OMIBench: Benchmarking Olympiad-Level Multi-Image Reasoning in Large Vision-Language Models

Qiguang Chen[1*], **Chengyu Luan**[1*], **Jiajun Wu**[2], **Qiming Yu**[1], **Yi Yang**[2], **Yizhuo Li**[1], **Jingqi Tong**[3], **Xiachong Feng**[4], **Libo Qin**[5,6†], **Wanxiang Che**[1†]

[1] Research Center for Social Computing and Interactive Robotics, Harbin Institute of Technology, [2] Central South University, [3] Fudan University, [4] The University of Hong Kong, [5] Harbin Institute of Technology (Shenzhen), [6] Text Computing and Cognitive Intelligence Ministry of Education Engineering Research Center, Guizhou University

{qgchen,car}@ir.hit.edu.cn, qinlibo@hit.edu.cn

Abstract

Large vision-language models (LVLMs) have made substantial advances in reasoning tasks at the Olympiad level. Nevertheless, current Olympiad-level multimodal reasoning benchmarks for these models often emphasize single-image analysis and fail to exploit contextual information across multiple images. We present OMIBench, a benchmark designed to evaluate Olympiad-level reasoning when the required evidence is distributed over multiple images. It contains problems from biology, chemistry, mathematics, and physics Olympiads, together with manually annotated rationales and evaluation protocols for both exact and semantic answer matching. Across extensive experiments on OMIBench, we observe meaningful performance gaps in existing models. Even the strongest LVLMs, such as Gemini-3-Pro, attain only about 50% on the benchmark. These results position OMIBench as a focused resources for studying and improving multi-image reasoning in LVLMs.

1 Introduction

Recent advances in large vision–language models (LVLMs) have enabled strong performance on demanding reasoning tasks, from elementary arithmetic to Olympiad-level problems that require deep domain knowledge and multi-step inference (Lu et al., 2022, 2024; Chen et al., 2024b; Wang et al., 2025d; Liu et al., 2025; He et al., 2024). A central driver of this progress is chain-of-thought (CoT) prompting (Wei et al., 2022), which elicits explicit intermediate reasoning steps in natural language (Wang et al., 2025d; Chen et al., 2024a, 2025a). In multimodal settings, these techniques enable LVLMs to fuse visual cues with textual information, yielding substantial gains on single-image Olympiad-level benchmarks (Zhang et al., 2024b; Cheng et al., 2025a).

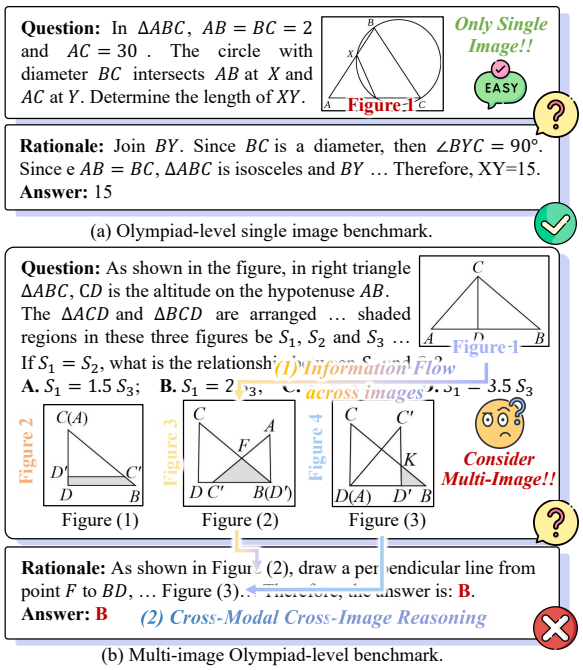


Figure 1: Comparison between existing single-image reasoning benchmarks (OlympiadBench) and our proposed Olympiad-level Multi-Image Reasoning Benchmark (OMIBench).

However, as illustrated in Figure 1 (a), existing multimodal Olympiad benchmarks largely remain restricted to single-image question settings (Zhao et al., 2024; Cheng et al., 2025c; Du et al., 2025). In real scientific and technical settings, however, problems often rely on multiple interdependent figures, diagrams, and experimental setups (Figure 1(b)) (Alampara et al., 2025; Chen et al., 2025b; Roberts et al., 2024; Liu et al., 2024a; Ji et al., 2025). Effective multi-image reasoning therefore requires not only interpreting each image, but also (1) **maintaining a coherent information flow across images**, and (2) **performing cross-image, cross-modal reasoning** that supports Olympiad-level problem solving. However, existing benchmarks (Zhang et al., 2024a; Lu et al., 2024; Małkiński and Mańdziuk, 2025; Cheng et al., 2025c)

*Equal contribution.
†Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45100–45135
July 2-7, 2026

Benchmark	Multi-Image	Maths	Physics	Chemistry	Biology	Difficulty	Rationale	Answer type	Question type
ScienceQA (Lu et al., 2022)	✘	✔	✔	✔	✔	H	~ 90%	Text, Num	MC
MME-CoT (Jiang et al., 2025)	✘	✔	✔	✔	✔	H	✔	Text, Num	MC, OE
M^3CoT (Chen et al., 2024b)	✘	✔	✔	✔	✔	COL	✔	Exp, Text, Num	MC
MMReason (Yao et al., 2025)	✘	✔	✔	✘	✘	COL	✔	Exp, Text, Num	OE
MathVista (Lu et al., 2024)	✘	✔	✘	✘	✘	COL	✘	Num	MC, OE
MathVerse (Zhang et al., 2024a)	✘	✔	✘	✘	✘	COL	✘	Num	MC, OE
MMMU (Yue et al., 2024)	<10%	✔	✔	✔	✔	COL	< 18%	Image, Text, Num	MC, OE
OlympiadBench (He et al., 2024)	<5%	✔	✔	✘	✘	COMP	✔	ALL	ALL
MuirBench (Wang et al., 2025a)	✔	✘	✘	✘	✘	H	✘	Text	MC
MMIU (Meng et al., 2024)	✔	✘	✘	✘	✘	H	✘	Text	MC
Blink (Fu et al., 2024)	✔	✘	✘	✘	✘	H	✘	Image, Text, Num	MC
ReMI (Kazemi et al., 2024)	✔	✔	✔	✘	✘	H, COL	✘	Image, Text, Num	MC, OE
OMIBench	✔	✔	✔	✔	✔	COMP	✔	ALL	ALL

Table 1: Comparison of representative multimodal benchmarks by image setting, subject coverage, difficulty, rationale, answer types and question types. For **difficulty**, H: high, COL: college, COMP: competition; For **answer or choice type**, Num: numeric value, Text: text expression answer or choice, Image: image choices; For **question type**, MC: multiple-choice, J: judgement, OE: open-ended.

only partially capture this multi-image context: they emphasize perception and cross-image reference resolution, but give limited attention to strong semantic and quantitative links across images and modalities in Olympiad-level reasoning. Hence, they offer an incomplete evaluation of multi-image Olympiad-level reasoning, especially in tasks requiring precise interpretation across visuals (Alampara et al., 2025; Cheng et al., 2025a).

To address this gap, as shown in Table 1, we introduce the Olympiad-level Multi-Image Benchmark (OMIBench), a large-scale suite for evaluating LVLMs' multi-image information integration and reasoning. OMIBench includes over 1,000 Olympiad-level problems in biology, chemistry, mathematics, and physics, with manually annotated rationales and answers. Each problem contains multiple images that jointly provide the evidence needed for multi-step reasoning and the final solution. OMIBench also offers reasoning-path annotations, enabling fine-grained analyses.

We benchmark representative LVLMs on OMIBench. The results reveal clear limitations, with accuracy below 51% and drops of up to 15% relative to single-image settings. Model outputs show recurring failures in visual perception, cross-image association, and cross-modal logical integration; compared with existing multi-image benchmarks, OMIBench produces performance decreases exceeding 20%. We also examine several strategies for improving performance, including long chain-of-thought, test-time scaling, ICL, and think-with-image approaches. Long CoT, parallel/sequential scaling, and ICL bring consistent but limited gains, while parameter scaling and think-with-image methods offer little benefit and sometimes reduce performance. These results suggest that progress will likely require more fundamental advances in model architecture and training.

In summary, our contributions are threefold:

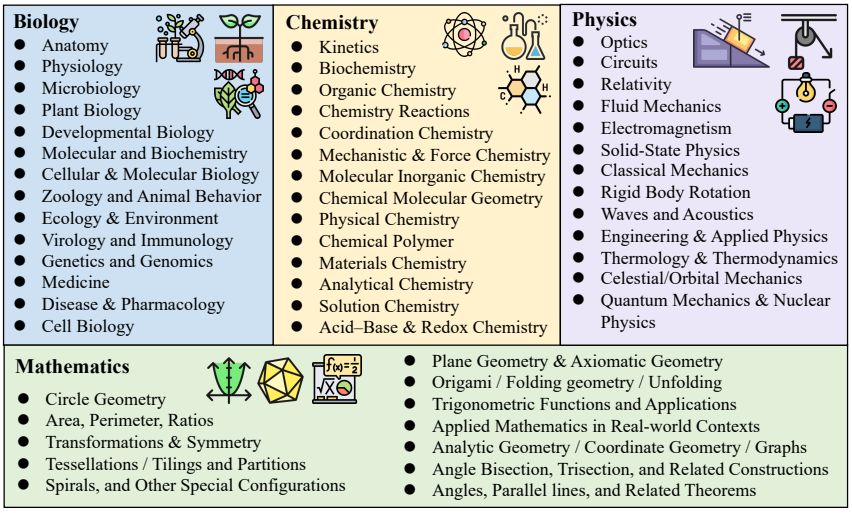
- We identify a critical gap in existing literature on evaluating multi-image Olympiad-level reasoning in LVLMs, a setting that requires autonomous cross-image alignment, selection, and integrative reasoning.
- We introduce OMIBench, a novel benchmark with over 1,000 Olympiad-level multi-image reasoning tasks spanning chemistry, physics, mathematics, and experimental design. We establish comprehensive baselines by evaluating state-of-the-art LVLMs, exposing major gaps in Olympiad-level multi-image reasoning.
- We provide diagnostic analyses and assess diverse enhancement techniques, including long CoT, test-time scaling, in-context learning, and think-with-image methods, to improve LVLM performance on OMIBench and identify promising directions.

The dataset and resources are available at https://github.com/LightChen233/OMIBench.

2 Task Definition

Unlike single-image multimodal CoT, multi-image CoT considers a set of images $\mathcal{I} = \{I_1, I_2, \ldots, I_n | n \geq 2\}$, a question Q, and a context C. The task is to answer Q by integrating evidence across multiple images, where different images may provide complementary information needed for the final answer. Specifically, OMIBench consists of the following two tasks:

Multiple-Choice Reasoning Task Given close set $\mathcal{O} = \{o_1, \ldots, o_n\}$ with n options, we first construct a textual prompt $\mathcal{T} = \texttt{Prompt}(Q, C, \mathcal{O})$.



Statistic	Number
Total Samples	1,322
Sample with Rationale	100%
Open-Ended Samples	748 (56.6%)
Multiple Choices Samples	574 (44.4%)
Biology size	251
Chemistry size	217
Mathematics size	430
Physics size	424
Average Image Number	3.07
* Avg. Biology Image	3.07
* Avg. Chemistry Image	4.00
* Avg. Mathematics Image	2.35
* Avg. Physics Image	3.34
Average question length	210.12
Average rationale length	420.18

Figure 2: Key statistics of OMIBench, encompassing diverse problem types across Biology, Chemistry, Mathematics, and Physics (over 1.3K samples; average 3.07 images per sample). Images are excluded from token counts.

The model then generates a stepwise rationale $\mathcal{R}_m = \{s_1, \ldots, s_m\}$, with each step s_i defined by:

$$s_i = \text{argmax}_{s_i \in \mathcal{R}_m} P(s_i | \mathcal{I}, \mathcal{T}). \quad (1)$$

Finally, the model selects the final answer $\mathcal{Y}$ from close option set $\mathcal{O}$, which is denoted as:

$$\mathcal{Y} = \text{argmax}_{o \in \mathcal{O}} P(o | \mathcal{R}_m). \quad (2)$$

Open-Ended Reasoning Task For open-ended problems, we first form an instruction prompt $\mathcal{T} = \texttt{Prompt}(Q, C)$, where $\texttt{Prompt}(\cdot)$ from the question and context, where $\texttt{Prompt}(\cdot)$ denotes the prompting procedure used to format the textual input. Conditioned on $\mathcal{I}$ and $\mathcal{T}$, the model produces a step-by-step rationale $\mathcal{R}_m = \{s_1, \ldots, s_m\}$, with each step generated as:

$$s_i = \text{argmax}_{s_i \in \mathcal{R}_m} P(s_i | \mathcal{I}, \mathcal{T}). \quad (3)$$

Finally, the model arrives at the final answer $\mathcal{Y}$ from open answer space $\mathcal{A}$, which is denoted as:

$$\mathcal{Y} = \text{argmax}_{\mathcal{A}} P(A | \mathcal{R}_m), \quad (4)$$

where $\mathcal{Y}$ derives from information in the images and question, requiring the model to integrate visual and textual cues for a coherent answer.

3 Olympiad-Level Multi-Image Reasoning Benchmark (OMIBench)

We build OMIBench to assess whether LVLMs can solve competition-grade scientific problems whose evidence is distributed across multiple images, with coverage across biology, chemistry, mathematics, and physics. Summary statistics are provided in Figure 2 and Table 3. Data construction details can be seen in Appendix A.

3.1 Design Principle

OMIBench targets the upper bound of Olympiad-level problem solving and supports research on LVLMs for multi-image reasoning in biology, mathematics, chemistry, and physics. Following He et al. (2024), it reflects the rigor of top competitions. Specifically, OMIBench includes:
Olympiad-level problems: Biology, mathematics, chemistry, and physics questions from international and national Olympiads for top students, in multiple-choice and open-ended formats, to assess advanced reasoning and intermediate steps.
Expert solutions and rationales: Each problem includes an expert solution with explicit reasoning. This lowers annotation and evaluation cost, strengthens correctness judgments, and provides supervision for analyzing model reasoning.
Multi-Image reasoning: Problems that require linking multiple images and their relations, testing cross-image and cross-modal reasoning and integration of visual evidence.

3.2 Data Annotation

The data annotation pipeline comprises four phases: data collection & selection, rationale annotation, quality control, and classification labeling.
Step 1: Data Collection & Selection. OMIBench comprises Olympiad-level problems in biology, chemistry, mathematics, and physics, with source distributions summarized in Table 3. The corpus integrates international Olympiads, national and regional contests, and mixed-complexity benchmarks, providing broad Olympiad-level coverage across subfields in these disciplines.

After collecting all PDF files, we use Mathpix OCR to convert problems into Markdown format, and team members manually verify each item to ensure accuracy. The Markdown texts are then normalized into a structured "Question–Rationale (if available)–Answer" format. For samples from mixed-complexity benchmarks, expert competitors further select and curate the items. Multilingual questions are translated with Google Translate and subsequently verified by human experts.

Step 2: Rationale Annotations. Most competition datasets omit solution rationales, which are essential for analyzing problem-solving. We therefore build expert-verified rationales via a two-stage pipeline combining LLM-assisted generation (Gemini-2.5-pro-thinking) and human verification.

Specifically, we use two-stage annotation: (1) LLM generates up to 16 candidate solutions per problem given the reference, retaining those with the correct final answer. If none succeed, we provide the ground-truth answer and regenerate a correct solution, reducing human effort by $\sim$20%. (2) Experienced annotators verify and refine the retained rationales by correcting errors, adding missing steps, removing redundancy, and standardizing notation. If a rationale is fundamentally flawed, annotators rewrite it while preserving valid core ideas. A final review ensures correctness, with dataset statistics in Figure 2.

Step 3: Quality Control To ensure dataset quality, we employ a dual-review protocol in which every problem receives at least one independent audit, complemented by weekly random sampling of 5% of examples for regression testing on key metrics, error rate, text-image alignment, and solution accuracy. Further, audit feedback is introduced to drive iterative updates to annotation guidelines and targeted retraining, forming a closed-loop quality assurance process that maintains a high-fidelity multimodal competition problem corpus.

Step 4: Classification Labeling. OMIBench problems in biology, chemistry, mathematics, and physics fall into two categories: open-ended and multi-choices. The combined Olympiad and high-stakes examination corpus covers a wide range of subfields, as shown in Figure 2. We first use GPT-4o to generate preliminary topic labels, and then manually assign final topic to ensure consistency and correctness across the corpus.

4 Main Experiments

4.1 Experiments Setup

We evaluate advanced open-source and closed-source LVLMs (see Appendix B for additional evaluation details). Each model generates answers using boxed-format ("\boxed{$\cdot$}") prompts, and open-source models are deployed on NVIDIA A800 or A100 GPUs. Temperatures are selected from $[0, 1]$. Model outputs are evaluated using exact-match accuracy and GPTScore, which assesses semantic equivalence under multimodal contextual constraints for open-ended answers (see Appendix C and Appendix J for more details on the metrics and their reliability). We report micro-averaged accuracy as the overall metric.

4.2 Main Results

Table 2 presents the overall experimental results, yielding two key findings:

OMIBench provides a more challenging evaluation framework than existing benchmarks. The highest-performing model (Gemini-3-Pro) achieves only 50.53% on OMIBench, substantially lower than on current benchmarks. This increased difficulty amplifies performance differences between models, enabling more precise capability comparisons.

Substantial gaps persist between leading closed- and open-source models, though model scale alone is insufficient. Gemini-3-Pro-Preview achieves about 15% higher accuracy than the best open-source models. However, GPT-4o, despite being closed-source and competitive on complex tasks, achieves accuracy only marginally above open-source models, suggesting that architecture and training strategies beyond parameter count determine performance on challenging benchmarks.

5 What's essential in OMIBench?

We analyze OMIBench in the context of existing multi-image benchmarks and isolate what makes its Olympiad-style multi-image reasoning distinct. More details are shown in Appendix D & E & F.

OMIBench needs deeper Olympiad-level cross-modal reasoning. To gauge Olympiad-level reasoning demands, we compare OMIBench with single-image OlympiadBench using the same LVLMs. Figure 3(a) shows only moderate Spearman correlation across models ($\rho = 0.614 < 0.7$), suggesting that multi-image inputs shift relative rankings even on similar Olympiad problems. Accordingly, Gemini-3.0-Pro drops from

Model	Biology		Chemistry		Mathematics		Physics		Total	
	ACC	Score	ACC	Score	ACC	Score	ACC	Score	ACC	Score
Instruction LVLMs										
InternVL3-1B (Zhu et al., 2025)	27.41	7.97	6.54	1.84	13.39	3.26	14.86	1.42	15.40	3.33
InternVL3-2B (Zhu et al., 2025)	31.47	20.72	12.35	11.98	17.88	7.67	14.81	9.43	18.57	11.42
Qwen2.5-VL-3B-Instruct (Bai et al., 2025b)	27.89	21.12	17.05	9.68	17.91	11.16	15.57	8.25	18.91	11.87
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	37.85	31.87	20.28	15.67	13.72	8.37	11.56	14.39	18.69	15.96
InternVL3-8B (Zhu et al., 2025)	43.57	33.07	15.80	16.13	23.30	9.30	19.53	12.74	24.71	16.04
InternVL3-14B (Zhu et al., 2025)	47.94	38.25	22.66	**20.74**	24.28	11.92	21.89	16.35	27.74	19.79
InternVL3-38B (Zhu et al., 2025)	50.92	43.03	13.36	17.97	24.35	15.81	**24.81**	16.75	27.74	21.63
InternVL3-78B (Zhu et al., 2025)	47.41	46.61	17.16	**20.74**	**27.30**	17.21	22.95	18.63	28.06	23.83
Qwen2.5-VL-32B-Instruct (Bai et al., 2025b)	48.21	48.61	22.93	19.82	24.01	21.40	23.54	**19.81**	28.28	25.80
Qwen2.5-VL-72B-Instruct (Bai et al., 2025b)	**51.79**	**53.39**	**23.45**	19.82	24.19	**27.67**	22.47	17.45	**28.76**	**27.99**
Long CoT LVLMs										
InternVL3.5-1B (Wang et al., 2025c)	32.46	17.93	14.29	6.91	15.12	4.65	23.58	8.02	21.56	8.62
Qwen3-VL-2B-Instruct (Bai et al., 2025a)	27.44	19.92	12.33	5.07	11.53	6.99	12.50	8.75	14.99	9.69
InternVL3.5-2B (Wang et al., 2025c)	34.66	25.10	16.13	13.82	20.00	11.86	22.64	12.74	23.00	14.98
InternVL3.5-30B-A3B (Wang et al., 2025c)	43.03	43.43	19.35	20.28	24.42	14.19	18.87	13.21	25.34	20.43
Qwen3-VL-4B-Instruct (Bai et al., 2025a)	43.03	36.65	17.05	11.06	27.91	23.72	18.40	13.92	25.95	20.95
InternVL3.5-14B (Wang et al., 2025c)	51.00	40.24	19.80	21.20	27.44	19.30	18.63	20.75	27.84	24.05
InternVL3.5-241B-A28B (Wang et al., 2025c)	52.58	45.42	**21.58**	**25.35**	31.16	20.47	19.58	21.23	29.94	26.25
InternVL3.5-8B (Wang et al., 2025c)	47.41	37.05	17.97	18.43	27.91	32.33	17.92	17.69	26.78	26.25
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	46.61	43.43	16.13	17.05	27.44	29.30	20.05	18.63	26.85	26.55
InternVL3.5-38B (Wang et al., 2025c)	49.40	41.83	20.28	22.12	22.33	25.12	24.53	23.58	27.84	27.31
Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025a)	48.51	48.61	13.29	12.90	31.42	32.33	20.02	20.99	28.03	28.59
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025a)	**60.41**	**63.20**	17.23	22.58	37.48	34.19	23.77	23.58	34.11	34.39
Qwen3-VL-32B-Instruct (Bai et al., 2025a)	57.62	58.57	14.09	20.74	**44.40**	**40.70**	**25.48**	**25.00**	**35.87**	**35.78**
Close-sourced LVLMs										
GPT-4o-mini (Hurst et al., 2024)	56.57	40.24	10.81	21.66	27.49	11.86	18.25	17.22	27.31	20.58
GPT-4o (Hurst et al., 2024)	60.96	53.00	15.67	22.58	29.69	15.49	17.10	18.92	24.05	24.88
Gemini-2.5-Flash (Comanici et al., 2025)	58.13	64.54	22.42	18.43	41.16	38.37	21.21	23.35	34.91	35.25
Gemini-2.5-Pro (Comanici et al., 2025)	59.53	66.14	22.88	23.96	42.28	53.49	22.90	31.84	36.16	44.10
OpenAI-o4-mini (OpenAI, 2025c)	51.41	57.37	23.67	**32.41**	41.28	56.61	19.03	35.38	33.19	45.97
GPT-5-mini (OpenAI, 2025b)	58.96	59.36	22.12	24.42	37.44	56.74	24.39	**43.63**	34.83	47.73
GPT-5 (OpenAI, 2025b)	**68.13**	62.55	23.96	29.03	39.30	56.51	20.52	40.80	36.23	48.11
Gemini-3-Pro-Preview (Google DeepMind, 2025)	64.51	**71.31**	**25.52**	25.35	**55.77**	**62.56**	**25.59**	38.92	**42.79**	**50.53**

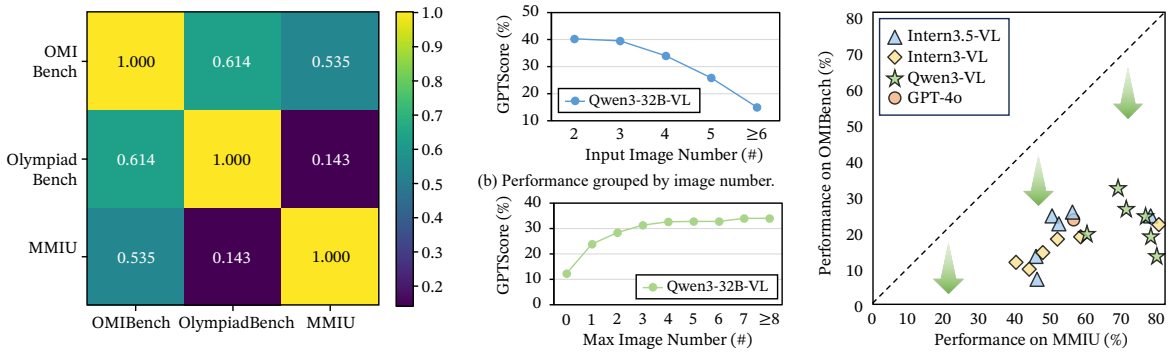
Table 2: Main results on OMIBench, where the **bold** content denotes the best performance. Here, " ": best performance, " ": second performance, " ": third performance. Rows are ordered by total average GPT-Score.

75.67% accuracy on OlympiadBench to 50.53% on OMIBench (>25% absolute), indicating the added difficulty of multi-image Olympiad reasoning. To further probe these demands, we sample 100 problems and rate rationales from o4-mini and Gemini-3.0-Pro. The human review finds logical errors in 46% of key steps, exposing a gap between fluent rationales and correct reasoning. This gap calls for stronger rationale generation to reach Olympiad-level reasoning depth.

OMIBench needs stronger awareness of the information flow across images. Generally, MMIU (Meng et al., 2024) targets basic multi-image understanding, whereas OMIBench demands complex reasoning across images. To confirm this, we compare model performance. As shown in Figure 3 (a), OMIBench shows moderate Spearman correlation with MMIU (< 0.7): it links to multi-image tasks but Olympiad-level difficulty alters model rankings. Beyond this, we further examine multi-image information integration. As shown in Figure 3 (b), single-image accuracy reaches 40%, dropping below 15% for inputs with $\geq$ 6 images. Restricting instances to one image (Figure 3 (c)) causes at least a 10% absolute performance drop versus using all images, revealing LVLMs' struggles with cross-image integration.

OMIBench needs combined Olympiad-level cross-image and cross-modal reasoning. As shown in Figure 3 (a), MMIU and OlympiadBench yield poorly aligned model rankings, whereas OMIBench is more consistent with both, suggesting that it more faithfully captures the joint demands of multi-image and Olympiad-style problems. The performance comparison in Figure 3 (d) further illustrates this distinction: MMIU primarily evaluates basic visual understanding (even the weakest models achieve >40% accuracy), whereas OMIBench requires substantially deeper reasoning (the best model reaches only around 40%). This pronounced performance gap highlights the increased difficulty of OMIBench and its value for stress-testing LVLMs on both multi-image integration and Olympiad-level reasoning.

Mistake Analysis. We further analyze GPTScore-annotated incorrect samples to



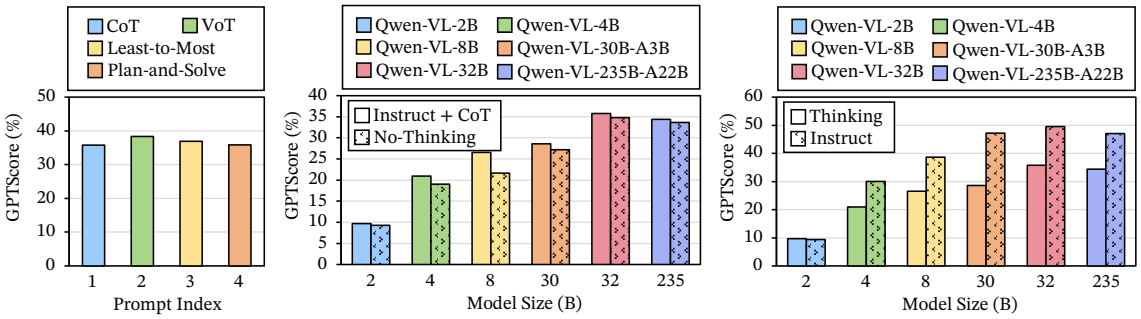
(a) The spearman correlation matrix of OMIBench, OlympiadBench, and MMIU benchmarks.

(b) Performance grouped by image number.

(c) Performance with different maximum numbers of input images.

(d) Comparison of performance and difficulty differences between MMIU and OMIBench.

Figure 3: The performance analysis for benchmark feature analysis and statistics.



(a) The performance of difference prompting strategies on Qwen3-VL-32B-Instruct.

(b) The performance comparison between CoT prompting and No-Thinking prompting strategies on Qwen3-VL-Instruction series.

(c) The performance comparison between Instruction and Thinking model variant on Qwen3-VL series.

Figure 4: Performance analysis on Long CoT Strategies. More details can be seen in Appendix G.

identify common failure modes, grouping them into three categories. As shown in Figure 7, visual perception failures account for 13% of errors, cross-image association failures for 29%, and logical reasoning fallacies for 41%. These distributions reveal persistent challenges in complex visual interpretation, multi-image integration, and logical consistency, highlighting the need for targeted advances to improve LVLM performance.

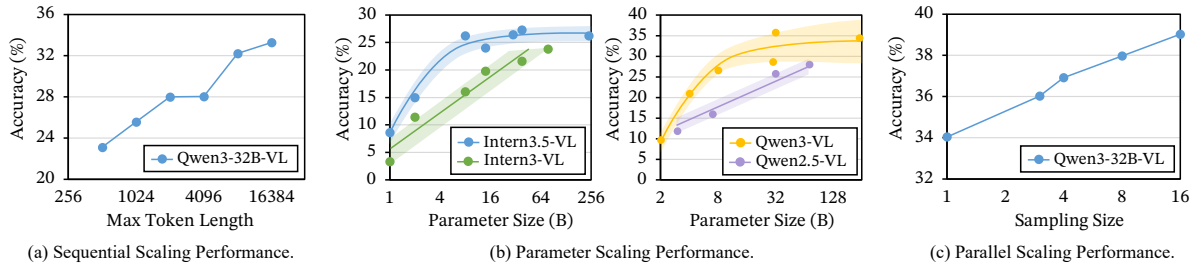
Cross-image reasoning remains the main bottleneck. To disentangle cross-image reasoning from confounders such as increased visual information, longer inputs, or OCR noise, we construct an *information-equivalent single-image* control by concatenating all images for each problem into a single composite image while keeping the question text and answer choices unchanged. This preserves the total visual and textual information, logical difficulty, input length, and OCR-related noise. Detailed results are provided in Appendix L. The information-equivalent single-image setting consistently outperforms the original multi-image setting, indicating that the multi-image *organization* itself, which requires models to autonomously align, filter, and integrate evidence across images, is the primary source of the observed performance gap rather than visual volume or OCR noise.

Human-expert baselines confirm the difficulty of OMIBench. To calibrate the absolute difficulty of OMIBench against human performance, we conduct an initial human-baseline study on a 52-problem subset. Human experts achieve above 80% accuracy, and trained non-experts exceed 57%, while the strongest current model (Gemini-3-Pro) reaches only 48.08%, leaving a gap of more than 30 points to experts and around 10 points even to trained non-experts. Additional results are provided in Appendix K.

6 How to improve on OMIBench?

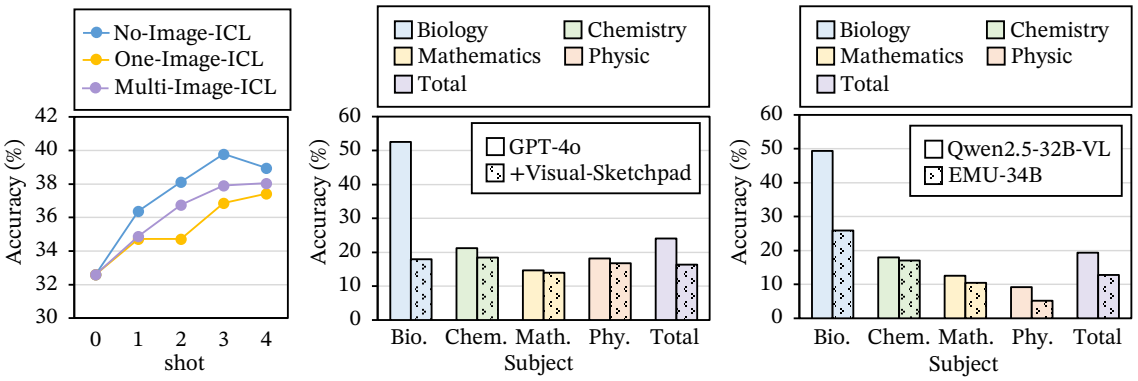
6.1 Can Long CoT Strategies Help?

Prompting methods that usually work do not significantly improve performance on OMIBench. Prior work has found that chain-of-thought (CoT) prompting can improve model performance on Olympiad-level reasoning tasks (Chen et al., 2025a; Li et al., 2025). We test whether these gains transfer to OMIBench by comparing widely used CoT prompting strategies, asking whether prompt en-



(a) Sequential Scaling Performance. (b) Parameter Scaling Performance. (c) Parallel Scaling Performance.

Figure 5: Performance analysis on 3-dimensional Test-Time Scaling paradigms. See Appendix H for more details.



(a) The Performance on In-Context Learning (b) The Performance on Tool-based Think-with-Image (c) The Performance on Internal Think-with-Image

Figure 6: The performance analysis on multimodal In-Context Learning and Thinking-with-Image paradigms.

gineering can narrow the gap or whether improvements remain marginal. As shown in Figure 4 (a), we observe no statistically significant differences across prompts, indicating that current CoT prompting yields only limited gains on OMIBench.

Long CoT thinking model variants significantly improve performance on OMIBench compared with prior benchmarks. We further evaluate a "no-thinking" prompting setting and find that both reasoning-oriented and non-reasoning LVLMs still struggle on OMIBench. As shown in Figure 4 (b), enabling or disabling explicit "thinking" yields little performance change. We then compare the "thinking" and "instruct" variants of Qwen3-VL to assess whether more advanced Long CoT paradigms improve multimodal reasoning. As shown in Figure 4 (c), the thinking variant outperforms the instruct variant on OMIBench by about 10%, substantially improving Olympiad-level performance. This gain is notably larger than that on earlier single-image reasoning benchmarks such as MathVista (<5%) (Bai et al., 2025a).

6.2 Can Test-Time Scaling Help?

To assess test-time scaling on OMIBench, we vary three orthogonal axes (parallel sampling, sequential reasoning depth, and model size), and obtain the following observations:

Sequential scaling remains effective in OMIBench. We test whether longer test-time reasoning improves performance by varying the maximum reasoning length. The resulting accuracy–reasoning-length curve characterizes the marginal returns (and potential saturation) of additional sequential inference computation. As shown in Figure 5(a), increasing the token budget from 512 to 16,384 yields a near-monotonic accuracy gain, indicating that such scaling remains effective on OMIBench, where MMIU and OlympiadBench performance becomes comparable.

Parameter scaling limits on OMIBench necessitate increased activated rather than total activable parameters. We evaluate test-time scaling versus parameter count using models from 1B to 235B parameters under identical inference. Unlike OlympiadBench and MMIU (Figure 5(b)), InternVL and QwenVL plateau on OMIBench at ∼25% and ∼35% GPTScore, respectively. This plateau is mainly associated with Mixture-of-Experts (MoE) models, where added capacity may not be activated at inference. In contrast, performance plotted against activated parameters shows a positive, near-linear trend, implying that multi-image Olympiad tasks require more concurrently

active parameters rather than a larger inactive pool.

Parallel scaling still works in OMIBench. With temperature set to 0.6 and self-consistency applied, accuracy in Figure 5(c) improves monotonically as k increases and exhibits approximately log-linear scaling in the number of samples. These results indicate that parallel scaling remains effective on OMIBench.

6.3 Can In-Context Learning Help?

Inspired by Multimodal In-Context Learning (MM-ICL) (Qin et al., 2024), we ask whether curated in-context examples can improve multimodal reasoning without parameter updates.

MM-ICL offers a limited logical connection of multi-image context. Figure 6(a) shows that multi-image ICL exceeds single-image ICL, yet both underperform No-Image-ICL, indicating that current LVLMs benefit more from textual context than from multi-image visual context for cross-image reasoning.

Multiple visual logical connections remain inferior to textual connections. This differs from prior single-image findings (Chen et al., 2024b; Qin et al., 2024), suggesting that MM-ICL can link multimodal context to some extent but remains weaker than text-based connection.

6.3.1 Can Thinking with Images Help?

Further, we examine whether models can effectively *think with images (TwI)* by generating or manipulating intermediate visual artifacts during reasoning, rather than producing only textual rationales (Cheng et al., 2025b,a; Su et al., 2025). Additional details are provided in Appendix I.

Current tool-based TwI works in single image domain but fails in OMIBench. Tool-based TwI with GPT-4o and VisualSketchpad still suffers substantial performance degradation on OMIBench in Figure 6 (b), indicating limited transfer from single-image tasks.

Current internal TwI works in single image domain but fails in OMIBench. Internal TwI with Emu-3.5-34B likewise incurs large performance degradation in Figure 6 (c) and even falls below Qwen-2.5-32B-VL, further underscoring its limited transfer from single-image tasks.

6.4 Can SFT or Tools Close the Gap?

Beyond prompting and test-time strategies, we ask whether (1) supervised fine-tuning (SFT) on existing multi-image data, or (2) integration of external visual tools, can substantially narrow the gap on OMIBench. Detailed numbers are reported in Appendix M (Tables 9 and 10).

SFT on existing multi-image data yields only limited gains. We fine-tune InternVL3.5-8B and Qwen3-VL-8B-Instruct on two representative multi-image SFT datasets: **CMMCoT** (Zhang et al., 2026) and **MMDU** (Liu et al., 2024b). Naive SFT on the simpler MMDU even slightly degrades performance ($\sim$0.5–1%), while SFT on the more reasoning-intensive CMMCoT consistently improves both backbones, but with average gains still below $\sim$1.5–2%. This shows that the field currently lacks training data tailored to *Olympiad-level* multi-image reasoning, and that closing the gap on OMIBench likely requires fundamentally new training resources.

External visual tools help only when paired with strong base models. We evaluate three external-tool integration frameworks, Visual Sketchpad (Hu et al., 2024), SlowPerception (Wei et al., 2024), and CogFlow (Chen et al., 2026b), on top of GPT-4o and GPT-5. When the base model is weaker (GPT-4o), tool augmentation generally *degrades* performance, as the model frequently fails to invoke visual tools correctly in multi-image scenarios. With a stronger backbone (GPT-5), tool augmentation yields modest additional gains in some subjects, but the overall improvement remains limited, indicating that reliably solving OMIBench requires stronger underlying intelligence to orchestrate tool use, rather than tools alone.

7 Related work

Competition Benchmarks. Several benchmarks assess LLM reasoning using competition-style questions. MATH (Hendrycks et al., 2021) and OlymMATH (Sun et al., 2025) collect high-school mathematics contest problems across diverse topics and require multi-step reasoning. AGIEval (Zhong et al., 2024) covers multiple subjects, including mathematics and physics, with questions drawn from competitive examinations. OlympiadBench (He et al., 2024) targets Olympiad-level mathematics and physics and provides per-problem rationales to enable deeper analysis of model reasoning. In addition, AIME (AIME, 2024, 2025) and AMC (AMC, 2023) use authentic contest problems that probe a broad range of mathematical skills and concepts.

Multimodal Benchmarks. To evaluate multimodal geometric reasoning, datasets such as Geometry3K (Lu et al., 2021), GeoQA (Chen et al., 2021), GeoQA+ (Cao and Xiao, 2022), and UniGeo (Chen et al., 2022) pair natural-language problems with diagrams (Ji et al., 2025). MathVista (Lu et al., 2024) and MathVerse (Zhang et al., 2024a) assess broader multimodal mathematical reasoning in vision-language models. ScienceQA (Lu et al., 2022), MMMU (Yue et al., 2024), and M3CoT (Chen et al., 2024b) extend evaluation across multiple disciplines (Jiang et al., 2025). OlympiadBench (He et al., 2024) targets advanced scientific reasoning through Olympiad problems in mathematics and physics, while Physics Big (Timur et al., 2024) provides a large-scale collection of physics competition problems for evaluating quantitative problem solving.

Recent work has also examined multi-image understanding in LVLMs. Mementos (Wang et al., 2024) studies narrative and temporal reasoning over image sequences; MC-Bench (Xu et al., 2025) evaluates multi-context visual grounding; and MANTIS (Jiang et al., 2024) offers interleaved vision–language instruction-tuning data with broad, general-domain difficulty. Although these benchmarks effectively test cross-image reference, alignment, and commonsense composition, they mainly emphasize narrative or perceptual integration rather than the tight semantic and quantitative coupling required by competition-level scientific problems.

Overall, existing benchmarks have advanced the evaluation of multimodal reasoning, but they still focus largely on single-image settings or relatively shallow multi-image perception, and seldom capture Olympiad-level multi-step reasoning. OMIBench addresses this gap by combining Olympiad-level scientific reasoning with evidence distributed across multiple interdependent images, requiring coherent image–image and image–text reasoning to derive the final answer.

8 Conclusion

This work introduced OMIBench, a large-scale multi-image Olympiad-level benchmark for evaluating LVLMs on complex multi-image reasoning. Experiments show substantial performance drops relative to single-image tasks, driven by failures in multi-image integration and grounded cross-modal reasoning. These findings establish multi-image reasoning as a central challenge and motivate advances beyond prompting.

Ethical Considerations

In this paper, we introduce OMIBench, a demanding multimodal benchmark for assessing mathematical and physical reasoning in current large models and future AGI systems. We outline the dataset construction pipeline, encompassing data collection from official sources only, OCR processing, cleaning, deduplication, and expert annotations.

Each problem includes rigorous annotations, with an evaluation script provided for reproducible model assessment. OMIBench thus supports advances in AI scientific reasoning. To ensure reproducibility and curb carbon-intensive redundant computation, we will release the dataset and scripts publicly. All experiments adhere to relevant model and data licenses.

Limitations

OMIBench still has several limitations in evaluation. First, some questions require open-ended textual reasoning, such as multi-part solutions or responses containing multiple valid scientific statements, and therefore cannot yet be evaluated fully reliably using symbolic tools such as SymPy. These cases still require model-based or human review. Second, even by GPTScore evaluation, it may still under-credit creative solutions or partially correct answers in open-ended settings. A further limitation concerns dataset construction. Due to the complexity and resource requirements of building a multimodal scientific reasoning benchmark, although OMIBench covers multiple disciplines and problem formats, it does not yet cover the full range of multi-image Olympiad-style reasoning found in real educational and scientific settings.

Acknowledgements

We gratefully acknowledge the support of the National Natural Science Foundation of China (NSFC) via grant 62236004, 62476073, 92570120 and 62306342. This work was supported by the Scientific Research Fund of Hunan Provincial Education Department (24B0001). This work was sponsored by the Excellent Young Scientists Fund in Hunan Province (2024JJ4070), the Science and Technology Innovation Program of Hunan Province under Grant 2024RC3024. This study was also funded by the Open Project of the Text Computing and Cognitive Intelligence Ministry of Education Engineering Research Center (No. TCCI250101).

References

AIME. 2024. American invitational mathematics examination (aime) aime 2024-i & ii.

AIME. 2025. American invitational mathematics examination (aime) 2025-i & ii.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, NM Anoop Krishnan, and Kevin Maik Jablonka. 2025. Probing the limitations of multimodal language models for chemistry and materials research. *Nature computational science*, pages 1–10.

AMC. 2023. American mathematics competitions.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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

Jie Cao and Jing Xiao. 2022. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523.

Qiguang Chen, Yantao Du, Ziniu Li, Jinhao Liu, Songyao Duan, Jiarui Guo, Minghao Liu, Jiaheng Liu, Tong Yang, Ge Zhang, et al. 2026a. The molecular structure of thought: Mapping the topology of long chain-of-thought reasoning. *arXiv preprint arXiv:2601.06002*.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025a. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Qiguang Chen, Libo Qin, Jiaqi Wang, Jinxuan Zhou, and Wanxiang Che. 2024a. Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. *Advances in Neural Information Processing Systems*, 37:54872–54904.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024b. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221, Bangkok, Thailand. Association for Computational Linguistics.

Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, et al. 2025b. Ai4research: A survey of artificial intelligence for scientific research. *arXiv preprint arXiv:2507.01903*.

Shuhang Chen, Yunqiu Xu, Junjie Xie, Aojun Lu, Tao Feng, Zeying Huang, Ning Zhang, Yi Sun, Yi Yang, and Hangjie Yuan. 2026b. CogFlow: Bridging perception and reasoning through knowledge internalization for visual mathematical problem solving. In *International Conference on Learning Representations (ICLR)*.

Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, et al. 2025a. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510*.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. 2025b. Comt: A novel benchmark for chain of multimodal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23678–23686.

Ziming Cheng, Binrui Xu, Lisheng Gong, Zuhe Song, Tianshuo Zhou, Shiqi Zhong, Siyu Ren, Mingxiang Chen, Xiangchao Meng, Yuxin Zhang, et al. 2025c. Evaluating mllms with multimodal multi-image reasoning benchmark. *arXiv preprint arXiv:2506.04280*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Hang Du, Jiayang Zhang, Guoshun Nan, Wendi Deng, Zhenyan Chen, Chenyang Zhang, Wang Xiao, Shan Huang, Yuqi Pan, Tao Qi, et al. 2025. From easy to hard: The mir benchmark for progressive interleaved multi-image reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 859–869.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Google DeepMind. 2025. Gemini 3: Technical report. Technical report. https://deepmind.google/.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can MLLMs reason in multimodality? EMMA: An enhanced multimodal reasoning benchmark. In *Forty-second International Conference on Machine Learning*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Jan Hrubes, Adam Tywoniak, Martin Balouch, Stanislav Chvíla, and Jan Hrabovsky. 2021. Chemistry race/-chemiklání: Team-based competition in chemistry. *Journal of Chemical Education*, 98(12):3878–3883.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Yiyan Ji, Haoran Chen, Qiguang Chen, Chengyue Wu, Libo Qin, and Wanxiang Che. 2025. Mpcc: A novel benchmark for multimodal planning with complex constraints in multimodal large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5188–5197.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. 2025. MME-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In *Forty-second International Conference on Machine Learning*.

Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Sreenivas Gollapudi, Dee Guo, et al. 2024. Remi: A dataset for reasoning with multiple images. *Advances in Neural Information Processing Systems*, 37:60088–60109.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. 2024a. Mibench: Evaluating multimodal large language models over multiple images. *arXiv preprint arXiv:2407.15272*.

Wentao Liu, Hanglei Hu, Jie Zhou, Yuyang Ding, Junsong Li, Jiayi Zeng, Mengliang He, Qin Chen, Bo Jiang, Aimin Zhou, et al. 2025. Mathematical language models: A survey. *ACM Computing Surveys*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024b. MMDU: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for LVLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Mikołaj Małkiński and Jacek Mańdziuk. 2025. Deep learning methods for abstract visual reasoning: A survey on raven's progressive matrices. *ACM Computing Surveys*, 57(7):1–36.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. 2024. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*.

OpenAI. 2025a. Evaluating AI's ability to perform scientific research tasks. OpenAI Blog. https://openai.com/index/frontierscience/.

OpenAI. 2025b. GPT-5 system card. Technical report. https://openai.com/.

OpenAI. 2025c. OpenAI o4-mini System Card. Technical report. https://openai.com/.

Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. 2024. What factors affect multimodal in-context learning? an in-depth exploration. *Advances in Neural Information Processing Systems*, 37:123207–123236.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *Advances in Neural Information Processing Systems*, 37:18695–18728.

Junhao Shen, Haiteng Zhao, Yuzhe Gu, Songyang Gao, Kuikun Liu, Haian Huang, Jianfei Gao, Dahua Lin, Wenwei Zhang, and Kai Chen. 2025. Semi-off-policy reinforcement learning for vision-language slow-thinking reasoning. *arXiv preprint arXiv:2507.16814*.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. 2025. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*.

Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2025. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. *arXiv preprint arXiv:2503.21380*.

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

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

Zaharov Timur, Konstantin Korolev, and Aleksandr Nikolich. 2024. Physics big.

Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, et al. 2025. Thinking with video: Video generation as a promising multimodal reasoning paradigm. *arXiv preprint arXiv:2511.04570*.

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, et al. 2025. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *arXiv preprint arXiv:2506.01713*.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2025a. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *The Thirteenth International Conference on Learning Representations*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. 2025b. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19541–19551.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025c. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. 2024. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 416–442.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025d. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Haoran Wei, Youyang Yin, Yumeng Li, Jia Wang, Liang Zhao, Jianjian Sun, Zheng Ge, Xiangyu Zhang, and Daxin Jiang. 2024. Slow perception: Let's perceive geometric figures step-by-step. *arXiv preprint arXiv:2412.20631*.

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

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind's eye of llms: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37:90277–90317.

Yunqiu Xu, Linchao Zhu, and Yi Yang. 2025. MC-Bench: A benchmark for multi-context visual grounding in the era of MLLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Huanjin Yao, Jiaxing Huang, Yawen Qiu, Michael K Chen, Wenzheng Liu, Wei Zhang, Wenjie Zeng, Xikun Zhang, Jingyi Zhang, Yuxin Song, et al. 2025. Mmreason: An open-ended multi-modal multi-step reasoning benchmark for mllms toward agi. *arXiv preprint arXiv:2506.23563*.

Fangchen Yu, Haiyuan Wan, Qianjia Cheng, Yuchen Zhang, Jiacheng Chen, Fujun Han, Yulun Wu, Junchi Yao, Ruilizhen Hu, Ning Ding, et al. 2025. Hipho: How far are (m) llms from humans in the latest high school physics olympiad benchmark? *arXiv preprint arXiv:2509.07894*.

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 Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. 2025. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*.

Yuheng Zha, Kun Zhou, Yujia Wu, Yushu Wang, Jie Feng, Zhi Xu, Shibo Hao, Zhengzhong Liu, Eric P Xing, and Zhiting Hu. 2025. Vision-g1: Towards general vision language reasoning with multi-domain data curation. *arXiv preprint arXiv:2508.12680*.

Guanghao Zhang, Tao Zhong, Yan Xia, Mushui Liu, Zhelun Yu, Haoyuan Li, Wanggui He, Fangxun Shu, Dong She, Yi Wang, and Hao Jiang. 2026. CMM-CoT: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. 2024a. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Yuan Zhang, Ming Lu, Junwen Pan, Tao Huang, Kuan Cheng, Qi She, and Shanghang Zhang. 2025. Chainv: Atomic visual hints make multimodal reasoning shorter and better. *arXiv preprint arXiv:2511.17106*.

Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. 2024b. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. 2024. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Appendix

A Data Construction Details

OMIBench was constructed via a rigorous multi-stage pipeline that aggregates high-quality, reasoning-intensive problems from diverse global sources. Key steps are outlined below.

A.1 Details of Data Collection

Our collection strategy prioritized three criteria: (1) difficulty, requiring multi-step chain-of-thought reasoning (Chen et al., 2026a); (2) diversity, spanning text-only and multimodal contexts; and (3) authority, drawing on established competitions and vetted academic benchmarks.

The data acquisition process followed two primary streams: (1) **Original contest archival.** We manually collected official problem sets from international and national Olympiad archives (e.g., ICHO, IMO, CPHO) and regional tournaments (e.g., ASO, CUPT). For sources available only as PDFs, we used optical character recognition (OCR) tools specialized for scientific notation (e.g., Mathpix) to extract textual content and LaTeX equations, while high-resolution figures were cropped and preserved for multimodal evaluation. (2) **Integration of existing benchmarks.** To broaden coverage, we adapted images and questions from recent open-source benchmarks, including OlympiadBench (He et al., 2024), Mv-MATH (Wang et al., 2025b), and EMMA (Hao et al., 2025). For these sources, we standardized the data format to unify the representation of questions, images, and ground-truth answers across disciplines.

Table 3 lists the sources for each discipline and the rationale for their inclusion.

Biology: Data were sourced from the *Australian Science Olympiads (ASO)* and the *Indian National Biology Olympiad (INBO)*. These contests feature long-context problem statements that require synthesizing facts with experimental data, testing LLMs' capacity for evidence-based reasoning.

Chemistry: This subset combines classical analytical problems from the *International Chemistry Olympiad (ICHO)* with interdisciplinary challenges. We include questions from the *British Biology Olympiad (BBO)* that overlap with biochemistry, alongside the *Chemistry Race* and modified *EMMA* instances, to diversify problem formats from organic synthesis planning to physical chemistry calculations.

Suject	Source
Biology	Australian Science Olympiads (ASO) [*] Indian National Biology Olympiad (INBO) [†]
Chemistry	British Biology Olympiad (BBO) [‡] International Chemistry Olympiad (ICHO) [§] Chemistry Race (Hrubes et al., 2021) EMMA (Hao et al., 2025)
Mathematics	International Mathematical Olympiad (IMO) [¶] Chinese Mathematical Olympiad (CMO) [‖] European Girls' Mathematical Olympiad (EGMO) [**] OlympiadBench (He et al., 2024) Mv-MATH (Wang et al., 2025b)
Physics	China Undergraduate Physics Tournament (CUPT) [††] Chinese Physics Olympiad (CPHO) [‡‡] OlympiadBench (He et al., 2024) Physics-Big (Timur et al., 2024)

Table 3: Source statistics of the constructed OMIBench dataset.

Mathematics: To capture high-level symbolic reasoning, we aggregate problems from the *IMO*, *CMO*, and *EGMO*. Supplemented by *Mv-MATH* and *OlympiadBench*, this component is heavily multimodal, with plane geometry diagrams and function graphs that require aligning visual perception with symbolic deduction.

Physics: We incorporate problems from the *China Undergraduate Physics Tournament (CUPT)* and the *Chinese Physics Olympiad (CPHO)*, selected for their emphasis on physical intuition and complex modeling of real-world phenomena. Data from *Physics-Big* provide additional coverage of mechanics and electromagnetism.

A.2 Details of Format Conversion and Data Selection.

To further ensure dataset quality, we apply additional filtering and translation checks beyond the main pipeline described in the paper.

Format Conversion. First, we convert all PDF files to Markdown using Mathpix OCR and normalize them to a "Question–Rationale (if available)–Answer" schema.

Format Filtering. Team members review each item to confirm that (i) the problem statement is complete and legible, and (ii) the answer is well defined and unambiguous. Items with severe OCR errors, missing essential information (e.g., truncated questions or missing answers), duplicated content, or unresolved formatting issues (such as unreadable formulas or diagrams) are removed from the final set.

Difficulty Curation. For benchmarks spanning multiple difficulty levels, experienced competition participants further curate the pool by excluding

trivial, overly domain-specific, or stylistically inconsistent problems, so that the remaining items better align with the intended reasoning skills and difficulty.

Multilingual Translation and Verification. For multilingual components, all non-English questions are first translated into English using Google Translate. Human annotators fluent in English then verify and, when necessary, correct the translations to preserve the original semantics, mathematical notation, and any subtle constraints or assumptions. During this process, annotators flag and resolve ambiguities (e.g., multiple plausible interpretations of a term or condition), and any items whose meaning cannot be reliably disambiguated are discarded.

A.3 Details of Rationale Annotations

Most existing competition-style datasets provide only final answers or brief solution sketches, which are insufficient for analysing model reasoning behaviours. To address this limitation, we construct expert-verified, step-by-step rationales for each problem using a two-stage pipeline that combines LLM generation with careful human verification.

A.3.1 LLM-Assisted Rationale Generation

We first use Gemini-2.5-pro-thinking to generate multiple candidate solutions for each problem.

Model Prompting: For every problem, the input to the model includes: the problem statement; all auxiliary information required to solve the problem (e.g., provided figures, tables, or input–output formats), when available. Concretely, we use a prompt template that instructs the model to "Let's think step-by-step!" and enforces a standardised answer format. Each candidate must include both an explicit reasoning trace and an explicit final answer. Specifically, the model prompt is structured as follows:

LLM Rationale Generation Prompt-P1

You are an expert problem solver. Given the following problem statement and any auxiliary information, provide a clear step-by-step solution that leads to a final answer. Explain each step of your reasoning, and format your response as: {Reasoning: ... Final Answer: ...}.

LLM Rationale Generation Prompt-P2

Problem Statement: `[Insert problem statement here]`
Auxiliary Information: `[Insert figures, tables, or other relevant data here]`

For each problem, Gemini-2.5-pro-thinking samples up to 16 candidate solutions at moderate temperature (0.6) to balance diversity and coherence.

Rationale Filtering: After generation, we automatically filter candidates, retaining only those whose final answer matches the known correct answer under the official evaluation protocol. If at least one such candidate exists, all matching candidates are kept as potential rationales. If none of the 16 candidates is correct, we append the corresponding reference solution sketch or official final answer to the prompt as "**Reference Solution:** `[Insert reference solution here]`" to trigger generation of a correct rationale. To avoid trivial rationales that merely restate the answer, we apply simple automatic heuristics to discard degenerate candidates, such as explanations shorter than 50 tokens or responses that only paraphrase the question or the reference answer without intermediate steps.

Overall, this LLM-assisted stage reduces human annotation effort by about 60% relative to fully manual authoring, while still providing a rich pool of candidate rationales for each problem.

A.3.2 Human Verification and Refinement

Based on the filtered candidate rationales, experienced annotators are required to verify and refine all LLM-generated rationales. Annotators are graduate-level students or domain experts with strong backgrounds in mathematics, computer science, or related fields, and receive detailed written guidelines and training examples before starting annotation.

Annotator Recruitment and Training. Annotators were selected from graduate students with prior experience in machine learning or related quantitative fields. Before using the annotation platform, they completed a written tutorial on task definitions, edge cases, and examples of acceptable and unacceptable annotations; passed a 30-item calibration test spanning diverse problem types and difficulty levels with at least 80% agreement with an expert gold standard; and signed a code-of-conduct and confidentiality agreement on responsible data han-

dling. We require all annotation experts to have at least one professional competition experience and be trained to at least a bronze medal level.

Annotation Guidelines. Each annotator uses an interface that displays the problem, the ground-truth final answer, the official or reference solution (when available), and one or more LLM-generated candidate rationales. When no rationale is correct, the interface instead shows three candidate incorrect rationales. The detailed guidelines to instruct annotators are as follows:

Annotator Guidelines-P1

These guidelines define how annotators should evaluate and edit solution rationales for reasoning tasks (e.g., mathematical or logical problems). The goal is to ensure that all accepted rationales are correct, complete, and stylistically consistent.

For every candidate rationale, annotators should perform the following checks:

1. **Conceptual correctness.**
   - Verify that all major inference steps are logically valid.
   - Check that all problem constraints are correctly interpreted.
   - Ensure that the rationale does not introduce unwarranted assumptions or ignore conditions.
2. **Computational correctness.**
   - Check all algebraic manipulations, arithmetic calculations, and case analyses for errors.
   - Confirm that intermediate results are correct and consistent across steps.
3. **Completeness of reasoning.**
   - Ensure that the solution does not skip non-trivial steps.
   - Include necessary intermediate results (e.g., key substitutions, simplifications, or case splits).
   - Check that each step is connected to the previous one and that the overall argument is coherent.

Annotator Guidelines-P1

4. **Consistency with the final answer.**
   - Confirm that the reasoning actually leads to the stated final answer.
   - Ensure that there is no mismatch between intermediate conclusions and the final result.
   - Verify that the final answer is stated explicitly, unambiguously, and in the required format.
5. **Clarity and style.**
   - Assess whether the rationale is easy to follow for a competent reader.
   - Avoid unnecessary repetition, irrelevant digressions, and overly verbose explanations.
   - Check adherence to the agreed notation, terminology, and formatting conventions.

After applying the evaluation criteria above, annotators should choose exactly one of the following actions for each candidate rationale:

1. **Accept with minor edits.** Use this option when the rationale is fundamentally correct and complete, but has small issues such as:
   - Minor wording problems (e.g., awkward phrasing, ambiguous pronouns).
   - Slightly unclear transitions between steps.
   - Cosmetic inconsistencies in notation, symbols, or formatting.

   In this case, annotators should directly edit the text to correct these minor issues without changing the core reasoning.
2. **Substantive revision.** Use this option when the rationale contains the correct core ideas but has more serious local problems, such as:
   - Missing but recoverable intermediate steps.
   - Redundant detours, digressions, or unnecessary case splits.
   - Local mistakes (e.g., a computation error in one step, a minor mislabeling) that can be fixed without changing the overall solution strategy.

Annotator Guidelines-P3

In this case, annotators should:

- Fix computational and logical errors.
- Fill in missing, non-trivial steps.
- Reorganize the structure for better flow, and remove superfluous parts.

3. **Complete rewrite.** Use this option when the rationale is not salvageable as a whole, for example:

   - The overall reasoning is conceptually flawed or contradicts the problem statement.
   - The logical structure is inconsistent or self-contradictory.
   - The explanation is so unclear, disorganized, or confusing that repairing it would be harder than rewriting.

   In this case, annotators should:

   - Discard the rationale as the primary solution.
   - Write a new rationale from scratch that is correct, complete, and clear.
   - Optionally reuse any locally correct insights from the original rationale (e.g., a correct formula, an accurate sub-case analysis), but only if they fit naturally into the new solution.

Annotators should choose the **least intrusive** action that yields a high-quality rationale. If in doubt between "Substantive revision" and "Complete rewrite," prefer "Complete rewrite" when the existing structure significantly impedes clarity or correctness.

Annotators should apply these guidelines consistently across all examples to ensure uniform quality and style in the final dataset.

A.4 Details of Quality Control

This section provides a detailed description of the quality control procedures summarized in the main paper, including the dual-review protocol (5$ per sample), weekly random sampling and regression testing, metric definitions, and the closed-loop feedback process for guideline refinement and model retraining.

A.4.1 Dual-Review Annotation Workflow

To ensure dataset quality, each problem instance (comprising the statement, associated images, and solution) underwent a dual-review protocol consisting of primary annotation followed by an independent audit. (1) A primary annotator first verified the content and assigned task-specific metadata; (2) subsequently, a blinded auditor assessed problem well-posedness, text–image alignment, and solution validity. The auditing interface presented problems in their final form, allowing reviewers to rate quality on a 5-point Likert scale and flag specific defects, such as ambiguous statements, misleading visual content, or erroneous solutions.

A.4.2 Disagreement Resolution and Escalation

Disagreements on critical fields, such as solution correctness or label assignment, triggered an automatic escalation to a senior reviewer. These discrepancies were identified via logical inconsistencies (e.g., opposing validity flags) or rating divergences of at least 2 points. A senior expert with domain experience then examined the full annotation history to issue a binding decision: retaining the primary annotation, adopting the auditor's revision, or rewriting the content entirely with a supporting rationale.

A.4.3 Weekly Random Sampling Procedure

To complement per-example dual review, we implemented weekly random sampling for quality assurance. Each week, 5% of annotated or modified examples were randomly selected using stratified sampling across problem type, difficulty level, and source (newly created vs. revised). Senior reviewers blind to original annotator identities re-evaluated these samples.

Finally, these strategies enabled the estimation of residual error rates and enforced strict quality control. The kappa value of our annotation correctness is close to 0.86, indicating good annotation quality. Table 2 summarizes the final coverage statistics and rationale distributions.

A.5 Details of Classification Labeling

Taxonomy construction. To ensure consistent topic annotations across domains, we build a unified three-level taxonomy for OMIBench: *domain* (biology, chemistry, mathematics, physics), *subfield* (e.g., algebra, combinatorics, organic chemistry, mechanics), and *fine-grained topic* (e.g., polynomial inequalities, graph coloring, nucleophilic substitution).

Fine-Grained Topic Annotation. Because the fine-grained topics are unknown a priori, we first prompt GPT-4o to perform open-ended topic analysis for each sample and to generate classification labels in the form {CLASS: X}. Specifically, the prompts are as follows:

Fine-Grained Topic Annotation Prompt

You are an expert annotator for multi-image, multi-discipline Olympiad-level problems. Your task is to assign the most specific subcategory label to each problem.
Each sample may contain:

- One or more images (diagrams, experiment setups, screenshots, etc.).
- Optional text (problem statement, description, notes).

Annotation rules:

- Choose the label by the core method and main concept, not by surface story.
- Always pick the most fine-grained label available (do not annotate coarse-grained subjects like "Optics").
- If it mixes multiple topics, use "#" as a divider.
- Use all images and text jointly; visual cues (rays, lenses, optical axes, diagrams) are as important as text.

Your output must be exactly one label string by {CLASS: X}. Do not output explanations or reasoning.

Subfield Annotation. We first use GPT-4o (Hurst et al., 2024) to assign preliminary fine-grained topic labels, and then aggregate these labels into subfield-level categories, following our taxonomy, via K-means clustering on RoBERTa (Liu et al., 2019) embeddings. For each subject area, we then construct a list of valid subfield labels from the clustering results and prompt GPT-4o to select the most appropriate subfield for each previously predicted fine-grained topic. For example, in mathematics, if the fine-grained topic is "Integral Calculation Function Area," the corresponding subfield is "Area, Perimeter, Ratios." The prompt template is as follows:

Subfield Annotation Prompt-P1

You are an expert annotator for multi-image, multi-discipline Olympiad-level problems.

Subfield Annotation Prompt-P2

Your task is to assign the most specific subcategory label to each problem.
`[Domain name]`
Classification Labels:

- `[Fine-grained topic list]`

Your output must be exactly one label string by {CLASS: X} from a fixed label list. Do not output explanations or reasoning.

Manual Verification and Correction. After automatic pre-labeling, all problems undergo manual review by expert annotators with relevant domain expertise. Annotators may retain the GPT-4o label, modify the topic within the same domain, or revise both domain and topic. Annotators are encouraged to propose taxonomy changes when they observe repeated, systematic mismatches between available topics and the actual problem content.

B Detailed Main Experiment

B.1 Model Inference & Evaluation Setting

In our main experiments, we evaluate a set of large vision-language models (LVLMs) on OMIBench in the zero-shot setting. The evaluated models include InternVL3 (Zhu et al., 2025), Qwen2.5-VL (Bai et al., 2025b), InternVL3.5 (Wang et al., 2025c), Qwen3-VL (Bai et al., 2025a), GPT-4o (Hurst et al., 2024), Gemini-2.5 (Comanici et al., 2025), OpenAI-o4-mini (OpenAI, 2025c), GPT-5 (OpenAI, 2025b), and Gemini-3 (Google DeepMind, 2025). To ensure fair comparison, we standardize input prompts across models, adapting only the minimal syntax or special tokens required by each interface. The prompt specifies the task description, any associated images, and the required output format. An example prompt template is shown below:

Chain-of-Thought Prompting-P1

Please reason step by step, and then provide the final answer in the exact format: "\boxed{ANSWER}".

[Question]
problem_text0 [IMAGE0] problem_text1 [IMAGE1] ... problem_textn

Chain-of-Thought Prompting-P2

[Choices] # only for multiple choices problems
A. option_A
B. option_B ...
Let's think step-by-step!

B.2 Rationale Quality Evaluation

Model Analyses We leverage the advanced reasoning capabilities of GPT-4o to assess the quality of generated rationales. Specifically, we employ two distinct sets of prompts to evaluate the intrinsic quality of the rationales and their alignment with human annotations, respectively. To assess the intrinsic quality of the rationales, we utilize a 5-point Likert scale, prompting the model to quantify the coherence and logical validity of the reasoning process. The specific prompts are detailed below:

Rationale Quality Evaluation Prompt-P1

You are given a question, a model answer, and a rationale (the step-by-step explanation or reasoning produced by the model). Your task is to evaluate the quality of the rationale, **not** the quality of the final answer itself.
Please read the rationale carefully and rate it along the following dimensions.

1. Logical correctness

Does the rationale follow a sound and coherent line of reasoning?

- 5 – Fully correct: Each step is logically valid, with no contradictions or clear mistakes.
- 4 – Mostly correct: Overall reasoning is sound, with only minor issues that do not affect the main conclusion.
- 3 – Partially correct: Some important steps are correct, but there are noticeable gaps or errors.
- 2 – Mostly incorrect: Reasoning is largely flawed or inconsistent, with only a few correct fragments.
- 1 – Completely incorrect: The rationale is illogical, self-contradictory, or entirely wrong.

Rationale Quality Evaluation Prompt-P2

2. Faithfulness to the answer

Is the rationale genuinely explaining how the given answer is (or would be) derived, instead of being disconnected or made-up?

- 5 – Fully faithful: The rationale clearly and directly supports the given answer. No hallucinated steps are needed to justify the answer.
- 4 – Mostly faithful: Slight mismatches, but the core reasoning is aligned with the answer.
- 3 – Partially faithful: Some parts support the answer, but other parts are irrelevant or inconsistent.
- 2 – Weakly faithful: The rationale only loosely relates to the answer, or relies heavily on speculation.
- 1 – Not faithful: The rationale does not explain the answer at all, or contradicts it.

3. Use of information

Does the rationale appropriately use the information provided in the input (question, context, passage, etc.)?

- 5 – Excellent: Uses all relevant information, does not ignore key facts, and does not add unsupported facts.
- 4 – Good: Uses most important information, with minor omissions or slightly extra but harmless details.
- 3 – Fair: Uses some relevant information, but misses several important points or includes noticeable unsupported content.
- 2 – Poor: Rarely uses the provided information or relies heavily on invented details.
- 1 – Very poor: Almost no connection to the provided information.

4. Clarity and readability

Is the rationale clear, easy to follow, and understandable to a careful reader?

- 5 – Very clear: Well-structured, concise, and easy to follow step by step.

Rationale Quality Evaluation Prompt-P2

- 4 – Clear: Mostly easy to understand, with only minor awkward wording or small jumps.
- 3 – Moderately clear: Understandable overall, but contains confusing segments or disorganized structure.
- 2 – Unclear: Hard to follow, with long, tangled, or repetitive explanations.
- 1 – Very unclear: Nearly impossible to understand the reasoning.

5. Level of detail

Is the rationale appropriately detailed for explaining the answer?

- 5 – Ideal detail: Includes all key steps, neither too brief nor overly long; no crucial step is skipped.
- 4 – Slightly off: Missing a minor step or slightly verbose, but still adequate.
- 3 – Mixed: Some important steps covered, but either too high-level or too verbose.
- 2 – Inadequate: Too short and high-level, or extremely verbose without adding real value.
- 1 – Very inadequate: Provides almost no meaningful explanation, or is overwhelmingly long and unfocused.

Overall rationale quality

After rating each dimension, give an **overall score** for the rationale from 1 to 5, considering all aspects together:

- 5 – Excellent rationale
- 4 – Good rationale
- 3 – Acceptable rationale
- 2 – Poor rationale
- 1 – Very poor rationale

Please base your judgment on the rationale text itself. Do not penalize a rationale just because the final answer is wrong, as long as the reasoning process is internally coherent and properly uses the provided information. Your output must be exactly one label string by {"correctness": NUMBER, "faithfulness":

Rationale Quality Evaluation Prompt-P3

NUMBER, "information-usage": NUMBER, "clarity": NUMBER, "detail": NUMBER, "overall": NUMBER}. Do not output explanations or reasoning.

Moreover, to evaluate the alignment between model-generated rationales and human-annotated ones, we prompt GPT-4o to compare the two texts and rate their similarity on a 5-point scale. The specific prompt is as follows:

Rationale Alignment Evaluation Prompt-P1

You are an expert evaluator specializing in natural language understanding and reasoning chains. Your task is to assess how well a model-generated rationale aligns with a human-annotated reference rationale.
Now, please evaluate the alignment between the **Model Rationale** and the **Human Reference** based on the provided **Question**. Focus on the underlying logic, the sequence of reasoning steps, and the key evidence used. Do not penalize for differences in writing style or length, provided the core logic remains identical.

Input Data:
[Question]
`[Question Content]`

[Human Reference]
`[Human Annotated Rationale]`

[Model Rationale]
`[Model Predicted Rationale]`

Scoring Criteria (1-5 Likert Scale):

- **5 (Perfect Alignment):** The model rationale uses the exact same logic, key evidence, and reasoning steps as the human reference. Differences are purely stylistic.
- **4 (High Alignment):** The model rationale captures all key logical points of the human reference but may include minor, non-contradictory extra details or slightly different sequencing.

Rationale Alignment Evaluation Prompt-P2

- **3 (Partial Alignment):** The model captures the main conclusion and primary evidence but misses a subordinate step or uses slightly different reasoning to arrive at the same result.
- **2 (Low Alignment):** The model rationale arrives at the correct answer but uses significantly different logic or misses critical evidence present in the human reference.
- **1 (No Alignment):** The model rationale contradicts the human reference, uses fallacious logic, or fails to address the specific constraints mentioned in the human text.

Output Format:
Provide your response in the following JSON format:

```
```json
{
  "justification":
  "[Brief rationale]",
  "score": [Integer 1-5]
}
```
```

Human Analyses To rigorously quantify the discrepancy between surface-level coherence and deep logical correctness in LVLMs, we conducted a fine-grained human evaluation. Similar to Appendix C.3, we randomly sampled 100 instances from the model outputs. Unlike coarse-grained scoring, our evaluation required expert annotators to inspect the generated rationale step-by-step and identify the *root cause* of the first fatal error encountered.

To decouple visual perception capabilities from reasoning engines, we developed a specific error taxonomy comprising five distinct categories: (1) **Visual Perception Failures**, where the model misinterprets explicit visual semantics; (2) **Cross-Image Association Failures**, indicating an inability to synthesize information across multi-view inputs; (3) **Logical Reasoning Fallacies**, covering invalid deductions and calculation errors despite correct perception; (4) **Instruction Comprehension Biases**, regarding format or constraint violations; and (5) **Other** for hallucinations or uncategorized failures. This taxonomy allows us to diagnose whether performance bottlenecks stem from the vision encoder, the cross-modal alignment, or the LLMs' reasoning core.

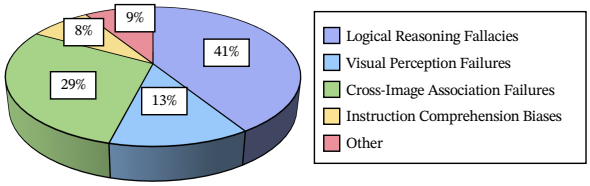


Figure 7: Distribution of different reasoning error types labeled by human.

Annotator Guidelines-P1

You are given a multi-image Olympiad-level problem and a model-generated rationale. Your task is to verify the correctness of the solution. If an error is found in a logical step, classify it into exactly one of the following five categories based on its primary cause.

1. Visual Perception Failures The model incorrectly recognizes or misses explicit visual information within an image. This includes misreading text, misidentifying geometry/objects, or failing to perceive attributes like color or position.

2. Cross-Image Association Failures The model fails to synthesize or track information across multiple images. This includes errors in understanding temporal sequences, matching objects between different views, or aggregating data from separate image panels.

3. Logical Reasoning Fallacies The model correctly perceives the visual data but fails in the subsequent reasoning process. This includes invalid deductions, calculation errors, misapplication of theorems, or flawed causal logic.

4. Instruction Comprehension Biases The model fails to adhere to specific constraints provided in the text prompt. This includes ignoring format requirements, violating negative constraints (e.g., "do not use calculus"), or misunderstanding the core question.

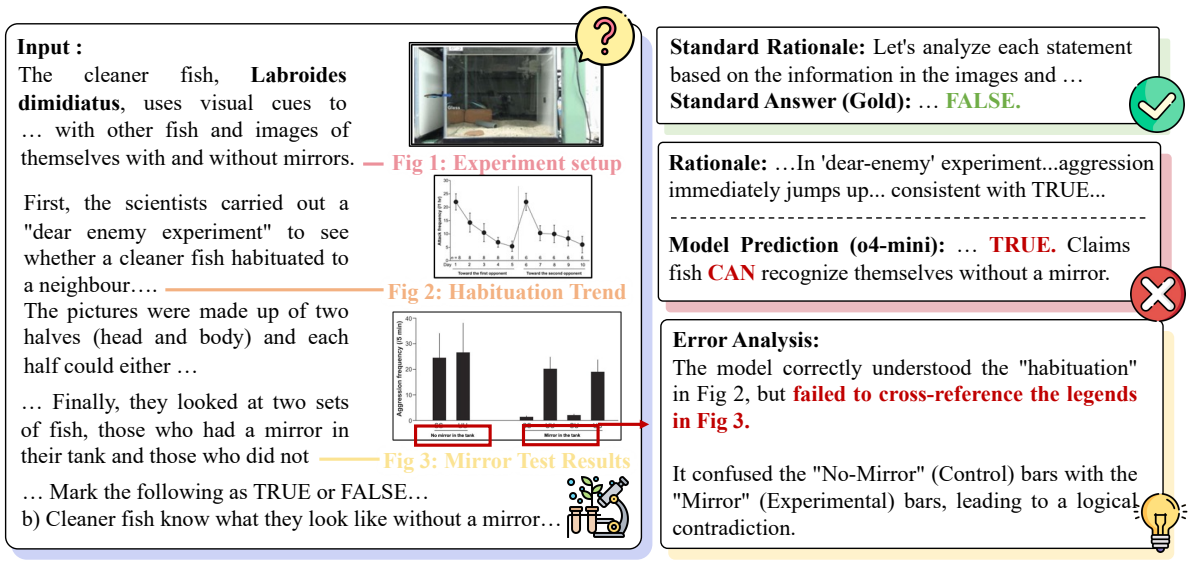


Figure 8: Analysis of reasoning error examples on o4-mini.

Annotator Guidelines-P2

5. Other Any error that does not fit the above categories. This includes hallucinating constraints that are absent from both text and images, or generating unintelligible content.

Case Studies of Reasoning Errors As illustrated in Figures 8 to 15, we present representative examples of reasoning errors identified in our human evaluation. These cases highlight common pitfalls in LVLM reasoning, such as multi-image information flow confusion, misapplication of physical laws, incorrect geometric interpretations, and flawed logical deductions.

C Details of Accuracy and GPTScore

C.1 Matching accuracy

For each OMIBench input x_i with gold answer y_i and the model produces prediction $\hat{y}_i$. Matching accuracy is computed at instance level:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[\text{match}(\hat{y}_i, y_i)\big], \qquad (5)$$

where N denotes total evaluated instances across all tasks.

Specifically, the matching function $\text{match}(\cdot,\cdot)$ applies consistent normalization across heterogeneous tasks:

Text normalization. Both $\hat{y}_i$ and y_i are lowercased, whitespace is trimmed and collapsed to single spaces, and common punctuation (periods, commas, question marks, trailing colons/semicolons) is stripped from English segments when not part of alphanumeric tokens.

Textual or equation answers. Matches are determined by longest common subsequence (LCS) ratio:

$$\text{match}(\hat{y}_i, y_i) = \begin{cases} \frac{\text{Len}(\text{LCS}(\hat{y}_i, y_i))}{\max(\text{Len}(\hat{y}_i), \text{Len}(y_i))} & \text{if ratio} \geq \alpha \\ 0 & \text{otherwise} \end{cases}, \qquad (6)$$

where $\text{LCS}(\cdot,\cdot)$ denotes the longest common subsequence, and $\text{Len}(\cdot)$ represents the length of given sequence. In our experiments we set $\alpha = 75\%$.

Numeric answers. For numeric references, we parse both $\hat{y}_i$ and y_i into floating-point numbers after removing units and non-numeric suffixes. The match accuracy is calculated as:

$$\text{match}(\hat{y}_i, y_i) = 1 \quad \Leftrightarrow \quad |\hat{v}_i - v_i| \leq \epsilon, \qquad (7)$$

where v_i and $\hat{v}_i$ enote parsed gold and predicted values, respectively, and $\epsilon = 10^{-4}$. Unparseable inputs default to normalized string matching.

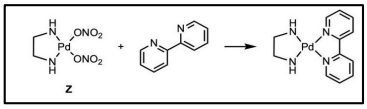
Multiple-choice questions. For predefined options, we extract predictions from model outputs via regex matching of boxed notation (e.g., $\backslash\text{boxed}\{A\}$). A prediction is deemed correct if the extracted symbol matches either the gold label directly or, when unavailable, the corresponding option text.

Unless otherwise specified, we report micro-averaged accuracy across all evaluated tasks.

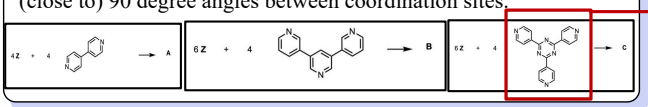
C.2 Model-based GPTScore evaluation

Beyond exact matching accuracy, we employ GPTScore, a semantic evaluation metric that captures partial correctness and alignment between

Input:
Metal complex **Z** can react with pyridine-based ligands to generate interesting structures. A simple 'chelating' ligand (the name derived from the Greek for claw), reacts with **Z** in a 1:1 ratio as shown, forming a new metal complex.



When the nitrogen atoms in the (poly)pyridine ligand do not chelate, very interesting structures can be formed. Write down/sketch the shape of the molecule formed when these ligands react with **Z**, in the stoichiometries indicated. You are not required to draw the organic ligands in full.
Hint: in all cases, palladium retains a square planar geometry with (close to) 90 degree angles between coordination sites.



Standard Rationale: To solve this problem, I need to analyze how each polypyridine ligand…
Standard Answer (Gold): … Octahedron.

Rationale: … each Pd remains square-planar... giving a **trigonal bipyramidal** geometry for C...

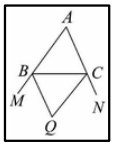
Model Prediction (o4-mini): … **Trigonal Bipyramidal.**

Error Analysis:
The model successfully identified the chemical components (Pd centers), but **failed the "Mental Folding" process from 2D schematics to 3D space.** Instead of recognizing the panel-folding logic leading to a closed Octahedron, it hallucinated an open or lower-symmetry Trigonal Bipyramidal structure.

Figure 9: Analysis of reasoning error examples on o4-mini.

Input:

Fig 1: Basic Model



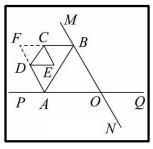


Fig 2: Variation Application

(1) **Basic Model:** As shown in Figure 1, $\angle CBM$ and $\angle BCN$ are exterior angles of ΔABC, and the bisectors of $\angle CBM$ and $\angle CBM$ intersect at point O. Please write the quantitative relationship between $\angle BOC$ and $\angle A$, and explain the reason.
(2) **Variation Application:** As shown in Figure 2, given that AB is not parallel to CD, AD and BC are the angle bisectors of $\angle BAP$ and $\angle ABM$ respectively, and DE and CE are the angle bisectors of $\angle ADC$ and $\angle BCD$ respectively.
… (2) If $AP \parallel DE$ and $BM \parallel CE$, find the degree measure of $\angle POM$.

Standard Rationale: … applies angle summation and bisector properties step-by-step…
Standard Answer (Gold): …65°.

Rationale: … hence their angles sum to 180°... The angle implies a fixed relationship of 90 + A …

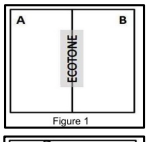
Model Prediction (o4-mini): 50°

Error Analysis:
The model exhibits "Heuristic Shortcutting". Instead of reasoning from geometric axioms for the complex Fig 2, it hallucinated a non-existent theorem based on visual similarity to Fig 1.
It treated the variation as a simple copy-paste task, ignoring the structural changes.

Figure 10: Analysis of reasoning error examples on o4-mini.

Input :
(2 points) An ecotone is a transition area between two biomes. It is where two communities meet and integrate. It may be narrow or wide, and it may be local or regional…

Fig 1: Straight



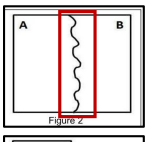
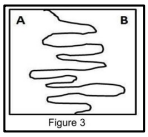


Fig2: Undulating Boundary

Fig 3: Intermingled



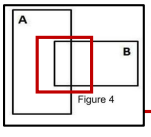


Fig4: Overlapping Zone

…
B. Figure 2 has marginal undulating boundary which suggest almost complete distinction. This can be …
… D. Figure 4 shows almost non overlapping ecotones ... distinct except for the overlapping regions…

Standard Rationale: Let's analyze each description about ecotones in relation to the figures: …
Standard Answer (Gold): Fig 2 (B) & Fig 4 (D)

Rationale: … Figure 1 shows two distinct areas separated by a sharp, straight line...

Model Prediction (Gemini-3-pro): **Fig4 (D)** selects ... Fig 4 and rejects … Fig 2…

Error Analysis:
The model was seduced by the "Visual Obviousness" of the overlapping region in Fig 4.

It reduced the "True/False"task to "Single Best Choice" task , failing to map the abstract symbol (wavy line) in Fig 2 to the concept of "River Ecotone".

Figure 11: Analysis of reasoning error examples on Gemini-3-Pro.

Input:
Enzymes rule over in biochemical processes. … ({phosphate} **NAD**$^+$ {**NADP**$^+$}.) or flavin adenine dinucleotide (FAD).

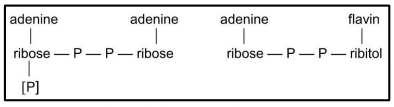


Fig 1 : Cofactor Structures

Decipher B1 to B4. Write down the balanced equations.

Under oxygen deficiency B1 partially converts into B5, an isomer of B2. This transformation requires a vitamin as the oxygen acceptor. Further oxidation of B5 leads to B6 characterized by 2 signals in ^{1}**H**-NMR spectrum.

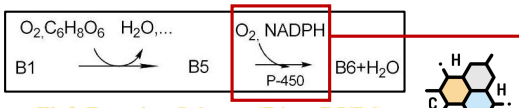


Fig2:Reaction Scheme (B1 -> B5/B6)

Standard Rationale: ... illustrates typical Cytochrome P-450 xenobiotic metabolism...
Standard Answer (Gold): … Anisole.

Rationale: "…shows two distinct areas separated by a sharp, straight line... The label 'ECOTONE' is placed..."

Model Prediction (Gemini-3-pro): … **Pyruvate .**

Error Analysis:
The model ignored the specific chemical constraints in the image (Visual Blindness).

Instead, it latched onto **the keyword " oxygen deficiency"** and retrieved a high-frequency textbook association (Pyruvate), forcing a biology answer onto a chemistry problem.

Figure 12: Analysis of reasoning error examples on Gemini-3-Pro.

Input:
As shown in the figure, in the rectangular paper…
Step 1: First, fold the rectangle $ABCD$ in half, with the crease line being MN, as shown in Figure (1);
Step 2: Then, fold point B onto the crease line MN, with the crease being AE, and the corresponding point of B on MN being B', resulting in the right triangle $\Delta AB'E$, as shown in Figure (2);
Step 3: Fold along EB' with the crease being EF, and AF intersecting the extension of $B'N$ at point G, as shown in Figure (3);
Then, in the figure formed by the folded paper, the area $S_{\Delta ABG}$ is _____.

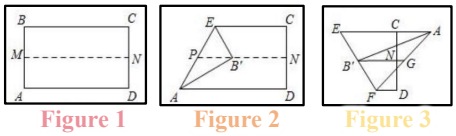


Figure 1 Figure 2 Figure 3

Standard Rationale: …reveals a critical $30° - 60° - 90°$ triangle geometry …
Standard Answer (Gold): …$3\sqrt{3}$.

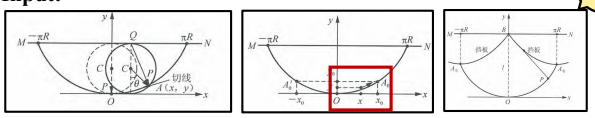
Rationale: ... The first fold is along the midline... The second fold... implies geometric relationships...

Model Prediction (Gemini-3-pro): **No Answer**

Error Analysis:
The model succeeded in Static Calculation (Fig 1 & 2) but crashed during Dynamic Simulation (Fig 3).
It lacks a "Mental Sandbox" to update the topology geometry after the final fold, leading to a disconnect between the visual state and its logical reasoning.

Figure 13: Analysis of reasoning error examples on Gemini-3-Pro.

Input:



When a ring rolls purely along a straight track, the trajectory of a point on the ring is called an epitrochoid, also known as a rolling curve or cycloid. (1) A ring of radius R rolls purely under a horizontal straight track MN, with the ring touching MN from below … When P is at O, the radius vector $\overrightarrow{CP}$ from C to P points straight downwards. As the radius vector $\overrightarrow{CP}$ rotates by an angle θ, the position of point P on the epitrochoid is denoted as A, with coordinates (x, y). At this moment, the point of tangency between the ring and MN is denoted as Q, and $\overrightarrow{QP}$ represents the instantaneous position vector of P relative to Q.
(1.1) Write down the relationships $x \sim \theta, y \sim \theta$ and describe in words the direction of the tangent line to the epitrochoid at point A.

Standard Rationale: To solve this problem, we need to understand the motion of a point on a …
Standard Answer (Gold): … $x = R(\theta + sin\theta)$.

Rationale: To find the relationships... let's analyze the motion... 1. Coordinate System: The origin O is ...

Model Prediction (Gemini-3-pro):
$x = R(\theta - sin\theta)$.

Error Analysis:
The model recognized the keyword "cycloid" and triggered a High-Frequency Pattern Retrieval (standard floor rolling).
It failed to "ground" its reasoning in the specific visual evidence (Counter-Clockwise Arrow), treating the image as a generic illustration rather than a source of physical constraints.

Figure 14: Analysis of reasoning error examples on Gemini-3-Pro.

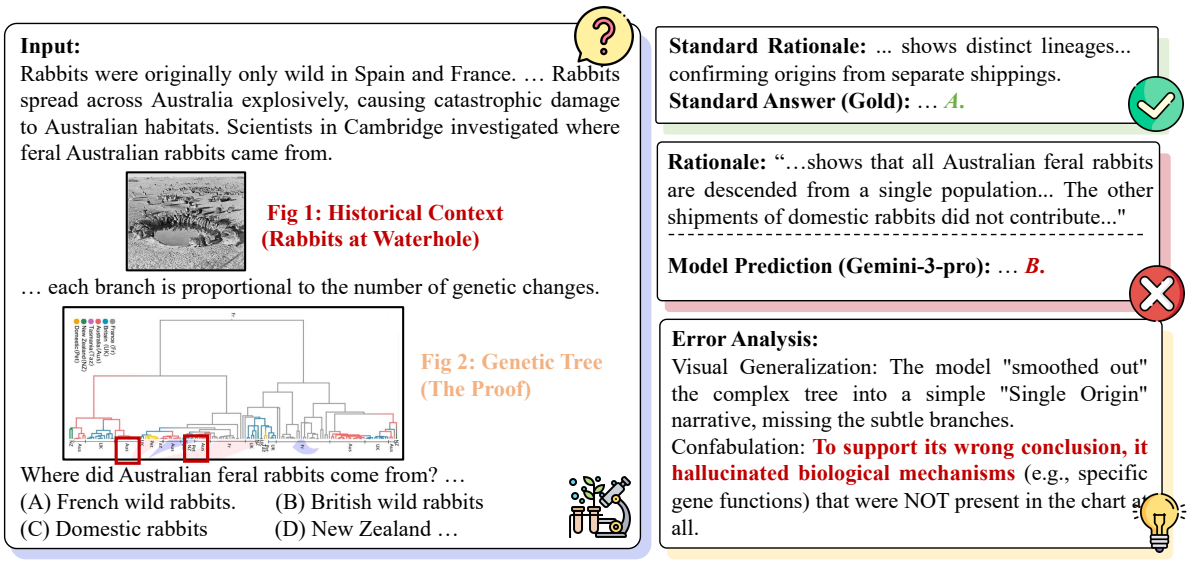


Figure 15: Analysis of reasoning error examples on Gemini-3-Pro.

predictions and references. This metric leverages a text-based language model (e.g., GPT-4-mini) as an automatic judge. For each input x_i, gold answer y_i, and model prediction $\hat{y}_i$, we prompt the judge model using:

GPTScore Evaluation Prompt-P1

You are a "Olympiad judger whose job is to decide whether two answers are equivalent in terms of their final conclusion and key reasoning.

Please follow these rules:

- If the model's final conclusion differs from the official answer, output "inconsistent".
- Some answers have multiple answers, and missing any one is a mistake.
- If the final conclusion matches but the reasoning has clear logical errors or does not rigorously justify the result, output "inconsistent".
- If the final conclusion matches and the reasoning is mathematically sound and sufficient to justify it, output "consistent".
- Different wording, order of steps, or using a different but correct method should still be treated as "consistent".

[QUESTION]

GPTScore Evaluation Prompt-P2

`[Multimodal Input Question]`

[GOLDEN ANSWER]
`[Golden Answer]`

[PREDICTED ANSWER]
`[Predicted Solution]`

Output only the following format (no extra text): "ANSWER: consistent" or "ANSWER: inconsistent".

To compute GPTScore, we first binarize each discrete score s_i as:

$$\tilde{s}_i = \begin{cases} 0 & \text{if return ANSWER: consistent} \\ 1 & \text{otherwise} \end{cases}, \quad (8)$$

GPTScore is then the mean across all N examples:

$$\text{GPTScore}(x_i) = \frac{1}{N} \sum_{i=1}^{N} \tilde{s}_i. \quad (9)$$

For per-task results, averaging is restricted to task-specific examples. Main results report micro-averaged GPTScore across all tasks.

C.3 Manual Analysis Protocol

To understand how performance diverges from GPTScore, we conducted manual analysis on cases where the two metrics disagree. For Gemini-3-Pro-Preview and InternVL3.5-1B, evaluation items were partitioned based on whether rule-based accuracy (match score>0.75 as correct) agreed with

GPTScore-based labels (GPTScore=1 as correct). From disagreement cases, 100 instances per model were randomly sampled, yielding 200 instances for detailed inspection.

We collected Model Judgment and Rule Judgment labels for 200 instances and calculated agreement rates across four categories: both correct, both incorrect, Model correct but Rule incorrect, and Model incorrect but Rule correct. Two annotators with LLM evaluation experience independently assessed discordant cases, categorizing disagreement sources as: No Effective Reasoning, Reasoning Error, Rule Match Error, and Others. The annotation guideline is detailed below:

Disagreement Source Guidelines-P1

To ensure rigorous and consistent classification, we established the following definitions to characterize the primary sources of metric divergence:

No Effective Reasoning

- **Definition:** The model outputs a final answer without providing a derivation or logical chain of thought, or the generated reasoning is incoherent/irrelevant to the query.
- **Key criterion:** Absence of a traceable cognitive process.

Reasoning Error

- **Definition:** The model attempts a step-by-step derivation but commits a logical fallacy, calculation mistake, or factual hallucination during the intermediate steps, leading to an incorrect conclusion.
- **Key criterion:** Flawed logic within a structured chain of thought.

Rule Match Error

- **Definition:** The model generates a semantically correct response that aligns with the ground truth, but the rule-based evaluation (exact match) judges it as incorrect due to formatting rigidity (e.g., synonym usage, "10.0" vs. "10", or verbose phrasing).
- **Key criterion:** Semantic correctness rejected by syntactic rigidity.

Others

Disagreement Source Guidelines-P2

- **Definition:** Disagreements arising from external factors such as ground-truth errors (label noise), ambiguous prompts, or unclassifiable multimodal alignment failures.
- **Key criterion:** Issues external to the model's reasoning capability or the extraction rule.

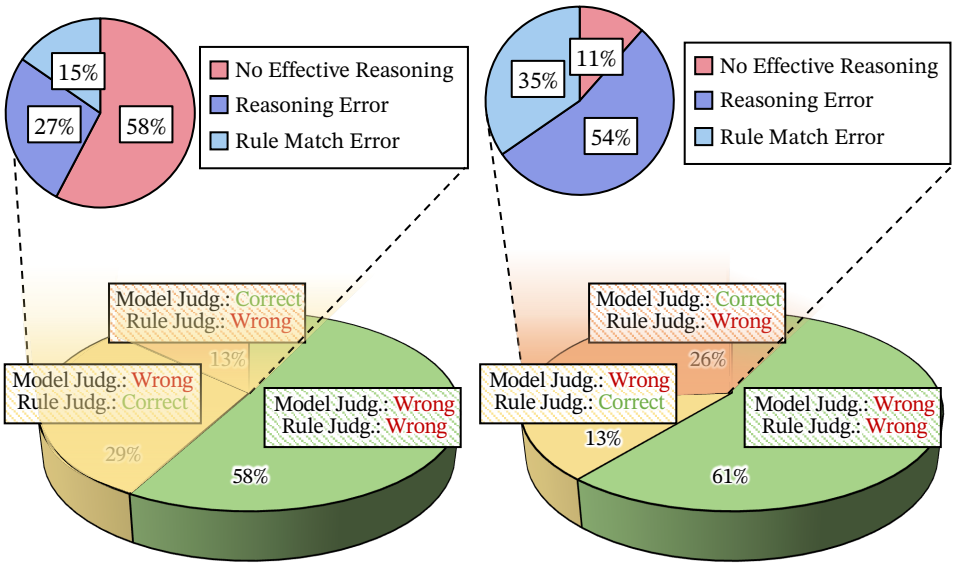
C.4 Performance Comparison between Accuracy and GPTScore

Accuracy overestimates weaker models and underestimates stronger ones relative to GPTScore, though model ranking is preserved. Models with higher GPTScore also achieve higher accuracy under GPTScore-based evaluation, but accuracy systematically overestimates weaker models and underestimates stronger ones, while largely preserving their relative ranking. To substantiate this, we manually analyze 100 instances where accuracy and GPTScore disagree for Gemini-3-Pro-Preview and InternVL3.5-1B.

As shown in Figure 16, weaker models often produce flawed reasoning yet output a definite option that rule judgment marks as correct; model judgment exposes these reasoning errors, revealing inflated accuracy. For stronger models, reasoning is typically correct, but open-ended questions lead to outputs that do not exactly match the reference; model judgment recovers these false negatives, showing that accuracy is deflated. Nonetheless, the induced model ranking remains largely consistent, indicating that accuracy is still a useful metric.

D Analyses for OMIBench Olympiad-Level Thinking Requirements

Correlation Analysis with Existing Multi-Image Benchmarks. To quantify how OMIBench relates to existing multi-image benchmarks while still reshaping model rankings, we perform a system-level correlation analysis with MMIU (Meng et al., 2024). To reduce reproduction cost, we reuse a subset of MMIU results from Wang et al. (2025c), yielding a score pair $(s_m^{\text{OMI}}, s_m^{\text{MMIU}})$ for each model m. As reported in the main text and shown in Figure 3(a), this analysis yields a moderate Spearman correlation of $\rho = 0.535$, which is well below the commonly used strong-correlation threshold of 0.7.



(a) Human annotation concerning about different judgement methods on Intern3.5-VL-1B (Weak LVLM).

(b) Human annotation concerning about different judgement methods on Gemini-3-Pro-Preview (Strong LVLM) .

Figure 16: Human annotation concerning different judgment methods (Model Judge: GPTScore, Rule Judge: Match Accuracy). Note: Correct/wrong indicates the automated method's output, not its actual accuracy.

E Detailed Analyses for OMIBench Multi-Image Requirements

E.1 Single- vs. Multi-image Olympiad-Level Benchmark Comparison

For the correlation analysis between OMIBench and OlympiadBench, we evaluate three representative LVLMs: GPT, Gemini, Qwen-VL, and InternVL. All models are used in their official API or checkpoint form without any additional fine-tuning or task-specific adaptation. In order to reduce evaluation cresource consumption, the partial results of OlympiadBench are taken from Wan et al. (2025); Wang et al. (2025c); Team et al. (2025a); Yue et al. (2025); Shen et al. (2025); Zhang et al. (2025); Zha et al. (2025); Team et al. (2025b); Yu et al. (2025).

Spearman correlation between benchmarks. To study whether Olympiad-level thinking transfers from the single-image to the multi-image setting, we compute the Spearman rank correlation coefficient between model performances on OlympiadBench and OMIBench. Concretely, for each model m, we obtain its accuracy on OlympiadBench, denoted by a_m^{Olympiad}, and its accuracy on OMIBench, denoted by a_m^{OMI}. We then rank the models separately according to a_m^{Olympiad} and a_m^{OMI} and apply the standard Spearman formula on these two rankings. The resulting coefficient is 0.614, which is below the commonly used threshold of 0.7 for strong monotonic correlation, showing that the relative ordering of LVLMs changes noticeably between the single-image and multi-image Olympiad settings.

Illustrative performance drop. As highlighted in the main text, a representative LVLM that reaches 75.67% accuracy on OlympiadBench drops to 50.53% accuracy on OMIBench when evaluated under the same decoding and prompting setup. This corresponds to an absolute decrease of over 25% and a relative performance drop of approximately 33%. Together with the moderate Spearman correlation, these findings substantiate the claim that OMIBench imposes systematically stronger demands on multi-image Olympiad reasoning than its single-image counterpart.

E.2 Relationship Between Number of Images and Accuracy

To quantify how multi-image complexity affects model performance, we analyze accuracy as a function of the number of images per instance. Each OMIBench question is annotated with its image count k. Instances are grouped into bins by k (e.g., $k = 1, 2, 3, 4, 5, \geq 6$), and the average accuracy for each bin is computed by aggregating predictions from all evaluated models. Concretely, for each bin B_k with k images, the bin-level accuracy is defined as:

$$\text{Acc}(k) = \frac{1}{|B_k|} \sum_{x \in B_k} \text{match}\{\hat{y}(x), y(x)\}, \quad (10)$$

where $y(x)$ is the ground-truth answer for instance x, $\hat{y}(x)$ is the model prediction, and match$\{\cdot,\cdot\}$ is the match function as mentioned in Appendix C.1. Figure 3 (b) reports the mean accuracy for each bin.

E.3 Ablation: Restricting Instances to a Limited Input Image

To verify that OMIBench necessitates multi-image integration rather than being solvable via limited input image cues, we constructed an ablation dataset wherein visual context was systematically restricted. For every instance containing $K > 1$ images, we retained the original question and ground truth but supplied only the primary image (the first in the canonical sequence). This approach isolates the impact of visual context reduction while maintaining the distributional properties of the query set.

We evaluated model performance across two conditions: (1) the standard multi-image setting, where models access the full visual complement ($k = K$ images); and (2) the partial-image ablation, where input is strictly limited to $k < K$ images per instance. As illustrated in Figure 3 (c), limiting visual input precipitates a marked decline in performance. Across all evaluated LVLMs, the mean accuracy drops by at least 10% in the single-image setting relative to the multi-image baseline. The degradation is particularly pronounced for inquiries requiring cross-referencing or comparative analysis, confirming that high performance on OMIBench relies on joint reasoning over multiple visual inputs.

F Detailed Analyses for "Combined Multi-Image and Olympiad-Level Thinking"

F.1 Correlation with MMIU and OlympiadBench

Figure 3 (a) examines model ranking consistency. MMIU prioritizes multi-image perception, whereas OlympiadBench emphasizes single-image reasoning; OMIBench integrates these demands by requiring reasoning across multiple images. Empirically, rankings between MMIU and OlympiadBench diverge, confirming their disparate foci. Conversely, OMIBench correlates more strongly with each baseline than the baselines do with one another. This suggests OMIBench effectively bridges the two domains, capturing the joint capabilities of multi-image understanding and complex problem-solving.

F.2 Detailed Error Analysis

In this section, we formally define the annotation protocol, the taxonomy of failure types, and the qualitative patterns observed for each category on OMIBench. The goal is to make the reported percentages in Figure 7 reproducible and interpretable.

F.2.1 Annotation Protocol

We analyzed OMIBench examples where LVLM responses fell below a predefined correctness threshold under GPTScore. These *candidate errors* underwent a rigorous secondary inspection to elucidate failure mechanisms. For each instance, annotators examined the visual inputs, instructions, reference solutions, and model outputs. A prerequisite validation step excluded false positives where GPTScore misclassified valid responses. Confirmed errors were then assigned a *primary* failure mode based on the fundamental deficit preventing a correct output.

Figure 7 illustrates the resulting distribution: visual perception failures (35%), cross-image association failures (30%), logical reasoning fallacies (25%), and instruction comprehension biases (10%). The following subsections detail the operational definitions of each category.

Visual Perception Failures occur when the LVLM misperceives basic visual facts in one or more images, yielding wrong or incomplete scene descriptions, even for simple questions, and thus unreliable downstream reasoning. We label an error as a visual perception failure when *at least one* of the following holds:

1. **Object misrecognition:** The model assigns an incorrect object category that the question depends on (e.g., bus vs. truck, dog vs. cat).
2. **Attribute misclassification:** The model gets the object class right but misperceives salient attributes such as color, number, relative size, pose, or state (e.g., open vs. closed, full vs. empty).
3. **Spatial relation errors:** The model misreads coarse spatial relations or layout (e.g., left vs. right, in front of vs. behind, above vs. below) that are visually clear and explicitly queried.
4. **Salient detail omission:** The answer would be correct if a clearly visible but critical detail did not exist (e.g., a small but prominent symbol, icon, or text overlay).

Cross-Image Association Failures occur when the LVLM parses each image reasonably well in

isolation but fails to correctly relate them when the question requires comparing, contrasting, or aggregating information across images. We assign this label when:

1. The model's descriptions of individual images (paraphrased or inferred from its answer) are largely accurate.
2. The question explicitly or implicitly involves multiple images (e.g., "between the first and second image", "across all panels").
3. The error stems from misalignment, confusion, or omission in how information from different images is combined.

Logical Reasoning Fallacies occur when the LVLM's basic visual understanding and cross-image mapping are adequate, but the chain of reasoning leading to the final answer contains flawed logical steps; fixing the reasoning alone would yield the correct answer. We annotate an error as a logical reasoning fallacy when: (1) The model's implicit or explicit description of relevant visual facts is broadly correct; perception is not the main source of error. (2) The natural-language explanation, if present, shows misapplied inference rules, unsupported assumptions, or inconsistent intermediate conclusions. (3) Adjusting the reasoning alone, without changing the perceived facts, would fix the answer.

Error distributions highlight complementary weaknesses in LVLM. Visual perception (35%) and cross-image association failures (30%) indicate persistent limits in fine-grained visual understanding and multi-image integration, while logical reasoning (25%) and instruction comprehension errors (10%) show that stronger visual encoders alone are insufficient without advances in structured reasoning and multimodal instruction following.

These results motivate future work on: (i) stronger low- and mid-level visual representations, (ii) explicit cross-image alignment and aggregation mechanisms, (iii) more reliable, verifiable reasoning procedures, and (iv) training schemes that sharpen sensitivity to multimodal instructions and output constraints.

G Details for Long CoT Experiments

This section provides the full experimental details for Section 6.1 (*Can Long Chain-of-Thought Strategies Help?*), including the exact prompting templates, decoding configurations, and the definition of the "thinking" and "no-thinking" settings used on OMIBench.

G.1 Prompting strategies on OMIBench

To provide the exact templates used for the comparison in Figure 4 (a), we systematically evaluate widely-used Chain-of-Thought prompting strategies on OMIBench, including Least-to-Most (Zhou et al., 2023), Plan-and-Solve (Wang et al., 2023), VoT (Wu et al., 2024). All prompts follow the generic instruction format below:

Least-to-Most Prompting (Zhou et al., 2023) is a strategy that decomposes complex problems into a sequence of simpler subproblems, solving them one by one. Specifically, the prompt used is as follows:

Least-to-Most Prompting

Please reason step by step, and then provide the final answer in the exact format: \boxed{ANSWER}.

[Question]
problem_text0 [IMAGE0] problem_text1 [IMAGE1] ... problem_textn

[Choices] # only for multiple choices problems
A. option_A
...
Let's break down this problem and solve it one by one.

Plan-and-Solve Prompting (Wang et al., 2023) first devises a high-level plan to tackle the problem, then executes the plan step by step. The specific prompt used is as follows:

Plan-and-Solve Prompting

Please reason step by step, and then provide the final answer in the exact format: \boxed{ANSWER}.

[Question]
problem_text0 [IMAGE0] problem_text1 [IMAGE1] ... problem_textn

[Choices] # only for multiple choices problems
A. option_A
...
Let's first understand the problem and devise a plan to solve it. Then, let's carry out the plan and solve the problem step by step.

Visualization-of-Thought (VoT) Prompting (Wu et al., 2024) encourages the model to visualize intermediate states after each reasoning step to enhance clarity and understanding. The specific prompt used is as follows:

Visualization-of-Thought (VoT) Prompting

Please reason step by step, and then provide the final answer in the exact format: \boxed{ANSWER}.

[Question]
problem_text0 [IMAGE0] problem_text1 [IMAGE1] ... problem_textn

[Choices] # only for multiple choices problems
A. option_A
B. option_B
...
Visualize the state after each reasoning step.

G.2 "Thinking" vs. "No-Thinking" Prompting Modes

We next describe how the "thinking" and "no-thinking" prompting modes in Fig. 4(b) are implemented for both reasoning-oriented and non-reasoning LVLMs. In the "no-thinking" mode, the prompt instructs the model to avoid step-by-step reasoning and output only the final answer:

No-Think Prompting

Return only the final answer in this exact format: \boxed{ANSWER}.

[Question]
problem_text0 [IMAGE0] problem_text1 [IMAGE1] ... problem_textn

[Choices] # only for multiple choices problems
A. option_A
B. option_B ...

G.3 "Thinking" vs. "Instruct" Model Variants

Finally, we describe the usage of the Qwen3-VL "thinking" and "instruct" variants whose comparison is summarized in Figure 4(c). We use two official checkpoints: "Qwen3-VL-Instruct," an instruction-following vision-language model optimized for general-purpose multimodal tasks; and "Qwen3-VL-Thought," a variant optimized for long-form reasoning that supports an explicit "thinking" mode with extended internal deliberation. As shown in Figure 4 (c), although the gains remain below Olympiad-level performance, they constitute the largest relative improvement among all tested Long CoT paradigms, indicating that specialized long-reasoning training can partially enhance multimodal Olympiad performance.

H Detailed Protocols for Test-Time Scaling Experiments

Unless otherwise specified, all results are reported on the OMIBench test split using GPTScore, and each configuration is evaluated on the full benchmark.

H.1 Sequential Scaling Protocol

To study sequential test-time scaling, we vary the maximum number of newly generated tokens per example, $L_{\max}$, while keeping all other hyperparameters fixed and re-evaluating the full OMIBench test set for each configuration. We sweep

$$L_{\max} \in \{512, 1{,}024, 2{,}048, 4{,}096, 8{,}192, 16{,}384\}, \tag{11}$$

a near-geometric progression that provides dense coverage on a log scale with a manageable number of runs.

H.2 Parameter Scaling Protocol

To analyze how test-time scaling interacts with model size, we evaluate two families of open-source multimodal language models: InternVL and

Model	Biology		Chemistry		Mathematics		Physics		Total	
	ACC	Score	ACC	Score	ACC	Score	ACC	Score	ACC	Score
Qwen3-VL-2B-Instruct	27.44	19.92	12.33	5.07	11.53	6.99	12.50	8.75	14.99	9.69
Qwen3-VL-2B-Thinking	24.86	9.56	10.46	1.38	22.25	15.81	10.75	6.84	17.12	9.38
Qwen3-VL-4B-Instruct	43.03	36.65	17.05	11.06	27.91	23.72	18.40	13.92	25.95	20.95
Qwen3-VL-4B-Thinking	43.58	36.65	18.94	11.52	50.80	44.65	20.41	20.75	34.45	30.03
Qwen3-VL-8B-Instruct	46.61	43.43	16.13	17.05	27.44	29.30	20.05	18.63	26.85	26.55
Qwen3-VL-8B-Thinking	52.39	52.59	15.96	17.05	49.90	49.30	21.97	30.66	35.84	38.65
Qwen3-VL-30B-A3B-Instruct	48.51	48.61	13.29	12.90	31.42	32.33	20.02	20.99	28.03	28.59
Qwen3-VL-30B-A3B-Thinking	54.44	55.78	28.13	32.26	63.70	59.30	27.92	37.50	44.63	47.20
Qwen3-VL-32B-Instruct	57.62	58.57	14.09	20.74	44.40	40.70	25.48	25.00	35.87	35.78
Qwen3-VL-32B-Thinking	65.72	64.14	29.83	24.88	60.66	60.93	31.92	41.98	47.34	49.54
Qwen3-VL-235B-A22B-Instruct	60.41	63.20	17.23	22.58	37.48	34.19	23.77	23.58	34.11	34.39
Qwen3-VL-235B-A22B-Thinking	55.00	61.35	34.14	33.18	53.71	48.84	26.81	43.87	42.12	47.05

Table 4: "Thinking" and "Instruct" results on OMIBench, where the bold content denotes the best performance in each category.

QwenVL. Within each family, we use checkpoints spanning roughly 1B to 235B parameters (see Table 2). The resulting performance–parameter-count curves are shown in Figure 5 (b), with parameter counts on a logarithmic scale. We mark the saturation region as the smallest parameter size beyond which all larger models yield less than 0.5 absolute GPTScore improvement. InternVL improves up to approximately the mid-sized checkpoints and then saturates around 25% GPTScore, whereas QwenVL continues to improve up to its largest public variant, plateauing around 35% GPTScore.

H.3 Parallel Scaling Protocol

To evaluate parallel test-time scaling, we fix the model and prompting scheme and vary the number of independent samples per example, denoted by k. For each configuration, we draw k stochastic reasoning trajectories and aggregate them via majority vote over the final answers. We sweep

$$k \in \{1, 3, 4, 8, 16\}. \tag{12}$$

For $k = 1$, this reduces to standard single-sample decoding. For $k > 1$, we keep all per-sample decoding hyperparameters fixed and only change the number of parallel draws. Since self-consistency is not well-defined for $k = 2$, we use $k = 3$ as the smallest multi-sample setting.

To enable self-consistency, we decode with temperature $T = 0.6$ and a maximum reasoning length $L_{\max} = 16{,}384$ tokens per sample. After decoding, we apply the common post-processing pipeline to extract and normalize the final answer from each of the k samples and then take the majority-voted answer. For each k, we compute the mean GPTScore on the full OMIBench test set using the majority-voted predictions, yielding the accuracy–sampling curve in Figure 5(c). Plotting GPTScore against $\log_2(k)$ reveals an approximately log-linear relationship between the number of samples and performance over the examined range of k.

H.4 Details of In-Context Learning Experiments on OMIBench

This section details the experimental protocol for the in-context learning (ICL) results reported in Figure 6(a), including the construction of in-context examples, the definition of the No-Image-ICL, Single-Image-ICL and Multi-Image-ICL conditions, and the control choices used to ensure fair comparison across conditions.

Prompt template and formatting. For each test instance, we randomly sample k demonstrations from OMIBench; these source problems are excluded when computing the evaluation metrics. For each instance, the sampled demonstrations are fixed across all ICL conditions to ensure a fair comparison. All ICL variants share a unified prompt template to isolate the effect of visual context. Each prompt is structured as follows:

In-Context Learning Prompt Template-P1

Please reason step by step, and then provide the final answer in the exact format: "\boxed{ANSWER}".

[EXAMPLE 1]

In-Context Learning Prompt Template-P2

[Question]
[Problem Text 0] [IMAGE0] ...
[Problem Text n_K]
[Choices] # only for multiple choices problems
[Solution]
[Example Solution]

[EXAMPLE 2]
[Question]
[Problem Text 0] [IMAGE0] ...
[Problem Text n_K]
[Choices] # only for multiple choices problems
[Solution]
[Example Solution]
...

[REQUEST]
[Question]
[Problem Text 0] [IMAGE0] ...
[Problem Text n_K]
[Choices] # only for multiple choices problems
A. option_A
B. option_B
...

No-Image-ICL Configuration. The No-Image-ICL condition tests whether models can exploit purely textual patterns in the demonstrations, even though OMIBench is a multimodal benchmark. Concretely, for each of the k demonstrations, we remove all images from the model input but keep the textual problem description (including any references to images, image indices, or regions) and the answer line. Thus, the model can only rely on textual information in the demonstrations, but still has access to the full visual information for the target question. As reported in the main text, adding these purely textual in-context examples substantially improves OMIBench performance compared to the zero-shot baseline.

Single-Image-ICL Configuration. The Single-Image-ICL condition evaluates whether attaching a *single* representative image to each demonstration offers additional benefits over text-only demonstrations. For each of the k demonstration instances: If the original OMIBench question contains a single image, we attach that image to the demonstration, exactly as in the dataset. All other aspects of the prompt (textual problem statement, options, and answer format) are identical to the No-Image-ICL case, and the test instance is again provided with its full set of images. This condition probes whether current LVLMs can effectively exploit minimal visual context in the demonstrations to further improve over text-only ICL.

Multi-Image-ICL Configuration. The Multi-Image-ICL condition provides the model with the full visual complexity of OMIBench within the in-context examples. For each of the k demonstrations, we attach *all* images associated with that OMIBench instance, preserving the original ordering. Thus, demonstrations in this condition mirror the multimodal structure of the test instance itself. The textual content and answer format again remain unchanged relative to the other ICL configurations, so that any performance differences can be attributed to how well the model leverages multi-image visual context in the demonstrations.

I Additional Details for "Thinking with Images"

I.1 Tool-based "Thinking with Images" (GPT-4o + VisualSketchpad)

Inspired by Cheng et al. (2025b); Tong et al. (2025), we instantiate the tool-based "Think-with-Image" paradigm by augmenting GPT-4o with VisualSketchpad. This integration enables the model to draw primitives and text, highlight or blur regions, and crop or zoom. We adapt VisualSketchpad's prompting strategy, originally optimized for single-image tasks, to our multi-image framework.

Qualitatively, the system exhibits distinct limitations on OMIBench: (1) fixating on isolated images while neglecting other crucial information for the task; (2) performing redundant edits that yield no new information; and (3) failing to spatially align objects across images. These behaviors, rare in single-image settings, highlight the limited transferability of existing visual tools to complex multi-image reasoning. More detailed case analyses are provided in Figure 17 & 18.

I.2 Internal "Thinking with Images" (EMU-3.5-34B)

In the internal "Think-with-Image" configuration, we employ EMU-3.5-34B, a unified multimodal generator designed to reason over inputs and synthesize visual content autonomously, eliminating the need for external APIs. This architecture enables the model to perform reasoning that integrates

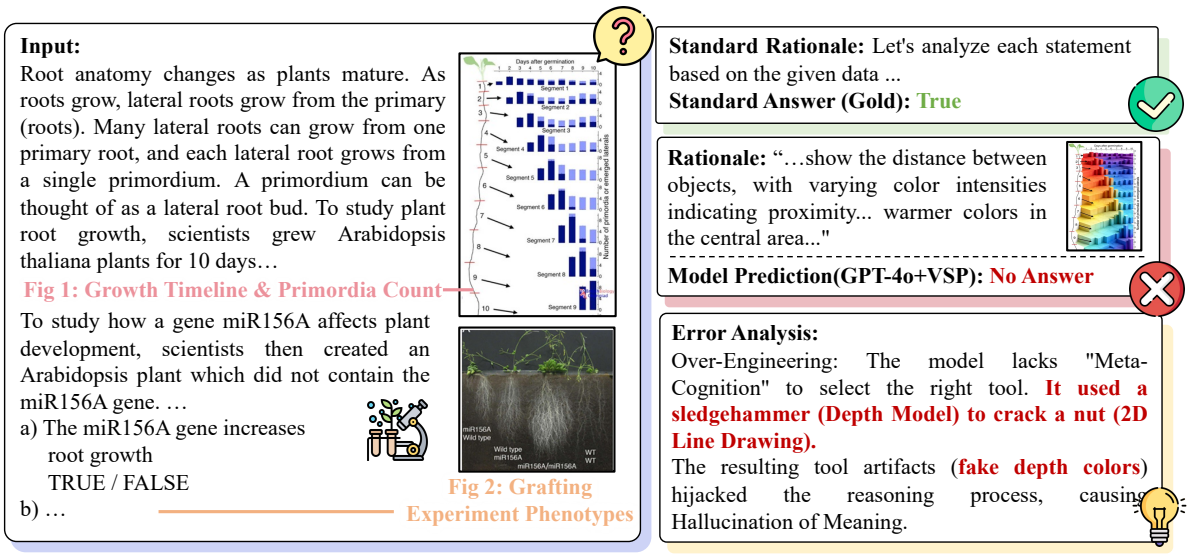


Figure 17: Analysis of reasoning error examples on GPT-4o+VisualSketchpad (VSP).

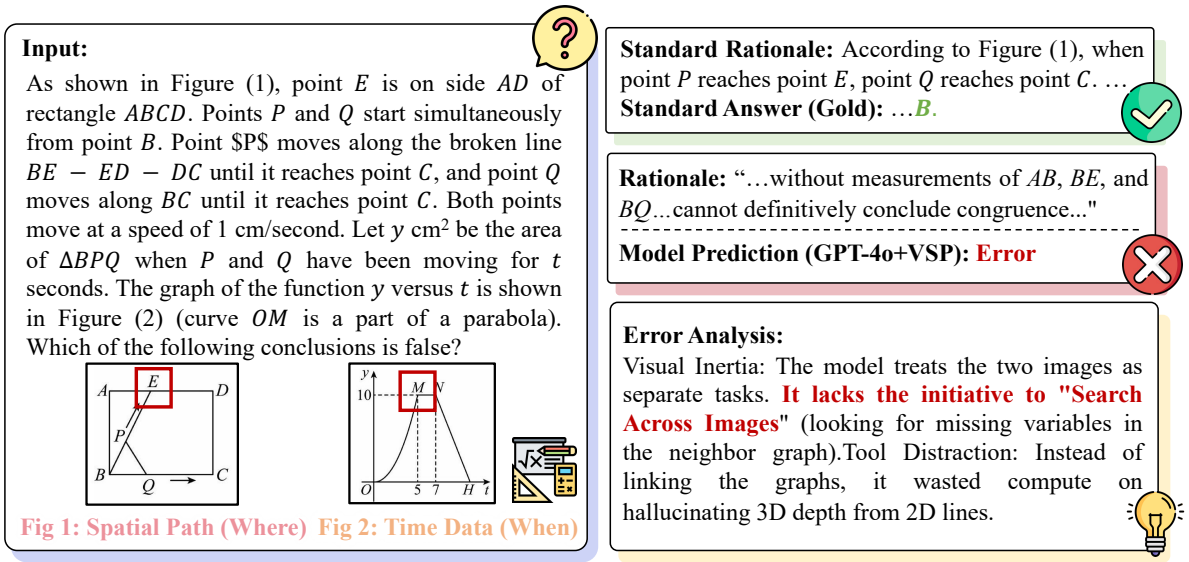


Figure 18: Analysis of reasoning error examples on GPT-4o+VisualSketchpad (VSP).

both textual analysis and self-generated visual aids. For inference, we adopt a mixed-precision protocol on 2 NVIDIA A100 80GB GPUs. We set the temperature to 0.3 and top-p to 0.9, a configuration empirically determined to optimally balance diversity and coherence during visual planning. More detailed case analyses are provided in Figure 19-21.

J Reliability and Stability of GPTScore

To address concerns about the trustworthiness of model-based evaluation, we report four complementary analyses of GPTScore.

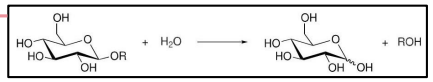
Complementarity with match accuracy. Match Accuracy captures exact symbolic agreement, while GPTScore evaluates semantic equivalence under the multimodal context (see Appendix C). The Spearman correlation between the two metrics over model-level scores is 0.93, indicating strong agreement and consistency.

Agreement with human ratings and cross-evaluator stability. We randomly sample 200 answers generated by Gemini-3-Pro-Preview and obtain human ratings. We then compare these human scores with GPTScore produced by three different evaluator models. As shown in Table 7, all Spearman correlation coefficients exceed 0.86 ($p < 0.05$), demonstrating both that model-based evaluation is well aligned with human judgment for logic-reasoning-style answers, and that GPTScore is stable across different evaluator models.

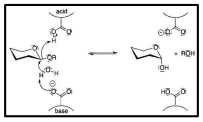
Sampling stability and statistical significance. We report the mean and standard deviation of both

Input:
… Glycosidases, … are enzymes that catalyze this reaction (scheme 1). … increase the speed of the C-O bond cleavage reaction by a factor of (10^{17})

Fig 1: General Hydrolysis

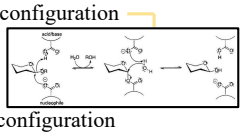


Scheme 1: Hydrolysis of the glycosidic bond. **Mechanism of hydrolysis catalyzed by glycosidase** ...



Scheme 2: Mechanism with inversion of configuration

Fig 2: Inversion Mechanism (The Reality)



Scheme 3: Mechanism with retention of configuration

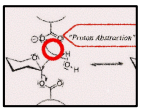
Chose the correct statement indicating the reactivity of sodium hydride:

Base / Acid/ nucleophile / electrophile

Standard Rationale: The image explicitly shows the residue abstracting a proton from water …

Standard Answer (Gold): …Base.

Rationale: … Retrieved the standard "Double Displacement" mechanism from memory, ignoring the specific "Inversion" diagram provided...



Model Prediction (Emu-3.5): Nucleophile

Error Analysis:

Visual Neglect: The model ignored the specific electron flow (arrow pushing) in the image.

Hallucination form Priors: Seeing "Glycosidase", it blindly applied a high-frequency textbook concept (Enzyme = Nucleophile), failing to adapt to the specific case shown (Enzyme = General Base).

Figure 19: Analysis of reasoning error examples on Emu-3.5.

Input:
In the rectangle $\square$ $ABCD$ shown in Figure (1), point E is on AD, and by folding point A to the right along BE, it forms the configuration shown in Figure (2). Then, $AF \perp CD$ is drawn at point F, as shown in Figure (3). If $AB = 2, BC = 3$, $\angle BEA = 60°$, then the length of AF in Figure (3) is:

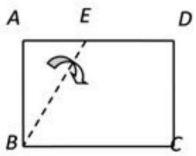


Figure (1)

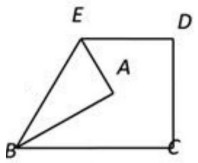


Figure (2)

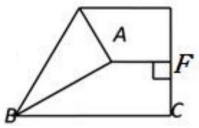


Figure (3)

Standard Rationale: … Strictly derived from given side lengths (AB=2, BC=3) and folding symmetry…

Standard Answer (Gold): …$3-\sqrt{3}$.

Rationale: Reasoning loop broke down…

Output: Infinite Loop Crash

Model Prediction (Emu3.5): ***Infinite Loop***

Error Analysis:

Regularization Bias: The model "regularized" a standard triangle into a "perfect" Equilateral Triangle (60°) simply because it looked tidy.

Logical Collapse: When the math derived from this fake angle didn't add up, the language model entered an infinite generation loop, unable to resolve the contradiction.

Figure 20: Analysis of reasoning error examples on Emu-3.5.

Input:
Rabbits were originally only wild in Spain and France. The Romans domesticated French rabbits, and introduced them to Britain (and the rest of Europe). During the Imperial era, rabbits were introduced to Australia. Rabbits spread across Australia explosively, causing catastrophic damage to Australian habitats. Scientists in Cambridge investigated where feral Australian rabbits came from.

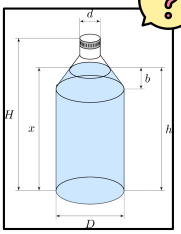


Fig 1: Geometric Constraints

… To work out where Australian feral rabbits came from, a genetic tree of the different rabbit populations was constructed. The length of each branch is proportional to the number of genetic changes.

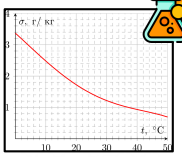


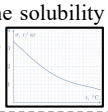
Fig 2: Chemical Distractor (No Use)

Does this tree prove feral rabbits in Australia came from multiple different shippings?

Standard Rationale: … Calculate volume of cone frustum + cylinder to find height…

Standard Answer (Gold): … $x \approx 20cm$.

Rationale: Step 1: Calculate the solubility of carbon dioxide in water... Henry's law...



Model Prediction (Emu3.5): … $x \approx 18.5cm$

Error Analysis:

Distractor Trap: The solubility graph (Fig 2) acted as a "Super-Stimulus", **hijacking the model's attention** away from the bottle's geometry (Fig 1).

Domain Drift: The model drifted from **Physics (Volume) to Chemistry (Solubility)**, performing complex but irrelevant thermodynamic calculations.

Figure 21: Analysis of reasoning error examples on Emu-3.5.

Metric	Evaluator	Biology		Chemistry		Mathematics		Physics		Total	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
GPTScore	Gemini-2.5-flash	72.49	1.01	26.61	0.95	61.57	0.88	40.33	1.17	50.98	0.42
GPTScore	GPT-5-mini	71.49	0.33	23.85	1.02	62.85	1.25	40.21	0.99	50.80	0.45
ACC	Gemini-2.5-flash	59.74	1.29	21.08	0.69	39.71	1.19	21.76	0.76	34.66	0.29
ACC	GPT-5-mini	60.14	1.20	21.43	0.88	36.92	0.79	24.47	0.89	34.75	0.22

Table 5: Stability of GPTScore and Accuracy (%) across three independent runs (mean and standard deviation).

Setting	Biology		Chemistry		Mathematics		Physics		Total	
	ACC	Score	ACC	Score	ACC	Score	ACC	Score	ACC	Score
Original multi-image	60.64	52.61	15.67	22.12	29.53	15.35	16.98	18.87	24.05	24.88
Inform.-equivalent single-image	**64.66**	**56.63**	**20.74**	**26.27**	**33.49**	**19.30**	**22.41**	**24.53**	**33.71**	**29.17**

Table 6: Performance comparison of GPT-4o between the original multi-image setting and the information-equivalent single-image setting.

Comparison	Spearman Cor.	p-value
Human vs. GPT-4o-mini	0.886	<0.05
Human vs. GPT-5-mini	0.874	<0.05
Human vs. Gemini-2.5-flash	0.869	<0.05

Table 7: Spearman correlation between human ratings and model-based GPTScore.

Evaluation subject	Accuracy
Human Expert A (avg.)	82.69%
Human Expert B (avg.)	80.77%
Human Non-expert A (avg.)	57.69%
Human Non-expert B (avg.)	61.53%
Gemini-3-Pro (best current)	48.08%

Table 8: Performance gaps between human experts/non-experts and the current strongest model on a 52-problem subset of OMIBench.

GPTScore and ACC across three independent sampling runs of Gemini-3-Pro-Preview, evaluated by Gemini-2.5-flash and GPT-5-mini. As shown in Table 5, the standard deviations remain small: for GPTScore, all subject-level deviations are below 1.3% and total-level Std is below 0.5%; for ACC, fluctuations are at most $\sim 1.3\%$ and total-level Std is below 0.3%. These variations are much smaller than the effective gains reported in our paper for techniques such as long-CoT thinking, sequential scaling, and in-context learning, which exceed 5% and in most cases exceed 10%. Moreover, paired comparisons for Qwen-VL-32B-Thinking vs. Instruct, Qwen-VL-32B-Instruct sequential scaling (1 vs. 16 samples), and Qwen-VL-32B-Instruct textual ICL (0-shot vs. 3-shot) over five repeated runs yield p-values all below 0.05, indicating that our conclusions are statistically stable.

Partial-credit evaluation (future work). Although GPTScore already provides graded semantic judgments under contextual constraints, we acknowledge that purely binary or near-binary judgments may underweight creative or partially correct solutions on highly open-ended OMIBench items. As a future extension, we plan to incorporate a rubric-guided partial-scoring mechanism following recent practice in scientific-reasoning evaluation (OpenAI, 2025a), and to release the corresponding rubrics and evaluator prompts together with the benchmark.

K Human Baseline Details

To calibrate the difficulty of OMIBench against human performance, we conduct a small-scale human study on a uniformly sampled 52-problem subset spanning biology, chemistry, mathematics, and physics. We recruit four participants: two domain experts (PhD/Master's students in STEM-related fields) and two trained non-experts (undergraduates without direct domain specialization). None of the participants has seen OMIBench before the study.

Before formal evaluation, all participants complete 10 familiarization problems that are not included in the reported subset. This stage is used only to acquaint them with the OMIBench input format, which often requires coordinating information across multiple images, textual problem statements, and sometimes auxiliary answer choices. During the actual evaluation, participants are allowed to inspect all provided images and text for each prob-

Model	Biology		Chemistry		Mathematics		Physics		Total	
	ACC	Score	ACC	Score	ACC	Score	ACC	Score	ACC	Score
InternVL3.5-8B	47.39	36.95	17.51	17.97	27.91	32.33	17.69	17.69	26.59	26.14
+ CMMCoT	**48.19**	36.55	**17.97**	**19.35**	**29.77**	**35.58**	**19.81**	**18.63**	**28.11**	**27.65**
+ MMDU	46.18	36.55	17.51	16.59	26.98	31.86	16.75	16.51	25.76	25.30
Qwen3-VL-8B-Instruct	46.59	43.37	16.13	16.59	27.21	29.07	**20.05**	**18.40**	26.74	26.29
+ CMMCoT	**49.00**	**49.40**	**18.43**	**18.43**	**31.63**	**30.70**	19.58	17.92	**28.86**	**28.11**
+ MMDU	45.78	41.77	16.13	15.67	25.81	29.53	19.34	18.16	25.91	25.91

Table 9: SFT results on OMIBench with different multi-image instruction-tuning datasets.

Method	Biology		Chemistry		Mathematics		Physics		Total	
	ACC	Score	ACC	Score	ACC	Score	ACC	Score	ACC	Score
GPT-4o	**60.64**	**52.61**	15.67	**22.12**	29.53	15.35	16.98	18.87	24.05	24.88
+ SlowPerception	51.41	52.21	7.83	17.97	18.14	6.98	1.89	14.39	17.50	19.70
+ CogFlow	47.39	40.56	7.37	13.36	8.84	4.42	6.37	2.36	15.08	12.05
+ Visual Sketchpad	18.88	17.67	**16.13**	17.97	9.53	13.72	15.33	16.75	14.24	16.14
GPT-5	68.13	**62.55**	**23.96**	29.03	39.30	56.51	20.52	40.80	36.23	48.11
+ SlowPerception	69.88	**63.45**	23.04	29.95	39.07	56.98	**20.99**	**42.22**	36.44	49.02
+ CogFlow	72.29	62.25	23.50	**33.64**	38.84	**57.44**	20.52	41.04	36.74	**49.17**
+ Visual Sketchpad	**73.90**	61.85	22.12	30.88	**39.53**	56.51	20.05	38.21	**36.89**	47.42

Table 10: Results of different external-tool integration frameworks on OMIBench.

lem and are asked to provide a final answer in the same problem setting as the benchmark.

We report average accuracy for each participant group in Table 8, together with the strongest current model on the same subset for reference. The results show that experts remain substantially above current LVLMs, while even trained non-experts still outperform the best model by a clear margin. This gap indicates that OMIBench is difficult but still solvable for humans with sufficient scientific background, making it a meaningful benchmark for measuring progress in multi-image Olympiad-level reasoning.

L Single-Image Control Details

To distinguish the effect of cross-image reasoning from confounders such as visual volume, input length, and OCR noise, we construct an information-equivalent single-image control. For each problem, all images are concatenated in their original logical order into a single composite image, while the question text and answer choices remain unchanged.

M Detailed Results for Training and External Tools

This appendix provides the full numerical results referenced in Section 6.4.

SFT on multi-image instruction-tuning data. Table 9 reports the per-subject ACC and Score for InternVL3.5-8B and Qwen3-VL-8B-Instruct fine-tuned on CMMCoT (Zhang et al., 2026) and MMDU (Liu et al., 2024b). CMMCoT consistently yields modest improvements on both backbones (Total ACC 26.59 $\rightarrow$ 28.11 and 26.74 $\rightarrow$ 28.86 respectively), while MMDU yields no improvement or slight degradation, suggesting that simple multi-image understanding data is insufficient for Olympiad-level reasoning.

External-tool integration. Table 10 reports the per-subject ACC and Score of three external-tool integration frameworks on top of GPT-4o and GPT-5. GPT-4o consistently degrades when augmented with any of the three tools, while GPT-5 shows modest, inconsistent gains.