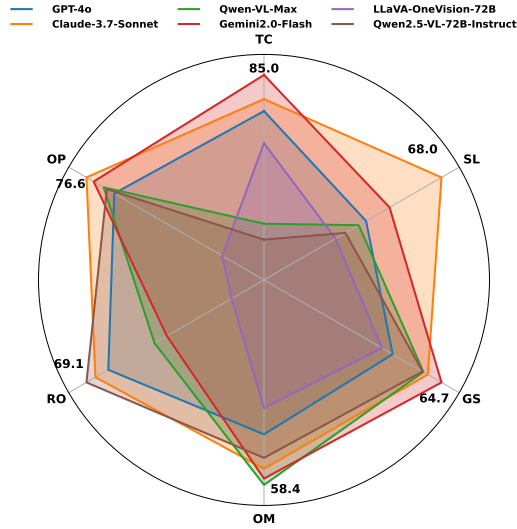


Figure 4: Comparative performance (%) of six various prominent LVLMs across six categories: Time and Calendar (TC), Space and Location (SL), Geometry and Shapes (GS), Objects and Motion (OM), Reasoning and Observation (RO), and Organization and Pattern (OP).

Table 2: Performance of various vision-language models (Close-Source, Open-Source, and Math Specialist categories) on a Multi-image setting across multiple tasks, including Time and Calendar, Space and Location, Geometry and Shapes, Objects and Motion, Reasoning and Observation, and Organization and Pattern.

Models	Time and Calendar		Space and Location			Geometry and Shapes					Objects and Motion		Reasoning and Observation		Organization and Pattern			Avg.
	Calendar	Clock	Direction	Location	Place	Angle	Quad	Rectangular	Shape	Triangle	Cube	Move	Reasoning	Observe	Organize	Pattern	Weight	
Random Guess	33.33	32.78	25.00	29.81	33.33	31.00	27.63	29.17	31.84	29.01	28.37	29.35	33.33	29.41	30.17	31.32	33.33	29.83
Human	100.00	96.00	100.00	93.85	96.67	95.60	96.84	95.00	94.02	94.07	97.67	94.63	100.00	100.00	93.20	95.52	100.00	93.30
Close-Source Models																		
GPT-4o-mini OpenAI et al. (2024)	80.00	60.66	0.00	38.46	53.33	38.40	21.05	33.57	37.99	55.56	32.19	38.24	0.00	28.68	60.00	41.38	100.00	34.88
GPT-4o OpenAI et al. (2024)	40.00	20.00	0.00	30.77	66.67	46.00	57.89	28.57	50.22	51.85	37.67	50.37	90.00	31.27	76.00	37.93	80.00	40.29
Claude-3.7-Sonnet Anthropic (2025)	100.00	50.00	100.00	53.85	50.00	58.00	63.16	57.14	60.70	59.26	40.41	67.28	100.00	31.27	76.40	53.45	100.00	46.63
Qwen-VL-Max Bai et al. (2023)	0.00	46.67	0.00	42.31	66.67	74.00	52.63	42.86	54.15	66.67	56.16	60.66	50.00	35.27	68.00	39.66	100.00	47.03
Gemini2.0-Flash DeepMind (2025)	100.00	70.00	20.00	57.69	66.67	70.00	68.42	53.57	61.14	70.37	44.52	68.75	40.00	35.53	74.00	46.55	100.00	49.77
Open-Source Models																		
Emu2-Chat Sun et al. (2024b)	0.00	13.33	0.00	3.85	0.00	4.00	10.53	10.71	12.66	3.70	8.90	6.99	0.00	3.62	0.00	3.45	20.00	6.05
Idelfe3-8B Laurencin et al. (2024)	0.00	3.33	20.00	15.38	33.33	11.60	10.53	17.86	23.14	3.70	9.59	16.91	0.00	9.69	8.40	15.52	0.00	12.91
DeepSeek-VL2 Wu et al. (2024)	0.00	23.33	0.00	23.08	16.67	14.00	10.53	14.29	29.69	14.81	6.85	18.38	10.00	9.43	44.00	20.69	0.00	15.47
Phi-3.5-vision-instruct Abidin et al. (2024)	0.00	23.33	100.00	19.23	66.67	16.00	15.79	28.57	27.07	33.33	22.60	22.79	0.00	21.32	34.00	12.07	0.00	22.73
InternVL2.5-8B Chen et al. (2024)	0.00	33.33	0.00	34.62	50.00	34.00	31.58	50.00	55.81	37.04	23.29	25.74	0.00	18.99	38.00	6.90	0.00	24.71
LLaMA-3.2-90B-Vision-Instruct AI (2024)	20.00	24.67	100.00	11.54	16.67	26.40	31.58	32.14	26.20	22.22	27.40	25.37	0.00	25.58	12.00	29.31	20.00	25.41
Qwen2.5-VL-72B-Instruct Bai et al. (2025)	100.00	13.33	0.00	19.23	50.00	20.00	31.58	25.00	30.13	51.85	32.19	40.81	0.00	25.19	30.00	27.59	0.00	29.24
Mistral-CLIP Jiang et al. (2024)	0.00	30.00	80.00	50.00	66.67	14.00	15.79	35.71	38.43	37.04	19.86	32.35	40.00	28.04	52.40	22.41	100.00	30.23
Mistral-Small-3.1-24B-Instruct Mistral (2025)	20.00	40.00	0.00	30.77	30.00	38.00	31.58	35.71	29.26	51.85	30.82	31.62	50.00	29.59	38.00	34.48	20.00	31.34
Kimi-VL-A3B-Thinking Team et al. (2025b)	100.00	26.67	100.00	30.77	33.33	48.00	36.84	28.57	49.78	33.33	30.14	41.91	0.00	25.32	68.00	27.59	100.00	34.13
LLaVA-Interleave-7B Li et al. (2024b)	0.00	36.67	20.00	19.23	83.33	46.00	26.32	57.14	39.74	29.63	30.82	33.46	50.00	33.46	62.00	31.03	100.00	35.47
LLaVA-OneVision-7B Li et al. (2024a)	0.00	40.00	0.00	11.54	83.33	44.00	36.84	32.14	37.99	48.15	30.82	46.69	50.00	32.56	58.00	29.31	100.00	36.63
Kimi-VL-A3B-Instruct Team et al. (2025b)	0.00	46.67	0.00	30.77	83.33	44.00	47.37	39.29	43.23	33.33	34.93	44.49	50.00	31.31	58.00	36.21	0.00	37.33
InternVL2.5-78B Chen et al. (2024)	20.00	31.33	100.00	42.31	66.67	54.00	47.37	46.43	53.28	55.56	33.56	40.44	50.00	28.04	76.00	31.03	100.00	37.56
Gemma3-27B-It Team et al. (2025a)	100.00	50.00	0.00	38.46	83.33	48.40	31.58	25.00	41.92	40.74	32.88	47.79	50.00	32.82	54.00	31.03	80.00	38.02
QVQ-72B-Preview Team (2024)	100.00	43.33	0.00	46.15	83.33	58.00	42.11	46.43	44.10	62.96	36.30	48.16	50.00	28.55	78.00	48.28	100.00	39.13
LLaVA-OneVision-72B Li et al. (2024a)	0.00	33.33	0.00	33.33	61.20	37.89	57.14	60.70	51.85	41.10	60.29	60.66	100.00	38.24	82.00	41.38	80.00	47.67
Qwen2.5-VL-72B-Instruct Bai et al. (2025)	0.00	40.67	0.00	53.85	50.00	68.00	68.42	53.57	55.02	74.07	58.22	60.66	60.00	35.53	76.00	43.10	100.00	48.08

4 EXPERIMENT

4.1 MAIN RESULTS

There are a total of 17 subtasks for the evaluation from the perspectives of Temporal Reasoning, Spatial Reasoning, Geometric Reasoning, Logical Reasoning, and Pattern Recognition abilities over 21 VLMs. Table 2 provides detailed evaluation results across six visual reasoning tasks. Human performance is near-perfect with an average score of 93.30, while random guessing achieves only 29.83, which emphasizes that these tasks, though inherently solvable by humans, pose substantial challenges to current AI systems.

Figure 4 shows the comparative performance of six various prominent LVLMs across six tasks. Their relative strengths lie particularly in tasks requiring spatial reasoning and observational interpretation, suggesting these models have better internal representations or more effective cross-modal alignment between visual and linguistic information. However, despite these advancements, even these top-performing closed-source models exhibit notable shortcomings relative to humans, particularly in high-complexity reasoning scenarios (e.g., Geometry and Objects and Motion), reflecting an ongoing gap in advanced spatial reasoning, logical reasoning and pattern recognition capabilities.

Open-source models present an even more heterogeneous and generally lower performance landscape, indicative of diverse model architectures, varying degrees of multi-modal integration sophistication, and potentially inconsistent data quality or quantity during training. For example, large open-source models, including Qwen2.5-VL-72B-Instruct (48.08%) and LLaVA-OneVision-72B (47.67%), demonstrate performance comparable to mid-tier closed-source models. Their comparatively stronger results, particularly in Geometry and Shapes and Organization and Pattern tasks, suggest these models benefit from scale and possibly more sophisticated visual encoders or pre-training strategies. However, they still encounter substantial difficulties in tasks requiring nuanced observation or reasoning about motion and object interactions, highlighting remaining challenges in achieving cognitive visual reasoning. The variability across different tasks, especially pronounced in Objects and Motion and Reasoning and Observation categories, points toward crucial areas requiring further research: enhancing temporal reasoning, improving dynamic visual understanding, and strengthening the integration of geometric and spatial cognition into visual-language models.

4.2 EVALUATION IN SINGLE-IMAGE SETTING

The evaluation is also conducted in a single-image setting for comparison. In single-image setting, we integrate visual and textual elements into a cohesive layout as shown in Figure 3. If a model performs well in single-image but poorly in multi-image, it suggests the model lacks compositional reasoning ability to link separate inputs.

The results in Table 3 reveal two key findings: First, most models perform significantly better in single-image settings compared to multi-image scenarios (average improvement of +42.3%), indicating a

Table 3: Performance comparison of vision-language models across different categories in single-image settings. The rightmost column shows the performance improvement ratio when switching from multi-image to single-image settings.

Models	Time and Calendar		Space and Location		Geometry and Shapes			Objects and Motion		Reasoning and Observation		Organization and Pattern		Avg.	Improvement Ratio			
	Calendar	Clock	Direction	Location	Place	Angle	Quad	Rectangular	Shape	Triangle	Cube	Move	Reasoning			Observe	Organize	Pattern
Random Guess	33.33	32.78	25.00	29.81	33.33	31.00	27.63	29.17	31.84	29.01	28.37	29.35	33.33	29.41	30.17	31.32	33.33	29.83
Human	100.00	96.00	100.00	93.85	96.67	95.60	96.84	95.00	94.02	94.07	97.67	94.63	100.00	93.59	93.20	95.52	100.00	93.30
Close-Source Models																		
GPT-4o-mini OpenAI et al. (2024)	100.00	20.00	0.00	30.77	100.00	42.80	26.32	50.00	56.77	40.74	34.93	43.01	90.00	32.43	72.00	37.93	60.00	39.65
GPT-4o OpenAI et al. (2024)	80.00	40.67	100.00	42.31	66.67	68.40	57.89	64.29	68.12	44.44	42.47	56.99	60.00	30.10	90.40	44.83	100.00	45.52
Claude-3.7-Sonnet Anthropic (2025)	100.00	54.67	80.00	65.38	83.33	61.20	68.42	78.57	68.56	77.78	43.84	69.12	100.00	34.37	92.00	63.79	100.00	51.69
Gemini2.0-Flash Deepmind (2025)	20.00	76.67	100.00	61.54	83.33	58.00	63.16	42.86	71.62	59.26	46.58	73.90	100.00	39.41	90.00	46.55	100.00	53.90
Qwen-VL-Max Bai et al. (2023)	0.00	53.33	100.00	73.08	83.33	80.00	52.63	75.00	69.87	66.67	57.53	72.43	100.00	43.54	91.60	41.38	80.00	57.03
Open-Source Models																		
Index3-8B Laurenon et al. (2024)	0.00	10.00	20.00	11.54	16.67	10.00	5.26	32.14	20.52	7.41	17.12	18.01	0.00	12.53	30.00	20.69	0.00	15.64
LLaMA-3.2-90B-Vision-Instruct AI (2024)	80.00	30.00	0.00	15.38	33.33	26.00	15.79	25.00	17.03	33.33	27.40	26.47	100.00	19.64	49.60	12.07	0.00	22.38
Emu2-Chat Sun et al. (2024b)	60.00	12.67	100.00	23.08	16.67	24.00	42.11	28.57	24.02	18.52	22.60	24.63	10.00	22.87	12.00	22.41	0.00	23.08
DeepSeek-VL2 Wu et al. (2024)	20.00	33.33	0.00	19.23	33.33	28.00	10.53	32.14	32.31	25.93	13.70	32.35	0.00	20.03	46.00	27.59	100.00	24.77
Mantis-CLIP Jiang et al. (2024)	0.00	35.33	80.00	23.08	0.00	28.00	42.11	46.43	31.88	11.11	26.03	25.00	0.00	27.52	12.00	34.48	0.00	27.50
LLaVA-Interleave-7B Li et al. (2024b)	00.00	30.00	100.00	30.77	0.00	26.00	36.84	42.86	33.19	14.81	31.51	26.47	50.00	29.07	28.00	25.86	0.00	29.24
Phi-3.5-vision-instruct Abdin et al. (2024)	0.00	13.33	80.00	19.23	16.67	24.00	10.53	42.86	34.50	22.22	32.19	29.78	20.00	31.40	46.00	25.86	100.00	30.93
LLaVA-One-Vision-7B Li et al. (2024a)	0.00	43.33	0.00	23.08	100.00	44.00	21.05	35.71	44.10	44.44	30.82	42.65	40.00	29.07	64.40	27.59	80.00	35.47
InternVL2.5-8B Chen et al. (2024)	0.00	33.33	0.00	26.92	50.00	46.40	31.58	39.29	51.53	48.15	31.51	42.65	30.00	28.42	60.80	29.31	80.00	36.16
Gemma3-27B-eTeam et al. (2025a)	80.00	40.00	0.00	26.92	33.33	48.40	21.05	57.14	45.85	33.33	33.56	45.22	100.00	30.10	66.80	20.69	60.00	36.80
Kimi-VL-A3B-ThinkingTeam et al. (2025b)	0.00	33.33	0.00	34.62	50.00	62.00	52.63	39.29	52.40	77.78	26.03	55.15	50.00	25.19	86.00	39.66	100.00	38.72
LLaVA-One-Vision-72B Li et al. (2024a)	20.00	53.33	0.00	30.77	33.33	38.00	47.37	39.29	51.53	55.56	39.73	41.54	100.00	32.95	32.00	55.17	100.00	39.24
Mistral-Small-3.1-24B-Instruct Mistral (2025)	20.00	40.00	0.00	38.46	50.00	64.00	57.89	46.43	56.77	70.37	30.14	50.74	100.00	31.65	82.00	43.10	80.00	42.21
QVQ-72B-Preview Team (2024)	80.00	41.33	80.00	61.54	50.00	64.00	68.42	39.29	58.95	81.48	32.19	64.34	50.00	35.01	90.00	50.00	100.00	47.44
InternVL2.5-78B Chen et al. (2024)	80.00	50.00	100.00	42.31	50.00	62.80	63.16	57.14	65.94	55.56	32.19	61.76	90.00	36.43	88.00	36.21	100.00	47.73
Kimi-VL-A3B-InstructTeam et al. (2025b)	0.00	70.00	100.00	50.00	66.67	50.00	31.58	35.71	59.39	51.85	46.58	62.13	50.00	38.11	82.00	46.55	100.00	48.37
Qwen2.5-VL-7B-Instruct Bai et al. (2025)	0.00	53.33	100.00	46.15	83.33	72.80	52.63	60.71	61.14	55.56	60.96	64.34	100.00	37.86	92.00	36.21	80.00	51.10
Qwen2.5-VL-72B-Instruct Bai et al. (2025)	20.00	53.33	100.00	73.08	83.33	80.00	52.63	75.00	69.87	66.67	57.53	72.43	90.00	43.54	92.00	41.38	80.00	57.03
Math Specialist Models																		
G-LLaVA-13B Gao et al. (2023)	0.00	40.00	0.00	23.08	33.33	20.40	31.58	32.14	26.64	25.93	15.75	26.10	0.00	26.49	24.00	24.14	20.00	25.47
G-LLaVA-7B Gao et al. (2023)	100.00	36.67	20.00	30.77	0.00	30.00	21.05	50.00	31.88	40.74	23.97	27.21	0.00	27.26	28.00	24.14	100.00	28.26
MathLlava Shi et al. (2024)	100.00	20.00	80.00	26.92	0.00	32.00	31.58	21.43	27.51	11.11	34.93	29.04	40.00	29.97	28.40	29.31	80.00	29.30

Table 4: Influence of Chain-of-Thought Wei et al. (2023a) on model performances.

Model	CoT	Time and Calendar		Space and Location			Geometry and Shapes			Objects and Motion		Reasoning and Observation		Organization and Pattern			Avg.		
		Calendar	Clock	Direction	Location	Place	Angle	Quad	Rectangular	Shape	Triangle	Cube	Move	Reasoning	Observe	Organize	Pattern	Weight	
GPT-4o OpenAI et al. (2024)	✗	100.00	40.00	20.00	30.77	66.67	46.00	57.89	28.57	50.22	51.85	37.67	50.37	90.00	31.27	76.00	37.93	80.00	40.29
	✓	100.00	40.00	0.00	38.46	66.67	52.00	63.16	32.14	53.71	66.67	33.56	52.57	100.00	30.75	82.00	58.62	100.00	42.03
		0.00	-20.00	-20.00	-7.69	0.00	+5.00	+5.27	+3.57	+14.82	-4.11	-4.52	+6.00	+20.69	+3.71	+6.00	+20.69	+20.00	+17.74
Qwen-VL-Max Bai et al. (2023)	✗	0.00	46.67	0.00	42.31	66.67	74.00	52.63	42.86	54.15	66.67	56.16	60.66	50.00	35.27	68.00	39.66	100.00	47.03
	✓	20.00	36.67	100.00	57.69	66.67	74.40	52.63	57.14	60.26	77.78	52.74	61.03	90.00	36.05	93.60	44.83	100.00	49.48
		+20.00	-10.00	+100.00	+15.38	0.00	+0.40	0.00	+14.28	+6.11	+11.11	-3.42	+0.37	+40.00	+0.78	+25.60	+5.17	0.00	+2.45
Gemini2.0-Flash Deepmind (2025)	✗	100.00	70.00	20.00	57.69	66.67	70.00	68.42	53.57	61.14	70.37	44.52	68.75	40.00	35.53	74.00	46.55	100.00	49.77
	✓	80.00	83.33	20.00	69.23	83.33	66.40	68.42	67.86	71.62	66.67	41.10	70.96	100.00	37.86	89.40	56.90	100.00	53.66
		-20.00	+13.33	0.00	+11.54	+16.66	-3.60	0.00	+14.29	+10.48	-3.70	-3.42	+2.21	+60.00	+2.33	+15.40	+10.35	0.00	+3.89

strong bias toward single-image optimization. For instance, Qwen-VL-Max shows a +21.3% gain in single-image performance, while models like Emu2-Chat exhibit dramatic improvements (+281.5%). Second, specialized multi-image models like LLaVA-Interleave-7B show the opposite trend (-17.6% in single-image mode), achieving higher accuracy in multi-image tasks than in single-image ones. This contrast suggests that unlike dedicated multi-image architectures, conventional models struggle to integrate visual information across multiple inputs, highlighting a critical limitation in current vision-language systems. Addressing this gap by effectively leverage cross-image cues for reasoning remains an essential challenge for future research.

4.3 RESULTS OF MATH SPECIALIST MODELS

The Math Specialist models, including G-LLaVA-13B, G-LLaVA-7B, and MathLlava, exhibit relatively low overall performance, with average scores from 25.47 to 29.30. Notably, G-LLaVA-13B records the lowest score at 25.47, while MathLlava achieves a slightly higher score of 29.30. Though these models are designed to focus on mathematical reasoning, their performance across diverse tasks-such as time and calendar, spatial reasoning, and geometric challenges-remains inconsistent. For example, while G-LLaVA-7B reaches a perfect score (100.00) on the Calendar sub-task, its scores in other categories, such as Clock and geometry-related tasks, are considerably lower.

Furthermore, the results indicate that these Math Specialist models struggle to match the performance of their general-purpose counterparts. Despite showing some strengths-for example, MathLlava scoring 34.93 on the Cube task-these models fall short on several key aspects, including Clock, Location, and reasoning tasks. This pattern underscores the challenge of integrating specialized mathematical capabilities with the broader spectrum of visual understanding.

5 ANALYSIS

5.1 INFLUENCE OF CHAIN-OF-THOUGHT ON MODEL PERFORMANCE

Chain-of-thought Wei et al. (2023a) reasoning generally enhances model performance, as the Table 4 shows stable improvements across several domains when CoT is enabled. For instance, Qwen-VL-Max exhibits a dramatic 40% boost in the “Reasoning” task, highlighting the significant impact of structured reasoning on spatial understanding. Gemini2.0-Flash also benefits substantially, with a

Table 5: Comparisons between existing visual math benchmarks for LVLMS.

Benchmark	Image Numbers	Question Numbers	Required Skills					Multi-Images	Answer Type
			Temporal	Spatial	Geometric	Logical	Pattern		
Olympiadbench He et al. (2024)	5,129	8,952	✗	✗	✓	✗	✗	✗	Free-form
GeoQA Chen et al. (2021)	4,998	4,998	✗	✗	✓	✗	✗	✗	Multiple Choice
MATH-Vision Wang et al. (2024a)	3,472	3,040	✗	✗	✓	✗	✗	✓	Free-form & Multiple Choice
MathVista Lu et al. (2023)	5,487	6,141	✓	✗	✓	✓	✗	✗	Free-form & Multiple Choice
MMU _{math} Yue et al. (2024)	577	540	✗	✗	✓	✗	✗	✓	Free-form & Multiple Choice
GeoMath Xu et al. (2024)	4,540	9,155	✗	✗	✓	✗	✗	✗	Free-form & Multiple Choice & Prove
U-Math Chernyshev et al. (2025)	225	1,100	✗	✗	✓	✗	✗	✗	Free-form
Blink Fu et al. (2024)	7,358	3,807	✗	✗	✗	✗	✓	✓	Multiple Choice
MM-MATH Sun et al. (2024a)	5,929	5,929	✗	✗	✓	✗	✗	✓	Free-form
MMIE _{math} Xia et al. (2024)	26,534	20,103	✗	✗	✓	✗	✗	✓	Free-form & Multiple Choice
Polymath Gupta et al. (2024)	5,000	5,000	✗	✓	✓	✗	✓	✗	Multiple Choice
NTSEBench Pandya et al. (2025)	4,642	2,728	✗	✓	✓	✗	✓	✓	Multiple Choice
BSA ¹ Xu et al. (2025)	312	312	✗	✓	✗	✗	✗	✓	Multiple Choice
MV-MATH Wang et al. (2025)	6,061	2,009	✗	✓	✓	✗	✓	✓	Free-form & Multiple Choice
Ours	6,697	1,720	✓	✓	✓	✓	✓	✓	Multiple Choice

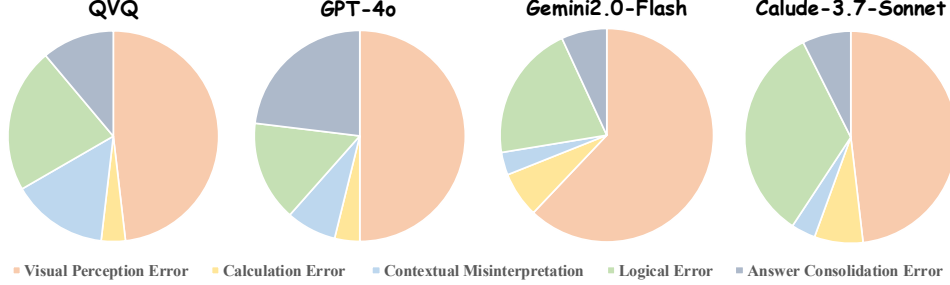


Figure 5: A comparison of error distributions among three model, GPT-4o, Gemini2.0-Flash, and Calude-3.7-Sonnet, across five error categories: visual perception errors, calculation errors, contextual misunderstandings, logical errors, and answer integration errors.

15.40 point increase in the “Pattern” category and a 16.66 point rise in “Place” suggesting that CoT particularly aids in tasks requiring complex organizational and geometric reasoning.

While improvements are evident, the efficacy of chain-of-thought (CoT) prompting exhibits strong task-dependent variation. CoT consistently enhances performance in multi-step reasoning tasks (e.g., Pattern and Reasoning tasks), where all models show gains. However, it proves neutral in perception-heavy tasks (e.g., Calender and Direction tasks) due to interference with low-level spatial or temporal processing. Nonetheless, the overall trend supports that incorporating CoT tends to enhance problem-solving abilities, especially in tasks that demand high-level reasoning and pattern recognition.

5.2 COMPARISON WITH OTHER BENCHMARKS

In comparison to existing visual math benchmarks, our dataset stands out in several important ways as shown in Table 5. While benchmarks such as Olympiadbench He et al. (2024) and GeoQA Chen et al. (2021) focus primarily on specific skills like geometry and logical reasoning, our benchmark includes a broader spectrum of required skills, including temporal, spatial, geometric, logical, and pattern recognition. This comprehensive skill coverage provides a more holistic evaluation of LVLMS. Additionally, our dataset supports multi-image tasks, a feature not widely supported by other benchmarks such as Blink Fu et al. (2024) and GeoQA Chen et al. (2021), enhancing its applicability for real-world tasks that require understanding across multiple visual inputs. Moreover, our benchmark boasts a higher image-question ratio than other benchmarks, meaning that on average, each question is associated with more images. Finally, our dataset offers multiple-choice answer types for easier evaluation, unlike other benchmarks that provide free-form answer format which is hard to evaluate, such as MM-MATH Sun et al. (2024a) and U-Math Chernyshev et al. (2025).

5.3 ERROR DISTRIBUTION FOR VCBENCH

We define five error types in this benchmark: Visual Perception Error indicates that the model misinterprets or fails to accurately perceive visual content; Calculation Error captures mistakes made during arithmetic computations; Contextual Misinterpretation occurs when the model misreads the textual conditions, such as treating unrelated information as relevant; Logical Error refers to flaws in the reasoning process; and Answer Consolidation Error encompasses failures to directly answer the question or instances where multiple, conflicting answers are provided. We conduct manual

error classification for all questions across four top-tier models, enabling precise identification of each model’s failure patterns and relative weaknesses across different error categories. As shown in Figure 5, Visual Perception Errors are predominant across all models, with Gemini2-Flash exhibiting the highest rate at about 62%. This persistent pattern across architectures suggests that enhancing visual perception capabilities remains the most critical challenge for multimodal models. Calculation Errors remain consistently low (ranging from about 4% to about 7%), indicating that basic arithmetic computation has become relatively robust in modern models. Contextual Misinterpretation errors are minimal, particularly for Gemini2-Flash (about 3%) and Claude (about 4%), which indicates a relatively robust understanding of textual context. However, QVQ’s comparatively higher rate (6%) may reflect its tendency toward over-reasoning, where excessive analysis leads to detachment from the original question context.

On the other hand, discrepancies are more apparent in the Logical and Answer Consolidation Error rates. Claude shows a significantly high Logical Error rate of about 33% compared to GPT-4o’s about 15% and QVQ’s about 22%, revealing the weaknesses in its deductive reasoning pipelines. Moreover, while Answer Consolidation Errors are generally low (QVQ at about 11% and both Gemini2-Flash and Claude at about 7%), GPT-4o presents a higher rate of about 23%, suggesting its advanced reasoning capabilities may come at the cost of response discipline, where the model sometimes generates multiple answers rather than a single one. This trade-off between exploratory reasoning and answer precision presents an important optimization target for future iterations.

5.4 ANALYSIS OF PROBLEM DIFFICULTY AND MODEL PERFORMANCE

All questions in our benchmark were sourced from established online question banks and annotated by editors with a difficulty coefficient ranging from 0.0 to 1.0. Questions with coefficients between 0.0 and 0.35 are categorized as easy, those between 0.35 and 0.75 as medium, and those from 0.75 to 1.0 as hard. Overall, 27.7% of questions are classified as easy, 41.6% as medium, and 30.7% as hard. Interestingly, the results in Table 6 reveal that questions annotated as hard tend to yield higher accuracy, while the easy and medium problems register lower accuracy. This counterintuitive outcome may be attributed to the fact that simpler questions, which primarily require the identification of patterns rather than intricate computations, pose a different challenge compared to the hard questions that demand complex calculation and structured reasoning. It is important to note that the difficulty levels in our benchmark are derived from the original textbook platforms, where they are based on large-scale student performance statistics and thus reflect authentic human difficulty. Upon closer analysis, we observed that models often perform worse on problems considered “easy” by human standards. These are typically perception-heavy tasks. For instance, in clock reading questions, models struggle to recognize the positions of hour and minute hands, particularly when the clocks have decorative or irregular designs. Similarly, in block counting tasks, the presence of stacked or overlapping cubes frequently confuses model predictions due to challenges in depth perception and object segmentation. Although both types of questions are suitable for elementary students and considered trivial by humans, they expose current models’ significant limitations in low-level visual understanding, leading to surprisingly low performance on these “easy” items.

Table 6: Accuracy comparison of various models on questions categorized by difficulty along with their average performance.

Models	Easy	Medium	Hard	Avg.
LLaMA-3.2-90B-Vision-Instruct	22.22	26.15	23.89	25.41
Mantis-CLIP	29.63	29.30	32.37	30.23
InternVL2.5-78B	25.93	36.03	41.62	37.56
QVQ-72B-Preview	18.52	36.71	45.66	39.13
LLaVA-OneVision-72B	29.63	45.32	53.76	47.62
Qwen2.5-VL-72B-Instruct	25.93	45.49	55.11	48.08

6 CONCLUSION

This paper introduces VCBENCH—a comprehensive evaluation framework designed to assess multimodal mathematical reasoning with explicit visual dependency. By addressing the limitations of existing datasets in multi-image integration and cross-modal relational reasoning, our benchmark provides a detailed analysis of 26 state-of-the-art LVLMS across six cognitive domains and 17 task categories. The evaluation reveals significant performance disparities, particularly in areas such as multi-step instruction following, basic visual perception, cross-image consistency, and vulnerability to visual hallucinations.

7 ETHICS STATEMENT

This research complies with the ICLR Code of Ethics. All Large Vision-Language Models (LVLMs) evaluated in this study are publicly available and widely adopted in academia and industry. VCBENCH was constructed using synthetic and open-access visual content, without involving any personally identifiable, sensitive, or proprietary information. Our work is limited to methodological exploration and model evaluation, with no direct experimentation on human subjects or real-world deployment. While the benchmark is designed to rigorously assess multimodal mathematical reasoning, we recognize that advancements in LVLMs may inadvertently reinforce existing biases embedded in training corpora. We encourage future researchers and practitioners to systematically take fairness, bias, and robustness into account when applying or extending our methodology. The authors declare no conflicts of interest associated with this submission.

8 REPRODUCIBILITY STATEMENT

We have taken extensive measures to ensure the reproducibility of our findings with VCBENCH. The benchmark, including all problems, images, and evaluation protocols, is described in detail in Sections 3 and 4. We provide comprehensive documentation on dataset construction, problem generation, cognitive domain taxonomy, and visual dependency design. Evaluation procedures for all 26 LVLMs—including model configurations, scoring metrics, and inference setups—are transparently reported in the main text and further clarified in the Appendix. All preprocessing, benchmark scripts, and evaluation code will be released alongside the paper, enabling independent replication of our results. In addition, the full problem set, associated images, and baseline evaluation outputs will be released as part of the anonymous supplementary materials, to facilitate benchmarking, comparison, and future work in multimodal mathematical reasoning.

REFERENCES

- Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacrose, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219, 2024.
- Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016. URL <https://arxiv.org/abs/1505.00468>.
- Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. <https://ai.meta.com/>, 2024.
- Anthropic. Claude 3.7 sonnet and claude code. <https://www.anthropic.com/claude-3-7-sonnet-system-card>, 2025.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL <https://arxiv.org/abs/2308.12966>.

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai 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. *CoRR*, abs/2502.13923, 2025.
- Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online, August 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.findings-acl.46/>.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *CoRR*, abs/2412.05271, 2024.
- Konstantin Chernyshev, Vitaliy Polshkov, Ekaterina Artemova, Alex Myasnikov, Vlad Stepanov, Alexei Miasnikov, and Sergei Tilga. U-math: A university-level benchmark for evaluating mathematical skills in llms, 2025. URL <https://arxiv.org/abs/2412.03205>.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021a. URL <https://arxiv.org/abs/2110.14168>.
- Karl Cobbe, Vineet Kosaraju, et al. Gsm8k: A dataset for grade school math word problems. *arXiv preprint arXiv:2110.14168*, 2021b.
- Google Deepmind. Gemini 2.0 flash. <https://deepmind.google/technologies/gemini/flash/>, 2025.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: multimodal large language models can see but not perceive. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIII*, volume 15081 of *Lecture Notes in Computer Science*, pp. 148–166. Springer, 2024. URL https://doi.org/10.1007/978-3-031-73337-6_9.
- Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-llava: Solving geometric problem with multi-modal large language model, 2023. URL <https://arxiv.org/abs/2312.11370>.
- Himanshu Gupta, Shreyas Verma, Ujjwala Anantheswaran, Kevin Scaria, Mihir Parmar, Swaroop Mishra, and Chitta Baral. Polymath: A challenging multi-modal mathematical reasoning benchmark, 2024. URL <https://arxiv.org/abs/2410.14702>.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiad-bench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024.
- Dan Hendrycks, Collin Burns, et al. Measuring mathematical problem-solving with the math dataset. *NeurIPS*, 2021.
- Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. MANTIS: interleaved multi-image instruction tuning. *CoRR*, abs/2405.01483, 2024.

- Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *CoRR*, abs/2408.12637, 2024.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *CoRR*, abs/2408.03326, 2024a.
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024b. URL <https://arxiv.org/abs/2407.07895>.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL <https://arxiv.org/abs/1405.0312>.
- Hongwei Liu, Zilong Zheng, et al. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*, 2024.
- Wentao Liu, Hanglei Hu, et al. Mathematical language models: A survey. *arXiv preprint arXiv:2312.07622*, 2023.
- Pan Lu, Hritik Bansal, et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.
- Mistral. Mistral small 3.1. <https://mistral.ai/news/mistral-small-3-1>, 2025.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispositi, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrew Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O’Connell, Ian O’Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra

- Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubei, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermeni, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiye Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.
- Pranshu Pandya, Vatsal Gupta, Agney S Talwarr, Tushar Kataria, Dan Roth, and Vivek Gupta. Ntsebench: Cognitive reasoning benchmark for vision language models, 2025. URL <https://arxiv.org/abs/2407.10380>.
- Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models, 2024. URL <https://arxiv.org/abs/2406.17294>.
- Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juan-Zi Li. MM-MATH: advancing multimodal math evaluation with process evaluation and fine-grained classification. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pp. 1358–1375. Association for Computational Linguistics, 2024a. URL <https://aclanthology.org/2024.findings-emnlp.73>.
- Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiyang Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14398–14409. IEEE, 2024b.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivi re, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Ga l Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi,

- Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025a. URL <https://arxiv.org/abs/2503.19786>.
- Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, and Ziwei Chen. Kimi-vl technical report, 2025b. URL <https://arxiv.org/abs/2504.07491>.
- Qwen Team. Qvq: To see the world with wisdom, December 2024. URL <https://qwenlm.github.io/blog/qvq-72b-preview/>.
- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024a.
- Peijie Wang, Zhong-Zhi Li, Fei Yin, Xin Yang, Dekang Ran, and Cheng-Lin Liu. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts, 2025. URL <https://arxiv.org/abs/2502.20808>.
- Yiqi Wang, Wentao Chen, Xiaotian Han, et al. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024b.

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023a. URL <https://arxiv.org/abs/2201.11903>.
- Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. Cmath: Can your language model pass chinese elementary school math test?, 2023b. URL <https://arxiv.org/abs/2306.16636>.
- Jiayang Wu, Wensheng Gan, Zhenlian Qi, et al. Multimodal large language models: A survey. *IEEE International Conference on Big Data (BigData)*, 2023.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *CoRR*, abs/2412.10302, 2024.
- Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, Lijuan Wang, and Huaxiu Yao. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models, 2024. URL <https://arxiv.org/abs/2410.10139>.
- Shihao Xu, Yiyang Luo, and Wei Shi. Geo-llava: A large multi-modal model for solving geometry math problems with meta in-context learning. In *Proceedings of the 2nd Workshop on Large Generative Models Meet Multimodal Applications*, MM '24, pp. 11–15. ACM, October 2024. doi: 10.1145/3688866.3689124. URL <http://dx.doi.org/10.1145/3688866.3689124>.
- Wenrui Xu, Dalin Lyu, Weihang Wang, Jie Feng, Chen Gao, and Yong Li. Defining and evaluating visual language models’ basic spatial abilities: A perspective from psychometrics, 2025. URL <https://arxiv.org/abs/2502.11859>.
- Yibo Yan, Shen Wang, et al. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509*, 2024.
- Zhen Yang, Jinhao Chen, et al. Mathglm-vision: Solving mathematical problems with multimodal large language models. *arXiv preprint arXiv:2409.13729*, 2024.
- Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, 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 Wenhui Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9556–9567. IEEE, 2024. URL <https://doi.org/10.1109/CVPR52733.2024.00913>.

A APPENDIX

A.1 EXPERIMENT DETAILS

Table 7: Generation parameters for LVLMS (with grouped configurations).

Model	Generation Setup
GPT-4o-mini & GPT-4o	API URL: https://api.openai.com/v1/chat/completions temperature = 0.2, max_tokens = 1024
Claude-3.7-Sonnet	API URL: https://api.anthropic.com/v1/messages , temperature = 0.2, max_tokens = 1024
Gemini2.0-Flash	API URL: https://generativelanguage.googleapis.com/v1beta/models/gemini-pro:generateContent , temperature = 0.2, max_tokens = 1024
Qwen-VL-Max	Use dashscope package, temperature = 0.2, max_new_tokens = 1024
Open-Source Models	<i>Same parameters for all below:</i> Deployed by vllm, with do_sample = True, temperature = 0.2, max_new_tokens = 1024 • Idefics3-8B • LLaMA-3.2-90B-Vision-Instruct • Emu2-Chat • DeepSeek-VL2 • Mantis-CLIP • LLaVA-Interleave-7B • Phi-3.5-vision-instruct • InternVL-2.5 • LLaVA-OneVision-7B/72B • Gemma3-27B-it • Mistral-Small-3.1-24B-Instruct • Qwen2.5-VL-7B/72B-Instruct
QVQ-72B-Preview	do_sample = True, temperature = 0.2, max_new_tokens = 2048
G-LLaVA-7B/13B	do_sample = True, temperature = 0.2, max_new_tokens = 1024
MathLlava	do_sample = True, temperature = 0.2, max_new_tokens = 1024

A.2 IMPACT OF IMAGE COUNT ON MODEL ACCURACY

Table 8: Model accuracy (%) across problems grouped by image count. The number of questions per image count is shown in the second row. As image count increases, accuracy generally decreases with some fluctuations.

Model / Image Count	2 images	3 images	4 images	5 images	6–7 images	8–10 images	11+ images
Question Count	316	544	515	221	76	34	14
GPT-4o	58.9	47.3	44.7	40.3	35.5	30.9	21.4
Claude-3.7-sonnet	63.5	54.8	50.1	46.7	40.8	37.2	23.1
Qwen-VL-max	61.9	52.1	48.0	45.2	39.7	36.7	18.7
Gemini-2.0-flash	65.2	57.1	53.8	51.2	44.1	39.8	26.2

Table 8 presents the accuracy of four leading multimodal models across problems grouped by the number of images provided in each question. We observe a general trend: as the number of images increases, average model accuracy tends to decrease. These results underscore the importance of developing models capable of robust visual reasoning, especially in scenarios requiring the integration of multiple visual inputs.

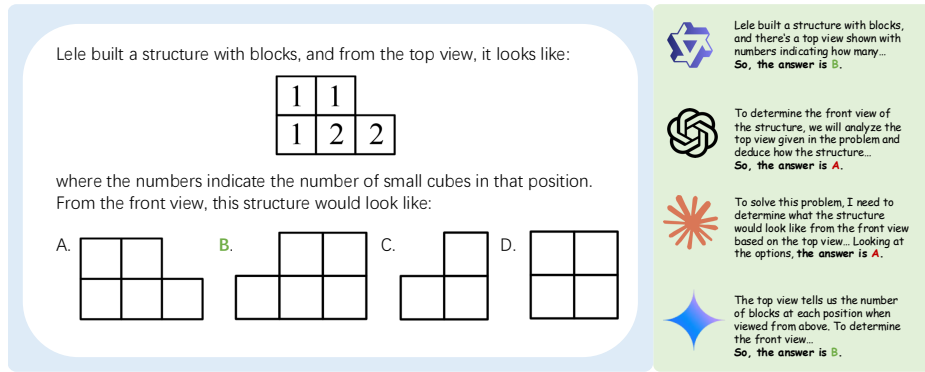


Figure 6: Case for Visual Perception Error.

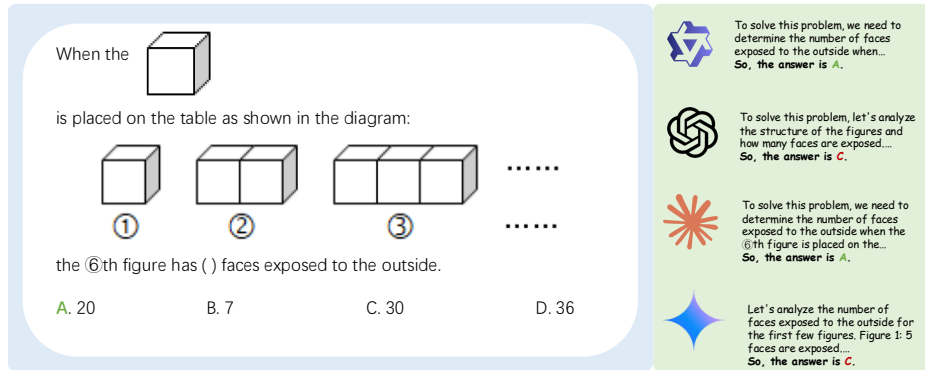


Figure 7: Case for Calculation Error.

A.3 CASE STUDIES

Individually, each of the five cases highlights a distinct type of reasoning error made by models. In Figure 6 (Visual Perception Error), models failed to accurately map the numbers from a top view to a front view, resulting in confusion about the arrangement of cubes in three-dimensional space. Figure 7 (Calculation Error) exposed mistakes in pattern recognition and extrapolation, with models incorrectly calculating the number of exposed faces as the sequence grew. Figure 8 (Contextual Misinterpretation) showed that models struggled to correctly interpret the scenario and geometric relationships required to form a triangle, often misapplying the cutting method or misreading the intent of the task. In Figure 9 (Logical Error), models made incorrect choices about physical stability, suggesting a lack of real-world reasoning about spatial structure and balance. Figure 10 (Answer Consolidation Error) demonstrated the models' challenges in mentally rotating and consolidating three-dimensional information to select the correct two-dimensional viewpoint.

These examples reveal that current models face significant difficulties in visual and mathematical reasoning. Their limitations extend beyond simple miscalculations, reflecting deeper struggles with tasks that require the integration of visual perception, contextual understanding, logical reasoning, and spatial manipulation. In many cases, models fail to synthesize information across multiple steps or modalities, leading to errors in connecting abstract concepts to specific visual details or consolidating reasoning into correct answers. Together, these challenges suggest that current models lack the comprehensive, multi-step cognitive processing abilities required for human-level performance in complex multimodal mathematical problems.

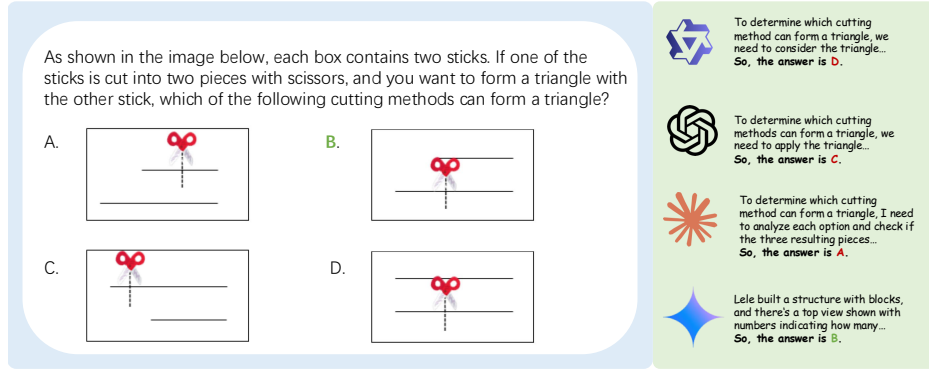


Figure 8: Case for Contextual Misinterpretation.

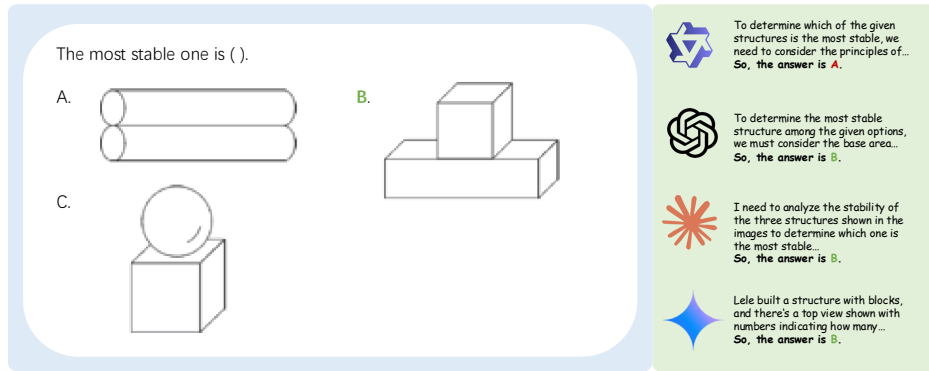


Figure 9: Case for Logical Error.

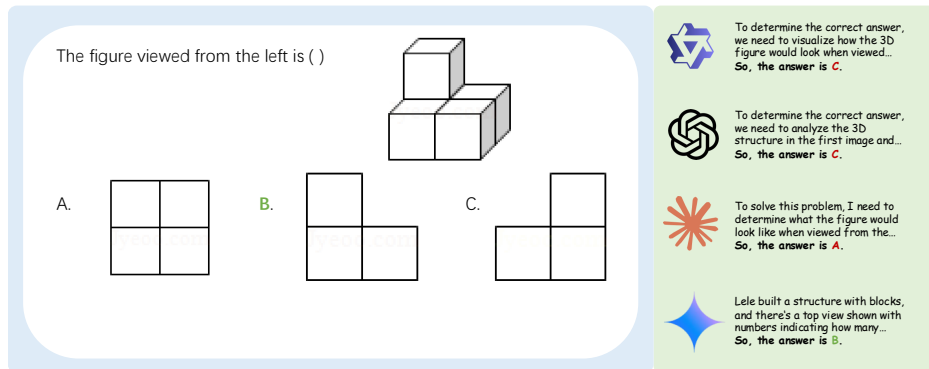


Figure 10: Case for Answer Consolidation Error.

A.4 PROMPT FOR EXPERIMENT

Table 9: Inference Prompt.

Inference Prompt
You are a helpful AI assistant. Please answer the following questions and output the answer options directly. <i>Question: { question }</i>

Table 10: Inference Prompt with Chain-of-Thought.

Inference Prompt with Chain-of-Thought
You are a helpful AI assistant. Please think step by step before answer the following questions and the output the answer. <i>Question: { question }</i>

Table 11: LLM-Based Evaluation Prompt.

LLM-Evaluation Prompt
You are an answer evaluator. I will give you a response and an answer. Please tell me whether this response is correct or wrong. Just answer yes or no. For example, Response: The figure that cannot be folded into a cube is: C. <image> Correct Answer: B So, you need to respond no only. Response: The unfolded shape of the cube is: B. <image> Correct Answer: B So, you need to respond yes only. Here is the response and correct answer I want you to evaluate. <i>Response: { model response }</i> <i>Correct Answer: { correct answer }</i>