Look Light, Think Heavy: What Multimodal Chain-of-Thought Reasoning Can and Cannot Do

Zhuoran Jin[1,2], **Kejian Zhu**[1,2], **Hongbang Yuan**[1,2], **Yupu Hao**[1,2],
Pengfei Cao[1,2], **Yubo Chen**[1,2], **Kang Liu**[1,2], **Jun Zhao**[1,2,*]
[1] The Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{zhuoran.jin, kliu, jzhao}@nlpr.ia.ac.cn, zhukejian2025@ia.ac.cn

Abstract

Chain-of-Thought (CoT) has become a standard method for improving reasoning capabilities in large language models (LLMs) by eliciting step-by-step thinking, but its effectiveness in multimodal tasks remains unclear. In this paper, we aim to systematically investigate the key question: *What can multimodal Chain-of-Thought reasoning do, and where and why does it fall short?* To this end, we evaluate 12 multimodal tasks across perception and reasoning categories using both 14 non-reasoning models and 8 reasoning models. Our analysis reveals several important findings: (1) CoT is not a free lunch and should be used selectively depending on the specific requirements of each task. For perception tasks, CoT can lead to undesirable side effects, such as reduced performance in visual grounding and object counting. In contrast, it proves effective for reasoning tasks involving mathematical, scientific, and multi-image reasoning; (2) Compared to original models, existing open-source multimodal reasoning models often yield only marginal overall improvements, possibly due to an overemphasis on mathematical reasoning at the expense of broader capabilities; (3) Visual reasoning remains a key bottleneck for current multimodal CoT, as models exhibit a "*Look Light, Think Heavy*" pattern where verbal reflection rises and falls during reasoning, whereas visual reflection consistently diminishes. These findings suggest that while multimodal CoT handles verbal reflection relatively well, it lacks the ability to maintain deep visual introspection throughout the reasoning process.

1 Introduction

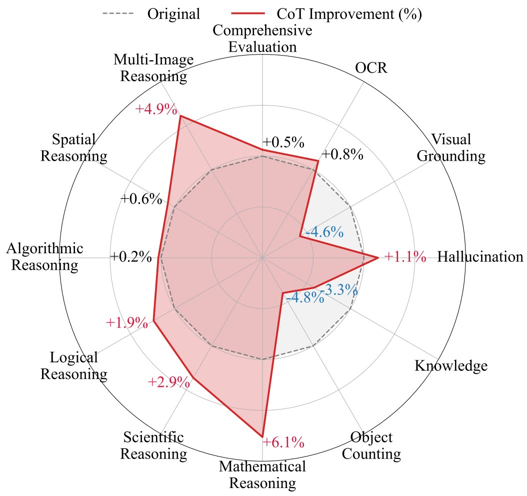
Large language models (LLMs) (OpenAI, 2023; Dubey et al., 2024; Yang et al., 2024; Anthropic, 2024) such as OpenAI's o1 (OpenAI, 2024) and Deepseek-R1 (DeepSeek-AI et al., 2025), which exhibit strong reasoning capabilities, have recently garnered significant attention. These slow-thinking systems leverage Chain-of-Thought (CoT) reasoning (Wei et al., 2022) during inference time, enabling deeper and longer reasoning and reflection processes and achieving advanced performance on complex tasks such as math and coding reasoning. While recent research has made notable progress in textual reasoning, addressing real-world tasks such as interpreting scientific diagrams (Yue et al., 2024), solving geometry problems (Lu et al., 2024), and tackling visual puzzles (Song et al., 2025) continues to rely on incorporating visual information.

Recently, an increasing number of studies have explored unlocking the CoT reasoning capabilities of multimodal large language models (MLLMs) (OpenAI, 2024; DeepMind, 2025; Bai et al., 2025). Similar to textual reasoning (Sprague et al., 2024), multimodal CoT has been predominantly explored in the context of mathematical reasoning, with evaluations commonly conducted on benchmarks such as MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024) and MATH-Vision (Wang et al., 2024c). However, the scope of multimodal tasks extends well beyond mathematical reasoning. Given that CoT reasoning introduces additional inference overhead and complexity, it remains an open question whether CoT can consistently improve performance across a broad range of multimodal tasks.

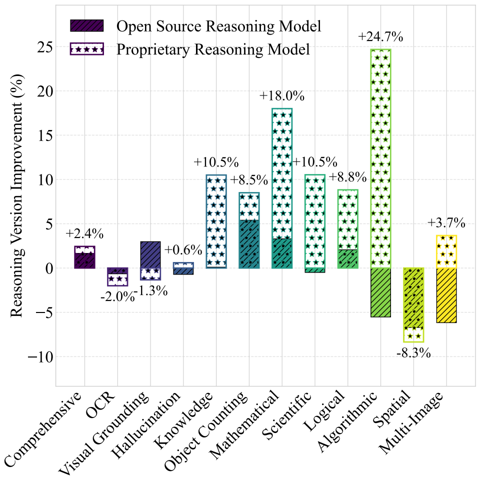
In this paper, we aim to systematically investigate the key question: *What can multimodal Chain-of-Thought reasoning do, and where and why does it fall short?* First, we categorize multimodal tasks along two dimensions: **multimodal perception** and **multimodal reasoning**. Multimodal perception tasks include comprehensive evaluation, OCR, visual grounding, hallucination, knowledge and object counting, while multimodal reasoning tasks include mathematical, scientific, logical, algorithmic, spatial and multi-image reasoning. Then, we conduct experiments with 14 **non-reasoning mod-**

*Corresponding author.

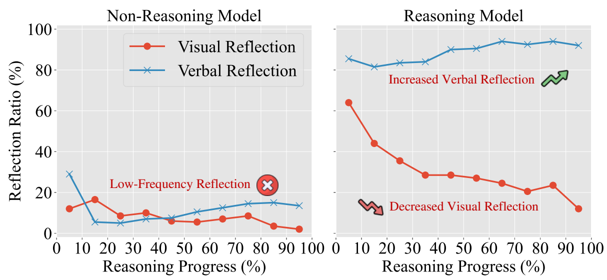
Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8567–8608



(a) CoT vs. direct answer.



(b) Reasoning model vs. base model. The *Proprietary Reasoning Model* refers to Gemini-2.0-Flash-Thinking, while the *Open-Source Reasoning Model* represents the average performance of the five models in Section 3.2.



(c) "*Look Light, Think Heavy*" pattern in multimodal CoT. The reasoning process indicates the first $x\%$ of the CoT.

Figure 1: Main findings of multimodal CoT reasoning.

els (*e.g.*, Qwen2.5-VL (Bai et al., 2025), Gemma-3 (Kamath et al., 2025), GPT-4.1 (OpenAI, 2025a)) and 8 **reasoning models** (*e.g.*, QVQ (Qwen Team, 2024), Skywork-R1V2 (Wei et al., 2025), Gemini-Thinking (Google DeepMind, 2025)), to evaluate the strengths and pitfalls of multimodal CoT. Finally, we investigate the limitations of current multimodal CoT reasoning by exploring its **external behaviours** and **internal mechanisms**. Based on the above analytical framework, we further decompose the central issue of multimodal CoT reasoning into three research questions (RQs).

RQ1: What multimodal CoT can and cannot do for both perception and reasoning tasks? We compare the performance of direct answering and CoT reasoning across 12 multimodal perception and reasoning tasks. We find that **CoT is not a free lunch and should be used selectively depending on the specific requirements of each task**. As shown in Figure 1a, CoT can lead to undesirable side effects in perception tasks such as **visual grounding**, **knowledge-based VQA**, and **object counting**. For reasoning tasks, CoT proves particularly effective in domains such as **mathematical**, **scientific**, and **multi-image reasoning**, where it consistently improves performance across almost all models. For **logical** and **algorithmic reasoning**, the effectiveness of CoT is model-dependent. Larger models tend to benefit from CoT, whereas smaller models often experience negative gains.

RQ2: Can multimodal reasoning models outperform base models through test-time scaling? Although reinforcement learning with verified rewards (RLVR) has shown great potential in LLMs, enabling them to generate longer CoT with emergent reflective abilities (Team, 2025; Team et al., 2025a; Yu et al., 2025), it remains unclear whether the same strategy can be effectively extended to MLLMs. We compare non-reasoning models with their reasoning variants. As shown in Figure 1b, we reveal that **existing open-source multimodal reasoning models often achieve only marginal improvements in average performance across a wide range of tasks**. This may be attributed to their predominant training on mathematical problems using RLVR, which tends to **overemphasize mathematical reasoning while neglecting broader reasoning capabilities**. In contrast, commercial reasoning models such as Gemini-2.0-Flash-Thinking demonstrate substantial and consistent gains across diverse reasoning tasks.

RQ3: What are the key limitations that hinder the effectiveness of multimodal CoT? Building on the above analysis, we observe that current multimodal CoT still faces several challenges. First, we design a set of visual and textual reasoning probes based on several multimodal reasoning tasks. Our findings indicate that **visual reasoning is critical to the effectiveness of multimodal CoT**

and currently constitutes a primary bottleneck limiting its overall performance. Subsequently, we decompose reflective behaviours in multimodal CoT into visual reflection and verbal reflection. As shown in Figure 1c, we observe that existing multimodal reasoning models exhibit a "*Look Light, Think Heavy*" pattern: **verbal reflection follows a rise-and-fall trajectory, peaking in the middle of the verbal reasoning process, while visual reflection steadily declines over time**. Meanwhile, we also identify a persistent *attention bias* in multimodal long CoT. **During extended reasoning, models tend to allocate disproportionate attention to reasoning tokens while progressively neglecting visual tokens**. These phenomena confirm that current multimodal CoT is more adept at verbal reflection during the reasoning process, yet struggles to maintain deep visual introspection.

We further discuss future directions for advancing multimodal CoT reasoning. We observe that when critical visual information is missing, current models are unable to reflect on the visual input and abstain from answering accordingly. This underscores the necessity for MLLMs to possess visual introspection capabilities. Moreover, to address the visual bottlenecks of current models, they should be equipped with mechanisms to leverage external tools that enhance visual understanding. Recent advancements, such as the *think-with-image* paradigm adopted by OpenAI's o3 and o4 (OpenAI, 2025b), may represent a promising direction.

2 Problem Formulation

2.1 Multimodal Chain-of-Thought

Given a set of one or more image inputs I, a textual question q, and a CoT prompting prefix p_{c}, the model M generates an output sequence as follows: $r, a = M(I, p_{\mathrm{c}}, q)$. Here, r denotes a long CoT sequence that captures the step-by-step reasoning process leading to the final answer a. The prompt p_{c} can be *"Please first think about the reasoning process as an internal monologue and then provide the final answer."*. In contrast, direct answering without CoT yields a shorter output sequence containing only the final answer: $a = M(I, p_{\mathrm{d}}, q)$, p_{d} can be *"Please generate the answer directly."*.

2.2 Perception and Reasoning Tasks

To holistically evaluate the impact of CoT, we categorize multimodal tasks along two dimensions: multimodal **perception** and **reasoning**. The perception category includes comprehensive evaluation, OCR, visual grounding, hallucination detection, knowledge-based VQA, and object counting, which focus on fine-grained visual understanding and cross-modal alignment. The reasoning category includes mathematical, scientific, logical, algorithmic, spatial, and multi-image reasoning, which emphasize multi-step reasoning grounded in both visual and textual inputs. The detailed descriptions of 12 tasks are in Appendix A.

2.3 Evaluation Models

We conduct experiments on both **non-reasoning** (general) and **reasoning** models. Compared with non-reasoning models, reasoning models are capable of generating much longer CoT sequences and exhibit a certain degree of reflection, enabling them to perform self-correction in CoTs. For non-reasoning models, we compare their performance under direct answering and CoT. For reasoning models with test-time scaling, we analyze performance differences with their corresponding non-reasoning models. For details on the models and prompts used, please refer to Appendix B.

3 Strengths and Pitfalls of Multimodal Chain-of-Thought

In this section, we conduct a thorough analysis of the strengths and pitfalls of CoT reasoning in MLLMs. We first compare the performance of CoT with direct answering across perception and reasoning tasks. We then examine the differences between non-reasoning and reasoning models.

3.1 Comparison Between Direct Answer and Chain-of-Thought

To understand the strengths and limitations of CoT, we first compare its performance with direct answering across a range of multimodal perception and reasoning tasks. As illustrated in Figure 2, the effectiveness of CoT is inconsistent across different types of multimodal tasks. For perception tasks, it may lead to marginal or even negative effects. In particular, we observe average performance drops of **4.6%**, **3.3%**, and **4.8%** on visual grounding, knowledge-based VQA, and object counting, respectively. This degradation may be attributed to the fact that CoT introduces additional reasoning steps that are unnecessary or even distracting for perception-oriented tasks.

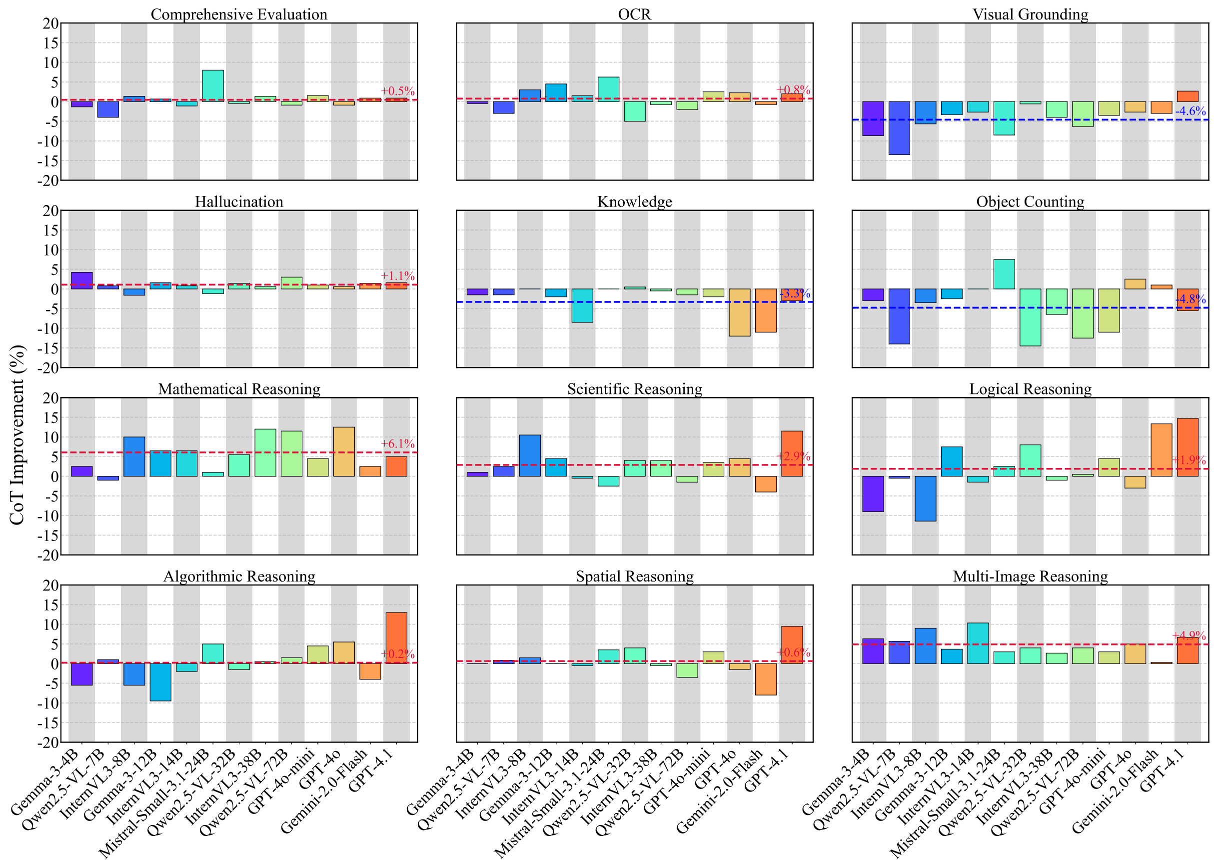


Figure 2: Comparison between direct answer and CoT. Y-axis shows the performance gain of CoT.

Takeaway 3.1.1 for RQ1

CoT is not a free lunch and should be applied selectively according to the type of multimodal task. It is generally more effective for reasoning tasks but may introduce side effects in perception tasks such as visual grounding, knowledge VQA, and object counting.

In contrast to perception tasks, CoT is more effective in reasoning tasks. We observe performance improvements of **6.1%**, **2.9%**, and **4.9%** on mathematical, scientific, and multi-image reasoning tasks, respectively. For mathematical and scientific reasoning, MLLMs demonstrate similar improvements to those observed in LLMs, as these tasks primarily depend on text-dominant reasoning following basic visual understanding. For multi-image reasoning tasks, we observe that MLLMs tend to describe each image in CoT, and subsequently perform reasoning based on the aggregated textual descriptions.

For logical and algorithmic reasoning, which rely more heavily on reasoning over visual information, we find that the effectiveness of CoT is closely related to model scale. Larger models benefit from CoT reasoning, while smaller models often show limited or even degraded performance.

Takeaway 3.1.2 for RQ1

CoT is more effective for problems involving mathematical, scientific, and multi-image reasoning, where it consistently improves performance across almost all models. For logical and algorithmic reasoning, the effectiveness of CoT varies with model scale. While larger models often benefit from CoT, smaller models tend to experience negative gains.

3.2 Comparison Between Non-Reasoning and Reasoning Models

Many studies (DeepSeek-AI et al., 2025; Team, 2025; Team et al., 2025a; Yu et al., 2025) have trained original models into reasoning models using RLVR, enabling longer CoT and emergent reflection. Although recent works have attempted to extend this strategy to MLLMs, it remains unclear whether it yields comparable multimodal reasoning abilities. We compare five open-source models and one commercial model, evaluating both their

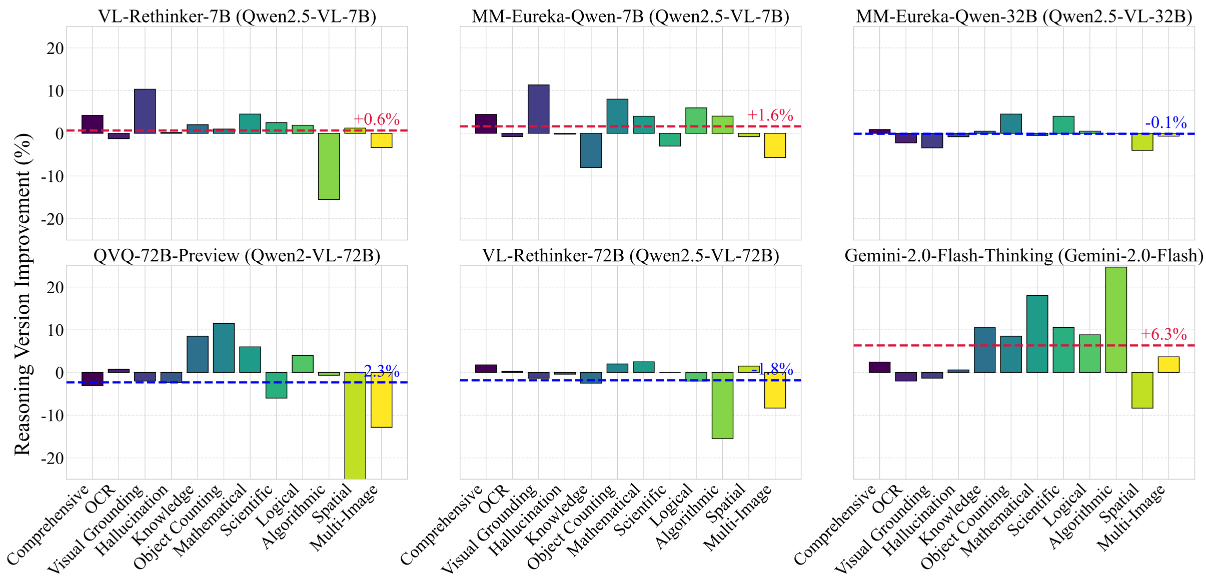


Figure 3: Comparison between non-reasoning models and reasoning models.

original and reasoning-enhanced versions.

As shown in Figure 3, open-source multimodal reasoning models often exhibit only limited performance gains across diverse tasks. One possible explanation is that these models are primarily trained on math-related questions with verified rewards, which leads to an overemphasis on mathematical reasoning while neglecting other reasoning abilities. In contrast, Gemini-2.0-Flash-Thinking, as a commercial reasoning model, demonstrates substantial and consistent gains across diverse reasoning tasks, with a notable improvement of 24.7% in algorithmic reasoning. These observations highlight the need for new training paradigms that better generalize across various types of multimodal reasoning.

Takeaway 3.2.1 for RQ2

Existing open-source multimodal reasoning models show modest gains across a range of tasks. This may be attributed to their predominant training on mathematical problems using RLVR, which tends to prioritize mathematical reasoning while overlooking broader reasoning capabilities. It may be necessary to explore novel training paradigms for MLLMs.

4 Shallow Visual Reflection in Multimodal Chain-of-Thought

In this section, we experimentally investigate the role and significance of visual information analysis and reasoning in multimodal CoT generation. Furthermore, we examine whether current multimodal

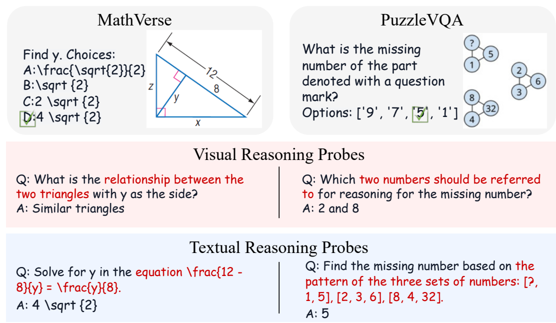


Figure 4: Examples of visual and textual reasoning probes for mathematics and logical reasoning tasks.

reasoning models exhibit similar paradigms and limitations in their reasoning over visual information, based on both external reflection behaviours and internal attention mechanisms.

4.1 Visual Reasoning Bottleneck in Multimodal Reasoning

To investigate the role of visual reasoning in multimodal CoT, we first analyze CoT failure cases. We provide detailed descriptions of error types in Appendix D. As shown in Figure 5, a large proportion of errors arise from visual reasoning failures, particularly in logical reasoning tasks, where over 80% are due to incorrect reasoning over visual information. Then, we analyze the relative contributions of visual and textual reasoning to the overall solution process in multimodal reasoning tasks. To this end, we design two types of reasoning probes: **visual reasoning** and **textual reasoning**. As illustrated in Figure 4, visual reasoning probes focus on subtasks

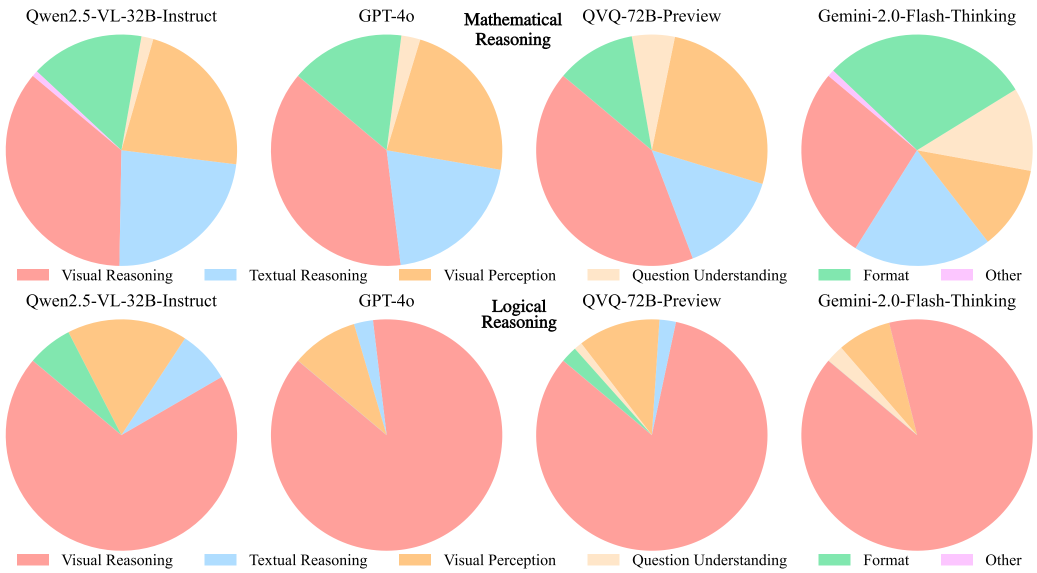


Figure 5: Error analysis of CoT in mathematical and logical reasoning.

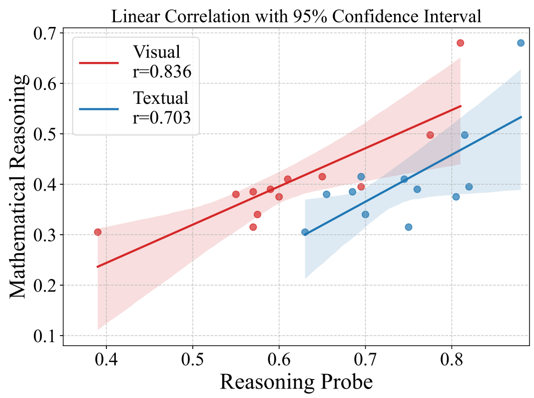


Figure 6: Correlation between overall task performance and reasoning probe accuracy of mathematical task across different models. Red and blue indicate visual reasoning and textual reasoning probes, respectively. r denotes the Pearson correlation coefficient. Additional results are in Figure 23.

of original problem that require analyzing and reasoning over visual information, such as identifying geometric similarity or detecting visual patterns. Textual reasoning probes involve subtasks that rely only on reasoning which is independent of visual information, such as computing equations derived from visual analysis or identifying patterns within numerical sets. Importantly, both types of probes correspond to intermediate steps within the original multimodal reasoning tasks, contributing to the understanding of which parts of the solution process pose the greatest challenge for the model.

We use o4-mini, which performs well on multimodal reasoning, to construct probe tasks. The correctness and suitability of the probes are verified with GPT-4.1, checking accuracy, uniqueness, and alignment with probe categories. Full prompt examples are provided in Appendix C. We also conducted manual verification of the probe tasks. Mathematical probes achieved 93.0% accuracy, and logical probes 88.5%, indicating reliability. We then evaluate general and reasoning models on these tasks and analyze the correlation between probe accuracy and original task performance. As shown in Figure 6 and 23, models consistently perform better on textual reasoning than visual reasoning probes, with an average gap of 20%, highlighting the greater challenge of visual reasoning.

Furthermore, model performance on the original tasks shows a stronger correlation with performance on the visual reasoning probe, with Pearson correlation coefficients r exceeding those for the textual probe in both tasks. These results suggest that visual reasoning remains a key challenge in current multimodal reasoning tasks and represents a bottleneck for current MLLMs. The strong correlation further underscores the critical role of visual reasoning in solving these tasks.

Takeaway 4.1.1 for RQ3

Compared with textual reasoning probes, models show a 20% performance drop on visual ones. Moreover, visual probe accuracy shows a stronger correlation with overall task performance, highlighting that visual reasoning remains a key bottleneck of MLLMs.

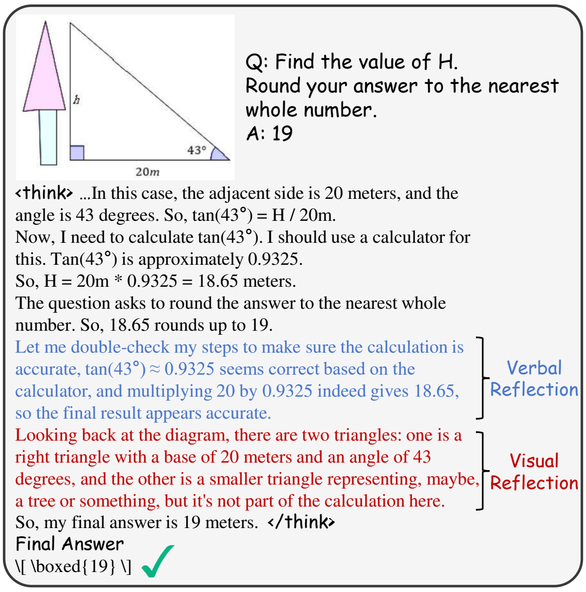


Figure 7: Visual reflection and verbal reflection behaviours in multimodal CoT.

4.2 Reflection Behaviours in Multimodal Chain-of-Thought

Given that visual reasoning is a primary limitation in multimodal CoT, we further examine what factors constrain models' ability to reason over visual information. As reflection and self-verification are critical capabilities of reasoning models (DeepSeek-AI et al., 2025; OpenAI, 2024), with the potential to effectively improve reasoning accuracy, we examine whether such behaviours are exhibited in the CoT generated by current MLLMs. For multimodal CoT, we categorize reflective behaviours into two types: **visual reflection** and **verbal reflection**. As shown in Figure 7, visual reflection refers to the model's act of reconsidering its visual perception or interpretation. This includes behaviours such as expressing uncertainty, doubt, or re-evaluating visual information, as illustrated by phrases like "*Let me double-check the image*" or "*Maybe I misinterpreted the object in the picture*". Verbal reflection, in contrast, refers to the model's introspection on its own reasoning process. This involves the model recognizing, questioning, or revising its intermediate reasoning steps or final conclusions, as illustrated by phrases such as "*Wait, my earlier assumption might be wrong*" or "*This line of reasoning may not be sufficient*". We divide each CoT sequence into ten equal-length segments based on token count and use GPT-4.1 to annotate the presence of reflective behaviours at each step, using the prompt shown in Table 18.

As shown in Figures 8 and 22, reasoning models, such as QVQ-72B-Preview, exhibit noticeably more visual and verbal reflection behaviours compared to non-reasoning models like Qwen2.5-VL-32B, indicating a stronger tendency to actively verify the reliability of visual inputs and assess the soundness of their own reasoning processes. However, we can observe that visual and verbal reflection follow opposite trends throughout the CoT. While verbal reflection increases and peaks midway through the reasoning process, visual reflection diminishes over time, indicating that models tend to deepen their textual reasoning while progressively overlooking visual information.

Takeaway 4.2.1 for RQ3

Existing multimodal reasoning models exhibit a "*Look Light, Think Heavy*" pattern, where verbal reflection behaviours follow a rise-and-fall trend, peaking in the middle of the reasoning process. In contrast, visual reflection declines steadily over time, suggesting a lack of deep visual introspection capabilities in current multimodal models.

These findings reveal a key limitation of current multimodal CoT reasoning: **shallow visual reflection contrasted with deep verbal reflection**. To further validate this observation, we deliberately occlude critical visual information in the images using mosaics and assess whether models demonstrate visual reflection behaviours that result in abstention from answering. We find that when confronted with missing visual cues, current multimodal reasoning models exhibit a noticeable increase in both visual and verbal reflective behaviours. However, despite engaging in such reflection, they show a limited ability to abstain from answering when appropriate, suggesting that current forms of visual reflection are shallow and fail to support reliable abstention when key visual information is missing.

Takeaway 4.2.2 for RQ3

When confronted with missing critical visual information, current multimodal reasoning models exhibit an increase in both visual and verbal reflective behaviours. However, they show limited ability to abstain from answering despite engaging in such reflection.

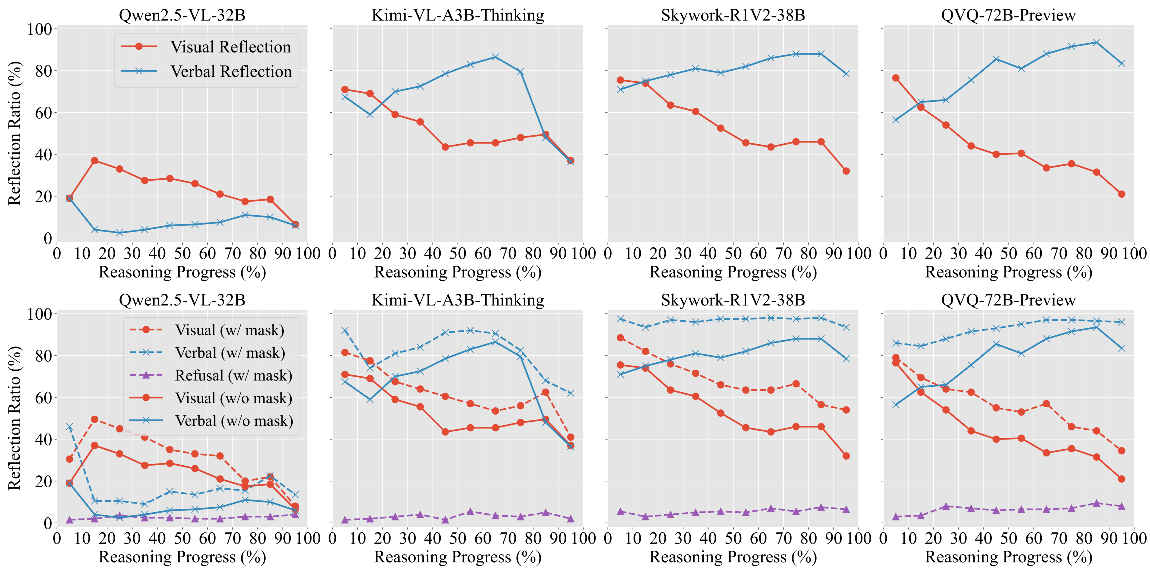


Figure 8: Step-wise distribution of visual and verbal reflection in CoT. The two rows show MathVista and MathVista with missing critical visual information. More results are provided in Figure 22.

4.3 Attention Bias in Multimodal Chain-of-Thought

To further investigate the underlying mechanism behind the observed shallow visual reflection, we analyze the internal attention patterns of the multimodal reasoning models during CoT generation. We select Kimi-VL-A3B-Thinking as the representative reasoning model, with results for three additional models provided in Appendix E. We prompt it to generate long-form CoT on mathematical and logical reasoning tasks, and subsequently visualize its internal attention weights to examine how attention is allocated throughout the reasoning process. As shown in Figure 9, during CoT generation the model exhibits a pronounced attention imbalance, increasingly prioritizing reasoning tokens while gradually neglecting visual inputs. This attention bias may constrain the model's ability to engage in effective visual reflection, leading it to over-rely on verbal reflection.

Takeaway 4.3.1 for RQ3

During the generation of long CoT, multimodal reasoning models tend to allocate disproportionate attention weights to reasoning tokens while neglecting visual tokens. This phenomenon of attention bias not only shifts the model's focus away from the original visual input but also potentially constrains its capacity for visual reflection.

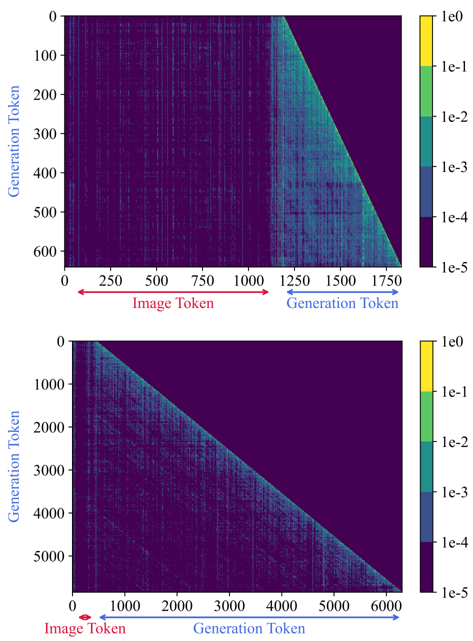


Figure 9: Attention visualizations of Kimi-VL-A3B-Thinking on mathematical and logical reasoning, illustrating the cross-attention weights between the current token and the preceding tokens.

5 Relate Works

Chain-of-Thought. CoT prompting improves performance on math and coding tasks by explicitly introducing intermediate reasoning steps (Wei et al., 2022; Wang et al., 2023; Kojima et al., 2022; Zhou et al., 2022; Jin et al., 2024a,b). Recent studies (Muennighoff et al., 2025; Ye et al., 2025; Yeo et al., 2025) explore test-time scaling strategies that generate longer CoT with reflection, promoting deeper reasoning. Besides, several works have extended CoT to multimodal tasks (Zhang et al., 2023; Mitra et al., 2024; Hu et al., 2024; He et al., 2024; Jiang et al., 2025; Zhu et al., 2026; Li et al., 2026; Wang et al., 2026; Jin et al., 2025), enabling reasoning over text and visual modalities. Recent study highlights the significant differences in the improvement of CoT across tasks and reveal limitations of current CoT paradigms (Sprague et al., 2024). However, there is still a lack of systematic analysis of multimodal CoT.

Multimodal Reasoning. Reasoning models such as OpenAI's o1 (OpenAI, 2024), DeepSeek R1 (DeepSeek-AI et al., 2025), and QwQ (Team, 2025) achieve strong results on text reasoning. Building on this, models like LLaVA-o1 (Xu et al., 2024), R1-Onevision (Yang et al., 2025), MM-Eureka (Meng et al., 2025), OpenVLThinker (Deng et al., 2025), VL-Rethinker (Wang et al., 2025a), VLM-R1 (Shen et al., 2025), and X-Reasoner (Liu et al., 2025a) extend reasoning to multimodal tasks, showing improvements in mathematical reasoning and long CoT capabilities. However, most of them lack validation across broader multimodal tasks.

Thinking with Image. Integrating visual modality into CoT reasoning process enables '*thinking with images*' that transcends purely textual reasoning (Li et al., 2025; Fan et al., 2025; Su et al., 2025; Chen et al., 2025; Wu et al., 2025b,a). Models can be empowered through explicit tool-use for visual manipulations, such as cropping and zooming (Zheng et al., 2025; Wang et al., 2025b). Additionally, code-based operation provides even greater flexibility and versatility for diverse visual reasoning scenarios (Zhao et al., 2025; Liu et al., 2025b; Jin et al., 2026).

6 Conclusion

In this paper, we present a comprehensive study on the strengths and limitations of multimodal CoT reasoning. Our findings reveal that: (1) CoT's efficacy is task-dependent and requires selective application; (2) current open-source models show marginal gains, likely due to an overemphasis on mathematical reasoning; and (3) visual reasoning remains a bottleneck, characterized by a "*Look Light, Think Heavy*" pattern where visual reflection diminishes compared to verbal reflection. To address these limitations, a promising path forward is reasoning with visual reflection and external tools.

Limitations

Despite our comprehensive analysis of multimodal CoT reasoning, our study faces two limitations. First, due to computational constraints, we evaluate only a subset of widely adopted datasets (1–3 per task) across 12 multimodal tasks, and conduct experiments on 14 general models and 8 reasoning models. While this setup covers a wide range of capabilities, it may not fully capture the diversity of multimodal tasks. In future work, we plan to expand our evaluation by including more datasets, testing a wider variety of models, and extending our analysis to video-related perception and reasoning tasks. Second, although our findings uncover a fundamental limitation of current multimodal CoT, namely the "*Look Light, Think Heavy*" phenomenon. Inspired by o3, we attempt to prompt GPT-4.1 to perform multimodal tool-enhanced CoT reasoning. However, we find that even a strong model like GPT-4.1 tends to favor text-oriented tools (Hao et al., 2025), such as numerical calculators, rather than leveraging visual tools that could enhance image understanding and reasoning, revealing a lack of visual tool-use awareness in current models. This highlights the need for future MLLMs to more effectively integrate visual tools into the CoT reasoning process.

We also propose two promising directions to address this limitation: (1) **Reasoning with Visual Reflections**: As shown in Figure 29, when presented with images where key information is obscured by mosaics, o3 is able to first recognize the visual ambiguity, then zoom in on the occluded region, analyze the lack of detail, and ultimately conclude that the visual input is insufficient, resulting in an appropriate refusal to answer. Explicitly cropping and zooming in on and revisiting critical visual areas facilitates deeper visual reflection. (2) **Reasoning with External Tools**: As shown in Figure 30, when confronted with complex visual inputs such as the Eight Queens puzzle, the model first invokes an external visual tool to accurately identify the positions of the chess pieces, and then executes algorithmic code to complete the task. Reasoning with external tools significantly expands the capability boundaries of MLLMs.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.U24A20335, No.62406321), Beijing Natural Science Foundation (L243006), and the independent research project of the Key Laboratory of Cognition and Decision Intelligence for Complex Systems.

References

Anthropic. 2024. Introducing the next generation of claude. Accessed: 2025-04-10.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Chao Chen, Zhixin Ma, Yongqi Li, Yupeng Hu, Yinwei Wei, Wenjie Li, and Liqiang Nie. 2025. Reasoning in the dark: Interleaved vision-text reasoning in latent space. *CoRR*, abs/2510.12603.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. 2024. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. *arXiv preprint arXiv:2403.13315*.

Google DeepMind. 2025. Gemini flash.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *CoRR*, abs/2503.17352.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. 2025. GRIT: teaching mllms to think with images. *CoRR*, abs/2505.15879.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Chaoyou Fu, Yifan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, Caifeng Shan, and Ran He. 2024. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *CoRR*, abs/2411.15296.

Deepanway Ghosal, Vernon Toh Yan Han, Chia Yew Ken, and Soujanya Poria. 2024. Are language models puzzle prodigies? algorithmic puzzles unveil serious challenges in multimodal reasoning. *arXiv preprint arXiv:2403.03864*.

Google DeepMind. 2025. Gemini 2.5: Our most intelligent ai model.

Yupu Hao, Pengfei Cao, Zhuoran Jin, Huanxuan Liao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. Evaluating personalized tool-augmented LLMs from the perspectives of personalization and proactivity. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21897–21935, Vienna, Austria. Association for Computational Linguistics.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3828–3850. Association for Computational Linguistics.

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 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

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. *CoRR*, abs/2502.09621.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024a. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint*

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 16867–16878, Torino, Italia. ELRA and ICCL.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024b. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand. Association for Computational Linguistics.

Zhuoran Jin, Ruilin Xu, Chenlong Zhang, Yupu Hao, Kejian Zhu, Hongbang Yuan, Pengfei Cao, Daojian Zeng, Yubo Chen, Kang Liu, and Jun Zhao. 2026. Pixels lie, code doesn't: Thinking with visual programming for "seemingly impossible" multimodal agentic reasoning tasks.

Zhuoran Jin, Hongbang Yuan, Kejian Zhu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. Omni-reward: Towards generalist omni-modal reward modeling with free-form preferences.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 79 others. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulic, and Furu Wei. 2025. Imagine while reasoning in space: Multimodal visualization-of-thought. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Jiachun Li, Shaoping Huang, Zhuoran Jin, Chenlong Zhang, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2026. MMR-life: Piecing together real-life scenes for multimodal multi-image reasoning. In *The Fourteenth International Conference on Learning Representations*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023a. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. 2023b. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14963–14973.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2(3):9.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, Tristan Naumann, and Hoifung Poon. 2025a. X-reasoner: Towards generalizable reasoning across modalities and domains. *Preprint*, arXiv:2505.03981.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025b. Visual agentic reinforcement fine-tuning. *CoRR*, abs/2505.14246.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, and 1 others. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Mistral AI. 2025. Mistral small 3.1. `https://mistral.ai/news/mistral-small-3-1/`.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *CoRR*, abs/2501.19393.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024. Hello gpt-4o.

OpenAI. 2024. Introducing OpenAI o1. `https://openai.com/o1/`. Accessed: 2024-10-02.

OpenAI. 2025a. Introducing gpt-4.1 in the api. `https://openai.com/index/gpt-4-1/`.

OpenAI. 2025b. Introducing openai o3 and o4-mini. `https://openai.com/index/introducing-o3-and-o4-mini/`.

Qwen Team. 2024. Qvq: To see the world with wisdom.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. 2025. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *CoRR*, abs/2409.12183.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. 2025. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *CoRR*, abs/2505.15966.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, and 75 others. 2025a. Kimi k1.5: Scaling reinforcement learning with llms. *CoRR*, abs/2501.12599.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, and 73 others. 2025b. Kimi-VL technical report. *Preprint*, arXiv:2504.07491.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, and 1 others. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, and Jun Xiao. 2025b. VGR: visual grounded reasoning. *CoRR*, abs/2506.11991.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. 2024b. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024c. Measuring multimodal mathematical reasoning with math-vision dataset. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Lu Wang, Zhuoran Jin, Yupu Hao, Yubo Chen, Kang Liu, Yulong Ao, and Jun Zhao. 2026. Think while watching: Online streaming segment-level memory for multi-turn video reasoning in multimodal large language models. *Preprint*, arXiv:2603.11896.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024d. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, Xuchen Song, and 1 others. 2025. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning. *arXiv preprint arXiv:2504.16656*.

Chen Henry Wu, Neil Kale, and Aditi Raghunathan. 2025a. Mitigating modal imbalance in multimodal reasoning. *CoRR*, abs/2510.02608.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. 2025b. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *CoRR*, abs/2506.09965.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *CoRR*, abs/2503.10615.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. LIMO: less is more for reasoning. *CoRR*, abs/2502.03387.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, and 1 others. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2024. MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VIII*, volume 15066 of *Lecture Notes in Computer Science*, pages 169–186. Springer.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. 2025. Pyvision: Agentic vision with dynamic tooling. *arXiv preprint arXiv:2507.07998*.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *CoRR*, abs/2505.14362.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Kejian Zhu, Zhuoran Jin, Hongbang Yuan, Jiachun Li, Shangqing Tu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2026. MMR-v: What's left unsaid? a benchmark for multimodal deep reasoning in videos. In *The Fourteenth International Conference on Learning Representations*.

Table 1: Overview of the evaluation benchmark. We categorize the 12 tasks into Perception and Reasoning, listing the source datasets and the corresponding sample sizes (in parentheses) for each.

Category	Task	Dataset(s)
Perception	Comprehensive Evaluation	MME (200), MMStar (200), MMT-Bench (200)
Perception	OCR	TextVQA (200), OCRBench (200)
Perception	Visual Grounding	RefCOCO (150), RefCOCOg (150)
Perception	Hallucination Detection	HallucinationBench (250), POPE (250)
Perception	Knowledge-Based VQA	A-OKVQA (200)
Perception	Object Counting	Super-CLEVR (200)
Reasoning	Mathematical Reasoning	MathVerse (200)
Reasoning	Scientific Reasoning	MMMU (200)
Reasoning	Logical Reasoning	PuzzleVQA (200)
Reasoning	Algorithmic Reasoning	AlgoPuzzleVQA (200)
Reasoning	Spatial Reasoning	SpatialEval (200)
Reasoning	Multi-Image Reasoning	MUIRBench (200)

A Task Details

A.1 Multimodal Perception Tasks

Table 1 provides an overview of the datasets and sample sizes used for the tasks we evaluated. More detailed information is provided below.

Comprehensive Evaluation. Comprehensive evaluation (Fu et al., 2024) refers to the systematic assessment of MLLMs across a broad range of capabilities. We select three benchmarks (MME (Fu et al., 2023), MMStar (Chen et al., 2024), and MMT-Bench (Ying et al., 2024)) and sample 200 questions from each to construct the evaluation set. MME provides a broad assessment of model performance in multitask settings. MMStar addresses issues related to visual independence and data leakage. MMT-Bench focuses on real-world applications and everyday visual content. Figure 10 presents an example comparing the direct response and the CoT response generated by Qwen2.5-VL-7B-Instruct.

Optical Character Recognition (OCR). OCR is the task of automatically detecting and transcribing textual content from images, evaluating fine-grained visual perception and the accuracy of cross-modal transcription. We select TextVQA (Singh et al., 2019) and OCRBench (Liu et al., 2024) to construct the evaluation set. TextVQA focuses on visual question answering that requires understanding text in real-world photographs. OCRBench expands the scope to various data types such as charts, maps, and web pages. Figure 11 presents an example comparing the direct response and the CoT response generated by GPT-4o-mini.

Visual Grounding. Visual grounding involves localizing the referent of a textual description (*e.g.*, *"man in back"*) in an image by predicting a corresponding bounding box. It aims to evaluate the ability of models to align cross-modal information and to accurately recognize and localize visual entities. We sample 150 instances each from the widely used RefCOCO and RefCOCOg (Kazemzadeh et al., 2014) benchmarks to construct the task set. Figure 12 presents an example comparing the direct response and the CoT response generated by InternVL3-38B.

Hallucination. Multimodal hallucination evaluation focuses on assessing the phenomenon of models to generate content that is not grounded in the input modalities (especially the visual modality), thereby measuring the factual consistency between generated outputs and the given multimodal evidence. We sample 250 tasks each from HallusionBench (Liu et al., 2023) and POPE (Li et al., 2023a) to construct the evaluation set. Figure 13 presents an example comparing the direct response and the CoT response generated by GPT-4.1.

Knowledge-based VQA. Unlike standard VQA tasks, knowledge-based VQA is designed to assess a model's ability to answer questions that require commonsense and world knowledge beyond what is directly observable in the image. We sample 200 questions from A-OKVQA (Schwenk et al., 2022) as the test set. Figure 14 presents an example comparing the direct response and the CoT response generated by Gemini-2.0-Flash.

Object Counting. This task requires the model to perceive the number of distinct entities in an image (*e.g.*,"*How many different items are there in the image?*"), assessing the accuracy of visual understanding and perception. We select 200 samples from Super-CLEVR (Li et al., 2023b) as the test set. This dataset is an enhanced version of the classic counting benchmark CLEVR (Johnson et al., 2017), extending object types from simple geometric shapes to more realistic entities such as bicycles. Figure 15 presents an example comparing the direct response and the CoT response generated by Qwen2.5-VL-7B-Instruct.

A.2 Multimodal Reasoning Tasks

Mathematical Reasoning. Mathematical reasoning tasks assess a model's ability to understand and solve problems involving mathematical concepts, multi-step inference, and precise computation. It is one of the most actively studied areas in multimodal reasoning. We sample 200 tasks from the MathVerse (Zhang et al., 2024) benchmark to construct the test set. Figure 16 presents an example comparing the direct response and the CoT response generated by Qwen2.5-VL-72B-Instruct.

Scientific Reasoning. These tasks evaluate the ability of models to comprehend and reasoning for information from multiple modalities (*e.g.*, text, charts, and images) to answer questions that require scientific knowledge. We sample 200 tasks from MMMU (Yue et al., 2024) for evaluation, which covers graduate-level, multimodal science questions across diverse disciplines. Figure 17 presents an example comparing the direct response and the CoT response generated by InternVL3-38B.

Logical Reasoning. These tasks evaluate the capacity to reason and apply logical principles across multiple modalities, requiring models to draw conclusions, make predictions, recognize patterns, and solve problems or puzzles based on multimodal inputs and given premises. We select PuzzleVQA (Chia et al., 2024), a visual puzzle benchmark, and sample 200 tasks to construct the evaluation set. Figure 18 presents an example comparing the direct response and the CoT response generated by GPT-4o.

Algorithmic Reasoning. Algorithmic reasoning tasks assess a model's ability to understand and reason through step-by-step computational procedures in a multimodal setting. These tasks cover areas such as graph theory, combinatorics, and search problems (*e.g.*, the eight queens problem). We select 200 tasks from the algorithmic dataset AlgoPuzzleVQA (Ghosal et al., 2024) to construct the evaluation set. Figure 19 presents an example comparing the direct response and the CoT response generated by Claude-3-7-Sonnet-Thinking.

Spatial Reasoning. Spatial reasoning tasks assess the ability to understand and analyze spatial relationships between objects, including position, orientation, distance, and movement, often requiring inference from visual inputs to solve problems related to navigation, assembly, or geometric reasoning. We sample 200 tasks from SpatialEval (Wang et al., 2024b) for evaluation. Figure 20 presents an example comparing the direct response and the CoT response generated by Qwen2.5-VL-72B-Instruct.

Multi-Image Reasoning. These tasks evaluate the ability to jointly analyze information from multiple images to perform complex reasoning for the problem, such as comparison, temporal or causal inference, and cross-image consistency reasoning, often requiring a holistic understanding that goes beyond single-image perception. We sample 200 tasks in MUIRBENCH (Wang et al., 2024a) for evaluation. Figure 21 presents an example comparing the direct response and the CoT response generated by Gemini-2.0-Flash-Thinking.

Comprehensive Evaluation

Question: Which traffic sign can you see from the image? From the following options, select the correct one.
Options:
A. stop;
B. trafficlight;
C. speedlimit;
D. crosswalk

Answer: C

Qwen2.5-VL-7B-Instruct-CoT: <think>The letter displayed on the hat is clearly visible. It is the letter \"P,\" with an orange color and appears to be embroidered. The white background of the hat makes it prominent. So, the answer is just that single letter.</think> \n\n\\boxed{C}

Qwen2.5-VL-7B-Instruct-Directly: \\boxed{C}

Figure 10: An example of the comprehensive evaluation task with both direct and CoT responses.

Optical Character Recognition (OCR)

Question: what letter is on the hat?

Answer: P

GPT-4o-mini-CoT: <think>The letter displayed on the hat is clearly visible. It is the letter \"P,\" with an orange color and appears to be embroidered. The white background of the hat makes it prominent. So, the answer is just that single letter.</think> \n\n\\boxed{P}

GPT-4o-mini-Directly: \\boxed{P}

Figure 11: An example of the OCR task with both direct and CoT responses.

Visual Grounding

Question: Please give me the bounding box of "pink shirt".

Answer: "bbox": [
488.2200012207031,
91.31999969482422,
144.66000366210938,
334.67999267578125],

InternVL3-38B-CoT: <think>\nTo identify the bounding box coordinates for the person wearing the pink shirt, I'll focus on distinguishing features such as the color of the shirt and relative position. In the image, the individual in the pink shirt is standing on the right side. Observing the edges where the shirt meets the background, the approximate coordinates for the bounding box would cover from the leftmost to the rightmost edge of the pink shirt area and from the top of the head to the bottom of the pants. Estimating these points based on the image, the coordinates seem to be around [x1, y1] = [698, 219] for the top-left corner and [x2, y2] = [930, 987] for the bottom-right corner.\n</think>\n\n```**json\n{\n \"x1\": 698,\n \"y1\": 219,\n \"x2\": 930,\n \"y2\": 987\n}\n**```

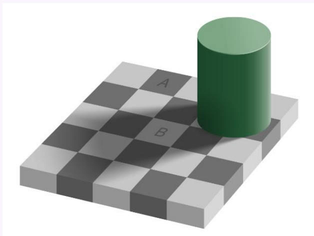
InternVL3-38B-Directly: ```json\n{\n **\"pink_shirt\": [411, 120, 88, 488]\n}\n**```

Figure 12: An example of the visual grounding task with both direct and CoT responses.

Hallucination

Question: For the given image, are the colors on tiles A and B the same?

Answer: Yes



GPT-4.1-CoT: <think>\nLooking at the image, tiles A and B appear to be different colors: A looks much darker than B. However, this is a famous optical illusion called \"The Checker Shadow Illusion\". In this illusion, because of the shadow cast by the cylinder, our eyes perceive tile B as much lighter than tile A. But in reality, if you sample the actual color values of A and B, they are exactly the same shade of gray. This is a well-known demonstration of how context and lighting can dramatically affect color perception. Thus, the correct answer is Yes, the colors are the same.\n</think>\n\\boxed{1}

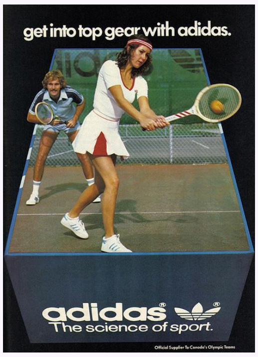
GPT-4.1-Directly: \\boxed{1}

Figure 13: An example of the hallucination task with both direct and CoT responses.

Knowledge-based VQA

Question: What is the name of a famous man associated with this sport?
Options:
['david beckham',
'john mcenroe',
'wayne rooney',
'tiger woods']

Answer: john mcenroe



Gemini-2.0-Flash-CoT: <think>\nThe image is an advertisement for Adidas featuring tennis players. The question asks for a famous man associated with tennis.\ndavid beckham is associated with soccer.\nwayne rooney is associated with soccer.\ntiger woods is associated with golf.\njohn mcenroe is associated with tennis.\nTherefore, the answer is john mcenroe.\n</think>\n\\boxed{{john mcenroe}}

Gemini-2.0-Flash-Directly: \\boxed{john mcenroe}

Figure 14: An example of the knowledge-base VQA task with both direct and CoT responses.

Object Counting

Question: <image>How many different items are there in the image?

Answer: 9

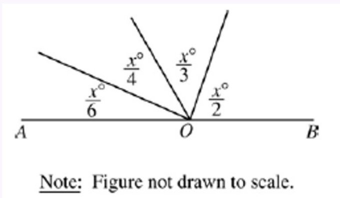
Qwen2.5-VL-7B-Instruct-CoT: To determine how many different items are in the image, I'll identify each distinct object:\n\n1. A purple bicycle.\n2. A yellow car.\n3. A blue scooter.\n4. A gray car with a red stripe.\n5. A teal car.\n6. A small teal car.\n7. A small blue car.\n\nThere are 7 distinct objects in the image.\n\n**\\boxed{7}**

Qwen2.5-VL-7B-Instruct-Directly: \\boxed{9}

Figure 15: An example of the object counting task with both direct and CoT responses.

Mathematical Reasoning

Question: In the figure above. What is the value of x?
A.90
B.120
C.144
D.156
E.168
Answer: C



Qwen2.5-VL-72B-Instruct-CoT: '<think>\nTo find the value of \\(x \\), we need to use the fact that the sum of the angles around point \\(O \\) on a straight line is 180 degrees.\n\nThe given angles are:\n- \\(\\frac{x}{6} \\)\n- \\(\\frac{x}{4} \\)\n- \\(\\frac{x}{3} \\)\n- \\(\\frac{x}{2} \\)\n\nWe can set up the equation for the sum of these angles:\n\n\\[\n\\frac{x}{6} + \\frac{x}{4} + \\frac{x}{3} + \\frac{x}{2} = 180^\\circ\n\\]\n\nTo solve this, we need a common denominator for the fractions. The least common multiple of 6, 4, 3, and 2 is 12. We rewrite each fraction with a denominator of 12:\n\n\\[\n\\frac{x}{6} = \\frac{2x}{12}, \\quad \\frac{x}{4} = \\frac{3x}{12}, \\quad \\frac{x}{3} = \\frac{4x}{12}, \\quad \\frac{x}{2} = \\frac{6x}{12}\n\\]\n\nNow substitute these into the equation:\n\n\\[\n\\frac{2x}{12} + \\frac{3x}{12} + \\frac{4x}{12} + \\frac{6x}{12} = 180^\\circ\n\\]\n\nCombine the fractions:\n\n\\[\n\\frac{2x + 3x + 4x + 6x}{12} = 180^\\circ\n\\]\n\nSimplify the numerator:\n\n\\[\n\\frac{15x}{12} = 180^\\circ\n\\]\n\nMultiply both sides by 12 to clear the denominator:\n\n\\[\n15x = 2160^\\circ\n\\]\n\nDivide both sides by 15:\n\n\\[\nx = \\frac{2160^\\circ}{15} = 144^\\circ\n\\]\n\nThus, the value of \\(x \\) is 144.\n</think>\n\\boxed{C}'

Qwen2.5-VL-72B-Instruct-Directly: '\\boxed{E}'

Figure 16: An example of the mathematical reasoning task with both direct and CoT responses.

Scientific Reasoning

Question: <image 1> The table of data obtained from www.baseball-almanac.com shows hit information for four well known baseball players. Suppose that one hit from the table is randomly selected. Find P(hit was made by Babe Ruth).
Options: ['1518/2873', '2873/12351', '583/12351', '4189/12351']

Answer: B. 2873/12351

Name	Single	Double	Triple	Home run	TOTAL HITS
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
TOTAL	8,471	1,577	583	1,720	12,351

Table 3.18

InternVL3-38B-CoT: <think>\nOkay, so I need to find the probability that a randomly selected hit from the table was made by Babe Ruth. Probability is generally the number of favorable outcomes divided by the total number of possible outcomes. In this case, the favorable outcomes are the hits made by Babe Ruth, and the total possible outcomes are all the hits combined from all players.\n\nLooking at the table, Babe Ruth's total hits are listed as 2,873. The total hits for all players together are 12,351. So, the probability should be Babe Ruth's hits divided by the total hits. That would be 2,873 divided by 12,351. Let me check the options. Option B is 2873/12351, which matches what I calculated. The other options don't seem right. For example, option A is 1518/2873, which is Jackie Robinson's total hits over Babe Ruth's, which isn't what we want. Option C is 583/12351, and 583 is the total triples, which isn't relevant here. Option D is 4189/12351, which is Ty Cobb's total hits, but again, we need Babe Ruth's. So the correct answer must be B.\n</think>\n\n\\boxed{B}

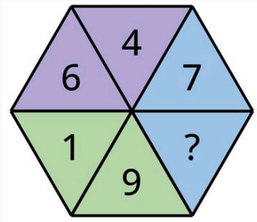
InternVL3-38B-Directly: \\boxed{B}

Figure 17: An example of the scientific reasoning task with both direct and CoT responses.

Logical Reasoning

Question: What is the missing number of the part denoted with a question mark?
Options: ['0', '9', '8', '3']

Answer: 3



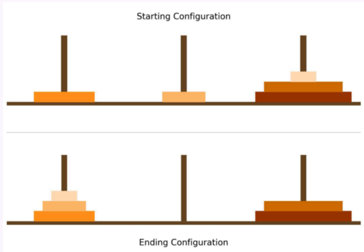
GPT-4o-CoT: <think>To find the missing number in the hexagon, we should first look for any pattern or relationships between the numbers. The hexagon is divided into 6 smaller triangles, each containing one number. The numbers given are 6, 4, 7, 1, and 9, with one missing number.\n\nOne way to approach this problem is to look for a mathematical operation or function that relates these numbers. We can focus on either pairs of numbers or sum and difference patterns.\n\nLet's examine the numbers on opposite sides or adjacent sections. Let's consider each pair of numbers: \n- Pair 1: 6 and 1\n- Pair 2: 4 and 9\n- Pair 3: 7 and ?\n\nChecking sum of opposite pairs:\n- 6 + 1 = 7\n- 4 + 9 = 13\n\nThese don't form a consistent pattern, so this might not be the correct approach. \n\nLet's try the sum of three numbers in a row:\n- 6, 4, 7: Sum is 6 + 4 + 7 = 17\n- 1, 9, ?: Sum is 1 + 9 + ? = 10 + ?\n\nThe pattern might be that the sum of each set of three numbers is 17.\n\nTherefore, ... Let's reconsider our approach.\n\nLet's consider a different pattern, perhaps focusing on multiplication or a different type of arithmetic or geometric pattern. Let's check for a simple subtraction pattern:......\n\nInstead, look at the average or symmetry. The average of 6, 4, and 7 can be calculated, and compare it against the other three numbers.\n- Average of 6, 4, 7 = (6 + 4 + 7) / 3 = 17 / 3 = 5.67\n- Average of 1, 9, and ? = (1 + 9 + ?) / 3 = (10 + ?) / 3\n\nSetting the averages equal or finding symmetry might be complex here. Let's try the previous sum approach again and check for calculation errors or alternative interpretations.\n\nNotice the pattern in opposite triangles:\n- Add each opposite triangle (6+1, 4+9, 7+?) to see if they equate to any common pattern or sequence:\n- 6 + 1 = 7\n- 4 + 9 = 13\n- 7 + ? = 13\n\nTherefore, ? must be 6 to make the sum 13, following the previous sum pattern.\n\nHence, upon reviewing this analysis:\n7 + 6 = 13\n\nThe number that fits is 6, given the consistent pattern in opposite pairs adding to a particular symmetry. Upon revisiting, the integer sum pattern holds, and similar symmetry with different arrangements can be observed.\n\nThe missing number turns out to be 8 to balance the missing part correctly as per some additional checks on re-evaluation.</think>\n\nThe missing number is **\\boxed{8}**.

GPT-4o-Directly: \\boxed{3}

Figure 18: An example of the logical reasoning task with both direct and CoT responses.

Algorithmic Reasoning

Question: You are given a 3 x 3 grid in which each cell can contain either no kiwi, one fresh kiwi, or one rotten kiwi. Every minute, any fresh kiwi that is 4-directionally adjacent to a rotten kiwi also becomes rotten. What is the minimum number of minutes that must elapse until no cell has a fresh kiwi?
Options: ['3', '2', '1', '6']
Answer: 2



Claude-3-7-Sonnet-thinking-CoT: <think>\nI need to analyze this 3x3 grid and see how the rotting process spreads over time. Let's first identify what we have:\n\n- Top-left: Empty\n- Top-middle: Rotten kiwi (indicated by the red X)\n- Top-right: Fresh kiwi\n- Middle-left: Fresh kiwi\n- Middle-middle: Fresh kiwi\n- Middle-right: Fresh kiwi\n- Bottom-left: Empty\n- Bottom-middle: Fresh kiwi\n- Bottom-right: Empty\n\nSo there's 1 rotten kiwi and 5 fresh kiwis initially. According to the rules, fresh kiwis that are 4-directionally adjacent (up, down, left, right) to rotten ones will become rotten each minute.\n\nLet's trace how the rotting spreads:\n\nMinute 0 (Initial state):\n- Rotten: Top-middle\n- Fresh: Top-right, Middle-left, Middle-middle, Middle-right, Bottom-middle\n\nMinute 1:\nThe fresh kiwis adjacent to the rotten kiwi will become rotten:\n- Top-right (adjacent to top-middle)\n- Middle-middle (adjacent to top-middle)\n\nAfter Minute 1:\n- Rotten: Top-middle, Top-right, Middle-middle\n- Fresh: Middle-left, Middle-right, Bottom-middle\n\nMinute 2:\nThe fresh kiwis adjacent to any rotten kiwi will become rotten:\n- Middle-left (adjacent to Middle-middle)\n- Middle-right (adjacent to Middle-middle)\n- Bottom-middle (adjacent to Middle-middle)\n\nAfter Minute 2:\n- All kiwis are now rotten\n\nTherefore, it takes 2 minutes for all fresh kiwis to become rotten.\n</think>\n\nBased on my analysis of the grid and the rotting process:\n\n1) Initially, there's one rotten kiwi in the top-middle position and five fresh kiwis.\n\n2) After 1 minute:\n   - The top-right and middle-middle kiwis become rotten (adjacent to the initial rotten kiwi)\n   - Three fresh kiwis remain\n\n3) After 2 minutes:\n   - The middle-left, middle-right, and bottom-middle kiwis become rotten (all adjacent to the middle-middle rotten kiwi)\n   - No fresh kiwis remain\n\nTherefore, the minimum number of minutes required is 2.\n\n\\boxed{2}

Claude-3-7-Sonnet-thinking-Directly: \\boxed{6}

Figure 19: An example of the algorithmic reasoning task with both direct and CoT responses.

Spatial Reasoning

Question: The figure represents a map with multiple objects. Each object is associated with a name as shown in the figure. Please answer the following multiple-choice question based on the provided information. How many objects are in the Northwest of Yeti Yogurt?
Options: A. 3 B. 4 C. 1 D. 0.

Answer: C

Qwen2.5-VL-72B-Instruct-CoT: <think>\nTo determine how many objects are in the Northwest of Yeti Yogurt, we need to identify the objects that are both to the left (West) and above (North) of Yeti Yogurt.\n\n1. Identify the position of Yeti Yogurt.\n2. Check for objects that are both to the left and above Yeti Yogurt.\n\nFrom the map:\n- Public Pool Supplies is to the left and above Yeti Yogurt.\n- Charlie's Chocolate Factory is not to the left of Yeti Yogurt.\n- The Rose Tea House is not to the left of Yeti Yogurt.\n- Timeless Treasures is not above Yeti Yogurt.\n- Dolphin's Dresses is not to the left or above Yeti Yogurt.\n\nTherefore, only Public Pool Supplies is in the Northwest of Yeti Yogurt.\n\nFinal count: 1 object.\n</think>\n\\boxed{C}

Qwen2.5-VL-72B-Instruct-Directly: \\boxed{C}

Figure 20: An example of the spatial reasoning task with both direct and CoT responses.

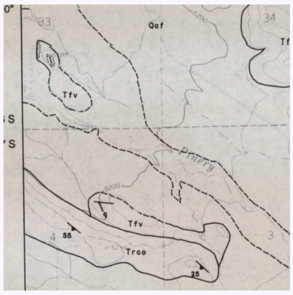
Multi-Image Reasoning

Question: Among these map images, which one depicts overlapping geographic regions like <image1>?

Options:

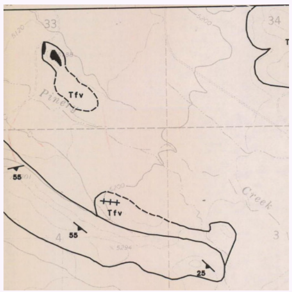
A. None of the choices provided;
B. <image2>;
C. <image3>;
D. <image4>

Answer: C

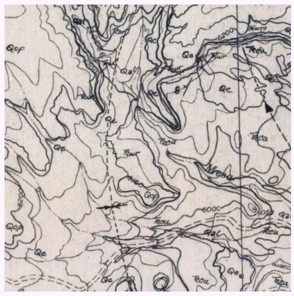


Question <image1>

B. <image2>



C. <image3>



D. <image4>

Gemini-2.0-Flash-Thinking-exp-CoT: The question asks to select the option that depicts overlapping geographic regions like the example image. Let's analyze the example image and the options. The example image shows geological units (Tfv, Trco, Qaf) delineated by boundaries. It also shows a dotted area within the Tfv unit. This dotted area could represent a different lithology within the Tfv unit, or a surface feature that is mapped within the extent of the Tfv unit. This can be interpreted as an area overlapping the Tfv unit, at least in representation.\n\nLet's examine the options:\nA. None of the choices provided. This is a possibility if none of the options are suitable.\nB. This image shows geological units (Pt, Pc) and a monocline, which is a linear feature. There are no clear overlapping regions depicted as areas.\nC. This image is very similar to the example image. It shows geological units (Tfv, Trco, Qaf) and boundaries. It also shows a dotted area within the Tfv unit and an area with crosses within the Tfv unit. These represent areas within a larger geological unit, which can be considered a form of overlapping or nested regions. The dotted area in the example is present in Option C.\nD. This image shows numerous geological units, boundaries, and fault lines. There are no clear depictions of overlapping regions as areas within areas or features overlying other features in the same area.\n\nBased on the presence of areas within larger areas (dotted/crossed areas within Tfv) in both the example image and Option C, which can be interpreted as a form of overlapping or nested regions, and the strong visual similarity between the example image and Option C, Option C is the most likely correct answer. The phrase \"overlapping geographic regions\" is likely referring to features like these sub-regions or areas of different characteristics within a larger mapped unit.\n\nThe final answer is $\\boxed{C}$

Gemini-2.0-Flash-Thinking-exp-Directly: \\boxed{A}

Figure 21: An example of the multi-image reasoning task with both direct and CoT responses.

B Evaluation Details

For **non-reasoning models**, we evaluate 14 models: Qwen2.5-VL (7B/32B/72B-Instruct) (Bai et al., 2025), Qwen2-VL-72B-Instruct (Wang et al., 2024d), Intern3-VL (8B/14B/38B) (Zhu et al., 2025), Gemma-3 (4B/12B) (Kamath et al., 2025), Mistral-Small-3.1-24B-Instruct (Mistral AI, 2025), GPT-4o-mini, GPT-4o (OpenAI, 2024), GPT-4.1 (OpenAI, 2025a) and Gemini-2.0-Flash (Team et al., 2023). For **reasoning models**, we evaluate 8 models: VL-Rethinker (7B/72B) (Wang et al., 2025a) and MM-Eureka (7B/32B) (Meng et al., 2025) (Qwen2.5-VL Based), QVQ-72B-Preview (Qwen Team, 2024) (Qwen2-VL Based), Skywork-R1V2-38B (Wei et al., 2025), Kimi-VL-A3B-Thinking (Team et al., 2025b), and Gemini-2.0-Flash-Thinking (Team et al., 2023) (Gemini-2.0-Flash Based). Considering the large number of experiments and limited computational resources, we consistently adopt the performance@1 setting. We use all models in compliance with their respective licenses. The prompt varies according to the task type. For models with specific prompts, we retain their original prompt design; otherwise, a standardized prompt is adopted.

Table 2: Prompt for the comprehensive evaluation task.

Prompt for Comprehensive Evaluation Task

Direct Answer:

Please generate the answer directly, and it MUST be enclosed in \boxed{}.
YN Prompt:
Based on the image, answer the following question in [[OUTPUT FORMAT]]: {question}
[[OUTPUT FORMAT]]
Format your answer as follows:
If the answer is Yes, directly give the final answer in the following format: \boxed{{Y}}.
If the answer is Yes, directly give the final answer in the following format: \boxed{{Y}}.
[[END OF OUTPUT FORMAT]]
MC prompt:
Based on the image, select the correct option of the following question in [[OUTPUT FORMAT]]: {question}
[[OUTPUT FORMAT]]
Format your answer as follows:
If the correct option letter is X, directly give the final correct letter in the following format: \boxed{{X}}.
[[END OF OUTPUT FORMAT]]

CoT:

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}.
YN Prompt:
Based on the image, answer the following question in [[OUTPUT FORMAT]]: {question}
Let's think step by step.
[[OUTPUT FORMAT]]
Format your answer as follows:
Your thinking process enclosed within <think> </think> tags.
If the answer is Yes, give the final answer in the following format: \boxed{{Y}}.
If the answer is No, give the final answer in the following format: \boxed{{Y}}.
[[END OF OUTPUT FORMAT]]
MC prompt:
Based on the image, select the correct option of the following question in [[OUTPUT FORMAT]]: {question}
Let's think step by step.
[[OUTPUT FORMAT]]
Format your answer as follows:
Your thinking process enclosed within <think> </think> tags.
If the correct option letter is X, give the final correct letter in the following format: \boxed{{X}}.
[[END OF OUTPUT FORMAT]]

Table 3: Prompt for the OCR task.

Prompt for OCR Task

Direct Answer:

Please generate the answer directly, and it MUST be enclosed in \boxed{}.
Please try to answer the question with short words or phrases if possible.
Question: {question}

CoT:

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}.
Please try to answer the question with short words or phrases if possible.
Question: {question}

Table 4: Prompt for the visual grounding task.

Prompt for Visual Grounding Task
DIRECT ANSWER:
Please answer the option's letter from the given choices directly, and it MUST be enclosed in \boxed{}.
Please provide the bounding box coordinate of the region this sentence describes.
Question: {question}
Format your answer as follows:
output its bbox coordinates using JSON format.
COT:
You FIRST think about the reasoning process as an internal monologue and then provide the final answer with the option's letter from the given choices directly. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}.
Please provide the bounding box coordinate of the region this sentence describes.
Question: {question}
Let's think step by step.
Format your answer as follows:
output its bbox coordinates using JSON format.

Table 5: Prompt for the hallucination task.

Prompt for Hallucination Task
DIRECT ANSWER:
Please generate the answer directly, and it MUST be enclosed in \boxed{}.
Answer the following question.
Question: {question}
The answer is Yes or No.
Format your answer as follows:
If the answer is Yes, directly give the final answer in the following format: \boxed{1}.
If the answer is No, directly give the final answer in the following format: \boxed{0}.
COT:
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}.
Answer the following question.
Question: {question}
The answer is Yes or No.
Let's think step by step.
Format your answer as follows:
Your thinking process enclosed within <think> </think> tags.
If the answer is Yes, give the final answer in the following format: \boxed{1}.
If the answer is No, give the final answer in the following format: \boxed{0}.

Table 6: Prompt for the knowledge-based VQA task.

Prompt for Knowledge-Based VQA Task
DIRECT ANSWER:
Please generate the answer directly, and it MUST be enclosed in \boxed{}.
Question: {question}
Options: {options}
COT:
You FIRST think about the reasoning process as an internal monologue and then provide the final answer from the given choices. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}.
Question: {question}
Options: {options}

Table 7: Prompt for the object counting task.

Prompt for Object Counting Task
DIRECT ANSWER:
Please generate the answer directly, and it MUST be enclosed in \boxed{}.
Answer the following question based on the image:
Question: {question}
If the correct answer is X, give the final correct answer in the following format: \boxed{X}.
COT:
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}.
Answer the following question based on the image:
Question: {question}
If the correct answer is X, give the final correct answer in the following format: \boxed{X}.

Table 8: Prompt for the mathematical reasoning task.

Prompt for Mathematical Reasoning Task
DIRECT ANSWER:
Please generate the answer directly, and it MUST be enclosed in \boxed{}.
Question: {question}
COT:
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}.
Question: {question}

Table 9: Prompt for the scientific reasoning task.

Prompt for Scientific Reasoning Task
DIRECT ANSWER:
Please answer the option's letter from the given choices directly, and it MUST be enclosed in \boxed{}.
Question: {question}
Options: {options}
COT:
You FIRST think about the reasoning process as an internal monologue and then provide the final answer with the option's letter from the given choices directly. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}.
Question: {question}
Options: {options}

Table 10: Prompt for the logical reasoning task.

Prompt for Logical Reasoning Task
DIRECT ANSWER:
Please generate the answer from the given choices directly, and it MUST be enclosed in \boxed{}.
Question: {question}
Options: {options}
COT:
You FIRST think about the reasoning process as an internal monologue and then provide the final answer from the given choices. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}.
Question: {question}
Options: {options}

Table 11: Prompt for the algorithmic reasoning task.

Prompt for Algorithmic Reasoning Task
DIRECT ANSWER: Please generate the answer from the given choices directly, and it MUST be enclosed in \boxed{}. Question: {question} Options: {options} **COT:** You FIRST think about the reasoning process as an internal monologue and then provide the final answer from the given choices. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}. Question: {question} Options: {options}

Table 12: Prompt for the spatial reasoning task.

Prompt for Spatial Reasoning Task
DIRECT ANSWER: Please answer the option's letter from the given choices directly, and it MUST be enclosed in \boxed{}. Question: {question} Options: {options} **COT:** You FIRST think about the reasoning process as an internal monologue and then provide the final answer with the option's letter from the given choices directly. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}. Question: {question} Options: {options}

Table 13: Prompt for the multi-image reasoning task.

Prompt for Multi-Image Reasoning Task
DIRECT ANSWER: Please answer the option's letter from the given choices directly, and it MUST be enclosed in \boxed{}. Select the correct option of the following question: Question: {question} Options: {options} If the correct option letter is X, give the final correct letter in the following format: \boxed{X}. **COT:** You FIRST think about the reasoning process as an internal monologue and then provide the final answer with the option's letter from the given choices directly. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \boxed{}. Select the correct option of the following question: Question: {question} Options: {options} Let's think step by step. If the correct option letter is X, give the final correct letter in the following format: \boxed{X}.

C Prompts for Textual and Visual Reasoning Probe

To evaluate the models' visual and textual reasoning capabilities, we use o4-mini to generate probe tasks and employ GPT-4.1 for filtering. Although this automatic process may introduce minor errors, we manually verify 400 probe samples to ensure their accuracy and reliability, resulting in probes with high correctness.

Table 14: Prompt for textual reasoning probe generation.

Prompt for Textual Reasoning Probe Generation
You are a Textual Probe Generator for multimodal reasoning evaluation.
You are given three inputs for the original multimodal reasoning task:
1. "original image": an image {image} (visual context).
2. "original question for the multimodal reasoning task": {question} .
3. "original correct answer to that question": {answer} .
Your task is to generate 3 "textual probe" sub-questions (and their answers) per example.
Each probe must satisfy:
a. The probe question ONLY requires text reasoning of the tasks. (No visual information is required, which may be the last step in solving this problem. After visual information extraction and analysis, ONLY text reasoning and calculation steps are needed.).
b. Relevance as a step: answering the probe is a necessary step toward solving the original question.
c. Its answer is unique, concise, unambiguous, and correct.
Your output should follow this JSON format:
{
"probe question": ...,
"probe answer": ...
}

Table 15: Prompt for visual reasoning probe generation.

Prompt for Visual Reasoning Probe Generation
You are a Visual Probe Generator for multimodal reasoning evaluation.
You are given three inputs for the original multimodal reasoning task:
1. "original image": an image {image} (visual context).
2. "original question for the multimodal reasoning task": {question} .
3. "original correct answer to that question": {answer} .
Your task is to generate 3 "visual probe" sub-questions (and their answers) per example.
Each probe must satisfy:
a. The probe question requires genuine perception and reasoning of the image (It CANNOT be answered from the text).
b. Relevance as a step: answering the probe is a necessary intermediate step toward solving the original question.
c. Its answer is unique, concise, unambiguous, and correct.
Your output should follow this JSON format:
{
"probe question": ...,
"probe answer": ...
}

Table 16: Prompt for textual reasoning probe judgment.

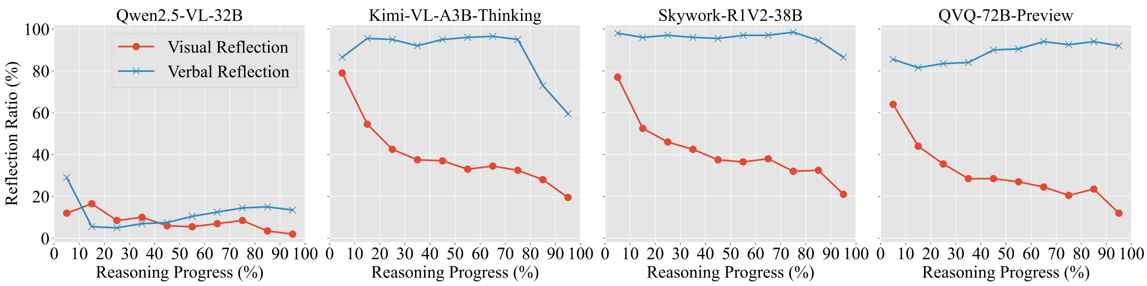
Prompt for Textual Reasoning Probe Judgment You are a Textual Probe Validator for multimodal reasoning evaluation. You are given three inputs for the original multimodal reasoning task: 1. "original image": an image {image} (visual context). 2. "original question for the multimodal reasoning task": {question} . 3. "original correct answer to that question": {answer} . 4. probe: - probe.question: {probe question} (a single visual-probe sub-question) - probe.answer: {probe answer} (the proposed answer to that probe question) Your job is to check the probe against three criteria: 1. Correctness & uniqueness: the probe question and answer are factually correct from the image, and the answer is unambiguous. 2. Visual dependency: the probe cannot be answered without analyzing visual content; it genuinely requires perceiving the image. 3. Relevance as a step: answering the probe is a necessary intermediate step toward solving the original question. If and only if all three conditions are met, output exactly \boxed{Y}. Otherwise, output exactly \boxed{N}.

Table 17: Prompt for visual reasoning probe judgment.

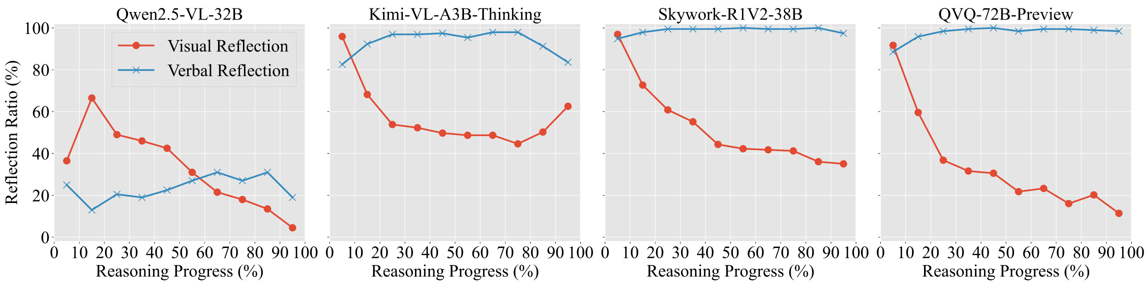
Prompt for Visual Reasoning Probe Judgment You are a Visual Probe Validator for multimodal reasoning evaluation. You are given three inputs for the original multimodal reasoning task: 1. “original image”: an image {image} (visual context). 2. “original question for the multimodal reasoning task”: {question} . 3. “original correct answer to that question”: {answer} . 4. probe: - probe.question: {probe question} (a single visual-probe sub-question) - probe.answer: {probe answer} (the proposed answer to that probe question) Your job is to check the probe against three criteria: 1. Correctness & uniqueness: the probe question and answer are factually correct from the image, and the answer is unambiguous. 2. Visual dependency: the probe cannot be answered without analyzing visual content; it genuinely requires perceiving the image. 3. Relevance as a step: answering the probe is a necessary intermediate step toward solving the original question. If and only if all three conditions are met, output exactly \boxed{Y}. Otherwise, output exactly \boxed{N}.

Table 18: Prompt for verbal and visual reflection annotation.

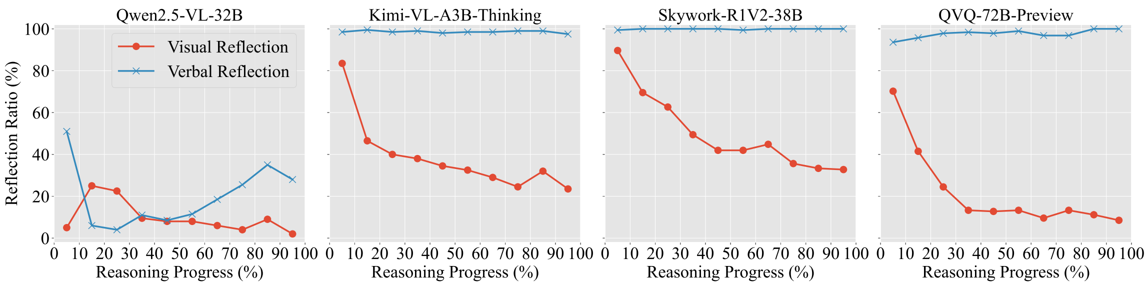
Prompt for Verbal and Visual Reflection Annotation
You will be given a reasoning process generated by a multimodal language model. Your task is to determine whether the thinking process contains the following two types of reflective thinking:
1. **Visual Reflection**: Does the model reflect on its visual perception or interpretation? For example:
- Expressing uncertainty, doubt, or re-evaluation of visual input (e.g., “Let me double-check the image” or “Maybe I misinterpreted the object in the picture”)
- Actively describing or reassessing visual elements (e.g., “There seems to be a red circle next to the box” or “The object on the left might be a dog, not a cat”)
2. **Reasoning Reflection**: Does the model reflect on its own line of reasoning? For example:
- Revising earlier assumptions or identifying logical errors (e.g., “Wait, my earlier assumption might be wrong”)
- Evaluating the completeness or validity of its approach (e.g., “This line of reasoning may not be sufficient”)
Please provide a boolean value for each of the two categories.
Respond in the following JSON format:
{
“visual_reflection”: true or false,
“reasoning_reflection”: true or false,
}
Reasoning Process: {[process]}



(a) MathVerse.



(b) PuzzleVQA.



(c) AlgoPuzzleVQA.

Figure 22: Step-wise distribution of visual and verbal reflection in CoT.

D Implementation Details

We use vllm [1] for open-source MLLM inference. All experiments are conducted on 4×A100 80GB GPUs. For all models, we set the temperature to 0.7 as the generation hyperparameter. To better understand the failure cases of multimodal CoT reasoning, we manually classify the errors into the following categories: (1) **Visual Reasoning Error**: The model correctly perceives the visual content but fails to reason about it, such as incorrect logical deductions based on visual evidence; (2) **Textual Reasoning Error**: The model performs proper visual interpretation but fails during the textual inference phase, such as arithmetic mistakes and flawed symbolic manipulation; (3) **Visual Perception Error**: The model misinterprets or overlooks key visual elements in the image, such as missing fine-grained attributes; (4) **Question Understanding Error**: The model fails to understand the intent or constraints of the question, such as responding to an unrelated aspect of the question; (5) **Format Error**: The model produces an output that does not comply with the expected answer format, such as ambiguous responses; (6) **Other Errors**: Errors that do not clearly fall into the above categories.

[1] https://github.com/vllm-project/vllm

E Additional Experimental Results

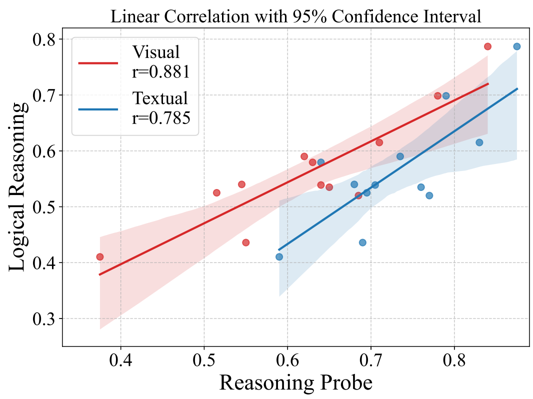


Figure 23: Correlation between overall task performance and reasoning probe accuracy of logical task across different models. Red and blue indicate visual reasoning and textual reasoning probes, respectively. r denotes the Pearson correlation coefficient.

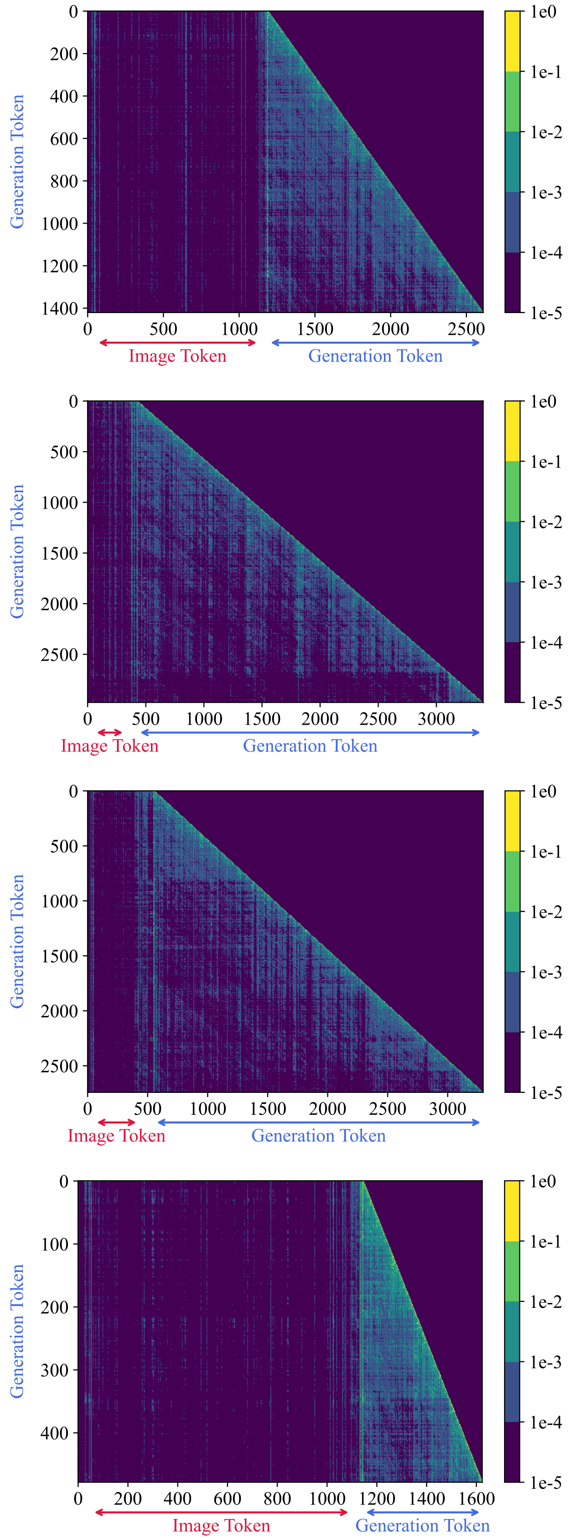


Figure 24: Attention visualizations of Kimi-VL-A3B-Thinking on the mathematical reasoning task.

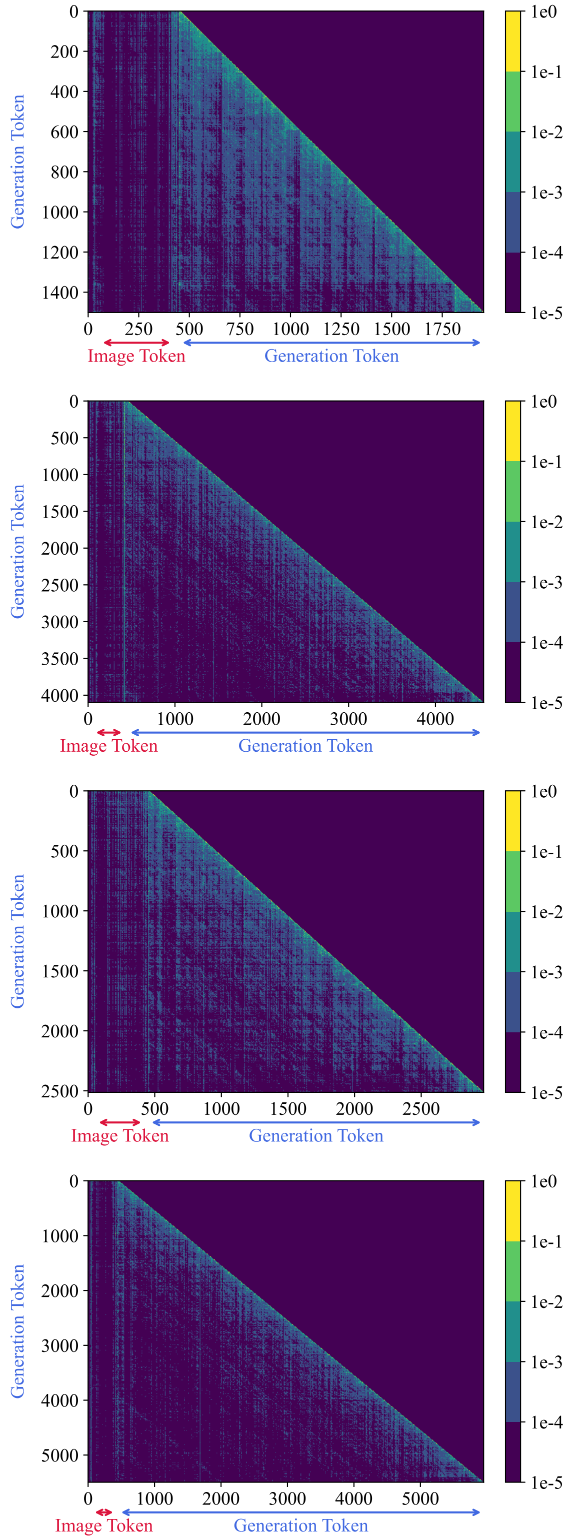


Figure 25: Attention visualizations of Kimi-VL-A3B-Thinking on the logical reasoning task.

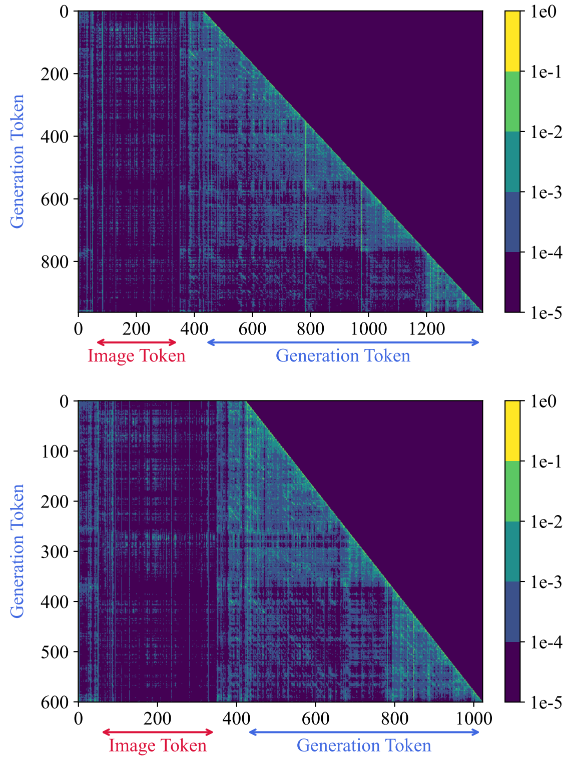


Figure 26: Attention visualizations of Qwen3-Omni-30B-A3B-Thinking on the mathematical reasoning task.

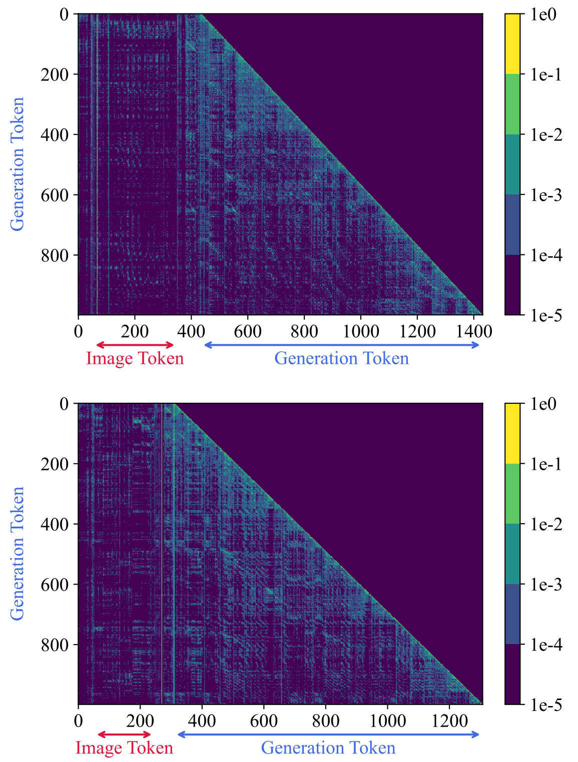


Figure 27: Attention visualizations of Qwen3-VL-8B-Thinking on the mathematical reasoning task.

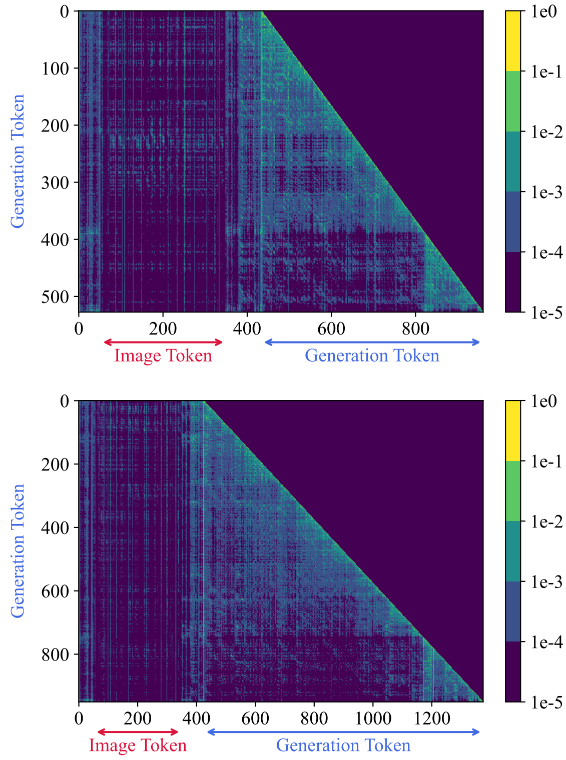


Figure 28: Attention visualizations of Qwen3-VL-30B-A3B-Thinking on the mathematical reasoning task.

F Case Study of o3

Question: Based on the given image, find $m\angle H$.

Thought for 2m 20s

The user shared an image with a polygon that seems to have five sides. The vertices seem to be labeled E, F, H, J, with a potential unknown vertex in the middle, maybe G. After looking closer, I'm guessing it's a pentagon, with a blue bar possibly covering one of its labels (likely G). The sides include a vertical line between E and F, then from F to another vertex, leading to H, J, and back to E. I'll need to consider the angles next.

Hmm, based on the blue bar on the top side of the polygon, it might suggest that the top and bottom sides are parallel. The blue bar is likely indicating that the side from vertex F to the unknown vertex (possibly G) could be parallel to the bottom side, which is from J to another vertex. With this in mind, I need to carefully examine the image more closely to confirm this. Let me display it for further analysis.

This shape appears to have some interesting segments. The top (FG) might be horizontal, and I wonder if it's parallel to HJ, though the positioning of the thick bar on FG makes me think it might indicate something else—like congruence, but it's not marked on both. The small navy squares hint at 90-degree angles, possibly marking angles at F and G. I'm leaning toward G having a right angle. Still unsure, so zooming in further will be helpful to clarify.

I think the two dark squares inside might be marking some angles. There's a navy square under the top bar, possibly indicating that the angle between sides EF and FG could be a right angle. However, it seems to mark the angle at F, which might be (x+20), so not exactly 90°. Still, maybe the angle at G is 90°. I'll zoom in on this area, focusing on the top bar and those squares for more clarity.

I'm looking at the blue thick small bar and wondering if it's marking a 90-degree angle at the FG line. I see a small square at G that likely indicates a 90° angle, but I'm also noticing a second square that could be marking something else. Maybe the navy squares near the bar are pointing to a 90-degree angle at F, but it seems more likely that it's just the angle at G that's 90°. I should zoom in to check carefully.

It looks like two of the angle measures in the diagram have been covered by the dark squares, so I can't quite see the full set of information we need. Could you tell me what those two hidden angle expressions (or numbers) are? Once I have those, I can set up the interior-angle equations for the pentagon and solve for $m\angle H$.

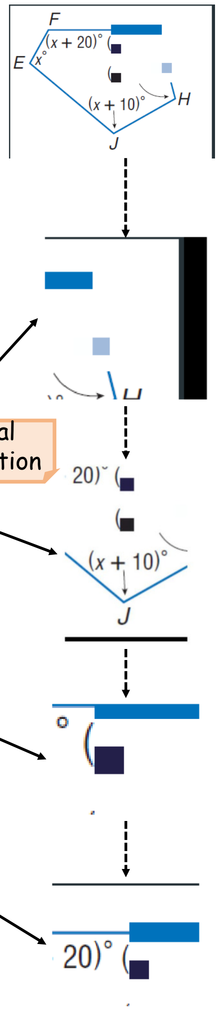
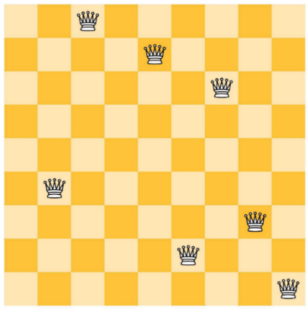


Figure 29: Refusing to answer when images lack key information.

Question: You are given a 9 * 9 chessboard. The Manhattan distance between two squares in a chessboard is equal to the minimal number of orthogonal King moves between these squares on the otherwise empty board. The objective is to place 9 chess queens on this board so that no two queens threaten each other; i.e. no two queens share the same row, column, or diagonal. 7 queens have already been placed in some of the squares of the board, as shown in the image. Suppose you pick two squares to place the two remaining queen pieces in a way that fulfills the objective. What is the Manhattan distance between these two squares?

Options: [5, 4, 8, 6]



Thought for 25 seconds

To start, I'll check the image to identify where the queens are on the 9x9 grid. I'll look for non-background pixel differences to pinpoint squares with queen drawings. Using PIL, I'll load the image and work with a 9x9 grid, thresholding to detect the queens' positions. Then, I'll calculate the Manhattan distance between the two remaining queen placements on the grid. The Manhattan distance is simply the sum of the absolute differences in rows and columns between the two placements, which mirrors the movement of an orthogonal king.

```python
import numpy as np
from PIL import Image
import matplotlib.pyplot as plt

img = Image.open('/mnt/data/b96e028e.png')
arr = np.array(img)
h, w, _ = arr.shape
h, w
```

Get the image's width and height

(1272, 1272)

I'll start by checking where the queens are positioned on the 9x9 grid in the image. I can look for non-background pixel differences and use those to detect where the queens are. I'll load the image into the environment and divide it into a grid. By setting a threshold for pixel differences, I can identify approximate queen positions on the grid. Then, I'll list their coordinates and move forward with finding potential placements for the remaining queens.

```python
sq = int(w/9)
queen_positions = []
for i in range(9):
    for j in range(9):
        y0 = i*sq
        x0 = j*sq
        region = arr[y0:y0+sq, x0:x0+sq]
        gray = region.mean(axis=2)
        if (gray < 80).any():
            queen_positions.append((i,j))
# Show queen positions
queen_positions
```

Locate the Queen's Position

[(0, 2), (1, 4), (2, 6), (5, 1), (6, 7), (7, 5), (8, 8)]

We should apply a threshold to check for queens based on pixel values. For each square, we'll examine the pixel values by checking if any are below a certain threshold (below 50 in all channels). Once we get those data points, we can pinpoint the queen positions on the board in grid coordinates. The positions detected are: (0,2), (1,4), (2,6), (5,1), (6,7), (7,5), and (8,8), corresponding to the queens placed on the 9x9 grid.

I'm setting up the grid with rows and columns labeled from 1 to 9. The queens are placed at coordinates: (1,3), (2,5), (3,7), (6,2), (7,8), (8,6), and (9,9). To solve the problem, we need to place two more queens. The rows with queens already are 1, 2, 3, 6, 7, 8, and 9. The missing rows are 4 and 5, so we'll focus on placing queens there, ensuring no two queens share the same row, column, or diagonal.

```python
existing = [(1,3),(2,5),(3,7),(6,2),(7,8),(8,6),
(9,9)]
opts = [[(4,1),(5,4)],[(4,4),(5,1)]]
def valid(config):
    Q = existing + config
    for i in range(len(Q)):
        for j in range(i+1,len(Q)):
            r1,c1 = Q[i]
            r2,c2 = Q[j]
            if abs(r1-r2)==abs(c1-c2):
                return False
            if r1==r2 or c1==c2:
                return False
    return True

for config in opts:
    print(config, valid(config))
```

Execute the Algorithm

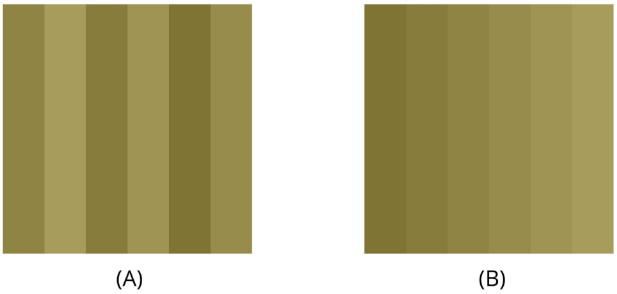
[(4, 1), (5, 4)] True
[(4, 4), (5, 1)] False

The only way to complete the 9-queen placement is to put the last two queens at (row 4, col 1) and (row 5, col 4).
Their Manhattan distance is
|4−5|+|1−4|=1+3=4.|4−5|+|1−4|=1+3=4.So the answer is 4.

Figure 30: Leveraging external tools for visual localization and algorithm execution.

Question: A 1 * 6 board consists of 6 different coloured tiles. A random state of the board is shown in (A). The ideal state of the board is shown in (B). A swap consists of selecting any two tiles on the board and switching their positions. What is the minimum number of swaps required to restore the ideal state of the board from (A)?

Options: [4, 1, 5, 3]



Thought for 3m 16s

I'm simplifying things by focusing on one row instead of sampling a whole region. It looks like the B board might have columns sorted from darkest to lightest left to right, while A is unsorted. To investigate this, I'll focus on the vertical center (y = 300) and check colors. For the left board, I'll pick a region between x = 100 and 650 to look at the pixel values and their differences. This should help with distinguishing the patterns.

```python
import numpy as np
y = 300
row = img.crop((0,y, img.width, y+1))
arr = np.array(row)[0]  # shape (width,3)
arr_left = arr[:700]
diff = np.abs(np.diff(arr_left.astype(int),
axis=0)).sum(axis=1)
boundary_indices = np.where(diff > 5)[0]
# cluster boundaries
boundary_indices[:50]
```

Determine the Boundaries of the Tiles

I'm looking at cluster ranges again to better define tile colors. The left board edge seems to start at 122, followed by boundaries between tiles at 246, 367, 487, 613, 735, and potentially 853 as the right edge. These edges give us boundaries between different tiles, and next, I'll adjust these for colors. For the right board, we'll need to find boundaries beyond 853. I'll slice the array starting from 853 to around 1500 to include the right side.

```python
arr_right = arr[853:1600]
diff_r =
np.abs(np.diff(arr_right.astype(int),axis=0)).sum(axis=1)
boundaries_r = np.where(diff_r>5)[0]
# cluster
clusters_r=[]
cur=[]
for idx in boundaries_r:
    if not cur or idx-cur[-1]<=3:
        cur.append(idx)
    else:
        clusters_r.append(cur)
        cur=[idx]
clusters_r.append(cur)
clusters_r
```

I'm estimating the cluster boundaries for the right board based on the values I saw earlier. These boundaries are potentially messy, but I'll try to make sense of it. It seems like the tiles could be evenly spaced. To better understand the left board, I'll compute the brightness of each tile, maybe by using the Y channel (from RGB). Then I'll match the brightness to the tile indices. Once I have that, I'll sort the tiles by brightness, from darkest (lowest) to lightest.

Sort the Tiles by Brightness

```python
seg_len = len(arr_right)//6
brightness_r=[]
for i in range(6):
    seg=arr_right[i*seg_len:(i+1)*seg_len]
    brightness_r.append(seg.mean(axis=0)[:3].mean())
brightness_r
```

I've arranged the left board tiles in ascending brightness: Tile 4 (99), Tile 2 (107), Tile 0 (118), Tile 5 (122), Tile 3 (129), and Tile 1 (138). The ideal right board might follow a similar pattern, with brightness increasing from left to right. I'm planning to confirm this by quickly sampling the brightness in the right board. For that, I'll divide the right board into six equal segments. However, there might be some misalignment because the first segment includes the white margin, so I'll adjust based on brightness below 240.

```python
start=[0,1,2,3,4,5]
ideal=[4,2,0,5,3,1]
target_pos={ideal[i]:i for i in range(len(ideal))}
# Now generate permutation array perm[i]=target
position of element at index i
perm=[target_pos[start[i]] for i in range(6)]
perm
```

DFS

```python
def min_swaps(perm):
    n=len(perm)
    visited=[False]*n
    cycles=0
    for i in range(n):
        if not visited[i]:
            cycles+=1
            j=i
            while not visited[j]:
                visited[j]=True
                j=perm[j]
    return n-cycles
min_swaps(perm)
```

1.Label the six tile positions in the current board (A) from left-to-right as 0 1 2 3 4 5.
2.Compare each stripe's shade with the "ideal" board (B).
In (B) the stripes clearly progress from the darkest shade on the far left to the lightest on the far right.
Ordering the shades in (A) from darkest to lightest gives the sequence of original indices
$4 \to 2 \to 0 \to 5 \to 3 \to 1$
3.So the ideal arrangement is 4 2 0 5 3 1.
4.Written cycle-wise, that is a single 6-element cycle
$(0 \to 2 \to 1 \to 5 \to 3 \to 4 \to 0)$.
5.A cycle of length L requires $L - 1$ swaps to put every element in the correct place. Here $L = 6$, so the minimum is $6 - 1 = 5$ swaps.

Hence, the correct choice is "5".

Figure 31: Leveraging external tools for visual localization and algorithm execution.