What, Whether and How? Unveiling Process Reward Models for *Thinking with Images* Reasoning

Yujin Zhou[1]*, Pengcheng Wen[1]*, Jiale Chen[2]*, Boqin Yin[1], Han Zhu[1], Jiaming Ji[3], Juntao Dai[3], Chi-Min Chan[1], Sirui Han[1]†

[1]Hong Kong University of Science and Technology
[2]Sun Yat-sen University
[3]Peking University
{yzhouha,pc.wen}@connect.ust.hk ertiam047@gmail.com siruihan@ust.hk

Abstract

The rapid advancement of Large Vision Language Models (LVLMs) has demonstrated excellent abilities in various visual tasks. Building upon these developments, the *thinking with images* paradigm has emerged, enabling models to dynamically edit and re-encode visual information at each reasoning step, mirroring human visual processing. However, this paradigm introduces significant challenges as diverse errors may occur during reasoning processes. This necessitates Process Reward Models (PRMs) for distinguishing positive and negative reasoning steps, yet existing benchmarks for PRMs are predominantly text-centric and lack comprehensive assessment under this paradigm. To address these gaps, this work introduces the first comprehensive benchmark specifically designed for evaluating PRMs under the *thinking with images* paradigm. Our main contributions are: (1) Through extensive analysis of reasoning trajectories and guided search experiments with PRMs, we define 7 fine-grained error types and demonstrate both the necessity for specialized PRMs and the potential for improvement. (2) We construct a comprehensive benchmark comprising 1,206 manually annotated thinking with images reasoning trajectories spanning 4 categories and 16 subcategories for fine-grained evaluation of PRMs. (3) Our experimental analysis reveals that current LVLMs fall short as effective PRMs, exhibiting limited capabilities in visual reasoning process evaluation with significant performance disparities across error types, positive evaluation bias, and sensitivity to reasoning step positions. These findings demonstrate the effectiveness of our benchmark and establish crucial foundations for advancing PRMs in LVLMs.

Introduction

The human visual system is dynamically heterogeneous in scale , freely switching between broad views and detailed focus as needed (Wandell 1995; Wang et al. 2025a). In contrast, conventional Large Vision Language Models (LVLMs) are fundamentally limited by their reliance on static tokenization of image information for language model comprehension—a process that introduces inevitable information loss and constrains the exploitation of fine-grained visual details (Bordes et al. 2024; Ghosh et al. 2024). Recent breakthroughs in LVLMs, exemplified by OpenAI's o3 (OpenAI 2025), have introduced the transformative *thinking with images* paradigm (Su et al. 2025c) to overcome these challenges. Unlike traditional approaches that isolate visual perception from reasoning, *thinking with images* paradigm mirrors the dynamic nature of human visual processing by enabling models to proactively edit and re-encode visual information at each reasoning step using integrated image processing tools, thereby mitigating static tokenization limitations and enhancing overall visual understanding (Zheng et al. 2025; Sarch et al. 2025; Su et al. 2025a; Hu et al. 2024; Sun et al. 2024; Cheng et al. 2025; Zhang et al. 2025; Fan et al. 2025; Zhu et al. 2025a).

However, despite these significant advances, LVLMs un-

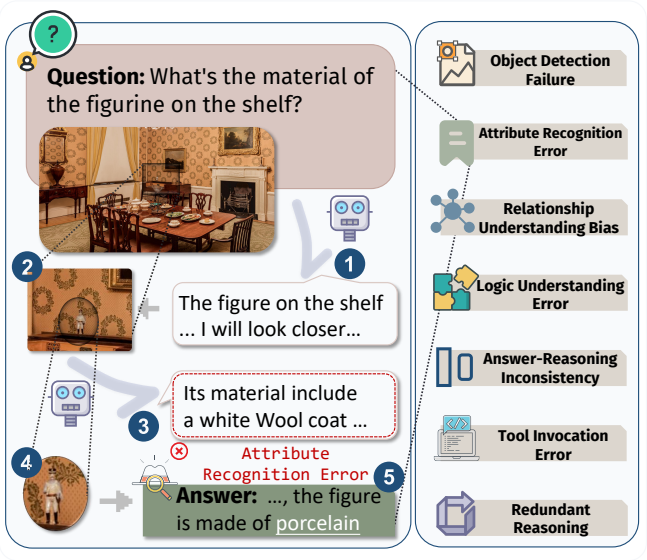

Figure 1: **Identified Error Types in *Thinking with Images* Paradigm.** We identify and categorize seven distinct error types from reasoning trajectories of current *thinking with images* models. Some are inherent LVLMs limitations and some are novel errors introduced by this paradigm.

*These authors contributed equally.
†Corresponding author.

der *thinking with images* paradigm still encounter several remaining issues that hinder their reasoning capabilities, as illustrated in Figure 1. Through comprehensive manual analysis of 7,558 reasoning trajectories generated by four *thinking with images* models across four established benchmarks, we systematically identified and taxonomized seven distinct categories of remaining reasoning errors under the *thinking with images* paradigm. These error categories highlight ongoing challenges that stem from both the inherent limitations of traditional LVLM and new issues introduced by the dynamic nature of the *thinking with images* paradigm, ultimately compromising the quality of the overall reasoning.

Inspired by the remarkable success of Process Reward Models (PRMs) in enhancing verbal reasoning through stepwise supervision (Chan et al. 2025a,b; Wang et al. 2025c; Lightman et al. 2023; Wen et al. 2025; Cao et al. 2025a; Shi et al. 2025; Zhang et al. 2024a; Cao et al. 2025b), we observe that reasoning trajectories within *thinking with images* paradigm can be naturally decomposed into discrete interleaved text-image steps. This structural compatibility naturally motivates us to investigate a meaningful question for the *thinking with images* paradigm:

Whether Process Reward Models Help?

In the absence of specialized PRMs for *thinking with images* paradigm, we explored prompting existing state-of-the-art LVLMs to function as PRMs for guided search evaluation across five benchmarks. Our preliminary experiments demonstrate that LVLMs can improve reasoning performance over baseline approaches.

To further analyze the abilities of existing models to identify errors within the paradigm of *thinking with images* and to facilitate future development of more effective specialized PRMs, **we introduce *ThinkWithImages*-PRMBENCH, a comprehensive and fine-grained benchmark designed to assess PRMs under the paradigm of *thinking with images***. Unlike existing process-level benchmarks that focus primarily on text-based trajectories, *ThinkWithImages*-PRMBENCH addresses the unique challenges of evaluating PRMs in *thinking with images* scenarios. Our benchmark comprises 1,206 meticulously curated instances spanning 4 major categories and 16 subcategories, with quality validated by five expert annotators. We implement controlled curation methodologies to maintain consistent difficulty levels across visual reasoning domains, while covering diverse scenarios including geometric analysis, spatial relationships, temporal dynamics, and multi-object interactions.

Equipped with *ThinkWithImages*-PRMBENCH, we conducted extensive evaluations across over 10 state-of-the-art models. Through quantitative analysis and qualitative observations, our key contributions are summarized as follows:

- **Fine-grained error type taxonomy**: Through extensive experiments and observations, **7** fine-grained error types are defined under *thinking with images* paradigm, systematically capturing and categorizing common visual reasoning failure modes.
- **Necessity and potential for improvements of PRMs**: Guided search experiments demonstrate that existing LVLMs can help when serving as PRMs but show significant room for improvement.
- **The first *thinking with images* PRM benchmark**: A curated collection of **1,206** manually annotated high-quality *thinking with images* reasoning trajectories spanning **4** categories and **16** subcategories, enabling fine-grained evaluation of existing LVLMs as PRMs.
- **Comprehensive experimental analysis and in-depth insights**: Reveals current LVLMs fall short as PRMs, demonstrating limited capability in capturing vision reasoning process evaluation with significant performance disparities across different error types. Analysis uncovers consistent positive bias in evaluations and notable step sensitivity with significant performance variations during multi-step reasoning processes.

Related Works

Large Vision Language Model Reasoning with Tools The integration of reasoning capabilities with visual understanding, enabling models to reason through interleaved text and image sequences, has emerged as a crucial frontier in multimodal reasoning known as *thinking with images* paradigm (Su et al. 2025c; OpenAI 2025). Current approaches for enhancing visual reasoning in LVLMs can be categorized into two directions: (1) Training-free methods: Early works explored prompt-based tool integration and other training-free approaches, enabling models to leverage external visual processing tools through carefully designed prompts, instructions, or predefined pipelines (Hu et al. 2024; Shen et al. 2024; Wang et al. 2025d; Li et al. 2025). (2) Training-based methods: A second line of work focuses on training models to acquire visual reasoning capabilities through data-driven approaches. This includes supervised fine-tuning with specialized datasets to enable models to invoke image processing tools during reasoning processes, as well as reinforcement learning approaches that employ reward-based optimization, typically computing rewards based on final answer correctness to incentivize models to follow *thinking with images* paradigm (Zheng et al. 2025; Su et al. 2025a; Cao et al. 2025c; Sarch et al. 2025; Jiang et al. 2025; Wu et al. 2025; Zhang et al. 2025; Su et al. 2025b; Liu et al. 2025). Our work focuses primarily on models obtained through training-based methods, analyzing their reasoning trajectories, and utilizing them as upstream *thinking with images* models for systematic investigation.

Benchmarks for Process Reward Models As PRMs have gained increasing attention for their ability to provide step wise supervision during reasoning, effectively and comprehensively evaluating their capabilities has become crucial (Luo et al. 2024; Feng et al. 2025). This has led to the development of several PRM benchmarks, including PRMBench (Song et al. 2025), VisualPRM (Wang et al. 2025c), MPBench (Xu et al. 2025), and ProcessBench (Zheng et al. 2024), which evaluate PRMs' ability to identify fine-grained errors in upstream models' reasoning trajectories. However, existing benchmarks primarily focus on evaluating PRMs for text-only reasoning trajectories. We

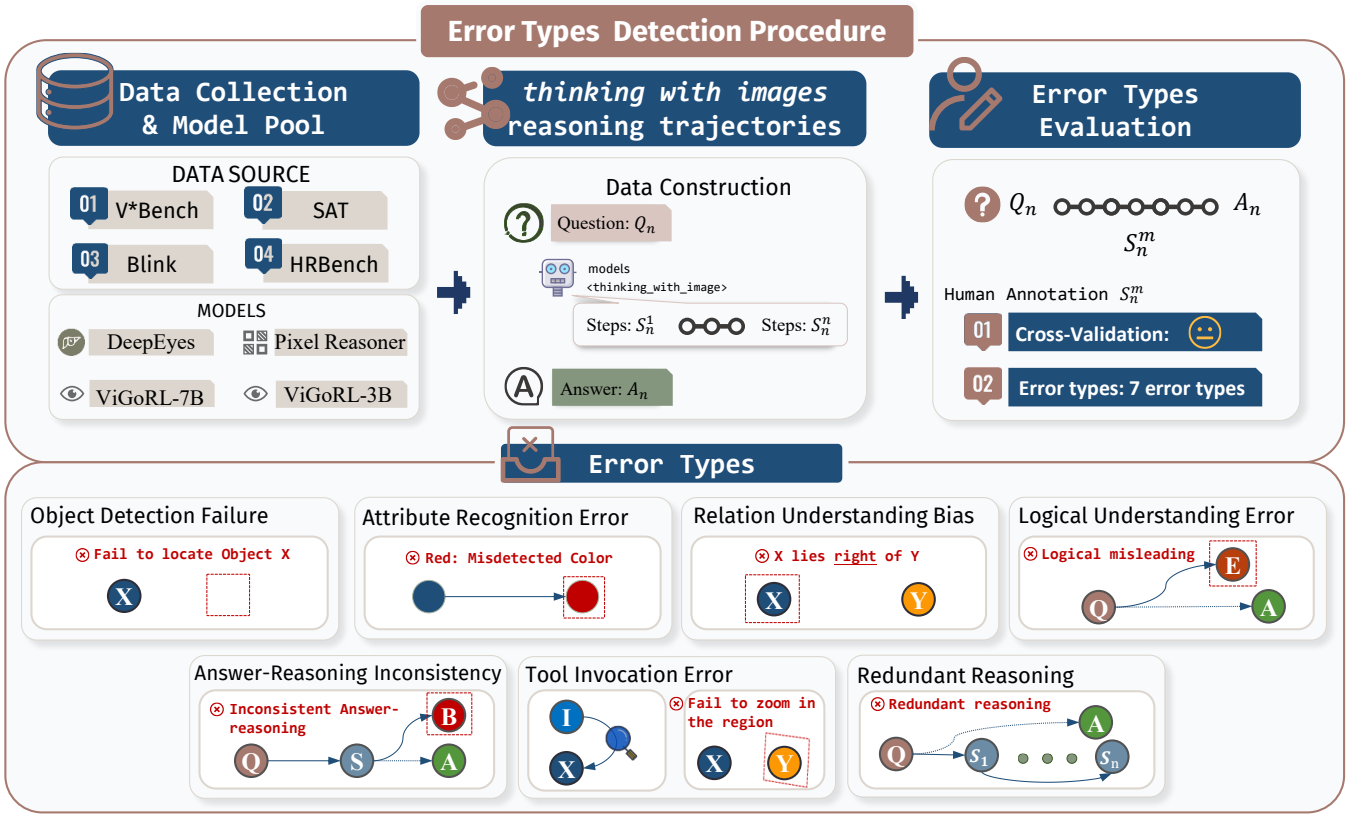

Figure 2: **Error Types Detection Procedure for *Thinking with Images* Paradigm.** We collect extensive reasoning trajectories by deploying four *thinking with images* models across four benchmarks. Through systematic analysis and categorization of these trajectories, we identify seven distinct errors that commonly occur in *thinking with images* paradigm.

introduce *ThinkWithImages*-PRMBENCH, the first benchmark specifically designed to evaluate PRMs' effectiveness in detecting errors within interleaved text-image reasoning trajectories under the *thinking with images* paradigm.

Study Setup

We begin by constructing a comprehensive collection of *thinking with images* reasoning trajectories, which serves as the foundation for our systematic investigation of reasoning errors, PRM effectiveness evaluation, and *ThinkWithImages*-PRMBENCHdevelopment. Using four representative *thinking with images* models across five benchmarks, we collect 7,558 reasoning trajectories.

Thinking with Images LVLMs

We select four state-of-the-art models that demonstrate representative capabilities in multimodal reasoning with integrated visual processing tools: ViGoRL-3B and 7B (Sarch et al. 2025), DeepEyes-7B (Zheng et al. 2025), and PixelReasoner-7B (Su et al. 2025a). These models are chosen based on their ability to perform dynamic visual processing operations such as cropping, zooming, which are fundamental characteristics of *thinking with images* paradigm.

Downstream Benchmarks

We evaluate our selected models across five established benchmarks that cover diverse visual reasoning scenarios: V*Bench (Huang et al. 2024), Blink (Fu et al. 2024), SAT-2 (Ray et al. 2024), HRBench (Wang et al. 2025b), and MME-RealWorld (Zhang et al. 2024b). These benchmarks collectively provide comprehensive coverage of visual reasoning tasks including geometric analysis, spatial relationships, temporal dynamics, and complex multi-object interactions, ensuring our analysis captures the full spectrum of challenges encountered in *thinking with images* applications.

What Issues Persist in Current *Thinking with Images* LVLMs?

Through extensive experiments and observation of large-scale trajectories, we identify and categorize seven distinct error types that serve as both our annotation guidelines and incorrectness classification criteria:

Object Detection Failure (ODF): The model fails to locate or detect objects within the image, resulting in missed visual elements critical to the reasoning process.

Attribute Recognition Error (AE): While the model successfully localizes objects, it fails to correctly identify their attributes. This includes confusion regarding colors (red vs.

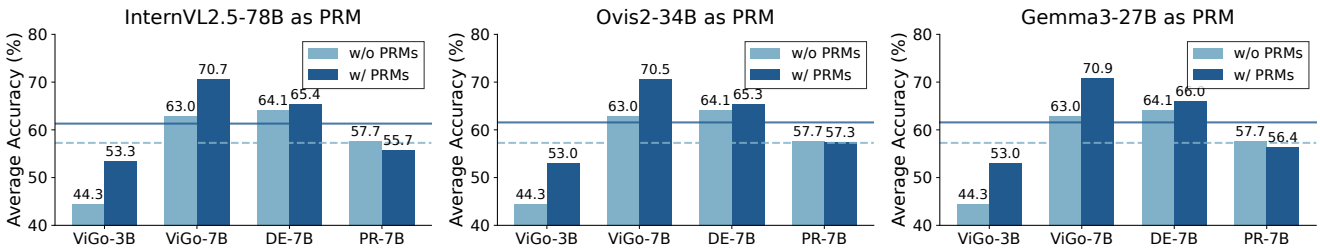

Figure 3: **Performance Comparison of *Thinking with Images* Models with and without LVLMs as PRMs.** Results demonstrate a general trend toward improved performance when using LVLMs as PRMs, though improvements vary across models.

orange), materials (metal vs. plastic), shapes, or states (open vs. closed).

Relationship Understanding Bias (RU): Inaccurate or ambiguous comprehension of spatial relationships ("left", "above", "between"), action relationships ("holding", "chasing"), or comparative relationships ("larger", "closer").

Logical Understanding Error (LU): The model demonstrates an inability to comprehend the fundamental question or task requirements, leading to irrelevant or illogical reasoning paths.

Answer-Reasoning Inconsistency (ARI): In the final step, the model produces a correct answer that does not align with the preceding reasoning process, often characterized by hasty conclusion-drawing without proper justification.

Tool Invocation Error (TE): Failure to correctly utilize available tools for object localization or visual analysis, resulting in inadequate visual processing.

Redundant Reasoning (RR): Situations where the answer is determined in an early step (e.g., step 1), making subsequent chain-of-thought steps superfluous and unnecessary for the final conclusion.

Whether Process Reward Models Can Help?

To investigate whether existing LVLMs can effectively serve as PRMs for *thinking with images* paradigm, we conduct preliminary experiments using various state-of-the-art models as surrogate PRMs.

Experimental Setup

Model Selection. We select three representative LVLMs spanning different model families and sizes to serve as surrogate PRMs: Gemma3-27B, Ovis2-34B, and InternVL2.5-78B. These models analyze reasoning trajectories generated under the *thinking with images* paradigm, where each trajectory consists of interleaved text-image steps with visual information processed throughout the workflow.

Evaluation Protocol. We employ **Guided Search** as our primary evaluation framework to assess PRM effectiveness. In this approach, candidate PRMs evaluate intermediate reasoning steps and provide real-time feedback to guide trajectory generation, as outlined in Algorithm 1.

Guided Search Implementation. For upstream models, we generate $k = 8$ step-by-step responses at each reasoning stage. PRMs score each candidate response, and the highest-scored response is selected (with ties broken by order). The process continues under a maxstep $= 10$ constraint.

Algorithm 1: PRM Guided Search for *Thinking with Images*

Input: Problem p, LVLM $\mathcal{M}$, PRM $\mathcal{R}$, Beam width k, Max steps T
Output: Best interleaved trajectory τ^*
$\mathcal{B} \leftarrow \{(\text{init\_state}, \emptyset, 0)\}$ {Initialize beam}
$\tau^* \leftarrow$ None, best_score $\leftarrow -\infty$
for $t = 1$ **to** T **do**
  $\mathcal{C} \leftarrow \emptyset$
  for each (s, τ, score) in $\mathcal{B}$ **do**
    steps $\leftarrow \mathcal{M}.\text{Generate}(s, k)$ {Generate text-image interleaved steps}
    for each step a in steps **do**
      $s' \leftarrow \text{Apply}(s, a)$, $\tau' \leftarrow \tau \cup \{a\}$ {a contains text and image}
      $r \leftarrow \mathcal{R}.\text{Score}(\tau')$ {PRM evaluates interleaved trajectory}
      Add $(s', \tau', \text{score} + r)$ to $\mathcal{C}$
    end for
  end for
  $\mathcal{B} \leftarrow \text{TopK}(\mathcal{C}, k)$ {Keep top-k candidates}
end for
return Best interleaved trajectory from $\mathcal{B}$

Evaluation Metrics. We measure performance using accuracy as our primary evaluation metric, and demonstrate PRM effectiveness by comparing LVLM performance before and after PRM integration.

Performance Analysis

Figure 3 demonstrates the overall effectiveness of LVLMs as PRMs in advancing *thinking with images* paradigm. While PixelReasoner-7B shows minimal to no improvement with PRMs, other models demonstrate consistent gains ranging from 1-8%, with PRMs generally outperforming non-PRM approaches across all evaluated models. These findings suggest that PRMs are effective under *thinking with images* paradigm, and there is still room for further improvement.

How Can We Build Better PRMs?

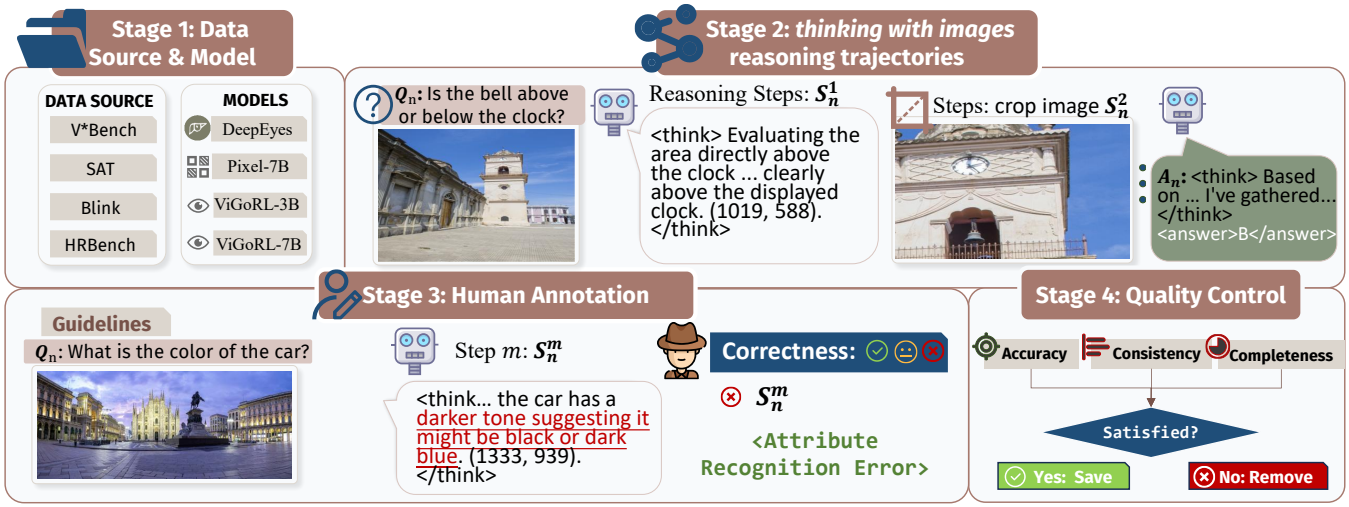

Figure 4: **Construction Pipeline for *ThinkWithImages*-PRMBENCH.** Building upon the reasoning trajectories collected as described in Figure 2, we perform step-by-step manual annotation for each trajectory and classify erroneous steps into seven error types. A comprehensive quality control process filters trajectories based on three key criteria: accuracy, consistency, and completeness, resulting in the curated *ThinkWithImages*-PRMBENCH.

Introducing *ThinkWithImages*-PRMBENCH

To build more effective PRMs, we introduce *ThinkWithImages*-PRMBENCH to support further advancements. In this section, we outline its construction pipeline and assess the performance of existing LVLMs as PRMs using our benchmark, aiming to highlight current challenges.

ThinkWithImages-PRMBENCH Construction Pipeline

ThinkWithImages-PRMBENCH is a meticulously designed multi-modal process reward benchmark tailored for *thinking with images* paradigm. Our construction comprises four key stages that ensure the scalability, accuracy, and reliability.

Stage 1: Data Collection We systematically collect question-image pairs from diverse open-source datasets that align with *thinking with images* paradigm. Our data collection strategy encompasses four primary sources: VBench, HRBench, SAT, and BLINK. These datasets are strategically selected to ensure comprehensive coverage across diverse visual reasoning scenarios and task complexities.

Stage 2: Trajectory Construction For each question-image pair collected from the open-source datasets, we employ our diverse model pool to generate multiple reasoning trajectories. This stage produces multiple reasoning trajectories for each query, providing a rich foundation for subsequent annotation.

Stage 3: Human Annotation The annotation process consists of two critical phases designed to ensure data quality and consistency:
Data Filtering: From the various reasoning trajectories generated in Stage 2, we first apply rule-based methods to deduplicate each query, retaining only one valid trajectory per question to ensure quality and uniqueness.

Manual Annotation: We utilize the open-source Label Studio platform with carefully designed annotation guidelines. Five annotators work over seven days to annotate each step based on our established guidelines and incorrect classification criteria. We filter out low-quality reasoning chains to maintain benchmark integrity.

Stage 4: Quality Assurance For the reasoning trajectories retained from Stage 3, three annotators conduct quality verification over three days in batches. Each question is evaluated across three dimensions: accuracy, consistency, and completeness. Only trajectories that satisfied all three criteria are retained. Through this rigorous quality assurance process, we finally collect 1,206 valid reasoning trajectories that make up our final *ThinkWithImages*-PRMBENCH.

This comprehensive four-stage pipeline ensures that our proposed bench provides a robust foundation for evaluating multi-modal PRMs in visual reasoning tasks. Our resulting benchmark systematically covers four fundamental dimensions: Recognition & Attributes, Space & Relationships, Dynamics & Actions, and Analysis & Reasoning, encompassing 16 specific categories of visual reasoning tasks as illustrated in Figure 5. The final dataset comprises 1,206 questions with corresponding image reasoning chains, with comprehensive statistics presented in Table 1. Figure 6 shows the distribution of error positions across reasoning steps.

Evaluation

Experimental Setup

Models To comprehensively evaluate LVLMs as PRMs on *ThinkWithImages*-PRMBENCH, we test open-source models including LLaVA-OneVision (7B, 72B) (Li et al. 2024), Qwen2.5-VL (7B, 72B) (Bai et al. 2025), InternVL2.5 (8B, 78B) (Chen et al. 2024; Zhu et al. 2025b), Gemma3 (4B, 27B) (Team et al. 2025), and Ovis2 (8B,

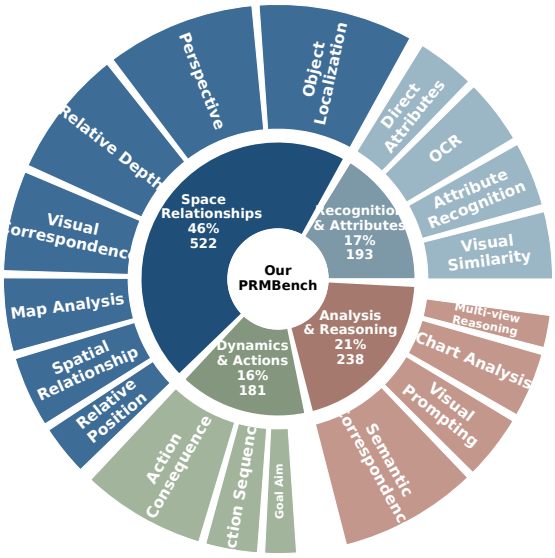

Figure 5: Composition of *ThinkWithImages*-PRMBENCH.

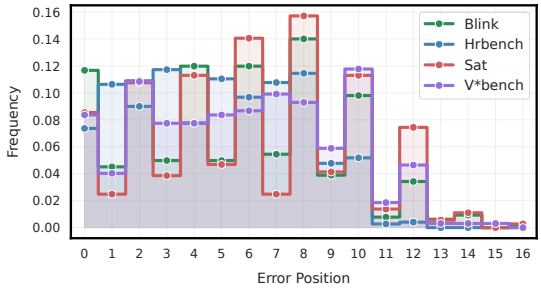

Figure 6: Distribution of Error Steps.

34B) (Lu et al. 2024), alongside proprietary models GPT-4o (Achiam et al. 2023) and Gemini-2.5-Flash (Comanici et al. 2025). Human evaluation results are in Table 2.

Evaluation Metrics We evaluate models using accuracy (ACC) and F1 scores. Accuracy measures the proportion of correctly classified steps. F1 mitigates class imbalance between correct and incorrect steps.

Main Results

The main results are shown in Table 2. Several key findings can be summarized as follows:

LVLMs as PRMs demonstrate limited capability in vision reasoning process evaluation. Our comprehensive evaluation reveals that current LVLMs struggle significantly when employed as PRMs for visual reasoning tasks. Among open-source models, Ovis2-34B demonstrates relatively strong performance with 41.74% ACC and 45.72% F1, slightly outperforming the proprietary Gemini-2.5-Flash model. However, even these best-performing models achieve considerably lower accuracy than human-level performance (83.61%). This substantial gap suggests that existing LVLMs are not yet competent to serve as reliable PRMs for visual reasoning tasks.

Benchmark Statistics	Count	Percentage
Total Samples	1,206	100.0%
• V*bench	134	11.2%
• Blink	478	39.7%
• Sat	284	23.4%
• HRbench	310	25.7%
Total Steps	12,714	100.0%
• Correct Steps	6,714	52.8%
• Incorrect Steps	5,233	41.2%
• Neutral Steps	767	6.0%
Total Images	6,544	–
Average Steps	10.41	–

Table 1: Statistics of *ThinkWithImages*-PRMBENCH.

Significant performance disparities exist across different error types. The results reveal dramatic variations in model performance across evaluation categories. While most models achieve ACC scores below 50% (ranging from 22-57%), they fail drastically when identifying specific complex reasoning errors such as TE and RU, where detection rates often drop below 5%. This indicates that LVLMs struggle particularly with detecting nuanced logical reasoning errors in multi-step processes.

Detailed Analysis

Now we present an in-depth examination of LVLMs serving as PRMs, grounded in our empirical results. Our analysis reveals the following:

Finding 1. LVLMs as PRMs show consistent positive bias in their reward assignments.

Table 3 demonstrates that both open-source and closed-source LVLMs exhibit significant reward bias during the evaluation process. Notably, the majority of models achieve accuracy rates below random chance (50%) when identifying incorrect samples, highlighting substantial deficiencies in existing LVLMs' ability to detect reasoning errors

Furthermore, the comparison between correct and incorrect sample performance reveals a consistent bias pattern across most models, with a pronounced tendency toward positive rewards (i.e., favoring correct classifications). For instance, Gemini-2.5-Flash achieves a remarkable 78.89% accuracy on correct samples, yet only 38.81% accuracy on incorrect samples, illustrating this systematic bias.

Findings 2. LVLMs as PRMs show notable step sensitivity and significant performance variations during the multi-step reasoning process.

Models↓ Metrics→	ACC	F1	TE	RU	AE	RR	ODF	ARI	LU
Proprietary Large Vision Language Models									
GPT-4o	31.45	41.34	4.61	40.10	52.94	33.33	64.02	79.31	58.21
Gemini-2.5-Flash	36.81	43.56	38.21	44.93	61.76	70.83	61.51	51.72	49.25
Open-Source Large Vision Language Models									
Qwen2.5-VL-7B	24.89	8.84	1.78	5.34	7.27	6.33	10.50	10.00	3.48
Qwen2.5-VL-72B	28.27	32.81	20.36	24.93	37.27	26.58	37.9	41.67	27.83
InternVL2.5-8B	22.77	13.30	2.55	8.4	7.79	1.41	18.18	20.45	10.94
InternVL2.5-78B	21.13	24.28	0.32	17.2	31.17	5.63	30.00	43.18	21.88
LLaVA-OneVision-7B	52.36	2.55	0.20	1.78	1.01	1.27	3.65	0.80	2.61
LLaVA-OneVision-72B	43.03	7.51	2.96	1.78	4.55	3.8	5.94	21.67	3.48
Gemma3-4B	47.10	39.32	66.54	21.05	31.91	22.73	41.21	48.15	40.40
Gemma3-27B	48.73	42.98	47.23	39.17	40.00	34.18	46.58	70.00	40.00
Ovis2-8B	30.56	19.77	9.76	10.63	11.76	4.17	15.90	37.93	12.69
Ovis2-34B	41.74	45.72	57.51	41.60	56.36	45.57	66.21	68.33	46.96
Human Performance	81.23	80.11	82.31	85.12	85.37	87.23	82.17	85.78	83.21

Table 2: **Performance of Current LVLMs as PRMs.** We evaluate models using accuracy (ACC) and F1 score as Overall metrics, along with performance on specific error type detection: TE, RU, AE, RR, ODF, ARI, and LU.

Models	Accuracy	
	Correct	Incorrect
Gemma3-27B	55.83	42.17
InternVL2.5-78B	62.50	47.63
LLaVA-OneVision-72B	63.61	56.52
Ovis2-34B	58.31	38.05
Gemini-2.5-Flash	78.89	38.81
Random	50.0	50.0

Table 3: **Model Performance Comparison.** Accuracy of different models on correct and incorrect samples.

All LVLMs exhibit severe volatility with accuracy fluctuating dramatically across consecutive steps, often dropping from high values to near-zero performance within one step. This instability is particularly pronounced in correct sample evaluation, where most models struggle to maintain consistent performance above 0.5. Performance disparities between text reasoning steps (even steps) and image extraction steps (odd steps) are evident, with most models achieving higher accuracy during text reasoning while experiencing degradation during image extraction. Gemini 2.5-Flash demonstrates superior stability with smaller fluctuation amplitudes, while Ovis2-34B stands out among open-source LVLMs for more consistent evaluation capabilities, maintaining smaller fluctuation ranges alongside Gemini.

Conclusion

In this work, we first investigate the problems existing in the *thinking with images* reasoning paradigm and define fine-grained error types. To address a core research question: Whether Process Reward Models Can Help?—We employ guided search experiments, revealing that while current LVLMs can help, they still have significant potential for improvement. Based on this finding, we introduce *ThinkWithImages*-PRMBENCH, a benchmark characterized by fine-grained evaluation categories and challenging error types. We construct 1,206 data samples through rigorous human filtering and annotation. *ThinkWithImages*-PRMBENCH serves as a comprehensive testbed for evaluating different LVLMs as PRMs in supervising vision reasoning under the *thinking with images* paradigm. Through comprehensive evaluation of various models, we reveal two key findings: LVLMs as PRMs demonstrate limited capability in evaluating vision reasoning processes, and performance varies dramatically between error types. These findings indicate the substantial gap between current LVLM capabilities and the requirements for reliable process supervision in visual reasoning tasks. We anticipate our work will promote development in LVLM capabilities as PRMs to help the visual reasoning process.

Limitations

Our analysis is based on reasoning trajectories from current *thinking with images* models, which primarily support basic visual operations such as cropping and zooming. More advanced models with richer visual capabilities and sophisticated reasoning tools may exhibit different error patterns and potentially address some of the identified error categories. Future work should extend this taxonomic analysis to incorporate trajectories from more advanced *thinking with images* models.

Acknowledgments

This work is funded in part by the HKUST Startup Fund (R9911), Theme-based Research Scheme grant (No.T45-205/21-N) and the InnoHK funding for Hong Kong Generative AI Research and Development Center, Hong Kong SAR.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bordes, F.; Pang, R. Y.; Ajay, A.; Li, A. C.; Bardes, A.; Petryk, S.; Mañas, O.; Lin, Z.; Mahmoud, A.; Jayaraman, B.; et al. 2024. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*.

Cao, C.; Li, M.; Dai, J.; Yang, J.; Zhao, Z.; Zhang, S.; Shi, W.; Liu, C.; Han, S.; and Guo, Y. 2025a. Towards Advanced Mathematical Reasoning for LLMs via First-Order Logic Theorem Proving. *arXiv preprint arXiv:2506.17104*.

Cao, C.; Zhu, H.; Ji, J.; Sun, Q.; Zhu, Z.; Yinyu, W.; Dai, J.; Yang, Y.; Han, S.; and Guo, Y. 2025b. Safelawbench: Towards safe alignment of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, 14015–14048.

Cao, M.; Zhao, H.; Zhang, C.; Chang, X.; Reid, I.; and Liang, X. 2025c. Ground-R1: Incentivizing Grounded Visual Reasoning via Reinforcement Learning. *arXiv preprint arXiv:2505.20272*.

Chan, C.-M.; Xu, C.; Ji, J.; Ye, Z.; Wen, P.; Jiang, C.; Yang, Y.; Xue, W.; Han, S.; and Guo, Y. 2025a. J1: Exploring Simple Test-Time Scaling for LLM-as-a-Judge. *arXiv preprint arXiv:2505.11875*.

Chan, C.-M.; Xu, C.; Zhu, J.; Ji, J.; Hong, D.; Wen, P.; Jiang, C.; Ye, Z.; Yang, Y.; Xue, W.; et al. 2025b. Boosting Policy and Process Reward Models with Monte Carlo Tree Search in Open-Domain QA. In *Findings of the Association for Computational Linguistics: ACL 2025*, 7433–7451.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Cheng, Z.; Chen, Q.; Xu, X.; Wang, J.; Wang, W.; Fei, H.; Wang, Y.; Wang, A. J.; Chen, Z.; Che, W.; et al. 2025. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Fan, Y.; He, X.; Yang, D.; Zheng, K.; Kuo, C.-C.; Zheng, Y.; Narayanaraju, S. J.; Guan, X.; and Wang, X. E. 2025. GRIT: Teaching MLLMs to Think with Images. *arXiv preprint arXiv:2505.15879*.

Feng, Z.; Chen, Q.; Lu, N.; Li, Y.; Cheng, S.; Peng, S.; Tang, D.; Liu, S.; and Zhang, Z. 2025. Is PRM Necessary? Problem-Solving RL Implicitly Induces PRM Capability in LLMs. *arXiv preprint arXiv:2505.11227*.

Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 148–166. Springer.

Ghosh, A.; Acharya, A.; Saha, S.; Jain, V.; and Chadha, A. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.

Hu, Y.; Shi, W.; Fu, X.; Roth, D.; Ostendorf, M.; Zettlemoyer, L.; Smith, N. A.; and Krishna, R. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37: 139348–139379.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.

Jiang, C.; Heng, Y.; Ye, W.; Yang, H.; Xu, H.; Yan, M.; Zhang, J.; Huang, F.; and Zhang, S. 2025. VLM-R ̂3: Region Recognition, Reasoning, and Refinement for Enhanced Multimodal Chain-of-Thought. *arXiv preprint arXiv:2505.16192*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, G.; Xu, J.; Zhao, Y.; and Peng, Y. 2025. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9098–9108.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Liu, Z.; Zang, Y.; Zou, Y.; Liang, Z.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025. Visual Agentic Reinforcement Fine-Tuning. *arXiv preprint arXiv:2505.14246*.

Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Luo, L.; Liu, Y.; Liu, R.; Phatale, S.; Guo, M.; Lara, H.; Li, Y.; Shu, L.; Zhu, Y.; Meng, L.; et al. 2024. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*.

OpenAI. 2025. Thinking with images. https://openai.com/index/thinking-with-images/.

Ray, A.; Duan, J.; Brown, E.; Tan, R.; Bashkirova, D.; Hendrix, R.; Ehsani, K.; Kembhavi, A.; Plummer, B. A.; Krishna, R.; et al. 2024. SAT: Dynamic Spatial Aptitude Training for Multimodal Language Models. *arXiv preprint arXiv:2412.07755*.

Sarch, G.; Saha, S.; Khandelwal, N.; Jain, A.; Tarr, M. J.; Kumar, A.; and Fragkiadaki, K. 2025. Grounded Reinforcement Learning for Visual Reasoning. *arXiv preprint arXiv:2505.23678*.

Shen, H.; Zhao, K.; Zhao, T.; Xu, R.; Zhang, Z.; Zhu, M.; and Yin, J. 2024. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. *arXiv preprint arXiv:2411.16044*.

Shi, W.; Zhu, H.; Ji, J.; Li, M.; Zhang, J.; Zhang, R.; Zhu, J.; Xu, J.; Han, S.; and Guo, Y. 2025. LegalReasoner: Step-wised Verification-Correction for Legal Judgment Reasoning. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7297–7313. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Song, M.; Su, Z.; Qu, X.; Zhou, J.; and Cheng, Y. 2025. PRMBench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*.

Su, A.; Wang, H.; Ren, W.; Lin, F.; and Chen, W. 2025a. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*.

Su, Z.; Li, L.; Song, M.; Hao, Y.; Yang, Z.; Zhang, J.; Chen, G.; Gu, J.; Li, J.; Qu, X.; et al. 2025b. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*.

Su, Z.; Xia, P.; Guo, H.; Liu, Z.; Ma, Y.; Qu, X.; Liu, J.; Li, Y.; Zeng, K.; Yang, Z.; et al. 2025c. Thinking with Images for Multimodal Reasoning: Foundations, Methods, and Future Frontiers. *arXiv preprint arXiv:2506.23918*.

Sun, G.; Jin, M.; Wang, Z.; Wang, C.-L.; Ma, S.; Wang, Q.; Geng, T.; Wu, Y. N.; Zhang, Y.; and Liu, D. 2024. Visual agents as fast and slow thinkers. *arXiv preprint arXiv:2408.08862*.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Wandell, B. A. 1995. *Foundations of vision.* Sinauer Associates.

Wang, A.; Chen, H.; Lin, Z.; Han, J.; and Ding, G. 2025a. LSNet: See Large, Focus Small. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9718–9729.

Wang, W.; Ding, L.; Zeng, M.; Zhou, X.; Shen, L.; Luo, Y.; Yu, W.; and Tao, D. 2025b. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7907–7915.

Wang, W.; Gao, Z.; Chen, L.; Chen, Z.; Zhu, J.; Zhao, X.; Liu, Y.; Cao, Y.; Ye, S.; Zhu, X.; et al. 2025c. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Wang, W.; Jing, Y.; Ding, L.; Wang, Y.; Shen, L.; Luo, Y.; Du, B.; and Tao, D. 2025d. Retrieval-augmented perception: High-resolution image perception meets visual rag. *arXiv preprint arXiv:2503.01222*.

Wen, P.; Ji, J.; Chan, C.-M.; Dai, J.; Hong, D.; Yang, Y.; Han, S.; and Guo, Y. 2025. Thinkpatterns-21k: A systematic study on the impact of thinking patterns in llms. *arXiv preprint arXiv:2503.12918*.

Wu, J.; Guan, J.; Feng, K.; Liu, Q.; Wu, S.; Wang, L.; Wu, W.; and Tan, T. 2025. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*.

Xu, Z.; Zhou, P.; Ai, J.; Zhao, W.; Wang, K.; Peng, X.; Shao, W.; Yao, H.; and Zhang, K. 2025. MPBench: A Comprehensive Multimodal Reasoning Benchmark for Process Errors Identification. *arXiv preprint arXiv:2503.12505*.

Zhang, D.; Zhoubian, S.; Hu, Z.; Yue, Y.; Dong, Y.; and Tang, J. 2024a. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37: 64735–64772.

Zhang, X.; Gao, Z.; Zhang, B.; Li, P.; Zhang, X.; Liu, Y.; Yuan, T.; Wu, Y.; Jia, Y.; Zhu, S.-C.; et al. 2025. Chain-of-Focus: Adaptive Visual Search and Zooming for Multimodal Reasoning via RL. *arXiv preprint arXiv:2505.15436*.

Zhang, Y.-F.; Zhang, H.; Tian, H.; Fu, C.; Zhang, S.; Wu, J.; Li, F.; Wang, K.; Wen, Q.; Zhang, Z.; et al. 2024b. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*.

Zheng, C.; Zhang, Z.; Zhang, B.; Lin, R.; Lu, K.; Yu, B.; Liu, D.; Zhou, J.; and Lin, J. 2024. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*.

Zheng, Z.; Yang, M.; Hong, J.; Zhao, C.; Xu, G.; Yang, L.; Shen, C.; and Yu, X. 2025. DeepEyes: Incentivizing" Thinking with Images" via Reinforcement Learning. *arXiv preprint arXiv:2505.14362*.

Zhu, H.; Dai, J.; Ji, J.; Li, H.; Cai, C.; Wen, P.; Chan, C.-M.; Chen, B.; Yang, Y.; Han, S.; and Guo, Y. 2025a. SafeMT: Multi-turn Safety for Multimodal Language Models. arXiv:2510.12133.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025b. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.