R-CoT: Reverse Chain-of-Thought Problem Generation for Geometric Reasoning in Large Multimodal Models

Anonymous authors
Paper under double-blind review

Abstract

Existing Large Multimodal Models (LMMs) struggle with mathematical geometric reasoning due to a lack of high-quality image-text paired data. Current geometric data generation approaches, which apply preset templates to generate geometric data or use Large Language Models (LLMs) to rephrase questions and answers (Q&A), unavoidably limit data accuracy and diversity. To synthesize higher-quality data, we propose a two-stage Reverse Chain-of-Thought (R-CoT) geometry problem generation pipeline. First, we introduce GeoChain to produce high-fidelity geometric images and corresponding descriptions highlighting relations among geometric elements. We then design a Reverse A&Q method that reasons step-by-step based on the descriptions and generates questions in reverse from the reasoning results. Experiments demonstrate that the proposed method brings significant and consistent improvements on multiple LMM baselines, achieving new performance records in the 2B, 7B, and 8B settings. Notably, R-CoT-8B significantly outperforms previous state-of-the-art open-source mathematical models by 16.6% on MathVista and 9.2% on GeoQA, while also surpassing the closed-source model GPT-4o by an average of 13% across both datasets.

1 Introduction

Large Language Models (LLMs) exhibit excellent reasoning capabilities and draw extensive attention from the artificial intelligence research community (Lu et al., 2023b) to mathematical problem-solving in textual form (Chen et al., 2024b; Liao et al., 2024; Zhou et al., 2024; Zhao et al., 2024b; Zhou & Zhao, 2024; Kim et al., 2024). However, LLMs still struggle to solve mathematical problems involving images that require visual comprehension. Geometry problems, as typical mathematical problems with images, play an important role in evaluating mathematical reasoning skills (Zhang et al., 2023c), requiring a high level of visual comprehension. Besides, even though some problems are not related to geometry on the surface, they require the same skills for models (e.g., fine-grained image comprehension skills and multi-step reasoning skills). With the appearance of o1 (OpenAI, 2024), GPT-4o (Islam & Moushi, 2024), Gemini (Team et al., 2023), and numerous Large Multimodal Models (LMMs) (Li et al., 2024a; Liu et al., 2024; Chen et al., 2024d; Bai et al., 2023), recent researches progressively investigate using LMMs to solve mathematical geometry problems.

Although LMMs show impressive results in general visual question-answering (VQA) tasks (Fan et al., 2024; Liu et al., 2024), they still face challenges in solving mathematical geometry problems. The main reason is that the training data for LMMs are mainly from natural scenes, which have a gap with geometric data, leading to poor performance. Additionally, the limited size of existing geometric datasets further limits the geometric reasoning performance of LMMs.

Existing approaches for generating Q&A pairs in geometric tasks can be broadly classified into three categories. The Rewording method (Gao et al., 2023) rewords Q&A pairs from open-source datasets using LLMs to increase the number of questions. But this method ignores the diversity of images and knowledge points, as shown in Fig. 1 (a). The Template-based method (Kazemi et al., 2023; Zhang et al., 2024) introduces a data engine generating accurate geometric images and Q&A pairs. However, the generated images often lack fidelity and the template-based Q&A pairs are

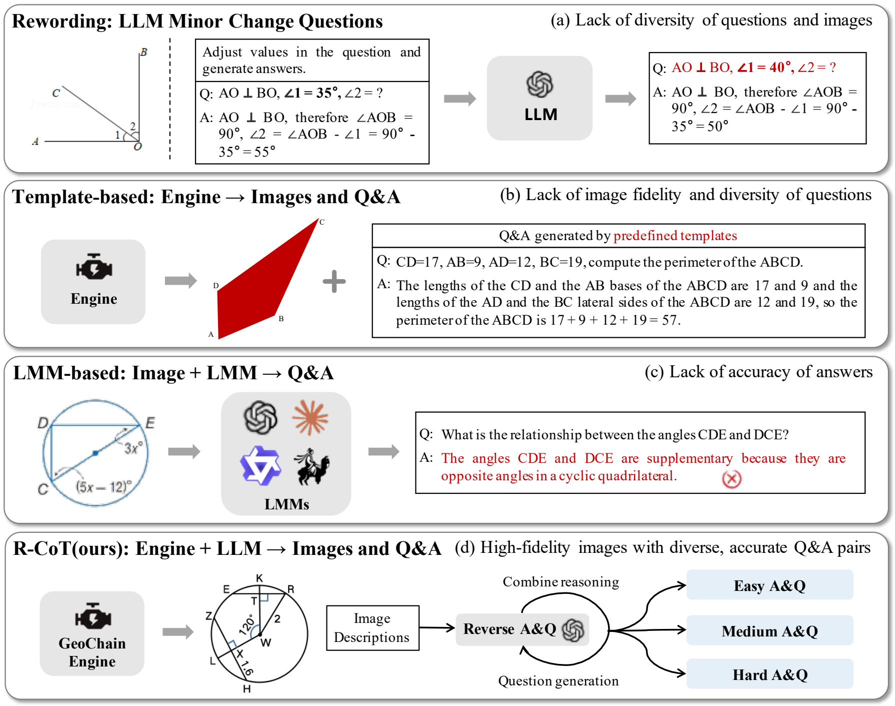


Figure 1: Comparison of R-CoT with existing data generation approaches. (a) Using LLMs to reword existing Q&A pairs without enriching the images and knowledge points. (b) Current geometry data generation engines produce low-fidelity images and template-based questions. (c) Due to limitations in visual perception and geometric reasoning of LMMs, Q&A pairs generated from images often have low accuracy. (d) We design GeoChain to generate high-fidelity geometric images with corresponding descriptions, followed by the Reverse A&Q, which uses an LLM to generate reasoning and questions from those descriptions.

limited in diversity, as shown in Fig. 1 (b). Lastly, the LMM-based method, which utilizes advanced LMMs to generate Q&A pairs from images, is widely used to generate high-quality training data for general VQA tasks (Chen et al., 2023; 2024a; Li et al., 2024b). However, they struggle with answer accuracy when generating geometric data due to limited reasoning capabilities, as illustrated in Fig. 1 (c).

To break through the data quality limitations on the geometric performance of LMMs, we propose a Reverse Chain-of-Thought (R-CoT) geometry problem generation pipeline, which combines the accuracy of the engine with the diverse geometry knowledge of LMMs (or LLMs), as shown in Fig. 1 (d). Specifically, we first design the GeoChain to generate high-fidelity geometric images step by step with corresponding descriptions focusing on relations between geometry elements, serving as priors for the following stage. Then, we introduce the Reverse A&Q to improve LLM-based geometric reasoning accuracy. The Reverse A&Q works in three steps, first segmenting the description for single-step reasoning, then progressively fusing the single-step reasoning to generate multi-step reasoning, and finally generating questions based on the multi-step reasoning results in reverse. Our method can significantly reduce incorrect answers by avoiding overly complex questions with an answer prior generation strategy. Using the R-CoT pipeline, we create a diverse GeoMM dataset containing geometric images with higher fidelity than existing synthetic data, along with accurate and diverse Q&A pairs.

R-CoT demonstrates consistent and significant improvements across multiple LMM baselines, achieving state-of-the-art (SOTA) results at 2B, 7B, and 8B model parameters. In particular, R-CoT-8B outperforms the closed-source model GPT-4o by an average of 13% and outperforms the previous SOTA open-source mathematical model by 16.6% and 9.2% on MathVista and GeoQA, respectively. Additionally, The R-CoT ensures greater training stability by generating accurate and high-fidelity data.

The main advantages of our method are summarized as follows:

- We introduce R-CoT, a novel reverse-process data generation pipeline for mathematical geometry that produces high-quality reasoning data. With R-CoT, we create GeoMM, a comprehensive dataset of high-fidelity geometric images and diverse Q&A pairs, offering better quality and lower variance compared to MAVIS and GeomVerse.
- We show that the proposed R-CoT can bring notable and consistent improvements across a range of LMM baselines such as LLaVA, Qwen, and InternVL. Using the recent LMM baselines, we achieve a new performance record in 2B, 7B, and 8B settings for solving geometry problems.
- We demonstrate state-of-the-art performance across both open-source and closed-source models. R-CoT-8B outperforms the leading open-source mathematical models and GPT-4o by 16.6% and 12.5% on MathVista, respectively, and by 9.2% and 14.5% on GeoQA, respectively.

2 Related Work

Recent research aimed at improving geometric reasoning in LMMs can be broadly divided into two categories. The first category focuses on inspiring geometric reasoning ability during the inference stage, while the other attempts to improve reasoning ability through targeted training.

Inspiring Model Potential During Geometric Inference. For the inference process, Zhao et al. (2024a) employs the chain of thought in visual and symbolic language modes to cross-validate and correct each other for the final result. Hu et al. (2024) utilizes code to generate images and solve problems through a visual chain of thought. Meanwhile, Mouselinos et al. (2024) uses a LLM as an agent to call external tools.

Improving Model Reasoning Ability During Geometric Training. For model training, symbolic geometry solvers like GeoS (Seo et al., 2015), Inter-GPS (Lu et al., 2021), and S2G (Tsai et al., 2021) aim to build formal language systems that use formal language for deductive reasoning on geometry problems. These systems are manually designed for formal languages with relatively small datasets, e.g. the GeoS dataset containing 186 problems and the Geometry3k (Lu et al., 2021) dataset containing about 3000 problems. The size of the datasets has increased slightly with the advent of neural geometric solvers, such as UniGeo (Chen et al., 2022), GeoQA (Chen et al., 2021), GeoQA+ (Cao & Xiao, 2022), and PGPS9K (Zhang et al., 2023a), with a total size of around 25k. The above datasets are collected manually, with high labeling costs and limited scale. With the rise of LMMs, these data scales are far from satisfactory for training, so many methods are devoted to building larger datasets. G-LLaVA (Gao et al., 2023) uses an LLM to reword original Q&A pairs in the GeoQA and Geometry3k dataset, resulting in 115k geometric Q&A data and 60k alignment data but does not increase the diversity of images and knowledge points. GeomVerse (Kazemi et al., 2023) uses a code-written engine to generate accurate geometric images and QA pairs, and there is still a certain gap between the generated images and real-world geometric images. Additionally, the questions generated by the template lack diversity.

To synthesize geometry data with both accuracy and diversity, we introduce R-CoT, a novel geometry data generation pipeline that addresses visual hallucinations and reasoning limitations in LMMs. This pipeline effectively generates the GeoMM dataset, featuring high-fidelity geometric images with accurate and diverse Q&A pairs.

3 Reverse Chain-of-Thought

The limited amount of high-quality mathematical geometry data restricts the geometric reasoning performance of existing LMMs. Current data generation methods possess two main limitations: (1) At the image level, their synthetic images have an appearance gap with real-world geometric images. (2) At the text level, their generated Q&A pairs lack accuracy and diversity, especially the relationships between geometry elements.

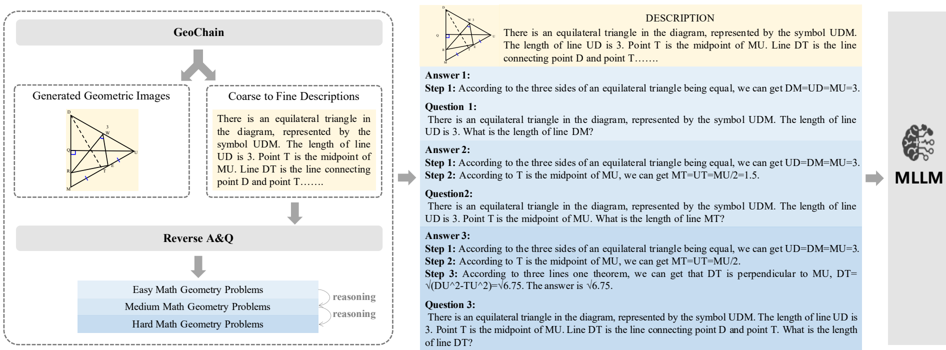


Figure 2: The GeoChain is utilized to obtain high-fidelity geometric images and corresponding descriptions. Subsequently, the Reverse A&Q is utilized to obtain accurate geometric Q&A pairs from descriptions.

To address these issues, we propose R-CoT, a two-stage mathematical geometry data generation pipeline. As shown in Fig. 2, in the first stage, to ensure the fidelity of the generated images, we develop GeoChain by referring to real-world mathematical geometry images. GeoChain can generate high-fidelity geometric images with multiple geometric elements in different relations. In the process of generating images, detailed image descriptions are also generated synchronously. These detailed image descriptions, which accurately describe the geometric elements and their relations, serve as priors for the second stage.

Algorithm 1: Pseudo-code of R-CoT

Input: Geometry substrates sampling rounds n, plot function f, image-description pair sets $\mathcal{S}$, line sampling rounds k, large language model $\mathcal{M}$
Output: Generated image $\mathcal{I}$, description $\mathcal{D}$, Question $\mathcal{Q}$; Answer $\mathcal{A}$

```
Initialization: I ← ∅, D ← ∅
for i ← 1 to n do
    Sample geometry substrate G_i and description D_i from image-description pair sets S
    Refresh I using plot function: I ← f(I, G_i)
    Refresh corresponding description: D ← D ∪ D_i
end
for j ← 1 to k do
    Select line drawing position P_j
    Draw line and label length: I ← f(I, P_j)
    Refresh corresponding description: D ← D ∪ P_j
    if j = k then
        Calculate all angle information R
        Draw angles and label degrees: I ← f(I, R)
        Refresh corresponding description: D ← D ∪ R
    end
end
Produce single-step reasoning result r_s using prompt P_s: r_s ← M(D, P_s)
Produce multi-step reasoning result r_c using prompt P_c: r_c ← M(r_s, P_c)
Generate answer A and its corresponding question Q using prompt P_q: A, Q ← M(r_c, P_q)
Return: I, D, Q, A
```

In the second stage, we design Reverse A&Q, which inputs only the image descriptions into an LLM to generate accurate and diverse Q&A pairs. This process successfully avoids the visual hallucinations caused by LMMs. Moreover, to break through the limitations of current LLMs in solving complex geometric problems, Reverse A&Q is designed to generate Q&A pairs step by step, inspired by the CoT reasoning framework (Wei et al., 2022). Firstly, an image description is segmented into several patches using LLMs. These description patches are inputted into Description Patch Reasoning to generate single-step reasoning results. Then, these single-step reasoning results are fused progressively to generate multi-step reasoning results in Chain-of-Thought Fusion. Finally, Ques-

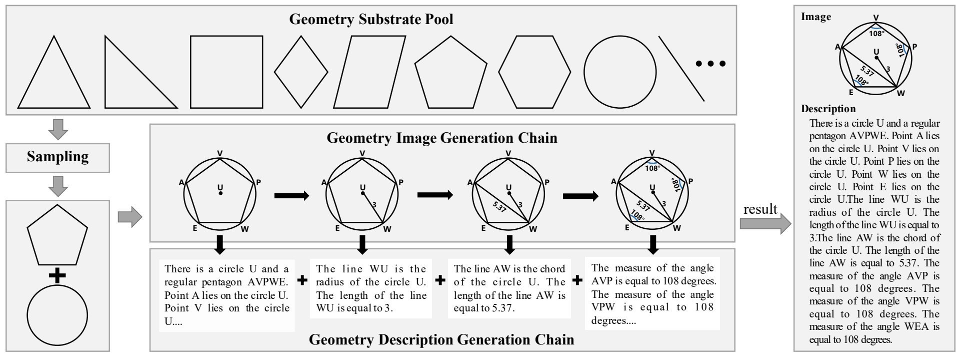


Figure 3: Overview of the GeoChain. We first construct a Geometry Substrate Pool containing various geometry substrates. Then one or more substrates are sampled from this pool and are inputted into the Geometry Generation Chain to generate the geometry image and corresponding description.

tion Generation generates questions based on the multi-step reasoning results. The pseudo-code of R-CoT is shown in Algor. 1.

3.1 GeoChain

To synthesize geometric images that are close to real-world geometric images, we design GeoChain, a chain of geometric images and descriptions generation engine that can generate both high-fidelity geometric images and their accurate descriptions. Only image descriptions will be used in the subsequent generation of geometric Q&A pairs.

As illustrated in Fig. 3, the GeoChain consists of three parts. Specifically, we first construct a geometry substrate pool that contains 20 different geometry substrates. Next, we randomly sample one or more substrates from this pool and input them into the Geometry Generation Chain. In the Geometry Generation Chain, the sampled substrates are combined into one geometric image step by step. Different from previous methods, our methods include many line operations (e.g., adding a line that connects midpoints of neighbor edges), which are common in real-world mathematical geometry images. Besides, at each step, we label the vertices with random letters (e.g., A, B, C) and annotate the geometric properties such as edge lengths and angles to create high-fidelity geometric images. Corresponding image descriptions are also generated step by step according to predefined templates. It is worth mentioning that these descriptions not only describe geometric shapes but also contain the relations between different geometric elements, such as points on which lines and whether two lines intersect. These relation descriptions are essential for the generation of relational geometry questions.

3.2 Reverse A&Q

Current LLMs still have limitations in solving complex geometric problems, using LLMs to directly generate Q&A pairs in one step may bring incorrect information. Inspired by CoT, we propose the Reverse A&Q (as shown in Fig. 4) to generate accurate and diverse Q&A pairs step by step using the generated image descriptions. This process consists of three steps: Description Patch Reasoning, Chain-of-Thought Fusion, and Question Generation.

Description Patch Reasoning. As the GeoChain can ensure the accuracy of the generated image descriptions, we want to maintain this data accuracy during the following steps. Hence, we design Description Patch Reasoning. First, image descriptions are segmented into patches to reduce the difficulty of reasoning. Then these description patches are inputted into an LLM to generate single-step reasoning results in a contextual learning manner.

Chain-of-Thought Fusion. To increase the complexity of the generated geometric problems, we introduce our Chain-of-Thought Fusion. In this step, single-step reasoning results are fused progressively, which means previous single-step reasoning results can provide necessary information

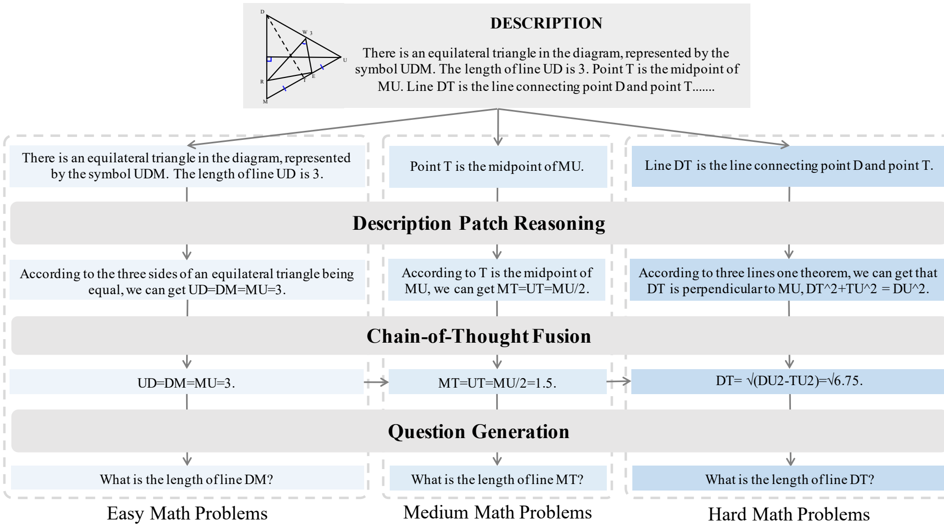


Figure 4: Overview of the Reverse A&Q. Image descriptions are segmented into patches and are used to generate single-step reasoning results. Then these single-step reasoning results are fused progressively to get multi-step reasoning results. Finally, questions are generated based on the multi-step reasoning results.

for later ones to get complex reasoning results. This method ensures that each reasoning step is logically connected.

Question Generation. When an LLM is directly tasked with generating geometric questions, it often fails to judge their difficulty accurately and thus produces incorrect answers. To address this issue, we employ our Question Generation to generate solvable questions of appropriate difficulty based on the generated multi-step reasoning results.

Detailed prompts for generation can be found in Appendix A.

3.3 GeoMM

Through the R-CoT pipeline, we construct a high-quality geometric dataset, GeoMM. Detailed statistical information regarding the images and text within GeoMM is presented in Fig. 5.

At the image level, GeoMM contains 20 geometric shapes, with the most common being triangles, quadrilaterals, and circles. To ensure the model can interpret geometric images of varying complexity, GeoMM includes images categorized into four complexity levels, determined by the number of geometric shapes present. Unlike previous-generation engines, which primarily focus on constructing geometric images through the combination of polygons or circles, we emphasize the critical role of lines in geometric figures. Lines with special properties, such as midlines or radii, are foundational to many geometric theorems (e.g., the midline theorems). To enhance the richness of geometric knowledge embedded in Q&A pairs generated at later stages, we integrate line elements with specific properties (e.g., radii) into the images. This approach significantly improves the fidelity of the generated images. A comparison of the synthesized images is shown in Fig. 6.

At the text level, the GeoMM dataset is composed of four major categories of geometric problems, with a particular emphasis on the relational question type, which is often underrepresented in existing synthetic datasets. The completeness of the generated geometric descriptions, which incorporate multiple relationships between geometric elements, facilitates the generation of relational questions. Such relational problems are intended to help the model better understand and process relative information among geometric components. Detailed examples of the different question types can be found in Appendix B.

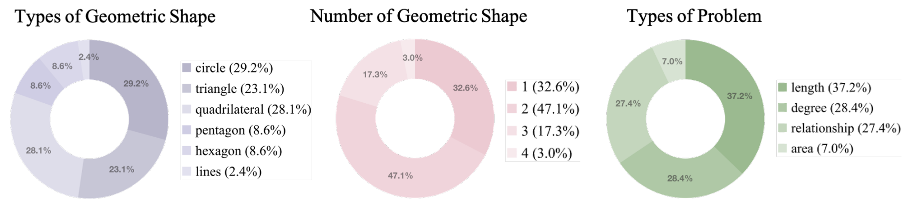
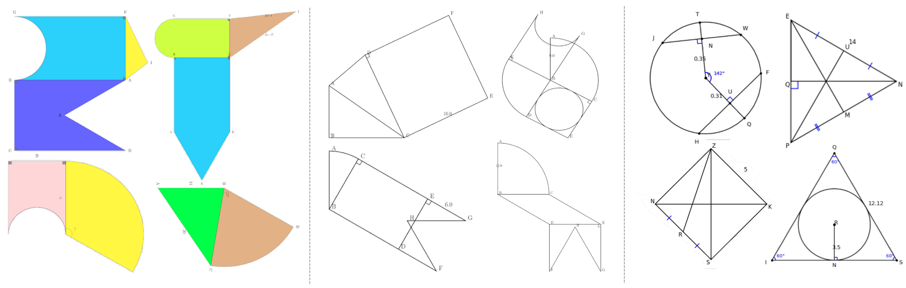


Figure 5: Statistical information about GeoMM.



(a) GeomVerse (b) MAVIS (c) GeoMM

Figure 6: Visualization comparison of recent geometry synthesis dataset.

4 EXPERIMENTS

4.1 SETUP

Our R-CoT pipeline utilizes ERNIE Bot 4.0 as the core LLM. We train several LMMs (Bai et al., 2023; Liu et al., 2024; Huang et al., 2024; Zhang et al., 2023b; Chen et al., 2024d) using geometric instruction data from the Geo170K (Gao et al., 2023) and our GeoMM dataset. Both the projected linear layer and the language model are trainable during training. The models are trained for two epochs with a batch size of one per NPU (Ascend910-65G). For evaluation, we compare these models with other LMMs on the geometry problem solving on the testmini set of MathVista (Lu et al., 2023a) and the test set of GeoQA (Chen et al., 2021) following Gao et al. (2023). We adopt Top-1 accuracy as the evaluation metric and employ the regular expression (Gao et al., 2023) to extract the predicted choices from the generated answers. The answer is considered incorrect if the regular expression fails to extract a valid answer.

Table 1: GeoMM effectiveness validation on different models. 'Geo-' indicates the model is fine-tuned only with geometric instruction data of Geo170K. Consistent and significant improvement without adding any additional parameters.

Model	MathVista	GeoQA
Geo-Qwen-VL-7B	47.6	53.9
R-CoT-Qwen-7B	51.0 (3.4↑)	55.7 (1.8↑)
Geo-LLaVA-1.5-7B	47.6	58.6
R-CoT-LLaVA-7B	49.5 (1.9↑)	61.3 (2.7↑)
Geo-Mini-Monkey-2B	55.3	61.8
R-CoT-Mini-Monkey-2B	57.7 (2.4↑)	62.6 (0.8↑)
Geo-InternLM-XC2-7B	58.2	63.8
R-CoT-InternLM-XC2-7B	62.0 (3.8↑)	67.8 (4.0↑)
Geo-InternVL-2.0-8B	71.1	74.2
R-CoT-InternVL-2.0-8B	73.1 (2.0↑)	75.9 (1.7↑)

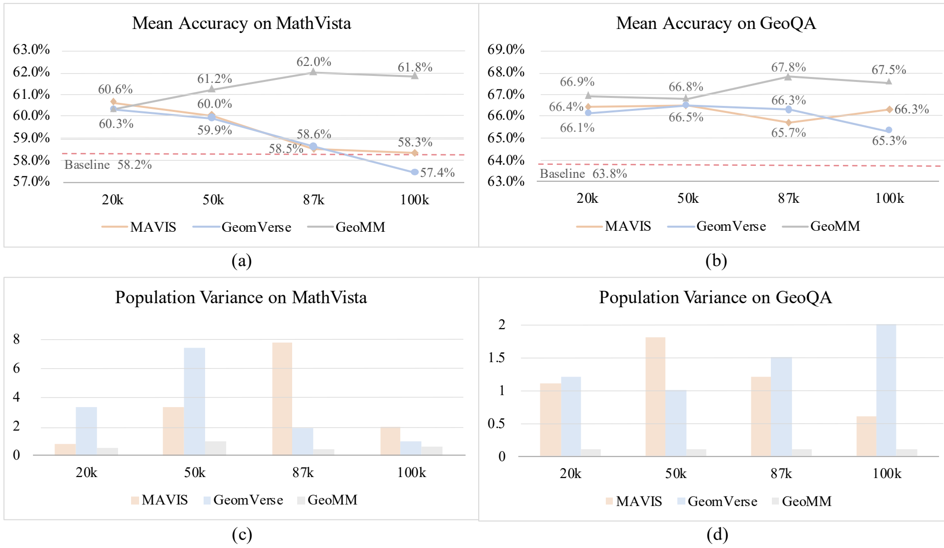


Figure 7: Compared with existing datasets at different data scales.

4.2 Effectiveness of GeoMM

Compared with existing datasets. We train our model using GeoMM and two recent synthetic datasets for geometric problems, *i.e.* MAVIS (synthesis part) (Zhang et al., 2024) and GeomVerse (Kazemi et al., 2023) at the same data scale for a fair comparison. Specifically, we sample the data from each dataset to different scales. As we observe clear performance fluctuations caused by the quality of train data, we train the models three times at each data scale and report the average Top-1 accuracy in Fig. 7 (a) and (b). In general, all three datasets can improve the geometry reasoning ability of the baseline model. The model trained using our GeoMM exhibits significantly superior performance in most settings, demonstrating the better quality of GeoMM. Moreover, as shown in Fig. 7 (c) and (d), the performance variance of our method is significantly lower. The more stabilized optimization also indicates the better quality of GeoMM since our method aims to improve the fidelity of images and the accuracy of Q&A pairs.

Influence of data scales. Intuitively, more training data can lead to better performance. As shown by the trend in Fig. 7 (a) and (b), the performance on both datasets can be further improved when increasing the scale of GeoMM. Surprisingly, the performance declines on MathVista when training with the other two datasets. We assume that the limited diversity and the gap between their data and real-world geometric problems would restrict their scalability. However, it is inevitable that the diversity of synthetic data is still constrained by the generation mechanism. When the data scale exceeds 87k, performance saturates. Therefore, we set the size of GeoMM to 87K and the following experiments are conducted using it.

Generalized effectiveness to other LMMs. We extend our method to several recent LMMs to verify its universality. Comparing the models trained only using Geo170K with using both Geo170K and our GeoMM, the latter exhibits consistent improvements in accuracy as shown in Tab. 1. Specifically, the baseline models are improved by at least 1.9% on MathVista and 0.8% on GeoQA, respectively. The performance difference is most obvious on Geo-InternLM-XC2-7B where R-CoT-InternLM-XC2-7B exhibits increases of 3.8% and 4.0%. The most advanced InternVL2.0-8B is still improved by 2.0% on MathVista and 1.7% on GeoQA. The results indicate that GeoMM not only has effective geometry knowledge but also can be widely applied to various advanced LMMs.

4.3 Ablation Study

Data generating procedures. To verify the effectiveness of detailed designs in our R-CoT, we set several variants shown in Tab. 2 by removing different proposed procedures. Each model is trained

Table 2: Ablation study on the data generating procedures.

Configurations			MathVista	GeoQA
Description Based	Reverse Generation	Step Reasoning		
✗	✗	✗	58.2	64.5
✓	✗	✗	60.4	65.1
✓	✓	✗	61.3	65.5
✓	✓	✓	62.0	67.8

Table 3: Ablation study on the robustness to polygonal distributions.

Method	Polygon Distribution					MathVista	GeoQA
	circle	triangle	quad	polygon	lines		
Group I	39.9%	15.3%	14.9%	14.9%	15.0%	60.1	66.6
Group II	29.2%	23.1%	28.1%	17.2%	2.4%	60.1	66.9
Group III	23.3%	18.1%	17.7%	20.5%	20.4%	60.6	67.1

using data generated by those variants and evaluated on MathVista and GeoQA. Introducing the description-based paradigm contributes to 2.2% and 0.6% on MathVista and GeoQA, respectively. Both step reasoning and reverse generation are designed to improve the accuracy of Q&A pairs. When using the reverse generation strategy, the accuracy on MathVista is improved by 0.9% and step reasoning can further boost performance by 0.7% on this basis. As a result, the full setting achieves the highest result on both datasets, demonstrating the effectiveness of each procedure in the data generation pipeline. Detailed examples of the impact of Reverse Generation and Step Reasoning on the accuracy of the generated data can be found in Appendix C.

Robustness to polygon distributions. As our dataset consists of several types of geometric shapes, we adjust the proportions of different polygon types and form three subsets of 20k data to train the model. Similar quantitative results within 0.5% in Tab. 3 show the impact of polygon distributions is almost negligible, demonstrating the strong robustness of our method to different polygon distributions. Therefore, the performance gain is mainly attributed to the diverse and accurate geometry representation and reasoning knowledge provided by our method.

4.4 Comparison with Previous State-of-the-Art

With the proposed method, we train three specialized models for geometry problem solving named R-CoT-2B, R-CoT-7B, and R-CoT-8B based on Mini-Monkey-2B, InternLM-XC2-7B, and InternVL-2.0-8B, respectively. We compare our models with both general and mathematical LMMs on the testmini set of MathVista and the test set of GeoQA. We use the same prompt prefix as G-LLaVA (Gao et al., 2023). As shown in Tab. 4, R-CoT-8B achieves the best performance on both datasets. Specifically, it significantly surpasses advanced closed-source GPT-4o by 12.5% on MathVista and 14.5 % on GeoQA. Compared to mathematical LMMs, it still outperforms SOTA open-source mathematical models by 16.6% on MathVista and 9.2% on GeoQA.

5 Discussion

To better understand why R-CoT leads to improvements, we conduct qualitative analysis by comparing the best-performing closed-source LMM GPT-4o with our model. Examples from different types of geometric images are shown in Fig. 8. Our model generates a more concise chain of thought and consistently arrives at the correct answer. In contrast, GPT-4o's problem-solving ability is primarily limited by its perceptual understanding of geometry; for instance, it often misinterprets angle relationships in these cases. We argue that our approach addresses this by introducing relational problems that were overlooked in previous datasets, thereby enhancing the model's fine-grained perceptual abilities, and allowing the model to produce a more streamlined reasoning process. The results also suggest that accurate comprehension of geometric components could be crucial for effective reasoning. More examples can be found in Appendix D. Due to the reasoning capabilities of current LMMs, we rely on LLMs to generate Q&A pairs. This can occasionally result in non-unique

Table 4: Top-1 Accuracy (%) on geometry problem solving on the testmini set of MathVista and the GeoQA test set. * represents the results from the existing papers.

Model	MathVista	GeoQA
Closed-source LMMs		
GPT-4o (Islam & Moushi, 2024)	60.6	61.4
GPT-4V	50.5*	-
Gemini Ultra	56.3*	-
Open-source LMMs		
LLaVA-LLaMA-2-13B (Liu et al., 2024)	29.3*	20.3*
mPLUG-Owl2-7B (Ye et al., 2024)	25.5	21.4
Qwen-VL-Chat-7B (Bai et al., 2023)	35.6	26.1
Monkey-Chat-7B (Li et al., 2024a)	24.5	28.5
Deepseek-VL-7B (Lu et al., 2024)	34.6	33.7
InternLM-XC2-7B (Zhang et al., 2023b)	51.4	44.7
InternVL-1.5-20B (Chen et al., 2024c)	60.1	49.7
Mini-Monkey-2B (Huang et al., 2024)	53.4	50.1
InternVL-2.0-2B	56.7	50.9
InternVL-2.0-8B	65.9	56.5
Open-source Mathematical LMMs		
Math-LLaVA-13B (Shi et al., 2024)	56.5*	47.8
G-LLaVA-7B (Gao et al., 2023)	53.4*	62.8*
MAVIS-7B (Zhang et al., 2024)	-	66.7*
R-CoT-2B	57.7	62.6
R-CoT-7B	62.0	67.8
R-CoT-8B	**73.1**	**75.9**

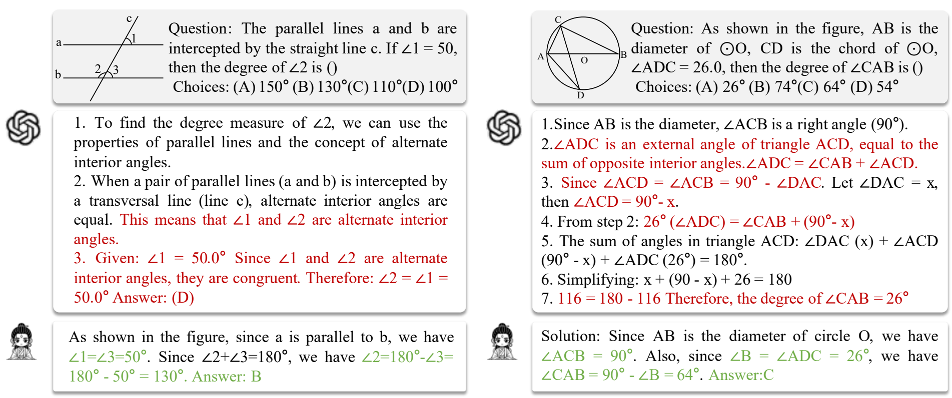


Figure 8: Problem-solving Comparison with GPT-4o.

images corresponding to the same descriptions. Although most of the generation results are correct, some errors still persist.

6 Conclusion

We propose R-CoT, a novel reverse generation pipeline that significantly enhances the quality and fidelity of geometry Q&A pair generation. The data produced by R-CoT offers obvious advantages over previous synthesis geometry datasets, such as MAVIS and GeomVerse. Our approach achieves consistent improvements over existing LMMs, setting new state-of-the-art results compared to both open-source and closed-source models. Our results highlight the critical role of high-quality data in improving the geometric reasoning capabilities of LMMs. We will extend this method to other types of mathematical questions while exploring strategies to mitigate LMM visual hallucinations and improve data accuracy, providing further insights for future research.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

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

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024a.

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

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3313–3323, 2022.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *CoRR*, 2023.

Xinyun Chen, Ryan Andrew Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models. In *Forty-first International Conference on Machine Learning*, 2024b.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024c.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024d.

Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Yang Zhao, Xinze Guan, and Xin Wang. Muffin or chihuahua? challenging multimodal large language models with multipanel vqa. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6845–6863, 2024.

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

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024.

Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. Mini-monkey: Alleviate the sawtooth effect by multi-scale adaptive cropping. *arXiv preprint arXiv:2408.02034*, 2024.

Raisa Islam and Owana Marzia Moushi. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*, 2024.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023.

Hyeonwoo Kim, Gyoungjin Gim, Yungi Kim, Jihoo Kim, Byungju Kim, Wonseok Lee, and Chanjun Park. Saas: Solving ability amplification strategy for enhanced mathematical reasoning in large language models. *arXiv preprint arXiv:2404.03887*, 2024.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26763–26773, 2024a.

Zhuowan Li, Bhavan Jasani, Peng Tang, and Shabnam Ghadar. Synthesize step-by-step: Tools templates and llms as data generators for reasoning-based chart vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13613–13623, 2024b.

Haoran Liao, Jidong Tian, Shaohua Hu, Hao He, and Yaohui Jin. Look before you leap: Problem elaboration prompting improves mathematical reasoning in large language models. *arXiv preprint arXiv:2402.15764*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: Towards real-world vision-language understanding. *CoRR*, 2024.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6774–6786, 2021.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*, 2023a.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. A survey of deep learning for mathematical reasoning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023b.

Spyridon Mouselinos, Henryk Michalewski, and Mateusz Malinowski. Beyond lines and circles: Unveiling the geometric reasoning gap in large language models. *arXiv preprint arXiv:2402.03877*, 2024.

OpenAI. Openai o1 system card. *preprint*, 2024.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1466–1476, 2015.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024.

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

Shih-Hung Tsai, Chao-Chun Liang, Hsin-Min Wang, and Keh-Yih Su. Sequence to general tree: Knowledge-guided geometry word problem solving. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 964–972, 2021.

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

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13040–13051, 2024.

Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. A multi-modal neural geometric solver with textual clauses parsed from diagram. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 3374–3382, 2023a.

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023b.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*, 2023c.

Xueliang Zhao, Xinting Huang, Tingchen Fu, Qintong Li, Shansan Gong, Lemao Liu, Wei Bi, and Lingpeng Kong. Bba: Bi-modal behavioral alignment for reasoning with large vision-language models. *arXiv preprint arXiv:2402.13577*, 2024a.

Zilong Zhao, Yao Rong, Dongyang Guo, Emek Gözlüklü, Emir Gülboy, and Enkelejda Kasneci. Stepwise self-consistent mathematical reasoning with large language models. *arXiv preprint arXiv:2402.17786*, 2024b.

Yongwei Zhou and Tiejun Zhao. Dual instruction tuning with large language models for mathematical reasoning. *arXiv preprint arXiv:2403.18295*, 2024.

Yue Zhou, Yada Zhu, Diego Antognini, Yoon Kim, and Yang Zhang. Paraphrase and solve: Exploring and exploiting the impact of surface form on mathematical reasoning in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2793–2804, 2024.

A DETAILS OF PROMPT IN REVERSE A&Q

We used ERNIE Bot 4.0 to implement Reverse A&Q. We describe the prompts used in Reverse A&Q, including the prompts for the Description Patch Reasoning (Fig. 9), the Chain-of-Thought Fusion (Fig. 10), and the Question Generation (Fig. 11). In these figures, the texts in blue are instructions, and in orange are the input information. Each prompt contains three contextual examples, and we show only one of them with the remaining examples replaced by ellipses.

B EXAMPLES OF GEOMM DATASET

Through the R-CoT, we construct a high-quality geometric dataset, GeoMM. In Fig. 12, we provide a detailed overview of specific cases from GeoMM. These cases demonstrate the variety of mathematical geometry question types covered by GeoMM, including solving for lengths, angles, areas, and geometry elemental relations. Each of these categories is critical for improving the geometric reasoning ability of LMMs.

Description Patch Reasoning Prompt

Use the mathematics you know to make simple inferences based on image descriptions. Make sure your reasoning is correct. You can ignore descriptions from which no relevant information can be inferred. Examples are as follows:

Input:
There is an equilateral triangle in the diagram, represented by the symbol ATW. The length of line AT is 9. Line AG is perpendicular to line TW. Point G lies on line TW.

Output:
Description 1: There is an equilateral triangle in the diagram, represented by the symbol ATW. The length of line AT is 9.
Reasoning 1: If the three sides of an equilateral triangle are equal, we can get AT = TW = WA = 9.
Description 2: Line AG is perpendicular to line TW. Point G lies on line TW.
Reasoning 2: Since AG is perpendicular to TW and G lies on TW, AG is the height of the equilateral triangle ATW. In an equilateral triangle, the height (h) can be calculated using the formula h = √3/2 * side length.
……
Input: [DESCRIPTION]
Output: [SINGLE-STEP REASONING]

Figure 9: The prompt of the Description Patch Reasoning.

Chain-of-Thought Fusion Prompt

Using contextual information to supplement conditions, combining related reasoning processes to obtain multi-step reasoning results. Examples are as follows:

Input: [
Description 1: There is an equilateral triangle in the diagram, represented by the symbol ATW. The length of line AT is 9.
Reasoning 1: If the three sides of an equilateral triangle are equal, we can get AT = TW = WA = 9.
Description 2: Line AG is perpendicular to line TW. Point G lies on line TW.
Reasoning 2: Since AG is perpendicular to TW and G lies on TW, AG is the height of the equilateral triangle ATW. In an equilateral triangle, the height (h) can be calculated using the formula h = √3/2 * side length.

Output:
Reasoning 1: If the three sides of an equilateral triangle are equal, we can get AT = TW = WA = 9.
Reasoning 2: Since AG is perpendicular to TW and G lies on TW, AG is the height of the equilateral triangle ATW. In an equilateral triangle, the height h can be calculated using the formula h = √3/2 * side length.
Therefore, AG = √3/2 * 9 = 9√3/2.
……
Input: [SINGLE-STEP REASONING]
Output: [MULTI-STEP REASONING]

Figure 10: The prompt of the Chain-of-Thought Fusion.

C THE CASE OF DIRECT GENERATION AND REVERSE A&Q GENERATION

The core idea of the Reverse A&Q is to improve the accuracy of Q&A pairs by first simplifying the reasoning based on descriptions and then generating corresponding questions from the answers in a reversed manner. A straightforward approach is directly prompting ERNIE Bot 4.0 to generate Q&A pairs from the input image description. However, as shown in the left of Fig. 13, this approach often fails to determine the correct answer. In contrast, the Q&A pairs produced by Reverse A&Q are correct for all three instances with our design.

D THE CASE OF GEOMETRIC REASONING FOR GPT-4O AND OUR MODEL

We conduct qualitative analysis by comparing the best-performing closed-source LMM GPT-4o with our model. Fig. 14 shows several examples from different types of geometric images. We highlight the incorrect key steps in red and the correct key steps in green. Obviously, our model generates a more concise chain of thought and arrives at the correct answer.

Question Generation Prompt

Generate a question based on description and reasoning and extract the chain of thought. Examples are as follows:

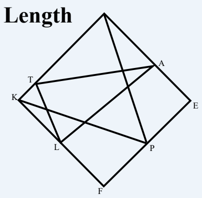
Input:
Description 1: There is an equilateral triangle in the diagram, represented by the symbol ATW. The length of line AT is 9.
Reasoning 1: If the three sides of an equilateral triangle are equal, we can get AT = TW = WA = 9.
Description 2: Line AG is perpendicular to line TW. Point G lies on line TW.
Reasoning 2: Since AG is perpendicular to TW and G lies on TW, AG is the height of the equilateral triangle ATW. In an equilateral triangle, the height h can be calculated using the formula h = √3/2 * side length. Therefore, AG = 9√3/2.

Output:
Question 1: There is an equilateral triangle in the diagram, represented by the symbol ATW. The length of line AT is 9. What are the lengths of the sides AW and TW?
Answer 1: Step 1: The three sides of an equilateral triangle are equal, we can get AT = TW = WA = 9.

Question 2: There is an equilateral triangle in the diagram, represented by the symbol ATW. The length of line AT is 9. Line AG is perpendicular to line TW. Point G lies on line TW. What is the length of the side AG?
Answer 2: Step 1: The three sides of an equilateral triangle are equal, we can get AT = TW = WA = 9. Step 2: Since AG is perpendicular to TW and G lies on TW, AG is the height of the equilateral triangle ATW. In an equilateral triangle, the height h can be calculated using the formula h = √3/2 * side length. Step 3: Therefore, AG = √3/2 * side length = 9√3/2.
……
Input: [MULTI-STEP REASONING]
Output: [QUESTION AND ANSWER]

Figure 11: The prompt of the Question Generation.

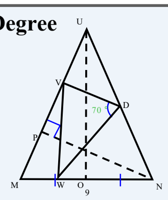
Length



Question:
There is a square in the diagram, represented by the symbol EIKF. The length of line KF is 7. Line KP is the line connecting point K and the midpoint P of line FE. What is the length of line KP?

Answer:
The length of KP can be calculated using the Pythagorean theorem since we have a right triangle KPF with KP as the hypotenuse. KP = √(KF^2 + PF^2) = √(7^2 + 3.5^2) = √(49 + 12.25) = √61.25. The answer is approximately 7.83 (rounded to two decimal places).

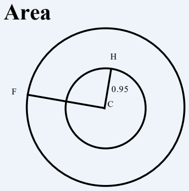
Degree



Question:
There is a triangle DWV. The measure of angle WDV is 70 degrees. What is the sum of the measures of angles DWV and VWD?

Answer:
Step1: According to the sum of the interior angles of the triangle being 180 degrees, we can get ∠WDV + ∠DWV +∠VWD = 180°.
Step2: Substituting the given value, ∠DWV + ∠VWD = 180 - 70 = 110°.

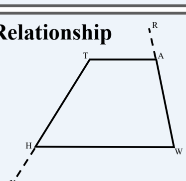
Area



Question:
The length of the line CF is equal to 1.9, and the length of the line CH is equal to 0.95. What is the difference in areas between the two concentric circles?

Answer:
Step1: Calculate the area of the larger circle using the formula A_large = π * CF^2 = π * 1.9^2.
Step2: Calculate the area of the smaller circle using the formula A_small = π * CH^2 = π * 0.95^2.
Step3: Subtract the area of the smaller circle from the area of the larger circle to find the difference: A_difference = A_large - A_small = 2.7075π.

Relationship



Question:
There is a trapezoid in the diagram, represented by the symbol HWAT. AR is the extension line of WA. HX is the extension line of TH. Which line is parallel to line AT?

Answer:
According to the definition of a trapezoid, in the trapezoid HWAT, the line HW is parallel to line AT. Therefore, the line that is parallel to line AT is line HW.

Figure 12: Examples of GeoMM dataset.

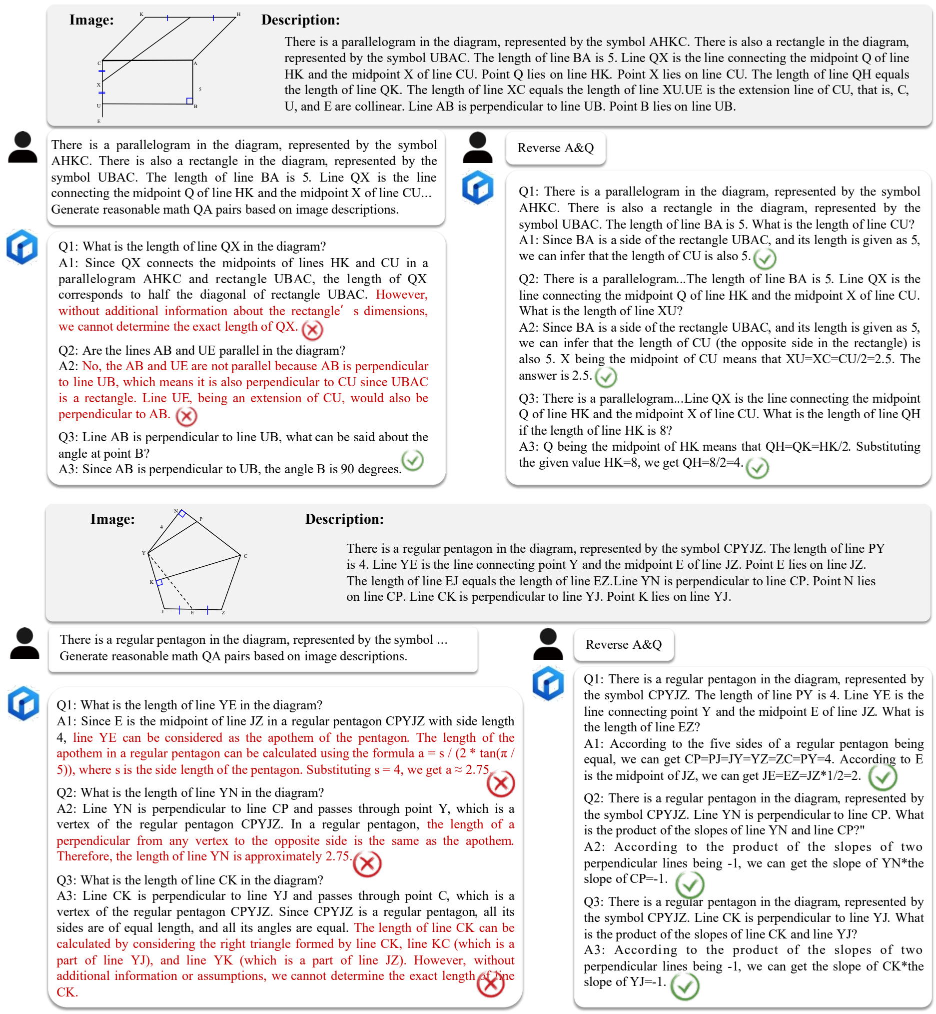


Figure 13: The Case of Direct Generation and Reverse A&Q Generation.

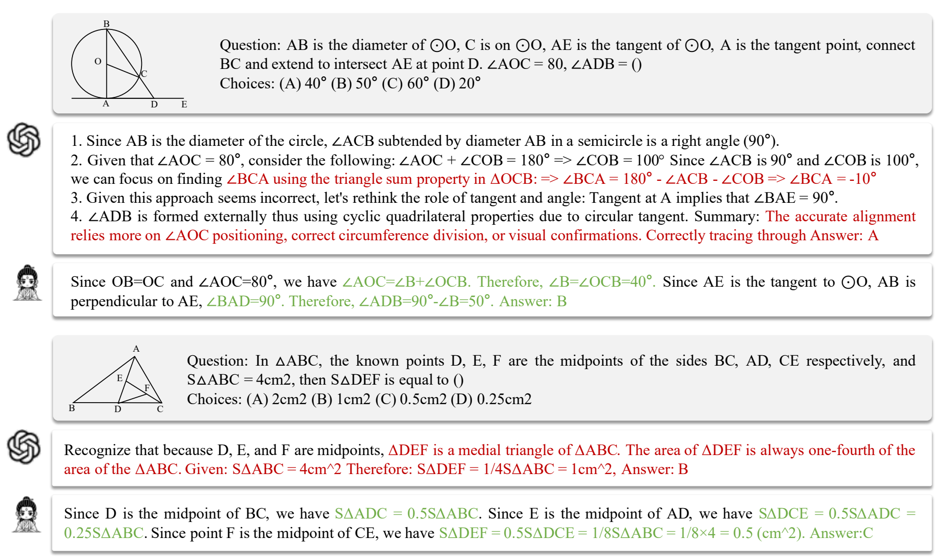


Figure 14: Problem-solving Comparison with GPT-4o.