CapGeo-Bench: Decoupling Visual Perception from Reasoning and Evaluating Geometric Understanding

Anonymous ACL submission

Abstract

While Multimodal Large Language Models (MLLMs) have achieved remarkable success in difficult purely textual mathematical reasoning tasks, even advanced closed-source models such as GPT-o3 still struggle with geometric problems. This discrepancy motivates us to investigate the root cause: is the bottleneck of multimodal geometric reasoning rooted in reasoning itself, or in the perception of geometric information from diagrams? To answer this question, we conduct an exploratory experiment and find that providing high-quality captions consistently and substantially boosts performance of MLLMs, empirically validating the visual perception bottleneck in geometric reasoning. However, MLLMs' capabilities in visual geometric perception remain insufficiently evaluated. To fill the gap, we propose **CapGeo-Bench**, a benchmark of 4,641 high-quality figure–caption pairs equipped with a fine-grained keypoint-based evaluation metric. Specifically, CapGeo-Bench encompasses three geometric classes and is devided into four difficulty levels according to the complexity of geometric primitives. Captions are scored at a fine-grained level by extracting keypoints along three dimensions-geometric elements, spatial relationships, and numerical attributes. CapGeo-Bench provides a reliable instrument and guideline for evaluating and advancing MLLMs' geometric perception capabilities. Our code and data are publicly available at: `https://anonymous.4open.science/r/CapGeo-2042/README.md`.

1 Introduction

Although MLLMs have achieved promising results across diverse tasks (Bai et al., 2025a; Yao et al., 2024b), multimodal reasoning remains a central challenge. Geometric benchmarks such as MathVerse (Zhang et al., 2024b), MathVista (Lu et al., 2023), and GeoQA (Chen et al., 2021) further underscore this limitation: even GPT-o3 and Gemini2.5-Pro (Gemini Team, Google DeepMind, 2025) often struggle to solve geometry-related problems. In contrast, they demonstrate much greater stability and accuracy in text-only reasoning tasks such as the International Mathematical Olympiad (IMO) (Huang and Yang, 2025) and International Physics Olympiad (IPhO) (Qiu et al., 2025). This stark discrepancy raises a critical question- *What exactly are the bottlenecks that constrain the multimodal geometric reasoning of the existing MLLMs?*

As shown in Figure 1, MLLMs frequently fail to obtain clear geometric conditions from visual input, producing hallucinated or mismatched relations that derail subsequent reasoning. In contrast, when accurate captions are supplied, GPT-o3 reasons correctly. Inspired by this observation and relevant research (Wei et al., 2024a; Chen et al., 2025a), we think that imperfect visual perception—rather than reasoning itself—is the dominant factor constraining geometric reasoning.

To verify this hypothesis, we design **CapGeo**, a caption-assisted reasoning setup that decouples perception from reasoning. Specifically, we systematically compare model performance under three input modalities: Image-only, Caption-only, and Caption+Image. As shown in Table 2, reasoning models equipped with GPT-o3-generated captions yields dramatic and consistent improvements over the Image-only baseline on MathVerse. We further conduct two ablation studies to rule out alternative explanations like textual prior knowledge and any specific template design. Together, these results empirically confirm that visual perception, rather than reasoning, is the primary bottleneck of multimodal geometric reasoning.

This finding also aligns with a recent trend in multimodal reinforcement learning, where researchers explicitly inject perception-level supervision or description-based reward signals into the training (Xing et al., 2025; Chen et al., 2025b;

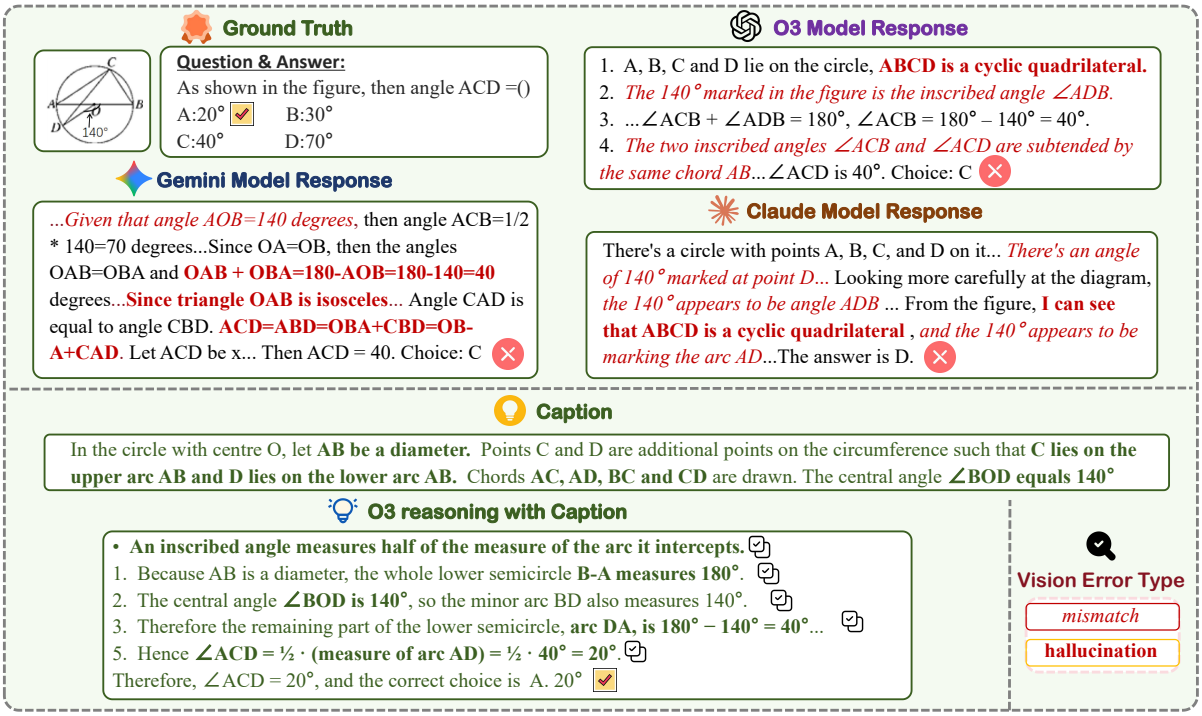

Figure 1: Bad Cases in Mathverse. The upper shows the incorrect results of MLLMs. *Mismatch* means generated relationship exists, but the relationship subjects mismatch. After caption assistance, GPT-o3 reasons correctly.

Xiao et al., 2025; Xia et al., 2025). However, *how systematically measure the quality of geometric captions and geometric perception capability of MLLMs* is still a blank area. Existing benchmarks and metric for general image captioning fall short for geometry, where descriptions must faithfully encode elements, spatial relations, and numerical constraints. To fill this gap, we introduce **CapGeo-Bench**, a benchmark of 4,641 high-quality geometric figure–caption pairs spanning three geometric classes and four difficulty levels. We further propose a keypoint-based fine-grained evaluation framework that scores captions along three dimensions—elements, spatial relationships, and numerical attributes—providing a principled metric for diagnosing MLLMs' geometric perception. Comprehensive experiments on CapGeo-Bench reveal that current MLLMs still struggle with geometric captioning, particularly in the numerical dimension. Moreover, a joint analysis shows that CapGeo-Bench scores are strongly correlated with downstream reasoning performance, confirming its reliability as a perception evaluation tool.

Our contributions are summarized as follows:

- We conduct an **exploratory experiment via CapGeo**, a caption-assisted reasoning setup that decouples perception from reasoning. Comprehensive comparisons across three input modalities verify that visual perception, rather than reasoning, is the primary bottleneck of multimodal geometric reasoning.
- Motivated by this finding, we release **CapGeo-Bench**, a high-quality benchmark of 4,641 geometric figure–caption pairs covering a broad spectrum of categories and difficulty levels, dedicated to evaluating MLLMs' geometric information extraction capability.
- We propose a novel **keypoint-by-keypoint geometric caption evaluation method** that scores captions along three fine-grained dimensions validated by multiple mathematics experts. This method provides a principled, interpretable, and reproducible metric for diagnosing MLLMs' geometric perception. Scores on CapGeo-Bench show a strong correlation with downstream reasoning performance, further confirming the reliability of our evaluation framework.

2 Related Work

2.1 MLLMs Reasoning

Recently, The performance of MLLMs shone brightly. High-performance MLLMs include OpenAI's o3 (OpenAI, 2025), Qwen2.5-VL series (Bai et al., 2025b), and LLaVA series (Liu et al., 2023).

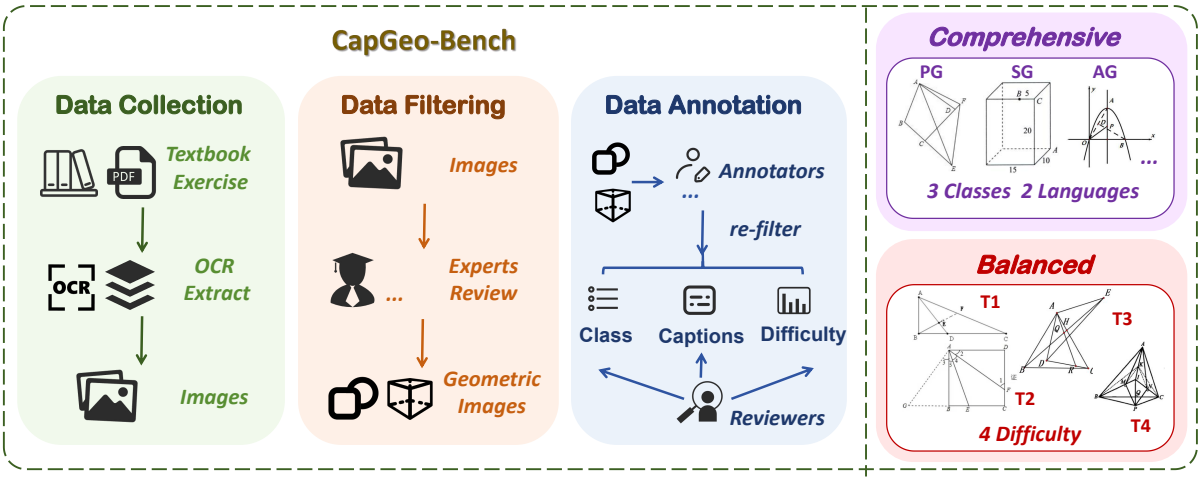

Figure 2: Overview of CapGeo-Bench. AG: Analytic Geometry, PG: Plane Geometry, SG: Solid Geometry

Current paradigms for enhancing multimodal reasoning capabilities of MLLMs encompass (Li et al., 2025): (1) Multimodal Chain-of-Thought (MCoT), which employs prompt engineering to guide LLs through step-by-step reasoning processes (Zhang et al., 2023; Jia et al., 2024b; Wei et al., 2024b; Gao et al., 2024); (2) Multimodal-o1 models (Zhang et al., 2024a; Yao et al., 2024a; Thawakar et al., 2025); (3) Multimodal-R1 models (Liu et al., 2025; Shen et al., 2025).

2.2 Geometric Reasoning

The research on multimodal geometric reasoning emerge constantly. G-LLAVA (Gao et al., 2023) trained the G-LLaVA model using the geometric synthesis dataset Geo170K; and MMGeoLM (Sun et al., 2025) constructed positive and negative sample pairs for contrastive learning training. Benchmarks such as MathVista (Lu et al., 2023) and MathVision (Wang et al., 2024) assess the geometric reasoning capabilities of MLLMs in multiple geometric domains. Some research explores the pseudo-visual phenomena in existing MLLMs, such as MathVerse (Zhang et al., 2024b).

Compared with studies like Describe-then-Reason (Jia et al., 2024a) framework, CapGeo is designed for exploratory experiment. It decouples perception from reasoning, providing conditions for a deep analysis of the performance differences under different input modalities to verify our view.

2.3 Image Captioning

Image captioning is a fundamental direction in Computer Vision. Broadly speaking, there are two main approaches: (i) direct caption generation. In the era of MLLMs, enhancing the captioning capabilities of MLLMs remains a significant focus for many researchers (Zhang et al., 2024c; Lee et al., 2024; Zheng et al., 2024). Meanwhile, the evaluation of image captions itself has also become an important research direction (Dong et al., 2024; Sarto et al., 2025); (ii) OCR-based methods. For MLLMs, OCR focuses on extracting pre-existing textual content from images. This capability not only emphasizes information extraction performance (Yang et al., 2021; Mathew et al., 2021; Fu et al., 2024; Liu et al., 2024), but also highlights the reasoning ability required in real-world scenarios (He et al., 2025; Huang et al., 2025).

Nevertheless, most studies target general natural images. Some geometric datasets include captions, but they lack evaluation (Peng et al., 2024). To address this gap, we propose *CapGeo-Bench*.

3 CapGeo: An Exploratory Experiment on the Perception Bottleneck

3.1 Experimental Design

CapGeo consists of two stages: a captioning stage that converts a geometric figure into a textual caption, and a reasoning stage that consumes the caption (and optionally the original figure) for problem solving.

Captioning stage. We instruct the captioner to describe all visual information in the figure using formal geometric language. see Figure 7 for the prompt template.

Reasoning stage. Given a question Q, the original image I, and the generated caption $C = \text{Caption}(I)$, the reasoning model outputs an answer

$$A = \text{MLLM}(Q, I, C).$$

Statistic	Value
Total images	4,641
- Difficulty 1	1,816
- Difficulty 2	2,133
- Difficulty 3	495
- Difficulty 4	197
- Plane Geometry	4,023
- Analytic Geometry	517
- Solid Geometry	101
Maximum word	586
Average word	61.4

Table 1: CapGeo-Bench Statistics

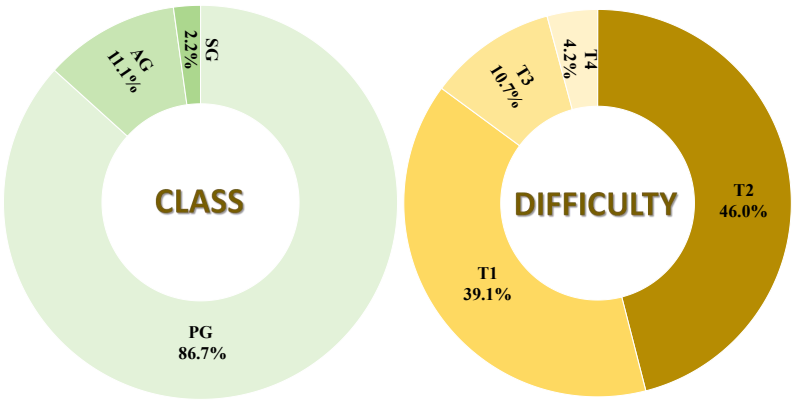

Figure 3: CapGeo-Bench Data Statistics Chart

By selectively masking I or C, we obtain three input modalities: (i) **Image only** (Img): the conventional vision-based setting; (ii) **Caption only** (Cap): pure text-based reasoning; (iii) **Caption + Image** ($Cap{+}Img$): joint multimodal input.

This controlled design allows us to disentangle the contributions of perception and reasoning.

3.2 Experimental setup and result

Experimental setup. We employ four captioning models (GPT-o3, GPT-4o, Qwen2.5-VL-72B-Instruct, Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)) and eight reasoning models spanning both MLLMs (Claude-Opus-4 (Anthropic, 2025), Gemini-2.5-pro (Gemini Team, Google DeepMind, 2025), GPT-o3, Qwen2.5-VL-72B-Instruct, Qwen2.5-VL-7B-Instruct) and LLMs (DeepSeek-R1 (Guo et al., 2025), Qwen2.5-72B-Instruct, and Qwen2.5-7B-Instruct (Bai et al., 2025a)). We evaluate on MathVerse (Zhang et al., 2024b) (Vision-Only and Vision-Intensive), MathVista (Lu et al., 2023) (geometry subset), and GeoQA (Chen et al., 2021) (200 sampled problems). All models use temperature 0.0 for deterministic output.

Verification. Table 2 presents results on MathVerse using GPT-o3 as the captioner, with full results across all four captioners in Appendix B and additional results on MathVista and GeoQA in Appendix C. All consistently show that caption-assisted reasoning substantially outperforms image-only baselines. For instance, on MathVerse (Vision-Only), Qwen2.5-VL-72B-Instruct surges from 8.6% to 66.4%, and Claude-Opus-4 from 44.8% to 73.0%. Crucially, even text-only LLMs (e.g., DeepSeek-R1, Qwen2.5-72B-Instruct) reach performance comparable to MLLMs once captions are supplied, indicating that captions carry sufficient geometric information for reasoning.

Ablation 1: Ruling out textual priors. A potential alternative explanation is that captions merely activate the model's textual prior knowledge rather than truly conveying visual information. To rule this out, we compare CapGeo's *Caption-only* setting against MathVerse's native *Text-only* setting, which provides only the textual question (without image or caption) and thus precisely isolates the pure textual prior. As shown in Table 4, Caption-only consistently and substantially outperforms Text-only—e.g., Qwen2.5-VL-72B-Instruct improves from 62.2% to 66.1% (Vision-Only) and 74.9% (Vision-Intensive). Since the Text-only setting provides neither image nor caption, the model can only rely on textual prior knowledge. The consistent and substantial gap between Caption-only and Text-only directly quantifies the contribution of extracted visual information, ruling out textual priors as the source of improvement.

Ablation 2: Ruling out template exploitation. Another concern is whether the gains depend on a specific captioning format. To address this, we replace our natural-style captions with a rigid JSON schema that explicitly enumerates geometric elements, relationships, and numerical information (Appendix D). We observe consistent performance degradation across all reasoning models, indicating that overly formalized representations disrupt the semantic continuity that LLMs naturally exploit. This confirms that CapGeo's improvements arise from accurate visual content extraction rather than from exploiting any particular template design.

Together, these results robustly confirm our hypothesis: **visual perception, rather than reasoning, is the primary bottleneck of multimodal geometric reasoning.**

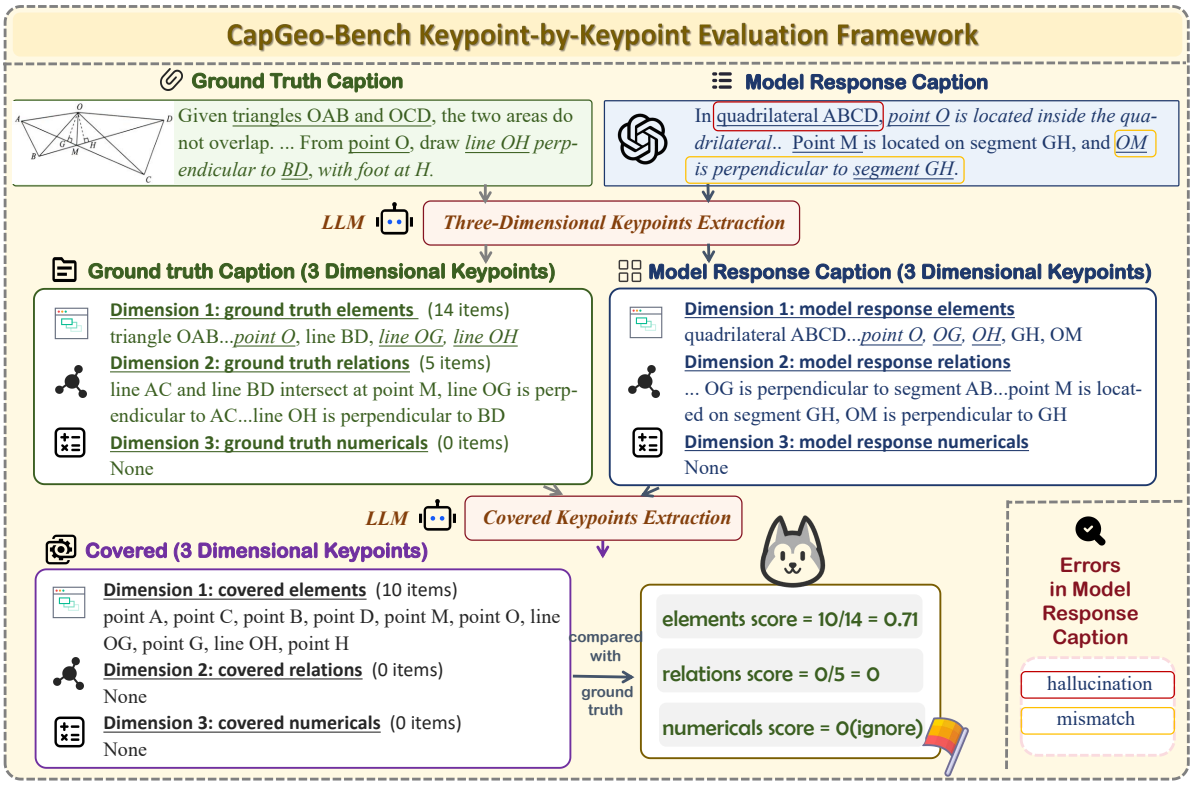

Figure 4: Overview of CapGeo-Bench Evaluation. Covered items in keypoints are marked by italics and underlining.

4 CapGeo-Bench: Grounding Geometric Captioning Ability

The exploratory experiment confirms that geometric perception, manifested concretely as captioning quality, is the dominant factor governing downstream reasoning accuracy. This finding motivates a natural follow-up question - how can we systematically measure the geometric captioning—i.e., perception—capability of MLLMs? Existing image-captioning benchmarks are inadequate for geometry, which demands faithful encoding of elements, spatial relations, and numerical constraints. To fill this gap, we introduce **CapGeo-Bench**.

4.1 Benchmark Construction

Overview. To bridge the gaps in evaluating the visual geometric information extraction capabilities of MLLMs, we designs CapGeo-Bench. As shown in Figure 3 and Table 1, CapGeo-Bench is a benchmark comprising high-quality 4641 geometric images with corresponding bilingual (Chinese and English) captions, covering 3 geometric classes and 4 difficulty levels. The entire benchmark design process is shown in Figure 2.

Data Collection. To minimize potential contamination from open-source data, we deliberately avoided using publicly available datasets. Instead, we curated a collection of geometric problems from K-12 textbooks in China, encompassing classic topics such as triangles, parallelograms, and circles, with difficulty levels ranging from middle school exercises to competition problems. The entire collection process was conducted in strict compliance with copyright and licensing regulations. We then employed the open-source detection tool DocLayout-YOLO (Zhao et al., 2024) to automatically extract geometric diagrams and manually filtered out non-geometric figures. In total, we obtained 5,000 high-quality geometric images.

Caption Annotation. To guarantee the accuracy and quality of captions, we recruited annotators with STEM backgrounds from Peking University, one of the top universities in China, to perform meticulous manual annotation. All annotators underwent rigorous assessments of their mathematical and geometric proficiency, followed by multiple rounds of training and pre-annotation exercises to ensure compliance with the required standards. During annotation, mathematics experts designed a unified annotation instruction, including geometric image re-filtering rules and geometric caption annotation principles. The complete procedures of the whole workflow, together with detailed instructions, are provided in Appendix F.

A review team composed of three professional

Models	MLLMs					LLMs		
	GPT-o3	Gemini-2.5	Claude-4	Qwen-VL-72B	Qwen-VL-7B	DS-R1	Qwen-72B	Qwen-7B
Vision Only								
Img	74.6	66.4	44.8	8.6	7.1	–	–	–
Cap	60.3	60.3	59.4	66.1	**61.9**	**68.2**	**66.1**	**63.3**
Cap+Img	**75.8**	**73.6**	**73.0**	**66.4**	60.0	–	–	–
Vision Intensive								
Img	73.0	79.6	61.8	58.0	43.3	–	–	–
Cap	**82.5**	83.0	79.3	**74.9**	**63.2**	**71.9**	**73.2**	**65.9**
Cap+Img	78.3	**85.4**	**85.5**	72.6	60.3	–	–	–

Table 2: Performance on **MathVerse** with GPT-o3 as captioner. Full results in Appendix B.

Model	Score Dimension	Class			Difficulty				Overall	Avg
		AG	PG	SG	T1	T2	T3	T4		
GPT-o3	Element	69.0	55.1	49.5	57.4	54.8	34.6	32.1	56.0	38.4
	Relation	41.0	38.9	26.8	41.1	35.7	20.7	20.3	38.9	
	Numerical	7.8	20.8	52.8	24.1	14.2	13.3	0.0	20.4	
GPT-4o	Element	64.0	53.2	49.4	54.6	54.6	51.6	50.8	54.2	30.4
	Relation	30.9	26.6	15.5	32.3	24.3	18.2	17.8	26.8	
	Numerical	19.2	9.3	24.5	12.2	10.1	3.5	3.6	10.2	
Qwen2.5-VL -72B-Instruct	Element	64.6	53.0	48.2	54.7	54.6	50.2	50.4	54.1	30.8
	Relation	33.2	28.5	20.0	33.4	26.9	20.8	18.4	28.9	
	Numerical	20.1	8.3	21.4	12.4	8.1	5.6	0.0	9.4	
Qwen2.5-VL -7B-Instruct	Element	58.1	44.6	31.9	47.0	46.3	45.8	44.3	46.5	24.2
	Relation	26.6	17.5	2.2	23.4	16.1	11.0	11.2	18.7	
	Numerical	17.7	5.9	42.9	8.7	7.5	2.6	0.0	7.4	

Table 3: Captioning Model Performance on **CapGeo-Bench** grouped by Class and Difficulty.

reviewers conducts strict cross-checks on all data annotation results. For annotations that fail the review, they will be re-annotated and re-evaluated. See Appendix F for details. We also adhere to labour ethics during the whole annotation process, ensuring all annotators were well-paid, with each receiving 10 dollars per geometric figure annotation. In the end, we obtained 4,641 high-quality, contamination-free geometric image–caption pairs.

4.2 Keypoint-by-Keypoint Geometric Caption Quality Evaluation

Although there are some metrics for evaluating image captions quality (Vedantam et al., 2015; Anderson et al., 2016), due to the peculiarities of geometric captions, there is still no scientific, quantitative and fine-grained metric for assessing geometric captions quality generated by MLLMs.

Geometric figures are composed of basic geometric elements through clear relationships. This structural feature determines that geometric captions can be systematically deconstructed into keypoints set. This cognition is highly consistent with the core idea of Geometric Formal Language (Lu et al., 2021). Based on this, we design a keypoint-by-keypoint geometric caption evaluation method with three dimensional evaluation metrics and corresponding evaluation framework.

Evaluation Metrics. The abstract combination of geometric points, lines, and letters conveys rich elemental, spatial, and numerical information. Thus, for MLLMs' captioning capability, we focus on the following three dimensional metrics:
(1) Elements: evaluate whether MLLMs can comprehensively identify all elements and their corresponding letter identifiers. The elements include basic shapes, lines, points, coordinate axes, etc. In Figure 4, Elements in Captions are underlined.
(2) Spatial Relations: evaluate whether MLLMs can comprehensively identify all spatial relations and their corresponding subjects. For specific details, See in Appendix G. In Figure 4, Spatial Relations in Captions are in italics.
(3) Numerical Relations: evaluate whether the MLLMs can comprehensively identify all the numerical relationships and their corresponding sub-

jects. Like length, angle, etc.

Reasoning Model	Text Only	Cap Only (GPT-o3)	
		V-Only	V-Intensive
Qwen2.5-VL-7B	51.3	61.9	63.2
Qwen2.5-VL-72B	62.2	66.1	74.9

Table 4: Caption-only vs. Text-only on MathVerse.

Evaluation Framework. As shown in Figure 4, we decompose the evaluation framework into three major steps: (1) Extract geometric elements keypoints set E, spatial relation keypoints set R, numerical relation keypoints set N from both model response captions T_c and ground truth T_g by LLM. This can be expressed by the formula as

$$E,\ R,\ N = \text{LLM}_{\text{extract}}(T)$$

See Figure 8 and Figure 9 for prompts for ground truth and model response caption three dimensional keypoints extraction.

(2) Identify covered items between ground truth keypoints and model response caption keypoints. Due to the varied expressions in natural language, performing a direct intersection operation is rough and insufficient. Therefore, we design prompt and employ LLM to perform semantically equivalent covered items extraction between the model response caption and ground truth keypoints set for each dimension. This process identifies semantically equivalent keypoints, which constitute the TP set for that dimension. This can be expressed by the formula as

$$TP_E = E_c \cap E_g = \text{LLM}_{\text{match}}(E_c, E_g)$$

See Figure 10 for prompt for covered three dimensional keypoints extraction of LLM_{match}.
(3) The score S for each dimension is calculated as macro F1. For a dimension $k \in \{E, R, N\}$, we first calculate Precision (P_k) and Recall (R_k):

$$P_k = \frac{|TP_k|}{|K_c|}, \quad R_k = \frac{|TP_k|}{|K_g|}$$

The final score S is expressed by the formula as:

$$S_k = \frac{2P_kR_k}{P_k + R_k}$$

We submitted the complete evaluation process of 200 geometric captions to three domain experts for independent review. The experts confirmed the validity and reliability of the evaluation processes, which verified the rationality of our method.

Reasoning Model	Natural	Structured
MLLMs		
Qwen2.5-VL-7B-Instruct	**32.4**	31.6
Qwen2.5-VL-72B-Instruct	**39.2**	36.0
LLMs		
Qwen2.5-7B-Instruct	**32.6**	31.9
Qwen2.5-72B-Instruct	**38.2**	35.5

Table 5: Ablation on caption style.

5 Experiments

To further evaluate the captioning capabilities of different MLLMs, we perform a systematic evaluation on CapGeo-Bench. We also conduct a joint analysis linking CapGeo-Bench scores with downstream reasoning performance in CapGeo.

5.1 Experimental Setup

Models In CapGeo-Bench experiments, we selected GPT-4o, GPT-o3, Qwen2.5-VL-72B-Instruct and Qwen2.5-VL-7B-Instruct as evaluation models and Gemini-2.5-pro as the as the model for extracting keypoints and matching in section 4.2.

Settings For local model inference, we employed the VLLM framework (Kwon et al., 2023). The model temperature was fixed at 0.0 to minimize randomness and ensure stable reasoning performance.

5.2 Experimental Analysis

Geometric captioning remains a significant challenge. Table 3 reports MLLMs' scores across three keypoint dimensions and four difficulty levels. GPT-o3 attains the best overall average (38.4%), while Qwen2.5-VL-72B-Instruct performs on par with GPT-4o. Several notable patterns emerge: (1) All models perform dramatically worse on the Numerical dimension, often producing mismatches between numerical values and their corresponding geometric elements—even GPT-o3 achieves only 20.4% on this dimension. We think that spatial relations rely on explicit physical connections, but numerical values are "floating" text near geometric components without explicit pointers, which make it difficult for MLLM. (2) Performance degrades sharply with increasing difficulty across all three dimensions . (3) Plane Geometry remains especially challenging despite being the most common category, underscoring the limited ability of current MLLMs to extract information from highly abstract and symbolic visual content.

These results reveal a substantial gap between

MLLM capabilities and the reliable geometric perception, establishing CapGeo-Bench as a challenging and discriminative benchmark.

Correlation - bridging CapGeo-Bench and downstream reasoning. To verify that CapGeo-Bench faithfully reflects perception capability, We conduct a correlation analysis between Three-dimensional average score of the Captioning Model and the performance of the Reasoning Models it assisted in Table 6. As shown in Figure 5, the two are strongly positively correlated—that is, captioners scoring higher on CapGeo-Bench consistently produce captions that lead to better downstream reasoning performance. This correlation (i) reconfirms our hypothesis from a different angle—better perception yields better reasoning—and (ii) Confirms CapGeo-Bench as a reliable evluation tool for geometric understanding.

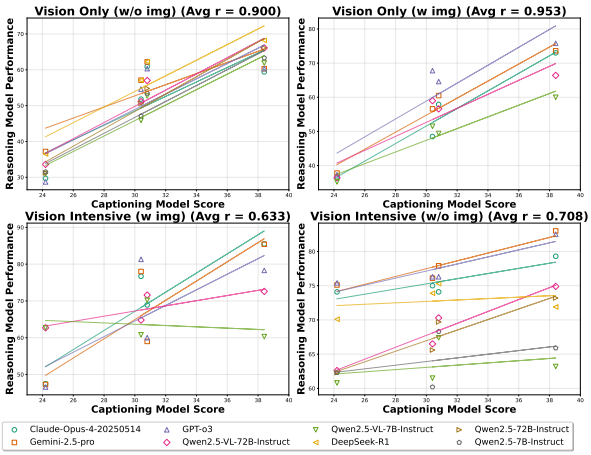

Figure 5: Correlation Analysis. w/o img means Cap Only, w img means Cap+Img, r means relative coefficient

Modality Conflict and Toxic Captions. Although high-quality captioner improve the performance, the observed abnormal drops need analysis.

Modality Conflict. As shown in 6, in Vision Intensive, when using low-quality Captioners, the performance of Cap+Img of the strong MLLMs is lower than Cap Only. For example, for GPT-o3 assisted by captioner Qwen2.5-VL-7B-Instruction, after adding the original image, the accuracy rate drops sharply from 75.4% to 46.6%; similarly, on MathVista, Claude-Opus's performance unexpectedly drops from 73.1% to 65.7% when assisted by GPT-4o. We think that when there is a contradiction between the caption and the image, MLLMs with stronger visual perception capabilities will suffer from modality conflict and experience a decline in performance. On the contrary, MLLMs with weaker visual capabilities are insensitive to this conflict, and their performance changes little.

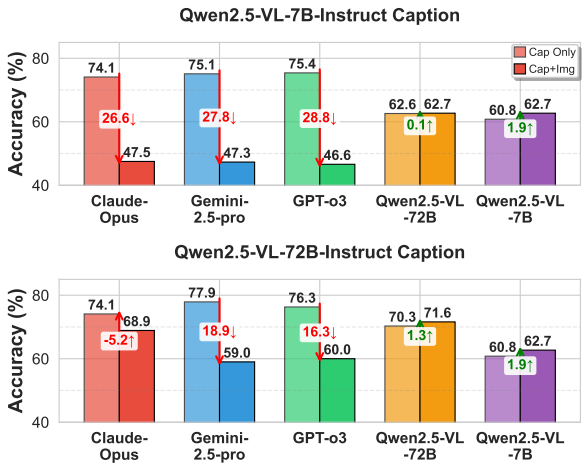

Figure 6: Modality Conflict Analysis

Toxic Captions. Under the Cap only setting, as revealed in our Case Study H, some "toxic" captions can mislead the reasoning model. For example, when using captions generated by Qwen2.5-VL-7B-Instruct in the Vision-Only tasks, some MLLMs exhibit performance degradation.

In short, by introducing CapGeo-Bench, we build a challenging benchmark that establishes a new standard for the geometric captioning community, encouraging future research to address these limitations. Beyond serving as an evaluation benchmark, CapGeo-Bench's fine-grained scoring naturally lends itself as a reward signal source for RL —an exciting direction for future work.

6 Conclusion

In this paper, we investigated the root cause of the persistent weakness of MLLMs in multimodal geometric reasoning. Using CapGeo as a controlled probe, we empirically verified that the primary bottleneck lies in visual perception rather than reasoning. Building on this finding, we proposed CapGeo-Bench, a high-quality benchmark of 4,641 geometric figure–caption pairs equipped with a fine-grained keypoint-based evaluation metric, designed to systematically assess MLLMs' geometric information extraction capability. Scores on CapGeo-Bench show a strong positive correlation with downstream reasoning performance, establishing it as a powerful tool for evaluating geometric perception capability. Together, CapGeo and CapGeo-Bench pave a new path for future advances in multimodal geometric perception and reasoning.

7 Limitations

Due to the influence and limitations of objective factors and actual conditions, despite our best efforts to avoid them, there are still some limitations in this study. Firstly, the limited time and computing resources prevent our experimental framework CapGeo from conducting a comprehensive evaluation on all SOTA models and datasets. Secondly, although CapGeo-Bench has taken strict measures in terms of quality control, such as the screening and training of annotators and the final review system, the inherent inaccuracy in the manual annotation process may still lead to minor mistakes and errors in the dataset.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Anthropic. 2025. Transparency hub: Model report. https://www.anthropic.com/transparency/model-report.

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

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

Guizhen Chen, Weiwen Xu, Hao Zhang, Hou Pong Chan, Deli Zhao, Anh Tuan Luu, and Yu Rong. 2025a. Geopqa: Bridging the visual perception gap in mllms for geometric reasoning. *ArXiv, abs/2509.17437, 2025a. URL https://arxiv.org/abs/2509.17437*.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*.

Yan Chen, Long Li, Teng Xi, Long Zeng, and Jingdong Wang. 2025b. Perception before reasoning: Two-stage reinforcement learning for visual reasoning in vision-language models. *arXiv preprint arXiv:2509.13031*.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. 2024. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. 2024. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al. 2024. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9096–9105.

Gemini Team, Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google DeepMind.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Haibin He, Maoyuan Ye, Jing Zhang, Xiantao Cai, Juhua Liu, Bo Du, and Dacheng Tao. 2025. Reasoning-ocr: Can large multimodal models solve complex logical reasoning problems from ocr cues? *arXiv preprint arXiv:2505.12766*.

Mingxin Huang, Yongxin Shi, Dezhi Peng, Songxuan Lai, Zecheng Xie, and Lianwen Jin. 2025. Ocr-reasoning benchmark: Unveiling the true capabilities of mllms in complex text-rich image reasoning. *arXiv preprint arXiv:2505.17163*.

Yichen Huang and Lin F Yang. 2025. Gemini 2.5 pro capable of winning gold at imo 2025. *arXiv preprint arXiv:2507.15855*.

Mengzhao Jia, Zhihan Zhang, Wenhao Yu, Fangkai Jiao, and Meng Jiang. 2024a. Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training. *arXiv preprint arXiv:2404.14604*.

Zixi Jia, Jiqiang Liu, Hexiao Li, Qinghua Liu, and Hongbin Gao. 2024b. Dcot: Dual chain-of-thought prompting for large multimodal models. In *The 16th Asian Conference on Machine Learning (Conference Track)*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Saehyung Lee, Seunghyun Yoon, Trung Bui, Jing Shi, and Sungroh Yoon. 2024. Toward robust hyper-detailed image captioning: A multiagent approach and dual evaluation metrics for factuality and coverage. *arXiv preprint arXiv:2412.15484*.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. 2025. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

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

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

OpenAI. 2025. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2025-09-28.

Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. 2024. Multimath: Bridging visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*.

Jiahao Qiu, Jingzhe Shi, Xinzhe Juan, Zelin Zhao, Jiayi Geng, Shilong Liu, Hongru Wang, Sanfeng Wu, and Mengdi Wang. 2025. Physics supernova: Ai agent matches elite gold medalists at ipho 2025. *arXiv preprint arXiv:2509.01659*.

Sara Sarto, Marcella Cornia, and Rita Cucchiara. 2025. Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. *arXiv preprint arXiv:2503.14604*.

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

Kai Sun, Yushi Bai, Zhen Yang, Jiajie Zhang, Ji Qi, Lei Hou, and Juanzi Li. 2025. Hard negative contrastive learning for fine-grained geometric understanding in large multimodal models. *arXiv preprint arXiv:2505.20152*.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

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

Lai Wei, Wenkai Wang, Xiaoyu Shen, Yu Xie, Zhihao Fan, Xiaojin Zhang, Zhongyu Wei, and Wei Chen. 2024b. Mc-cot: A modular collaborative cot framework for zero-shot medical-vqa with llm and mllm integration. *arXiv preprint arXiv:2410.04521*.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. 2025. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. 2025. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. *arXiv preprint arXiv:2506.07218*.

Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. 2025. Caprl: Stimulating dense image caption capabilities via reinforcement learning. *arXiv preprint arXiv:2509.22647*.

Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. 2021. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8751–8761.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. 2024a. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024b. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, et al. 2024a. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. *arXiv preprint arXiv:2410.02884*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. 2024b. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*.

Xian Zhang, Haokun Wen, Jianlong Wu, Pengda Qin, Hui Xue', and Liqiang Nie. 2024c. Differential-perceptive and retrieval-augmented mllm for change captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4148–4157.

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*.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. 2024. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *arXiv preprint arXiv:2410.12628*.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. 2024. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, pages 73–90. Springer.

A Prompt Appendix

In this section, we have detailedly listed all the prompts used in the experiments of the articles. In CapGeo, prompts related to captioning and reasoning are involved in Figure 8. In CapGeo-Bench, the prompt related to captioning is involved in Figure 8, the prompt for keyword extraction during the evaluation process is involved in Figure 9, and the prompt for covered keypoints extraction is involved in Figure 10. All prompts have been carefully and scientifically designed and iterated through experiments, and have been guided and verified by domain experts, resulting in excellent final effects.

B Full CapGeo Results with Additional Captioners

For brevity, the main paper (Table 2) reports CapGeo results using GPT-o3 as the captioner, since it produces the strongest captions and thus most clearly exposes the perception bottleneck. Here we provide the full results with three additional captioners (GPT-4o, Qwen2.5-VL-72B-Instruct, Qwen2.5-VL-7B-Instruct) for completeness in Table 6. The trends across all captioners are consistent with the conclusions in the main text: stronger captioners yield larger reasoning gains, while weaker captioners can introduce modality conflict and toxic captions.

C Full Results on MathVista and GeoQA

For brevity, the main paper focuses on MathVerse to verify the perception bottleneck. We additionally report CapGeo results on MathVista and GeoQA in Table 7. Trends are consistent with the conclusions in the main text: stronger captioners yield substantial reasoning gains, and the gap between Image-only and Caption-assisted settings reconfirms the perception bottleneck.

D Ablation Study on Caption Style

A natural concern is whether the gains of CapGeo stem from a specific captioning template rather than from genuine visual information extraction. To address this, we ablate two distinct caption styles while keeping all other factors fixed:

- **Natural Problem Reformulation** (used in our main experiments): the captioner is simply asked to "describe the complete geometric math problem in a clear and structured format", without enforcing any rigid schema. The output naturally takes the form of a standard geometry problem statement (see Figure 7).
- **Explicit Structured Decomposition**: the captioner is forced to output a symbolic JSON schema that explicitly enumerates three fields—*Geometric Elements*, *Element Relationships*, and *Numerical Information*.

We re-run CapGeo with both caption styles and report results in Table 5.

We observe a consistent degradation when using the highly structured JSON captions across all four reasoning models. We attribute this to two factors: (i) overly formalized JSON representations disrupt the semantic continuity that (M)LLMs naturally exploit during step-by-step geometric reasoning; and (ii) rigid schemas can fragment relational information that is more naturally expressed in mathematical prose. This ablation provides two important takeaways: (1) the gains observed in CapGeo arise from the *accurate extraction of visual information*, not from any specific captioning template; (2) given the inherently formal nature of geometry, asking the captioner for a natural mathematical-style description is itself a mild and well-motivated requirement, not a hand-crafted trick.

E Difficulty Classification

During the annotation process, the classification of difficulty levels was determined by **the complexity of geometric diagrams**, which aligns closely with the core motivation of CapGeo-Bench: to evaluate the capability of extracting geometric information.

Specifically, after reviewing all pre-screened geometric diagrams, we categorized them into four difficulty levels—**Simple (T1), Moderate (T2), Complex (T3), and Challenging (T4)**—based on the intricacy of their elements and structural composition. Representative diagrams for each difficulty level were provided to annotators to facilitate consistent and accurate labeling.

F Annotation Instructions

To ensure the reliability and accuracy of the annotated data, we adopted a rigorous selection and training process for annotators. Specifically, we recruited candidates from top universities and ranked them based on their high school mathematics performance and scores on geometry reasoning tests,

Reasoning Model \ Captioning Model	w/o	GPT-o3		GPT-4o		Qwen2.5-VL -72B-Instruct		Qwen2.5-VL -7B-Instruct	
	Img	Cap	Cap+Img	Cap	Cap+Img	Cap	Cap+Img	Cap	Cap+Img
Vision Only									
GPT-o3	74.6	60.3	**75.8**	54.6	67.8	60.3	64.6	28.7	37.3
Gemini-2.5-pro	66.4	60.3	**73.6**	57.1	56.6	62.2	60.5	37.2	37.8
Claude-Opus-4-20250514	44.8	59.4	**73.0**	51.8	48.5	60.9	57.9	29.7	36.7
Qwen2.5-VL-72B-Instruct	8.6	66.1	**66.4**	51.1	59.0	57.0	56.6	33.6	36.4
Qwen2.5-VL-7B-Instruct	7.1	**61.9**	60.0	45.9	51.5	52.7	49.4	31.0	35.2
DeepSeek-R1	–	**68.2**	–	57.1	–	62.3	–	36.5	–
Qwen2.5-72B-Instruct	–	**66.1**	–	50.6	–	54.8	–	31.2	–
Qwen2.5-7B-Instruct	–	**63.3**	–	47.1	–	53.4	–	31.5	–
Vision Intensive									
GPT-o3	73.0	**82.5**	78.3	76.3	81.3	76.3	60.0	75.4	46.6
Gemini-2.5-pro	79.6	83.0	**85.4**	76.1	78.0	77.9	59.0	75.1	47.3
Claude-Opus-4-20250514	61.8	79.3	**85.5**	75.0	76.7	74.1	68.9	74.1	47.5
Qwen2.5-VL-72B-Instruct	58.0	**74.9**	72.6	66.5	64.8	70.3	71.6	62.6	62.7
Qwen2.5-VL-7B-Instruct	43.3	63.2	60.3	61.5	60.8	**67.4**	70.2	60.8	62.7
DeepSeek-R1	–	71.9	–	73.9	–	**75.3**	–	70.1	–
Qwen2.5-72B-Instruct	–	**73.2**	–	65.6	–	69.7	–	62.3	–
Qwen2.5-7B-Instruct	–	65.9	–	60.2	–	**68.3**	–	62.3	–

Table 6: Full CapGeo results on MathVerse with all four captioners across three input modalities.

Model	MathVista					GeoQA				
	-	GPT-o3	GPT-4o	Qwen-72B	Qwen-7B	-	GPT-o3	GPT-4o	Qwen-72B	Qwen-7B
Claude-Opus-4-20250514	76.1	89.4	65.7	**91.7**	88.4	57.0	92.0	91.0	**92.5**	91.5
Gemini-2.5-pro-preview-05-06	94.9	96.3	95.8	**98.1**	97.7	90.0	95.0	96.5	97.0	**97.5**
GPT-4o	61.6	**65.7**	62.0	63.9	62.5	61.0	**63.0**	**63.0**	62.5	57.5
GPT-o3	93.1	94.9	94.0	94.4	**95.8**	92.5	94.5	94.5	93.5	**95.0**
Qwen2.5-VL-72B-Instruct	75.9	**92.6**	91.7	92.1	70.4	68.5	**73.5**	72.0	70.0	**73.5**

Table 7: Comparison of MLLMs on **MathVista** and **GeoQA** benchmarks with different captioning models.

ensuring sufficient mathematical reasoning and geometric spatial understanding capabilities.

Before undertaking the actual annotation tasks, candidates went through two rounds of training and pre-annotation. In the first stage, we provided comprehensive instruction covering the annotation objectives, guidelines, workflow, and quality standards, illustrated with detailed examples. Candidates were then required to perform pre-annotations on 10 geometric figures, which were reviewed by expert examiners. The examiners identified both individual and common issues, which were subsequently addressed in a second training session. Afterward, candidates completed another round of pre-annotation and review. Only those who successfully passed both evaluations were qualified as final annotators.

During the annotation process, annotators were required to follow a rigorous **Workflow**.

- First, they assessed the quality and content of the image to determine whether it was a valid geometric figure; images that did not meet the criteria were discarded.
- Next, annotators evaluated the type and difficulty level of the geometric figure.
- Finally, annotators carefully examined the construction logic, spatial relations, and numerical information of the figure, and produced a logically structured caption that covered all information in the image without introducing subjective hallucinations.
- For highly complex figures or when necessary, annotators were instructed to consult the original descriptions in the source PDF to en-

sure the accuracy and reliability of the final annotations.

Our **Annotation instructions** are as follows:

- **Strict adherence to expert guidelines**: Annotators were required to strictly adhere to unified guidelines designed by mathematics experts.
- **Quality control of visual materials**: Redundant, undescribable, or low-quality images were to be discarded.
- **Visual grounding of captions**: All captions had to be strictly grounded in the visual content of the figure, without introducing subjective interpretation or speculation.
- **Consistent geometric terminology**: Geometric terminology was required to remain consistent, ensuring standardization and reproducibility across annotations.
- **Beyond surface-level description**: Captions were expected to go beyond surface-level elements and capture essential geometric attributes or spatial relations (e.g., for a midpoint, not merely stating that a point lies on a line, but explicitly indicating its geometric significance).
- **Hierarchical structure**: Annotations were required to follow a hierarchical structure, prioritizing information according to geometric construction logic rather than presenting items indiscriminately.
- **Fine-grained detail annotation**: Fine-grained details such as dashed lines, colored segments, and area representations were also mandated to be annotated, ensuring comprehensive and precise descriptions.

Annotation Principles and Ground Truth Definition. In our study, the geometric elements constituting the Ground Truth are strictly derived from their presence in the caption descriptions. To ensure objectivity, the annotation process precludes subjective interpretation of the diagrams. Instead, annotators are instructed to adhere to the **logical construction** of the geometric diagrams as described in the original source textbooks. This ensures that all textual descriptions are firmly grounded in the visual information provided by the images.

Selection Criteria for Geometric Elements. Based on the construction logic, we categorize labeled elements into two types: (1) **Constructive Elements**, which are directly involved in the geometric construction, and (2) **Derivable Elements**, which can be inferred from existing logic. **Only the former are required to be labeled.** The specific protocols are as follows:

- **Explicitly Marked Elements:** Any element that is explicitly marked in the diagram or plays a direct role in the geometric construction sequence must be explicitly mentioned in the caption.
- **Implicit but Essential Elements:** If an unmarked point is essential to the subsequent construction logic (e.g., the intersection point of extended segments AB and CD, used as a reference for further construction), it must be described in the caption (e.g., denoted as "the intersection point of AB and CD").
- **Incidental Intersections:** Points that result merely from the intersection of existing segments without participating in the logic of geometric construction, and are not explicitly marked, are excluded from annotation.

Protocol for Spatial Relationships Given the interdependence of geometric relationships, our annotation principle strictly follows the original textbook's logical framework rather than providing an exhaustive enumeration of all deducible information:

- **Explicit Definition:** If the textbook explicitly states a relationship (e.g., "A is the midpoint of BC") and utilizes this for construction, it is included in the caption.
- **Logical Deduction:** If the textbook describes a condition that implies a relationship but does not explicitly state it (e.g., "AD is the median of BC with intersection point A"), we annotate the median relationship but do not require labeling A as the midpoint, even though it is logically deducible.

Examination. To further ensure the quality and consistency of the annotations, we implemented a rigorous cross-review mechanism. All annotation

results were independently examined by a review panel consisting of three expert evaluators. For samples that failed the initial review, re-annotations were conducted by different members of the annotation team, and the revised results were resubmitted for review. This iterative process of cross-annotation and review effectively ensured both accuracy and consistency of the final annotations.

We adhere to labour ethics during the whole annotation process, ensuring all annotators were well-paid, with each receiving 10 dollars per geometric figure annotation.

G Spatial Relations

To comprehensively extract the spatial relations in the caption, we have classified the following spatial relationship classification system based on the composition of the relationship subjects. We believe that the following contents all fall within the scope of spatial relations. This system encompasses five core types of relationships:

(1) The Point-Line Relations describe the positional dependence between points and lines: Midpoint, Foot of the perpendicular, Intersection point, Trisection point, Endpoint;

(2) The Line-Line Relations depict the relationship between line elements: Perpendicular, Parallel, Oblique intersection, Coincidence;

(3) The Line-Shape Relations characterize the structural connection between a straight line and complex geometric shapes: Angle bisector, Diagonal, Median, Altitude, Chord, Tangent, Diameter;

(4) The Point-Shape Relations define the characteristic position of a point relative to a complex figure: Centroid, Orthocenter, Circumcenter, Incenter, Vertex, Circle center;

(5) The Shape-Shape Relations describe the macroscopic spatial configuration among complete geometric figures: Disjoint, Tangency, Intersection, Containment, Congruence, Similarity, Concentric, Inscribed, Circumscribed.

H Case study

The following case is derived from image_73.png in the MathVerse dataset:

toxic caption case

Correct Example (GPT-o3)
Caption ($\surd$):
Points A, B, C lie on circle with center O. $\angle ABC = 70°$. Find central angle $\angle AOC$.
Options:
A) 140° B) 130° C) 120° D) 110°
Reasoning ($\surd$):
Central angle = 2× inscribed angle
Calculation:
$\angle AOC = 2 \times 70° = 140°$
Answer: A) 140°

Error Case (GPT-4o)
Caption ($\times$):
Incorrect: "$\angle BOC = 70°$" (should be $\angle ABC = 70°$)
Misled Reasoning ($\times$):
Based on wrong caption, calculates: $180° + 70° + \angle AOC = 360° \implies \angle AOC = 110°$
Answer:D) 110°

CapGeo and CapGeo-Bench Captioning prompt

MathVerse Vision Intensive:
Based on the provided geometric figure and the following question text, please describe the complete geometric math problem in a clear and structured format.

Question text: {question_text}

Please combine the visual information from the figure with the question text to provide a comprehensive description of the geometric math problem.

MathVerse Vision Only:
The geometric figure and the text problem is provided above, please directly describe the problem with geometric figure in a geometric math problem format.

Only output the geometric math problem and make sure the math problem is complete.

GeoQA and MathVista：
Based on the provided geometric diagram and the following question text, please describe the complete geometric problem in a clear and structured format.

Question text: {question_text}
Do not provide analysis or solution steps, only the problem statement itself.

CapGeo-Bench：
Based on the provided geometric figure, please describe the geometric figure in a clear and structured mathematical format. Mine as much visual information as possible from geometric figure.

Only output the geometric figure description.
The following are output examples.

Example: In quadrilateral PRSU, side PR is parallel to SU. Connect diagonals PU and RS, which intersect at point D. Point T lies on SU, and lines PT and PU are trisectors of angle SPR.Choose a point E on PT and draw auxiliary line DE such that DE is parallel to PR (and SU). Select points F and G on segments PD and SR respectively. Connect EF and EG so that EF is parallel to DG and EG is parallel to DF.Connect SE and extend it to intersect PR at point Q. Connect FG and extend it to intersect PR and SU at points K and L respectively.

CapGeo Reasoning prompt

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer>

Figure 7: CapGeo and CapGeo-Bench captioning prompt for Captioner Model and reasoning prompt for Reasoning Models during the experiments

Ground Truth 3 dimensional keypoints extraction prompt

You are a strict expert in geometric description evaluation. Please strictly extract the corresponding elements from the geometric description in accordance with the following requirements.

Evaluation Requirements:

1. All extractions must be strictly based on the given text - no fabrication or addition of information
2. Only extract the content explicitly mentioned in the original geometric description

Evaluation Steps:

1. From original problem description extract:

- ground_truth_elements: Explicitly mentioned geometric elements (like points A, lines AB, triangle ABC etc.)
- ground_truth_relations: Explicitly mentioned geometric relations (like line AB is parallel to line CD, line AB is perpendicular to line CD, line AB intersects line CD at point O, line AB is tangent to circle O at point C etc.)
- ground_truth_numericals: Explicitly mentioned numerical values (like AB = 7, angle ABC = 60°, AB = 0.5 * BC etc.)

Required Response Format (strict JSON):

```
{{
  "ground_truth_elements": [],
  "ground_truth_relations": [],
  "ground_truth_numericals": []
}}
```

===== original geometric description (ground_truth) =====

{ground_truth}

Please note that the returned output must be in valid and correct json format. Pay attention to the commas between group elements and key-value pairs, as well as the colons between key-values

Figure 8: Ground Truth 3 dimensional keypoints set extraction prompt for LLM during CapGeo-Bench evaluation

Model response caption 3 dimensional keypoints extraction prompt

You are a strict expert in geometric description evaluation. Please strictly extract the corresponding elements from the geometric description in accordance with the following requirements.

Evaluation Requirements:

1. All extractions must be strictly based on the given text - no fabrication or addition of information
2. Only extract the content explicitly mentioned in the original generated geometric description

Evaluation Steps:

1. From generated description extract:

- ocr_elements: Explicitly mentioned geometric elements (like points A, lines AB, triangle ABC etc.)
- ocr_relations: Explicitly mentioned geometric relations (like line AB is parallel to line CD, line AB is perpendicular to line CD, line AB intersects line CD at point O, line AB is tangent to circle O at point C etc.)
- ocr_numericals: Explicitly mentioned numerical (like AB = 7, angle ABC = 60°, AB = 0.5 * BC etc.)

2. Return required Response Format (strict JSON):

```
{{
   "ocr_elements": [],
   "ocr_relations": [],
   "ocr_numericals": []
}}
```

===== Generated Geometry Description =====

{generated_description}

Please note that the returned output must be in valid and correct json format. Pay attention to the commas between group elements and key-value pairs, as well as the colons between key-values

Figure 9: Model Response Caption 3 dimensional keypoints set extraction prompt for LLM during CapGeo-Bench evaluation

Covered 3 dimensional keypoints extraction prompt

You are a strict expert in geometric description evaluation. Please search for overlapping values in the original geometric description list and the generated geometric description list as per the following requirements

Evaluation Requirements:

1. Overlapping items must be strictly determined based on the original geometric description list and the generated geometric description list with no fabrication or addition of information
2. Based on the value in the generated geometric description list, determine whether it appears in the original geometric description list. If it appears, it is judged as overlapping.

Evaluation Steps:

1. Find overlaps:
- covered_elements: Intersection of ground_truth_elements and ocr_elements
- covered_relations: Intersection of ground_truth_relations and ocr_relations
- covered_numericals: Intersection of ground_truth_numericals and ocr_numericals
2. Return Required Response Format (strict JSON):

```
{{
  "covered_elements": [],
  "covered_relations": [],
  "covered_numericals": []
}}
```

===== original geometric description list (JSON) =====

{ground_truth}

===== generated geometric description list (JSON) =====

{generated_description}

Critical Reminders for Overlaps:

1. All angles must match both value and vertex points (e.g., "angle ABC=45°" and "angle ABD=45°" do not overlap)
2. Lengths must match both value and segments (e.g., "AB=5cm" and "AC=5cm" do not overlap)
3. Semantically identical items can be considered matches (e.g., "triangle ABC", "triangle BAC" and "triangle CAB" are considered to overlap, "line AB is parallel to line CD" and "AB parallel to CD" are considered to overlap, "red triangle ACE" and "triangle ACE" are considered to overlap)
4. Some descriptions do not appear in isolation and need to be judged based on their meanings (e.g."points D, M, and N" and ["point D", "point M", "point N"] are considered to overlap)

Please note that the returned output must be in valid and correct json format. Pay attention to the commas between group elements and key-value pairs, as well as the colons between key-values

Figure 10: Covered 3 dimensional keypoints extraction prompt for LLM during CapGeo-Bench evaluation