GeoPQA: Bridging the Visual Perception Gap in MLLMs for Geometric Reasoning

Guizhen Chen[1,2,*] **Weiwen Xu**[2] **Hao Zhang**[2] **Hou Pong Chan**[2]
Deli Zhao[2,3] **Anh Tuan Luu**[1] **Yu Rong**[2,3]
[1] Nanyang Technological University [2] DAMO Academy, Alibaba Group [3] Hupan Lab

Abstract

Recent advancements in reinforcement learning (RL) have enhanced the reasoning abilities of large language models (LLMs), yet the impact on multimodal LLMs (MLLMs) is limited. Particularly in vision-intensive tasks like geometric reasoning, MLLMs hallucinate frequently, leading to inaccurate reasoning. We attribute this to the *perceptual bottleneck* in MLLMs, which caps the benefits of reasoning training. To quantify this, we design a Geo-Perception Question-Answering (GeoPQA) benchmark, targeting basic geometric concepts and spatial relationships. Experiments on GeoPQA reveal significant shortcomings of MLLMs in visual perception, constraining RL reward signals for training. To address this bottleneck, we propose a two-stage RL training framework by first enhancing the visual perception of geometric structures, then fostering reasoning capabilities. Applied to Qwen2.5-VL-3B-Instruct, our two-stage training improves geometric reasoning by 9.7% and problem-solving by 9.1%, compared to the direct reasoning training approach. Our method also generalizes to other vision-intensive domains like figure understanding, highlighting the importance of perceptual grounding in effective MLLM reasoning.[1]

1 Introduction

Recent advances in reasoning models such as DeepSeek-R1 (DeepSeek-AI, 2025) have demonstrated that reinforcement learning with verifiable reward (RLVR) can markedly strengthen the reasoning abilities of large language models (LLMs; Team (2025); OpenAI (2024b, 2025)). Motivated by these successes, several studies have applied similar RL training recipes to multimodal LLMs (MLLMs; Guo et al. 2025; Yang et al. 2025; Shen et al. 2025; Team et al. 2025b; Yuan et al. 2025; Xu et al. 2025; Leng et al. 2025). However, the performance gains on vision-intensive reasoning benchmarks, such as MathVerse (Zhang et al., 2024b) and MathVista (Lu et al., 2024), remain relatively limited. A closer examination suggests that these limitations often originate from more foundational issues in visual understanding, even before complex reasoning is attempted. Figure 1 shows an example of a model struggling with identifying the rotation angle – a task that is easy for humans. Such fundamental errors in vision understanding affect subsequent logical deductions, preventing the model from being rewarded.

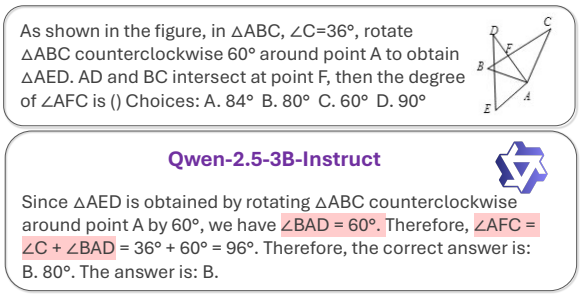

Figure 1: Illustration of an MLLM's perceptual errors leading to flawed reasoning. The model misidentifies rotation angles and misinterprets angle composition.

We hypothesize that *RL's efficacy in MLLMs is upper-bounded by their underlying visual perception ability*. Inadequate perception restricts the attainable reward signal and, consequently, the scope of reasoning improvement in RL training. To quantify this perception bottleneck, we curate visual perception QAs to assess models' understanding of basic geometric concepts and relationships. Empirical results indicate that MLLMs frequently fail to perceive geometric information, in contrast to the near-perfect human performance.

To overcome the perceptual bottleneck and promote effective reasoning training, we propose a two-stage RL framework comprising a perception stage followed by a reasoning stage, as shown in

[*] Guizhen is under the Joint PhD Program between Alibaba and NTU.

[1] https://github.com/DAMO-NLP-SG/GeoPQA

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25680–25688
November 4-9, 2025

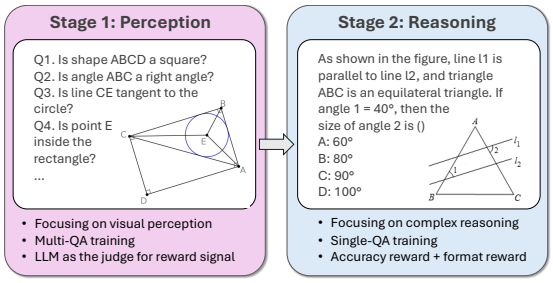

Figure 2: Overview of our two-stage RL framework.

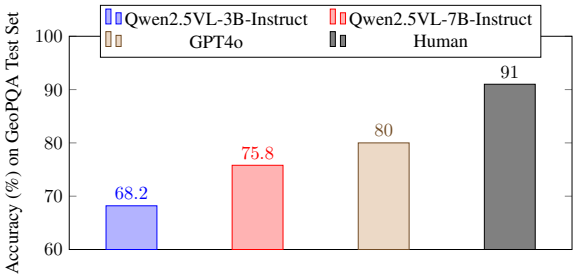

Figure 3: Perception capability of MLLMs against human on GeoPQA test set.

Figure 2. The first stage enhances models' visual understanding of basic geometric concepts and relationships using our curated perception-oriented QA dataset derived from both real and synthetic geometric diagrams. Building on the improved perceptual foundation, the second stage focuses on reasoning-oriented training, enabling the model to leverage its enhanced visual understanding and concentrate more effectively on the logical deduction process. Experiments on MathVista show that our two-stage framework is superior to the direct single-stage reasoning training, with 9.7% and 9.1% accuracy improvement in geometric reasoning and problem-solving respectively. It also outperforms larger open-source MLLMs and previous mathematical visual specialist models. Beyond geometric tasks, we further demonstrate that the perception-first paradigm generalizes to other vision-intensive tasks like figure and textbook understanding.

Our main contributions are threefold: (1) We reveal and quantify the perceptual limitations of current MLLMs in geometric tasks through targeted perception QAs, which are often overlooked by approaches focused solely on reasoning. (2) We introduce a two-stage reinforcement learning framework that first enhances visual perception before training for complex reasoning. (3) We validate the effectiveness of our approach on challenging geometric reasoning benchmarks, outperforming direct reasoning training and demonstrating potential generalization to other vision-intensive tasks.

2 Methodology

In this section, we first systematically assess the perceptual capabilities of MLLMs in the geometric domain. Next, we develop a two-stage RL framework that first enhances perception and subsequently boosts reasoning capabilities of MLLMs.

2.1 Preliminary Analysis

While existing research suggests that MLLMs struggle with geometric images, few studies comprehensively assess their perceptual abilities in this domain. MathVerse (Zhang et al., 2024b) infers perceptual ability based on models' capabilities to answer geometric questions with varying levels of visual descriptions in the text format. VisOnly (Kamoi et al., 2024) and Geoperception (Zhang et al., 2024a) directly evaluate geometric perception, but uses a limited range of templates, hindering the evaluation of diverse perceptual skills.

To thoroughly evaluate models' geometric perception, we construct a test set from the image-caption pairs in the Geo170K dataset (Gao et al., 2025). Specifically, we prompt Gemini-2.0-Flash-Thinking (Gemini-FT; DeepMind (2025)) to generate questions that require recognizing basic visual elements and spatial relationships directly from the image descriptions (see Appendix A.1). These questions cover: (1) **basic geometric elements** such as identifying shapes (*e.g.,* triangles, circles), comparing lengths, and recognizing angles (*e.g.,* right, acute, obtuse); (2) **geometric relationships** such as intersection, parallelism, perpendicularity, and tangency. To facilitate automatic evaluation, the answers are designed for easy verification, restricting them to yes/no, numerical values, or simple strings (*e.g.,* "ABC").

We assess the performance of several representative MLLMs in Figure 3. The results reveal significant deficiencies in models in answering these basic visual questions, even for state-of-the-art models like GPT4o, while humans easily attain over 90% accuracy in these questions. This highlights a critical perceptual gap that limits the effectiveness of subsequent reasoning training via RL.

2.2 Framework Overview

To overcome the perceptual bottleneck, we introduce a two-stage RL framework: perception followed by reasoning. Stage 1 focuses on improving the model's ability to accurately perceive and in-

terpret geometric information, while stage 2 leverages the enhanced perceptual foundation to develop more complex, multi-step reasoning capabilities.

2.3 Stage 1: Perception-oriented Training

Training data. To enhance the geometric perception of MLLMs, we curate a comprehensive **Geo-Perception Question-Answering (GeoPQA)** training dataset, comprised of both real-world and synthetic figures. For real-world images, we employed the same methodology as described in our preliminary study: leveraging Gemini-FT to generate perception-focused QAs. We further augment this data with synthetically generated geometric figures to cover a wider range of scenarios. Following methodologies in AlphaGeometry (Trinh et al., 2024) and AutoGen (Wu et al., 2023), we create basic shapes and composite shapes. Additionally, we use geometric annotations to visually enrich the diagrams (see Appendix A.2). The perception QAs for the synthetic figures are generated similarly to those for real images, concentrating on elements and relationships present in the generated diagrams. Since each image contains rich visual information, we generate a set of perception questions $Q = (q_1, \ldots, q_n)$ per image ($n \leq 7$). Dataset statistics are presented in Appendix A.3.

Quality control. To ensure quality, we use GPT-4o (OpenAI, 2024a) to filter out questions where: (1) the image does not explicitly contain the information required to answer the question, or (2) the provided ground-truth answer contradicts the information evident in the image description. To validate the quality of the dataset after GPT-4o filtering, we perform a human inspection on 100 random samples: 92% are valid and high-quality. More details are shown in Appendix A.4.

RL training. The input to the model at this stage is formulated as

$$x = (I, q_1, \ldots, q_n)$$

where I is the instruction, and q_i is a visual perception question. Given an input x, the policy π_θ generates a free-form textual response $y \sim \pi_\theta(y|x)$. This response y is expected to contain the answers to all n questions. Let the ground-truth answers be

$$A = (a_1, \ldots, a_n)$$

where a_i is the ground-truth answer to q_i. To evaluate the correctness of the generated response y, we employ GPT-4o-mini (OpenAI, 2024a) as the judge, denoted as J. The judge J parses the response y to extract the model's predicted answers for each perception question, yielding $J(y) = \hat{A} = (\hat{a}_1, \ldots, \hat{a}_n)$. The accuracy reward $R(x, y)$ for a given input x and generated response y is defined as:

$$R(x, y) = \begin{cases} 1, & \text{if } \hat{a}_i = a_i, \forall i \in \{1, \ldots, n\} \\ 0, & \text{otherwise.} \end{cases}$$

This strict reward function grants a positive reward only if all perception questions are answered correctly. To mitigate reward hacking (*e.g.,* the model learning to always output "yes"), we downsample training instances where all ground-truth answers are "yes". Other training details are kept the same as in the original Group Relative Policy Optimization (GRPO; DeepSeek-AI (2025)).

2.4 Stage 2: Reasoning-oriented Training

With the improved perceptual capabilities, the MLLM proceeds to stage 2, where it is trained on geometric reasoning tasks. We use the QA tuning subset from Geo170K (Gao et al., 2025). We follow the standard GRPO setup to apply RL training at this stage.

2.5 Implementation

We train our models based on Qwen-2.5-VL-3B-Instruct and Qwen-2.5-VL-7B-Instruct (Bai et al., 2025). Besides the backbone models, we compare against several baselines, including five proprietary MLLMs (Bai et al., 2024; Team et al., 2025a; OpenAI, 2023, 2024a) and six open-source MLLMs (Liu et al., 2024; Gao et al., 2025; Shi et al., 2024; Zhang et al., 2025; Dong et al., 2024; Chen et al., 2025). We evaluate models on geometry reasoning (GR) and geometry problem-solving (GPS) on MathVista (Lu et al., 2024). Training and evaluation details are provided in Appendix B and C respectively.

3 Results and Analyses

Main results. Table 1 shows that our two-stage approach significantly outperforms the reasoning-only training approach by a large margin of 9.7% in GR and 9.1% in GPS. Notably, the reasoning-only approach scores even slightly lower than the original baseline. This suggests that directly applying RL for reasoning without addressing underlying perceptual limitations can be ineffective or even

Model	GR	GPS
Proprietary MLLMs		
Qwen-VL-Plus	39.3	38.5
GeminiPro	–	40.4
GPT-4V	51.0	50.5
GeminiUltra	–	56.3
GPT-4o	**74.1**	**75.0**
Open-source MLLMs		
SPHINX-MoE (8 × 7B)	30.5	31.2
G-LLAVA* (13B)	–	56.7
Math-LLAVA* (13B)	56.5	57.7
MAVIS* (7B)	63.2	64.1
InternLM-XC2 (7B)	62.3	63.0
InternVL2.5 (4B)	64.4	67.3
Qwen2.5-VL-3B-Instruct	63.2	63.9
w/ Reasoning	62.3	63.0
w/ Perception and Reasoning mixed	65.7	65.9
w/ Perception followed by Reasoning	**72.0**	**72.1**

Table 1: Results of MathVista-*testmini* (Lu et al., 2024) on geometry reasoning (GR) and geometry problem solving (GPS). * denotes visual specialists in mathematics. The highest scores for proprietary and open-source MLLMs are **bolded**.

detrimental, and our perception training effectively bridges the perception gap.

In addition, we observe that incorporating perception data, whether mixed or sequential, boosts reasoning performance over reasoning-only RL. This validates the effectiveness of our perception QA dataset. Moreover, a structured, sequential approach of perception followed by reasoning training yields greater benefits than simply mixing perception and reasoning data, validating the effectiveness of our two-stage framework.

Compared to other leading open-source MLLMs and models specialized in mathematics, our method establishes new state-of-the-art performance on geometric tasks, even with a much smaller model size. While GPT-4o remains the top performer among proprietary models with 74.1% on GR and 75.0% on GPS, our two-stage framework, applied to the Qwen2.5-VL-3B-Instruct, considerably narrows the performance gap to just around 2%.

Enhancement on visual perception. Table 2 directly quantifies the benefits of our approach on models' visual perception, measured by GeoPQA. The results provide several key insights:

- **Effectiveness of perception training:** Perception training significantly improves the performance on GeoPQA by 21.6%, which validates its effectiveness in directly enhancing geometric visual perception.
- **Necessity of a staged approach:** Reasoning training degrades performance on GeoPQA by 15.1%, suggesting that training directly on high-level reasoning can cause the model to neglect or unlearn basic perceptual abilities, which justifies our two-stage approach.
- **Balanced two-stage approach:** Our two-stage approach maintains a high perception score of 83.2%, with a 15% gain over the baseline while also achieving the significant reasoning gains reported in our main results.

Model	GeoPQA
Qwen2.5-VL-3B-Instruct	68.2
w/ Perception	89.8
w/ Reasoning	53.1
w/ Perception followed by Reasoning	83.2

Table 2: Performance of Qwen2.5-VL-3B-Instruct on GeoPQA.

Impact of training strategy on tasks of different vision intensity. To further understand the benefits of our training framework, we analyze its performance on MathVerse (Zhang et al., 2024b) Plane Geometry problems, which includes 5 vision intensities: Text Dominant (TD), Text Lite (TL), Vision Intensive (VI), Vision Dominant (VD), and Vision Only (VO). As shown in Table 3, across all categories, both perception-involved methods generally outperform the reasoning-only approach and the base model. Notably, our two-stage approach excels in the Vision Only scenario compared to the single-stage mixed approach. The results suggest that while mixing perception and reasoning data can be beneficial, a dedicated initial stage focused purely on perception, as in our two-stage framework, is crucial for tasks where the model cannot rely on textual cues to compensate for perceptual weaknesses.

Model	TD	TL	VI	VD	VO
Qwen2.5-VL-3B-Instruct	46.5	39.6	37.5	38.4	37.1
w/ Reasoning	49.8	44.3	38.0	41.8	43.9
w/ Perception and Reasoning mixed	56.1	51.6	48.6	48.2	39.8
w/ Perception followed by Reasoning	55.3	52.5	47.5	47.6	45.5

Table 3: Results of MathVerse-*testmini* (Lu et al., 2024) on the plane geometry subset.

Training	GR	GPS	GeoPQA
Single QA	52.7	52.9	95.4
Multiple QAs	62.3	63.5	94.0

Table 4: Effects of perception training with single QA per image vs. multiple QAs per image.

Impact of multiple perception QAs per image. In stage 1, we formulate training samples by concatenating multiple perception questions for a single image. To evaluate the effectiveness of this setting, we conduct an ablation study comparing it against the conventional method, where each perception question is treated as an individual training sample. The results in Table 4 show that training with multiple QAs per image demonstrates a substantial advantage for downstream reasoning tasks, with an improvement of 9.6% for GR and 10.6% for GPS. While the multi-QA setup exhibits slightly lower performance on our perception task, this is potentially attributable to the stricter reward in the multi-QA training, in which the model only receives a positive reward if all sub-questions associated with an image are answered correctly. This more demanding training setup, however, encourages the model to learn more robustly on perception, ultimately leading to superior performance on downstream reasoning tasks.

Results on a larger scale. To demonstrate that our method remains effective at a larger scale, we extend our experiments to a 7B model. The results in Table 5 are consistent with our observations with the 3B model, with 2.6% gain in GR and 4.8% improvement in GPS. Notably, our 7B model surpasses all other models, including the strong proprietary baseline GPT-4o (74.1% GR, 75.0% GPS). These results reinforce our central claim: even for a more capable base model, enhancing foundational visual perception is a critical prerequisite for unlocking further gains in high-level reasoning and problem-solving.

Model	GR	GPS
Qwen2.5-VL-7B-Instruct	74.1	75.5
w/ Reasoning	73.6	75.0
w/ Perception followed by Reasoning	**76.2**	**79.8**

Table 5: Performance of Qwen2.5-VL-7B-Instruct on geometry reasoning (GR) and geometry problem solving (GPS) from MathVista-*testmini* (Lu et al., 2024).

Generalization to other tasks. To assess the broader impact of our two-stage training, we evaluate a diverse set of other tasks from MathVista, comparing our perception-then-reasoning approach against the reasoning-only baseline. The results are presented in Appendix D. Performance gains are observed in visually grounded tasks, including figure question answering (+1.5%), textbook QA (+2.6%), and scientific reasoning (+2.5%), indicating that improved visual perception from stage 1 training facilitates more effective reasoning for tasks that involve interpreting diagrams. The impact of geometry-focused perception training is less pronounced or slightly negative on tasks that are more text-reliant or require different types of visual understanding than geometry, such as numerical commonsense (-2.8%) and math word problems (-1.1%).

4 Conclusion

We investigate the fundamental challenge of improving geometric reasoning capabilities in MLLMs. Our analysis reveals that the effectiveness of reinforcement learning for reasoning is significantly constrained by MLLMs' visual perception, a critical bottleneck that is often not directly measured in previous work. By developing a targeted assessment of geometric perception and introducing a two-stage RL framework that explicitly enhances visual perception prior to reasoning training, we achieved substantial improvements on challenging benchmarks. The success of our perception-first training approach highlights an important principle for future work in multimodal reasoning: strong perceptual foundations are prerequisites for effective higher-level reasoning. Future directions include exploring whether our approach can enhance the performance of recent thinking-with-images approaches (OpenAI, 2025; Su et al., 2025; Ma et al., 2024) and generalizing our framework to other vision-intensive tasks like chart understanding (Huang et al., 2024).

Limitations

In our experiments, the accuracy of stage 1 training relies on an LLM judge (GPT-4o-mini) to parse free-form answers and determine the correctness for the perception QAs. This introduces extra cost and time calling the APIs.

Acknowledgment

This research is jointly supported by DSO grant DSOCL23216, DAMO Academy Research Intern Program, and Alibaba-NTU Singapore Joint Research Institute. We would also like to extend our gratitude to Interdisciplinary Graduate Programme and College of Computing and Data Science of NTU, for their support.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond.

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.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Google DeepMind. 2025. Introducing Gemini 2.0: our new AI model for the agentic era. Blog post on Google DeepMind website.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, and 4 others. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *Preprint*, arXiv:2401.16420.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. 2025. G-llava: Solving geometric problem with multi-modal large language model. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, and 1 others. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Kung-Hsiang Huang, Hou Pong Chan, Yi R Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Transactions on Knowledge and Data Engineering*.

Ryo Kamoi, Yusen Zhang, Sarkar Snigdha Sarathi Das, Ranran Haoran Zhang, and Rui Zhang. 2024. Visonlyqa: Large vision language models still struggle with visual perception of geometric information. *CoRR*, abs/2412.00947.

Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Hang Zhang, Yuming Jiang, Xin Li, Deli Zhao, Fan Wang, Yu Rong, Aixin Sun, and Shijian Lu. 2025. Mmr1: Advancing the frontiers of multimodal reasoning. https://github.com/LengSicong/MMR1.

Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, Wenqi Shao, Chao Xu, Conghui He, Junjun He, Hao Shao, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. 2024. SPHINX-x: Scaling data and parameters for a family of multi-modal large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32400–32420. PMLR.

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*.

Jingkun Ma, Runzhe Zhan, Derek F. Wong, Yang Li, Di Sun, Hou Pong Chan, and Lidia S. Chao. 2024. Visaidmath: Benchmarking visual-aided mathematical reasoning. *Preprint*, arXiv:2410.22995.

OpenAI. 2023. Gpt-4v(ision) system card.

OpenAI. 2024a. GPT-4o. https://openai.com/gpt-4o. Accessed: 2024-05-20.

OpenAI. 2024b. Openai o1 system card. *CoRR*, abs/2412.16720.

OpenAI. 2025. Introducing *o3* and *o4-mini*.

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

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4663–4680, Miami, Florida, USA. Association for Computational Linguistics.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, Linjie Li, Yu Cheng, Heng Ji, Junxian He, and Yi R. Fung. 2025. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *Preprint*, arXiv:2506.23918.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025a. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Kimi Team. 2025. 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 74 others. 2025b. Kimi-vl technical report.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. *CoRR*, abs/2308.08155.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, and 1 others. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.

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. *arXiv preprint arXiv:2503.10615*.

Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan, Deli Zhao, Tingyang Xu, Zhongyu Wei, and 1 others. 2025. Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *arXiv preprint arXiv:2507.22607*.

Jiarui Zhang, Ollie Liu, Tianyu Yu, Jinyi Hu, and Willie Neiswanger. 2024a. Euclid: Supercharging multimodal llms with synthetic high-fidelity visual descriptions. *Preprint*, arXiv:2412.08737.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024b. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186, Berlin, Heidelberg. Springer-Verlag.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Shanghang Zhang, Peng Gao, and Hongsheng Li. 2025. MAVIS: Mathematical visual instruction tuning with an automatic data engine. In *The Thirteenth International Conference on Learning Representations*.

A Details of GeoPQA

A.1 Prompt to generate perception QAs

The following prompt is used with Gemini-2.0-Flash-Thinking to generate the initial set of perception question-answer pairs.

```
Create perception questions based on the provided image description. The questions should be formulated such that:
1. They involve recognizing basic visual elements and spatial relationships directly from the image.
2. They are answerable from the image description.
3. Answers must be "yes/no", a number, or a simple string like "AB" (no spaces).
4. No reasoning should be provided with the answer.
5. Avoid rephrasing the same question.
6. Output the results as a JSON array of objects. Each object should have keys "question" and "answer". If no meaningful question can be generated, return an empty array. If the image is too simple, for example, only contains a single point or a line segment, return an empty array.
7. No more than seven questions should be generated.

Image Description: Triangle ABC is a right angle isosceles triangle, with ∠BAC as the right angle. The circle that passes through points A, C, and B has center D.
Questions: [{"question": "Is triangle ABC a right triangle?", "answer": "Yes"}, {"question": "Which vertex has the right angle in triangle ABC?", "answer": "A"}, {"question": "Does the circle pass through point D?", "answer": "No"}, {"question": "What is the measure of angle BAC?", "answer": "90"}, {"question": "Are sides AB and AC equal in length?", "answer": "Yes"}]

Image Description: <description>
Questions:
```

A.2 Synthetic geometric diagram generation

We create the following to generate the synthetic geometric diagrams:

- **Basic shapes**: Line segments, circles, triangles, quadrilaterals, and pentagons.
- **Composite shapes**: Combinations of 2-4 random basic shapes with predefined spatial relationships (e.g., a circle tangent to a triangle).
- **Annotations**: Diagrams are explicitly annotated with special geometric symbols, such as right-angle symbols and markings for equal sides/angles, which are commonly understood by humans but potentially ambiguous for MLLMs.

A.3 Dataset statistics

The dataset is split into 659 test samples and 5420 training samples. The training set contains a balanced mix of real-world and synthetic images. The distribution is shown in Table 6.

Image Type	# Images	# Questions
Real	2548	7038
Synthetic	2872	9303

Table 6: Distribution of real vs. synthetic images.

To provide an estimate of sample complexity, we analyse the number of perception questions associated with each image, which serves as a proxy for its visual complexity. The distribution is shown in Table 7.

# Questions per Sample	% Share
1	9.56
2	23.28
3	35.15
4	22.38
5+	9.63

Table 7: Percentage share of the number of questions per sample.

The created perception questions cover a range of geometric concepts, including (1) **basic geometric elements** such as identifying shapes (*e.g.,* triangles, circles), comparing lengths and recognising angles (*e.g.,* right, acute, obtuse); (2) **geometric relationships** such as intersection, parallelism, perpendicularity, and tangency. Table 8 shows the distribution of question types in GeoPQA.

Category	Sub-category	Count
Geometric Elements	Shapes	4387
	Angles	1737
	Lengths	1405
	Area/Perimeter	46
	Others	243
Geometric Relationships	Relative Position	5662
	Intersection	1108
	Perpendicularity	500
	Parallelism	234
	Tangency	432
	Congruence/Similarity	410
	Transformation	177

Table 8: Distribution of question types.

A.4 Quality control of generated perception QAs

We prompt GPT-4o (OpenAI, 2024a) to filter out low-quality questions. The following prompt is used.

```
Your task is to evaluate the correctness of a user's answer based on an image, its description, and a given question. The user's answer is considered incorrect if:
- The image does not explicitly contain the information needed to answer the question.
- The answer contradicts the information presented in the image description.

**Input**:
- Image Description: <description>
- Question: <question>
- User's Answer: <response>

**Output Format**:
Provide your reasoning and judgment (0 = correct, 1 = incorrect) in the following format:
<think>{{Your concise reasoning, including consideration of the image description and question, and how it relates to the user's answer.}}</think> <judge>{{0 or 1, 0 if the user's answer is correct, 1 if the user's answer is incorrect.}}</judge>
```

To validate the quality of the dataset after GPT-4o filtering, we perform a human inspection on 100 random samples: 92% are valid and high-quality. The 8% invalid samples comprised 2% from the synthetic subset and 6% from the real-world image subset. While there is a slight error rate, our main results show that perception training on this dataset still yields a significant benefit over reasoning-only training. This demonstrates the practical effectiveness of our dataset. Furthermore, since most errors are from the real-world subset, we can further improve the dataset quality by increasing the proportion of high-quality synthetic data in future iterations.

B Training setup

Table 9 shows the hyperparameter configuration in our training. We adopt the same settings across all experiments to ensure a fair and direct comparison.

C Evaluation

All evaluation is conducted using the VLMEvalKit toolkit[2], ensuring standardized and reproducible evaluation metrics.

[2] https://github.com/open-compass/VLMEvalKit

Hyperparameter	Configuration
Max Prompt Length	2048
Max Response Length	2048
Max Image Pixels	1,048,576
Min Image Pixels	65,536
Global Batch Size	128
Rollout Batch Size	512
Learning Rate	1e-6
Optimizer	AdamW
N Rollouts	5
Training Episodes	10

Table 9: Hyperparameters used in training.

D Performance on other tasks

Table 10 shows the performance of our method vs. the reasoning-only method on other MathVista tasks.

MathVista Task Category	Reasoning-only	Perception + Reasoning
Figure Question Answering	68.0	69.5
Textbook Question Answering	62.0	64.6
VisualQA	58.1	57.0
Scientific Reasoning	59.8	62.3
Numeric Commonsense	43.1	40.3
Arithmetic Reasoning	56.7	58.1
Algebraic Reasoning	62.6	69.4
Math word problem	62.9	61.8
Logical Reasoning	29.7	37.8

Table 10: Performance (%) on MathVista other tasks. "Perception + Reasoning" refers to our two-stage approach.