GENERALIZABLE GEOMETRIC IMAGE CAPTION SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models have various practical applications that demand strong reasoning abilities. Despite recent advancements in this area, these models still struggle to solve complex geometric problems. A key challenge stems from the lack of high-quality image-text pair datasets for understanding geometric images. Furthermore, most template-based data synthesis pipelines typically fail to generalize to questions outside their predefined templates. In this paper, we mitigate this issue by introducing a complementary RLHF process into the data generation pipeline. By adopting RAFT to adjust captions for image-text pairs generated from fewer than 50 templates and using reward signals derived from downstream mathematical problem-solving tasks, our pipeline successfully captures the key features of geometry problem-solving. This enables better task generalization and yields non-trivial improvements. Furthermore, the generated dataset also enhances the general mathematical reasoning capabilities of multimodal large language models beyond the domain of geometric mathematical problems, yielding accuracy improvements of 2.8%–5.3% in arithmetic, algebraic, and numerical tasks with even non-geometric input images.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities across a variety of vision-related tasks , including Visual Question Answering (VQA), vision grounding, and image captioning. Recent MLLMs, such as Qwen2.5-VL, Intern2.5-VL, and LLaVA Next (Bai et al., 2025; Chen et al., 2024; Liu et al., 2024), have shown superior performance compared to dedicated vision models across a wide range of visual tasks, highlighting the potential of unified multimodal architectures. As the field progresses, there has been growing interest in enhancing the reasoning capabilities of MLLMs (Jaech et al., 2024; Shao et al., 2024), which is seen as a crucial factor in pushing the performance limits of these models. Among various reasoning tasks, math reasoning (Zhang et al., 2024a) has received particular attention due to its structured problem-solving nature, offering a clear pathway for MLLMs to learn and improve their reasoning skills.

Research from MathVerse (Zhang et al., 2024a) has shown that MLLMs perform best when the input is purely textual, but their performance significantly drops when the input is visual-only. This underscores the pressing need for MLLMs to develop strong cross-modal reasoning capabilities, which involves accurately and comprehensively transferring information from the image to the text. Although numerous geometry and math datasets have been introduced (Lu et al., 2023; Zhang et al., 2024b; Wang et al., 2024) to boost various facets of model performance, high-quality datasets explicitly tailored for cross-modal reasoning remain scarce. That is because in existing datasets, the alignment between images and captions is often asymmetrical. For example, in geometric problems, two lines of equal length can be easily described in text but may not be correspondingly annotated or clearly visible in the image. Such discrepancies hinder the model's ability to learn robust cross-modal reasoning.

Meanwhile, Reinforcement Learning (RL) has been shown to significantly enhance model reasoning and generalization capabilities (Guo et al., 2025). Its reward-driven framework is particularly effective for cross-modal reasoning, allowing models to optimize decision-making through interactive feedback (Deng et al., 2025; Peng et al., 2025; Huang et al., 2025a). Building on these insights, we employ the RAFT method (Dong et al., 2023), designing a reward function that incorporates both

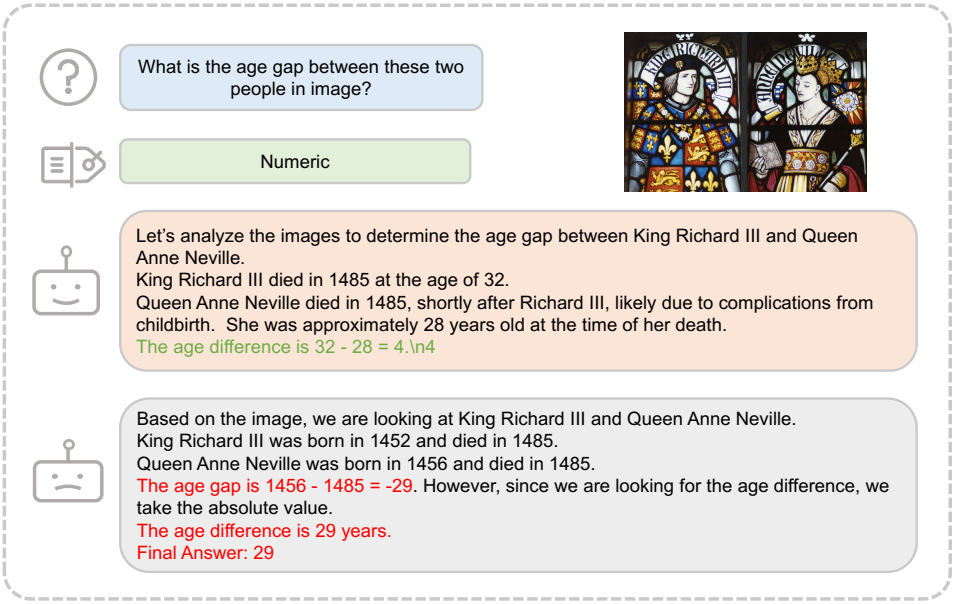

Figure 1: MLLMs learn from our synthetic geometric mathematical problems and generalize to algebraic cases with even non-geometric input images.

reasoning and caption rewards. This facilitates the alternating optimization of dataset quality and model reasoning abilities, thereby improving performance on complex multimodal tasks.

To bridge the gap between vision and language modalities, we propose an RL-based data refinement engine that iteratively enhances data quality. Built on top of the pipeline, a novel geometry dataset comprising 10,000 image-caption pairs is introduced. To the best of our knowledge, this is the first high-quality dataset in which visual and textual information are fully aligned, making it a valuable resource for improving cross-modal reasoning. Experimental results show that our dataset significantly improves models' cross-modal reasoning abilities and improves their performance on geo-image textualization tasks. Models trained on our dataset exhibit strong generalization to other math-focused benchmarks, including MathVerse and MathVista. In summary, our main contributions are summarized as follows:

- We introduce **GeoReasoning-10K**, a new dataset containing 10,000 carefully constructed image-caption pairs where visual and textual information are fully equivalent. This dataset serves as a high-quality resource for training cross-modal reasoning models.
- We propose **Geo-Image-Textualization**, a scalable RL-based data generation and refinement framework for producing high-quality synthetic image-caption pairs in geometry field. Our iterative RL-based optimization significantly improves data alignment and semantic accuracy, and generalizes to out-of-domain tasks in geometry.
- Extensive experiments show that models trained on GeoReasoning-10K generalize well to questions other than geometry problems in MathVerse and MathVista benchmarks.

2 Related Works

2.1 Data Generation

Recent studies have highlighted the shortage of high-quality geometry image–caption datasets, which limits fine-grained cross-modal reasoning in geometry tasks. Following AlphaGeometry (Trinh et al., 2024), Autogeo (Huang et al., 2025b) proposed an automatic data generation engine to generate image-caption pair and construct a 100K dataset called AutoGeo-100k. MATHGLANCE (Sun et al., 2025) intruduced GeoPeP, a perception-oriented dataset of 200K structured geometry image-text pairs

explicitly annotated with geometric primitives and precise spatial relationships. MagicGeo (Wang et al., 2025) formulates diagram synthesis as a coordinate optimization problem, ensuring formal geometric correctness via a solver before performing coordinate-aware text generation.

In parallel, numerous studies have investigated strategies to enhance cross-modal information transfer. A classical task within natural language processing (NLP) is Geometry Problem Solving (GPS), which has been extensively studied (Lu et al., 2021; Zhang et al., 2022). In this task, both the question and the answer are in formatted textual language. Solving geometry problems in this way often requires a specialized vision module, such as an object detector, to extract visual information. This information is then transformed into ostructured symbolic by the language model. Recent works have attempted to utilize LLMs and MLLMs (Xia et al., 2024) to solve the GPS task. While such methods offer partial solutions for cross-modal translation, they encounter two major limitations. First, they demand substantial pre-training on structured or formatted language inputs that diverge significantly from natural language, necessitating extensive pre-processing of geometry questions. Second, this discrepancy often hampers the model's ability to generalize across domains, limiting its broader applicability.

Despite these advances, existing pipelines still fail to guarantee full modality equivalence: captions often omit or abstract away visual details, and images lack exhaustively aligned text descriptions. In contrast, our Geo-Image-Textualization framework produces fully equivalent image–caption pairs through RL-based iterative refinement, thereby substantially improving cross-modal alignment and reasoning performance.

2.2 Image Captioning

Image captioning, which lies at the intersection of computer vision and natural language processing, aims to generate comprehensive descriptions of visual content and has been extensively studied in the context of natural images. While general-purpose multimodal large language models (MLLMs) such as mPLUG-Owl2 (Ye et al., 2024), Minigpt-4 (Zhu et al., 2023) and BLIP-3 (Xue et al., 2024) can perform this task to some extent, their effectiveness is often limited by insufficient fine-grained cross-modal alignment. Image-Textualization (Pi et al., 2024) addresses this challenge by incorporating multiple vision experts to produce more detailed and accurate descriptions.

However, applying image captioning to geometry-related tasks remains underexplored, primarily because most MLLMs are pre-trained on natural image datasets. OminiCaptioner (Lu et al., 2025) propose a unified visual captioning framework that converts diverse image types, including natural scenes, structured charts and geometric diagrams, into rich, fine-grained textual descriptions. However, its geometric annotations are drawn from AutoGeo and MAVIS, which themselves rely on synthetic or loosely aligned diagram–caption pairs rather than fully equivalent visual-textual representations. Moreover, the lack of high-quality geometry image-caption pairs makes it particularly difficult for MLLMs to extract and align geometric information accurately. Consequently, current models fail to extract and align intricate geometric information accurately and thus underperform on Geo-Image-Textualization tasks compared to their success on natural image captioning and general visual reasoning benchmarks

2.3 Reinforcement Learning in Multimodal Learning

Reinforcement learning (RL) has recently expanded from pure NLP applications into the multimodal domain, where agents must optimize policies across heterogeneous sensory inputs (e.g., images, text, audio). In these settings, RL agents learn to make decisions by maximizing long-term rewards rather than focusing solely on one-step prediction accuracy, a property particularly valuable for tasks requiring persistent cross-modal integration and reasoning (Zhou et al., 2025; Rocamonde et al., 2023). The central challenge in multimodal RL lies in effectively combining information from multiple modalities into a unified representation that can guide the selection of actions (Zhou et al., 2025). Recent research has made significant progress in addressing this challenge through various innovative approaches. Rocamonde et al. (2023) introduced zero-shot reward modeling (VLM-RMs) that leverages pretrained vision-language models such as CLIP as off-the-shelf reward functions, enabling RL agents to follow high-level natural language prompts without manual reward engineering. Further advances include the work of Li et al. (2024), which developed Diffusion Policy Gradient (DDiffPG) which combines unsupervised clustering with intrinsic motivation mechanisms and mode-

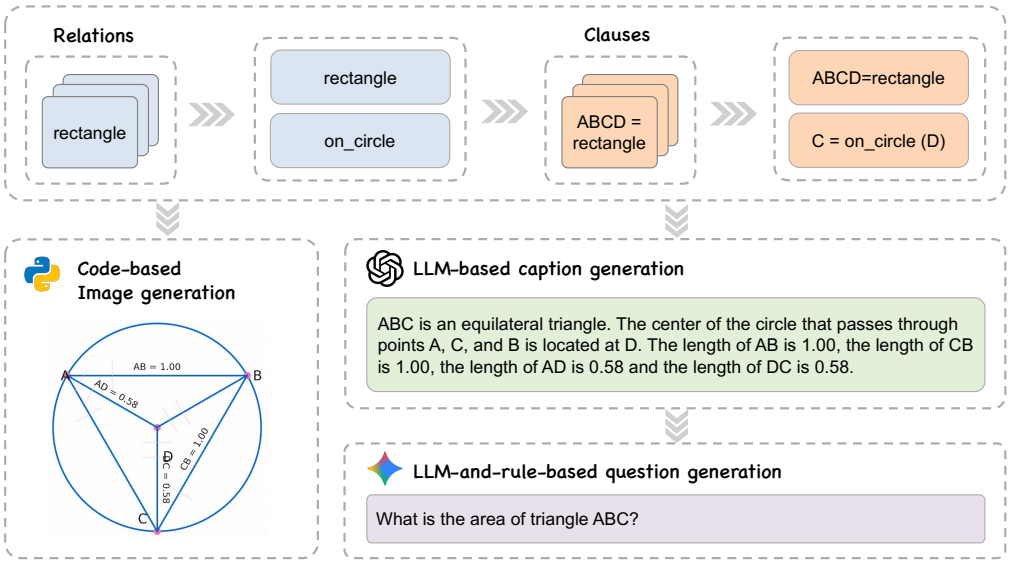

Figure 2: The geometry data synthesis pipeline, where a graph-based representation similar to AutoGeo (Huang et al., 2025b) is employed for generating the final geometry images. Our relation library contains more than 50 basic relations to cover

specific Q-learning techniques to prevent mode collapse while maintaining representational richness across modalities. Zhang et al. (2025) proposed StableReinforce for Multimodal Reward Models (R1-Reward), which reformulates reward modeling as a rule-based RL problem and introduces critical stability improvements in loss design and advantage estimation, yielding substantial performance gains on multimodal reward benchmarks.

Collectively, these methods demonstrate that incorporating RL strategies into multimodal training pipelines significantly enhances models' capacity to integrate and reason over diverse inputs modalities. Our work builds upon these insights by applying specialized RL techniques to the unique challenges of geometric reasoning, where precise cross-modal alignment is essential for effective problem-solving and comprehension.

3 METHODS

In this section, we introduce our Geo-Image-Textualization data generation process first, followed by our RAFT data engine framework used for refining the initial dataset.

3.1 SYNTHETIC DATA ENGINE

3.1.1 CLAUSES SAMPLING AND PROBLEM GENERATION

To generate the initial image-caption pairs, we adopt the generation procedures proposed in Alphy-Geometry. First, we use the Clause Generator to randomly sample relations based on complexity level. The complexity is defined as the number of definitions sampled; a higher number corresponds to greater complexity. Complexity is set to range from 1 to 3. A total of 45 distinct definitions are designed to ensure sufficient diversity in the generated image-caption pairs.

In the framework, *Relations* act as fundamental construction operations that systematically generate diverse yet semantically coherent geometric premises for subsequent theorem synthesis. Each relation (e.g., `angle_mirror`, `circumcenter`, etc.) encodes a precise geometric procedure—such as reflecting a point across an angle bisector or locating the circumcenter of a triangle. In addition to the construction rule, each definition maintains dependency metadata, specifying which primitive objects (points, lines, circles) and prior constructions it depends on. This enables the symbolic engine to get the minimal sets of premises required to derive a given theorem.

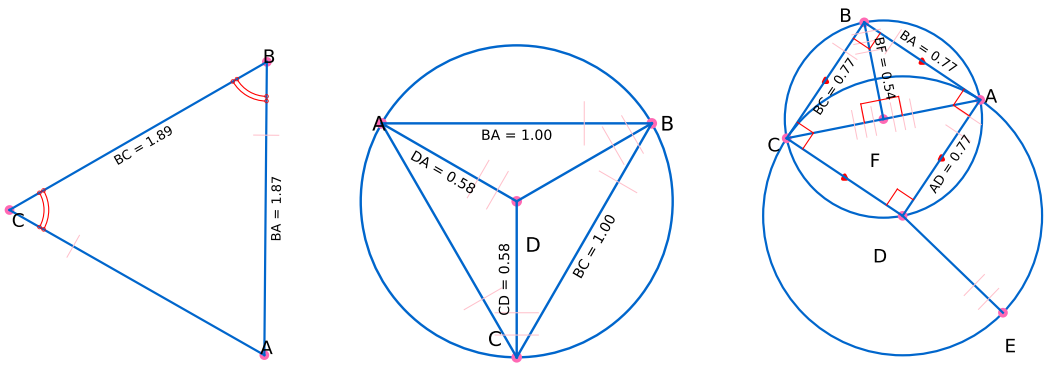

Figure 3:

I delete the original description of relations because it makes people confuse about relations and clauses.

After sampling relations, each relation is converted into a clause, with associated point variables. For instance, the relation `angle_mirror x a b c` denotes: given points `a`, `b`, and `c`, construct point `x` as the reflection of `c` across angle $\angle ABC$. Finally, the system constructs a graph-based representation in AutoGeo (Huang et al., 2025b) to model geometric problems. Each clause, corresponding to either a geometric construction or a relational assertion, is incorporated into the graph by instantiating nodes for geometric entities (e.g., points, lines, circles) and establishing edges that encode their interdependencies. Before adding each clause, the system verifies that it complies with a set of predefined geometric rules to ensure logical correctness.

3.1.2 Image-Caption Pair Generation

The geometric graph encodes all relevant entities, including points, lines, and circles, enabling the straightforward rendering of basic geometric elements, similar to AutoGeo. However, a fundamental limitation of AutoGeo is that the captions cannot be directly inferred from the rendered images because the visual content and the textual description are not semantically aligned. ~~For instance, as shown in Figure **??**, which is generated by AutoGeo, the caption "Point C is such that the triangle formed by connecting points A, B, and C is equilateral." cannot be visually verified from the image alone.~~ We argue that this misalignment constitutes a critical bottleneck in cross-modal reasoning.

To address this issue, we introduce a set of visual augmentation strategies that explicitly encode semantic relationships within the image, including:

1. **Segment Equality Representation:** We use short perpendicular ticks to indicate equal-length line segments. When multiple pairs of equal segments exist, we distinguish them using a different number of ticks (e.g., one tick, two ticks).
2. **Angle Annotations:** For angles that are integer multiples of 15° within the range $[15°, 165°]$, we directly annotate the angle values within the image.
3. **Parallel and Perpendicular Indicators:** Parallel lines are marked using matching directional triangles, and right angles are indicated using a small square symbol at the vertex.
4. **Equal Angle Representation:** Equal angles are denoted by marking them with the same number of arcs, consistent with conventional geometric notation.
5. **Intersection and Collinearity:** Dashed lines are used to explicitly indicate intersections and collinearity relationships among points or segments.

These symbolic augmentations ensure that every semantic relation mentioned in the caption is visually grounded in the image, thereby enabling faithful and interpretable cross-modal learning.

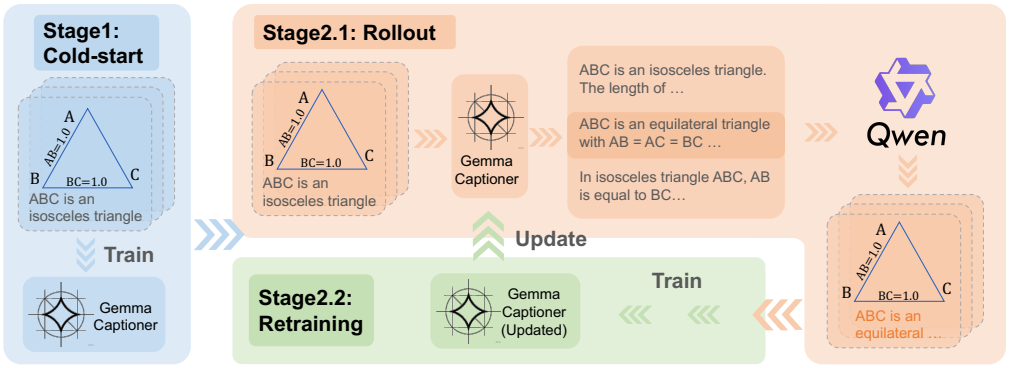

Figure 4: The workflow of Geo-Image-Textualization method.

3.1.3 Detailed Caption

For each clause in the symbolic representation, we apply a refined, rule-based template to convert it into natural language. These captions accurately describe the geometric diagram, including object relationships, special angle values, and other visual properties. Additionally, the captions explicitly state the lengths of specific line segments when such information is visually annotated in the image.By ensuring that all semantic content present in the image is mirrored in the caption, we achieve full cross-modal alignment.

Building upon these strategies, we construct a comprehensive image-caption generation engine. The engine is entirely rule-based, offering a fast, rigorous, reliable, and cost-effective solution for producing high-quality image-caption pairs—serving as a solid foundation for downstream multi-modal reasoning tasks.

3.2 RAFT Data Engine

Our proposed RAFT framework iteratively optimizes both the model and the dataset through a novel alternating paradigm. The method consists of two phases: (1) a cold-start phase to bootstrap initial captioning capabilities, and (2) a RAFT (Dong et al., 2023) optimization phase that cyclically refines the dataset and model via reinforcement learning. The overall framework is shown in Figure 4.

3.2.1 Cold-Start Phase

To initialize the model's ability to generate geometrically aligned captions, we first perform supervised fine-tuning (SFT) on the base vision language model using the GeoReasoning-10K dataset. This phase minimizes the standard cross-entropy loss:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(I,c^\star)\sim\mathcal{D}_\iota}[\log P_{\theta_0}(c|I)] \tag{1}$$

where I denotes an input geometric image, $c^\star$ and c indicate the ground-truth caption and the predicted caption respectively, and D_0 represents the initial dataset. The model parameter θ_0 is optimized to establish basic image-to-text mapping capabilities.

3.2.2 RAFT Optimization Phase

The RAFT optimization phase operates in alternating stages, as shown in Figure 4.

Rollout Experience Generation Suppose the current iteration is t. For each image I in the current dataset $\mathcal{D}_t$, we first generate N candidate captions $\{c_i\}(i = 1, 2, \cdots, N)$ using the current vision language model with parameter θ_t. Then, we utilize a specifically designed reward function $R(c_i, Q_i, c^\star)$ (detailed introduced in Section 3.2.3) to score each caption. Last, we retain the top-K caption $c_{\text{best}} = \arg\max_{c_i} R(c_i, Q_i, c^\star)$ to update the current dataset and construct the refined dataset $\mathcal{D}_{t+1}$.

Model retraining We update the model by training on $\mathcal{D}_{t+1}$ for one epoch, which is:

$$\theta_{t+1} = \arg\max_{\theta_t} \mathbb{E}_{(I, c_{\text{best}}) \sim \mathcal{D}_{t+1}} [\log P_{\theta_t}(c_{\text{best}}|I)] \tag{2}$$

This iterative process continues for $T = 5$ epochs, progressively enhancing both dataset quality and model performance.

3.2.3 REWARD FUNCTION

The composite reward $R(c, q, c^\star)$ balances task correctness and caption-image alignment, as shown in Eq. equation 3:

$$R(c, I) = \underbrace{w_r \cdot R_{\text{reasoning}}(c, q)}_{\text{Reasoning Reward}} + \underbrace{w_c \cdot R_{\text{caption}}(c, c^\star)}_{\text{Caption Reward}} \tag{3}$$

Reasoning reaward To evaluate a candidate caption's utility for solving downstream tasks, we leverage a frozen large language model of Qwen2.5-7B-Instruct (Yang et al., 2024) to generate an answer $a \sim P_{\text{LLM}}(a|q, a^\star, c)$ where $q, a^\star$ is the geometric question and its groundtruth answer generated by a reasoning model (GPT-o1) corresponding to the caption $c^\star$ in advance. As encouraged by mainstream RL process, we check both the format and correctness of the answer, which is:

$$R_{\text{Reasoning}} = s_c \cdot \mathbb{I}(a = a^\star) + s_f \cdot \mathbb{F}(a) \tag{4}$$

where $\mathbb{F}(\cdot)$ denotes the format checking function, and s_c, s_f indicate the weight of correctness and format, set as 0.9 and 0.1 in the experiments, respectively.

Caption reward To prevent reward sparsity during early training, we measure semantic relevance between c and the ground-truth caption $c^\star$ using ROUGE and BLEU-4, as shown in Eq. 5:

$$R_{\text{caption}} = \lambda_r \cdot ROUGE(c, c^\star) + \lambda_b \cdot BLEU(c, c^\star) \tag{5}$$

where λ_r, λ_b represent the weight of ROUGE score and BLEU-4 score, set as 0.7 and 0.3 in the experiments, respectively.

4 EXPERIMENTS

We conduct a comprehensive evaluation of our proposed framework. First we show the experimental setup in detail in Section 4.1. Section 4.2 illustrates the effectiveness of our Geo-Image-Textualization method. Section 4.4 evaluates the high quality of GeoReasoning-10K. And Section **??** shows case studies to explain the superiority of the proposed dataset.

4.1 EXPERIMENTAL SETUP

Our experiments utilize Gemma3-4B (Farabet & Warkentin, 2025), a commonly used lightweight multimodal architecture with strong mathematical reasoning capabilities as our base model. All the models are evaluated on MathVista (Lu et al., 2023) and MathVerse (Zhang et al., 2024a), two established mathematical reasoning benchmarks focusing on visual and mathematical problem-solving. The training and optimization pipeline contains two stages:

1. Cold start phase:
   In this stage we train each base model on the initial GeoReasoning-10K dataset for 1 epoch using standard supervised fine-tuning (SFT) with cross-entropy loss. The original learning rate of SFT is 1e-5, utilizing a cosine learning rate scheduler. We set 3% of the training steps as warm-up steps, where the learning rate grows linearly to the original learning rate 1e-5.
2. RAFT optimization phase:
   We conduct 5 refinement epochs, alternating between two sub-phases:
   - Caption Refinement: The current model generates 8 candidate captions for each image. Then we select the top 1 caption per image based on a composite reward considering both reasoning correctness and caption alignment. In the experiments, we set the weight of reasoning reward as 0.7 while that of caption reward as 0.3.
   - Model Retraining: Fine-tune the model on the updated dataset for 1 epoch using identical hyperparameters as the cold start phase.

4.2 EFFECTIVENESS OF GEO-IMAGE-TEXTUALIZATION METHOD

We validate the effectiveness of our Geo-Image-Textualization method through systematic experiments. Specifically, we implement the Cold Start and RAFT optimization pipelines on Gemma3-4B, generating refined models and datasets at each optimization stage. We then evaluate models of various stages on MathVista, a commonly used benchmark. The skills across all domains like geometry and arithmetic are summarized in Table. 1.

Table 1: Accuracy of Gemma3-4B models at various stages tested on MathVista

	baseline	**cold-start**	**raft-1**	**raft-2**	**raft-3**	**raft-4**	**raft-5**
all	46.2	47.6	48.7	48.1	49.2	49.0	**50.0**
geometry	60.7	62.3	63.2	64.0	63.6	60.3	**64.0**
arithmetic	42.5	45.0	44.8	45.3	45.9	**47.6**	46.5
algebraic	59.1	60.5	62.3	62.3	62.3	59.1	**63.3**
numeric	26.4	31.9	29.9	31.3	31.3	31.9	**31.9**

Table 2: Accuracy of Gemma3-4B models at various stages tested on MathVerse

	baseline	**cold-start**	**raft-1**	**raft-2**	**raft-3**	**raft-4**	**raft-5**
all	25.2	25.9	25.7	25.8	25.5	26.5	**27.4**
text dominant	32.0	35.5	35.1	35.2	35.1	36.5	**36.5**
text lite	25.9	27.4	28.2	**28.5**	27.4	26.6	26.3
vision intensive	24.0	24.8	24.4	24.4	23.1	26.1	**26.5**

It can be observed from Table 1 and Table 2 that the model after RAFT stages outperforms the base model across all domains. Specifically, the model achieves significant performance improvements across the arithmetic, algebraic, and numeric domains, with respective gains of 5.1%, 4.2%, and 5.5%. These results demonstrate the effectiveness of our approach in enhancing model performance as well as its generalization capability across different domains.

4.3 GENERALIZATION PROPERTY OF THE RAFTED MODEL AND GEOREASONING-10K

In this section, we evluate the generalization property of GeoReasoning-10K. Specifically, we evaluate the accuracy of baseline model and the RAFTed model on a commonly-used benchmark MMMU (Yue et al., 2023). The testing accuracy on all domains are shown in Table 3: Observed from Table 3,

Table 3: Accuracy of models evaluated on all domains of MMMU, where 'A&D', 'Busi', 'Sci', 'H&M', 'Human', 'Tech' are short for 'Art and Design', 'Business', 'Science', 'Health and Medicine', 'Humanities and Social Science', 'Tech and Engineering', respectively.

	Overall	**A&D**	**Busi**	**Sci**	**H&M**	**Human**	**Tech**
Gemma3-4B	42.1	50.0	**45.3**	34.0	44.0	**60.8**	29.0
Gemma3-4B-RAFT	**44.2**	**63.3**	40.0	**36.0**	**46.0**	60.0	**31.9**

the RAFTed model outperforms the baseline on most of the domains, indicating the generalization capacity of the model as well as the generalizaiton property of the proposed dataset.

We also conduct qualitative analyses on representative examples from MathVista to illustrate how our method enhances the model's ability across various domains. In particular, we select cases of different domains and compare the response of the RAFT optimized model and that of the base model. Figure 1 and Figure 5 present example cases from the geometry and algebraic domains, respectively. Other example cases of arithmetic and numeric domains are shown in Appendix B. In these cases, RAFT not only improves the reasoning ability on the geometry domain, but also generalize to other

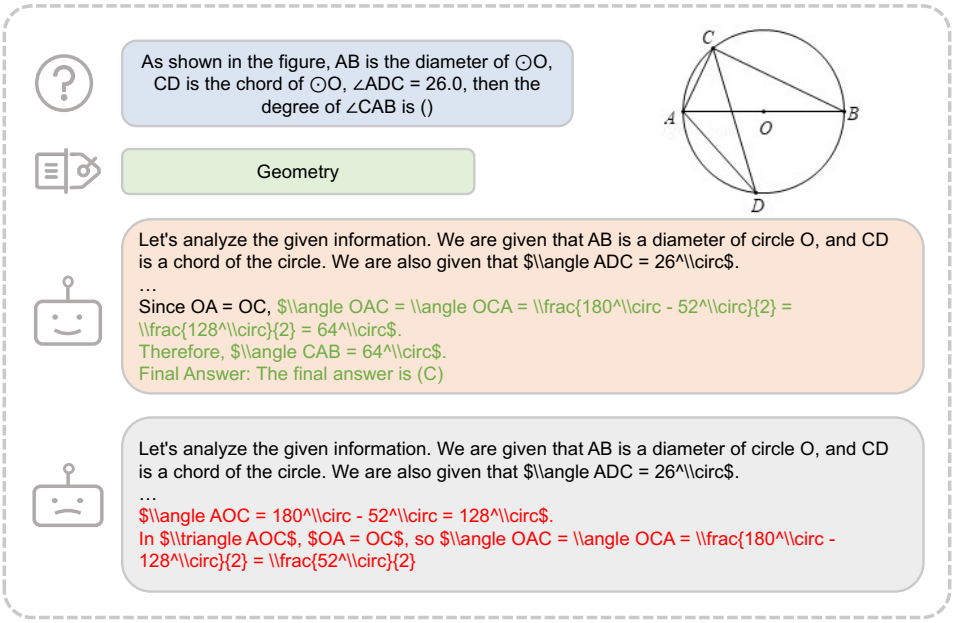

Figure 5: A geometric mathematical QA problem.

domains like arithmetic, algebraic, and numerical. This performance gain is hypothesized to stem from the mathematical reasoning abilities learned through the subtasks in geometric problems.

4.4 Quality Evaluation of GeoReasoning-10K

This section first tests the scaling property of our proposed dataset, then evaluates the data quality compared to other commonly-used datasets.

4.4.1 The Scaling Property

In this section, we test the scaling property of our proposed dataset. Specifically, we randomly sample 10%, 30%, 50%, 70%, and 100% subsets from the original dataset. Then we train the baseline models on these datasets, after which evaluating on various benchmarks. The testing accuracy of models trained on various sizes of datasets tested on MathVista and MathVerse are shown in Table 4 and Table 5, respectively.

Table 4: Accuracy of Gemma3-4B models trained with various sizes of datasets tested on MathVista

	0	**1k**	**3k**	**5k**	**7k**	**10k**
all	46.2	47.9	48.1	47.9	48.6	**48.9**
geometry	60.7	65.3	61.9	62.3	61.9	**67.4**
arithmetic	42.5	42.8	44.2	45.0	45.0	**45.6**
algebraic	59.1	62.3	60.9	61.2	60.5	**64.4**
numeric	26.4	25.7	29.2	29.2	29.2	**29.2**

As shown in Table 4, the model's performance across all domains of MathVista improves progressively with increasing dataset size. Notably, on larger datasets, the model achieves significant gains of 6.7% and 5.3% over the base model in the geometry and algebraic, respectively. Similarly, as presented in Table 5, the model's performance across most domains of MathVerse also exhibits a consistent improvement with increased training data. The positive correlation between model performance on different benchmarks and dataset scale demonstrates that our dataset possesses favorable scaling properties.

Table 5: Accuracy of Gemma3-4B models trained with various sizes of datasets tested on MathVerse

	0	1k	3k	5k	7k	10k
all	25.2	25.0	25.2	24.8	25.5	**26.3**
text dominant	32.0	33.2	32.9	32.3	32.8	**33.7**
text lite	24.0	24.8	25.3	24.7	26.7	**27.6**
vision intensive	24.0	23.4	23.8	23.3	24.1	**24.5**
vision dominant	24.0	23.1	24.1	23.9	24.7	**25.9**

4.4.2 Property for Reasoning Capacity Enhancement

In this section, we test the property for enhancing reasoning capacity of our proposed dataset, compared to other commonly-used datasets (AutoGeo (Huang et al., 2025b), GeoPeP (Sun et al., 2025), GeoGPT4 (Cai et al., 2024), and Geo170K (Gao et al., 2023)). Specifically, we randomly sample 10k samples of these baseline datasets, Then we train the baseline models (Gemma3-4B) on these datasets, after which evaluating on MathVista and MathVerse benchmarks. The testing accuracies are shown in Table 6. We also test the models trained on various sizes of AutoGeo and GeoPeP, with the results shown in Table 9 in Appendix C.

Table 6: Accuracy of Gemma3-4B models trained on 10k random samples of AutoGeo, GeoPeP, GeoGPT4, Geo170K, and our dataset.

	MathVerse	MathVista
AutoGeo	24.3	47.5
GeoPeP	24.0	48.2
GeoGPT4	24.5	46.9
Geo170K	23.7	48.1
ours	**26.3**	**48.9**

Observed from Table 6, the model trained on GeoReasoning-10K has better reasoning ability compared to that trained on other caption datasets and even outperforms other multimodal reasoning datasets. This observation demonstrates a more pronounced ability of GeoReasoning-10K to enhance the model's reasoning capabilities.

4.5 Ablation Study on Models

This section exhibit ablation studies on various training stages, and hyperparameters of the reward function.

4.5.1 Ablation Study on Training Stages

In this section, we conduct RAFT to the base model to test the effectiveness of RAFT method, as shown in Table 7:

Table 7: Accuracy of Gemma3-4B models of various stages evaluated on MathVista and MathVerse.

	MathVerse	MathVista
Gemma3-4B	25.2	46.2
Gemma3-4B-Coldstart	25.9	47.6
Gemma3-4B-RAFT	26.1	49.4
Gemma3-4B-Coldstart-RAFT	**27.4**	**50.0**

It can be concluded that RAFT is effective for both the base model and the one after cold-start, and cold-start is essential for enabling the model to acquire task-related capabilities.

4.5.2 Ablation Study on Hyperparameters

We evaluated the RAFTed models with various hyperparameters on MathVista and MathVerse, as shown in Table 8: As shown in Table 8, the reasoning reward plays an more important role in

Table 8: Accuracy of RAFTed models with various hyperparameters evaluated on MathVista and MathVerse, where w_r stands for the weight of reasoning reward.

	MathVerse	**MathVista**
w_r = 1	49.8	**27.5**
w_r=0.7	**50.0**	27.4
w_r=0	48.9	27.5

MathVista than MathVerse, indicating that the gain of genelization comes more from the helpness in solving the question other than comparison with captions.

In addition, it is observed in the result that the performance is not very sensitive to the selection of this hyperparameter, indicating the robustness of our RAFT method.

5 Conclusion

In this paper, we propose **Geo-Image-Textualization**, a novel reinforcement learning-based framework designed to generate high-quality, geometry-centered multimodal data. Leveraging this framework, we construct **GeoReasoning-10K**, a new dataset aimed at bridging the gap between visual and linguistic modalities in the geometry domain. Extensive experiments on the MathVista and MathVerse benchmarks demonstrate that our framework significantly enhances the cross-modal reasoning capabilities of MLLMs when fine-tuned on **GeoReasoning-10K**, with improvements generalizing to other domains, even with non-geometric input images. These results highlight the potential of rule-based generation for cross-modal and cross-domain learning, pointing to a promising direction for future research at the intersection of multimodal reasoning.

Limitations

Despite the fact that our data engine is capable of producing a rich and diverse variety of outputs—including detailed descriptions of a wide range of geometric primitives as well as the explicit representation of the relationships among these elements—this variety is inherently limited by the set of relations that have been predefined during the system's design phase. Consequently, while the engine is effective within the boundaries of its designed relational schema, it currently lacks the capacity to autonomously synthesize novel, creative, or out-of-scope relations that were not explicitly encoded beforehand. This limitation restricts the engine's ability to generate more complex, abstract, or unconstrained data that may require a broader, more flexible understanding of geometry or cross-domain reasoning. Overcoming this limitation and enabling the generation of such advanced data types remains an important direction for future research.

Acknowledgments

References

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

Shihao Cai, Keqin Bao, Hangyu Guo, Jizhi Zhang, Jun Song, and Bo Zheng. GeoGPT4V: Towards Geometric Multi-modal Large Language Models with Geometric Image Generation. arXiv e-prints, art. arXiv:2406.11503, June 2024. doi: 10.48550/arXiv.2406.11503.

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, 2024.

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

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and T. Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. ArXiv, abs/2304.06767, 2023. URL https://api.semanticscholar.org/CorpusID:258170300.

Clement Farabet and Tris Warkentin. Introducing gemma 3: The most capable model you can run on a single gpu or tpu. https://blog.google/technology/developers/gemma-3/, 2025.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. arXiv e-prints, art. arXiv:2312.11370, December 2023. doi: 10.48550/arXiv.2312.11370.

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

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749, 2025a.

Zihan Huang, Tao Wu, Wang Lin, Shengyu Zhang, Jingyuan Chen, and Fei Wu. Autogeo: Automating geometric image dataset creation for enhanced geometry understanding. IEEE Transactions on Multimedia, 2025b.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. arXiv preprint arXiv:2412.16720, 2024.

Zechu Li, Rickmer Krohn, Tao Chen, Anurag Ajay, Pulkit Agrawal, and Georgia Chalvatzaki. Learning multimodal behaviors from scratch with diffusion policy gradient. arXiv preprint arXiv:2406.00681, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

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. arXiv preprint arXiv:2105.04165, 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. arXiv preprint arXiv:2310.02255, 2023.

Yiting Lu, Jiakang Yuan, Zhen Li, Shitian Zhao, Qi Qin, Xinyue Li, Le Zhuo, Licheng Wen, Dongyang Liu, Yuewen Cao, et al. Omnicaptioner: One captioner to rule them all. arXiv preprint arXiv:2504.07089, 2025.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. arXiv preprint arXiv:2503.07536, 2025.

Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. Image textualization: An automatic framework for creating accurate and detailed image descriptions. arXiv preprint arXiv:2406.07502, 2024.

Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision–language models are zero-shot reward models for reinforcement learning. In NeurIPS, 2023.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems, 37:8612–8642, 2024.

Yanpeng Sun, Shan Zhang, Wei Tang, Aotian Chen, Piotr Koniusz, Kai Zou, Yuan Xue, and Anton van den Hengel. Mathglance: Multimodal large language models do not know where to look in mathematical diagrams. arXiv preprint arXiv:2503.20745, 2025.

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

Junxiao Wang, Ting Zhang, Heng Yu, Jingdong Wang, and Hua Huang. Magicgeo: Training-free text-guided geometric diagram generation. arXiv preprint arXiv:2502.13855, 2025.

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

Renqiu Xia, Mingsheng Li, Hancheng Ye, Wenjie Wu, Hongbin Zhou, Jiakang Yuan, Tianshuo Peng, Xinyu Cai, Xiangchao Yan, Bin Wang, et al. Geox: Geometric problem solving through unified formalized vision-language pre-training. arXiv preprint arXiv:2412.11863, 2024.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. arXiv preprint arXiv:2408.08872, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115, 2024.

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.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. arXiv e-prints, art. arXiv:2311.16502, November 2023. doi: 10.48550/arXiv.2311.16502.

Ming-Liang Zhang, Fei Yin, Yi-Han Hao, and Cheng-Lin Liu. Plane geometry diagram parsing. arXiv preprint arXiv:2205.09363, 2022.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In European Conference on Computer Vision, pp. 169–186. Springer, 2024a.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. arXiv preprint arXiv:2407.08739, 2024b.

Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, Haojie Ding, Jiankang Chen, Fan Yang, Zhang Zhang, Tingting Gao, and Liang Wang. R1-reward: Training multimodal reward models through stable reinforcement learning. arXiv preprint arXiv:2505.02835, 2025.

Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. arXiv preprint arXiv:2504.21277, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592, 2023.

A Technical Appendices and Supplementary Material

A.1 Experimental Setup

To ensure consistency, we adopt the official evaluation codebases of both MathVerse and MathVista, using the GPT-4o-mini API to evaluate the performance of our MLLM. Specifically, following each benchmark's official setup, we use GPT-4o-mini to extract and assess the correctness of answers for MathVerse, and to extract answers for MathVista.

We evaluate MLLMs on Geo, MathVerse and MathVista using A100 by vllm. We employ Qwen2.5-VL-3B-Instruct as our base model and fine-tune it on Georeasoning-10K using 4 L20 GPUs. The training process is distributed using torchrun with the DeepSpeed ZeRO-3 optimization strategy. The Hyperparameters are as follows:

A.2 Data Source

GeoReasoning-10K dataset is generated through rule-based methods and further refined using the RAFT framework. The reward question and the Geo benchmark are generated by DeepSeek-R1 with specific prompt.

A.3 License

- **GeoReasoning-10K** is released under the MIT License.
- **MathVerse** and **MathVista** are evaluated using their official codebases, which are publicly available under the Apache 2.0 License and MIT License, respectively.
- Our use of the **GPT-4o-mini API** for evaluation complies with OpenAI's API usage policies.
- All third-party datasets and models used in this work are under their respective licenses, and we ensure compliance with their terms of use.

B Case Studies

Additional cases that further demonstrate the generalization capability of our RAFT method are presented here. As shown in Figure 6 and Figure 7, these examples correspond to the arithmetic and numeric domains, respectively.

C Comparision to Other Datasets

We test the model performance trained on various sizes (1k, 10k, total 100k) of AutoGeo, GeoPeP, with the accuracy of MathVista and MathVerse shown in Table 9:

We can conclude from the Table 9 that GeoReasoning-10K demonstrates a more pronounced capacity to enhance models' reasoning ability and a much better scaling properties compared to other baselines.

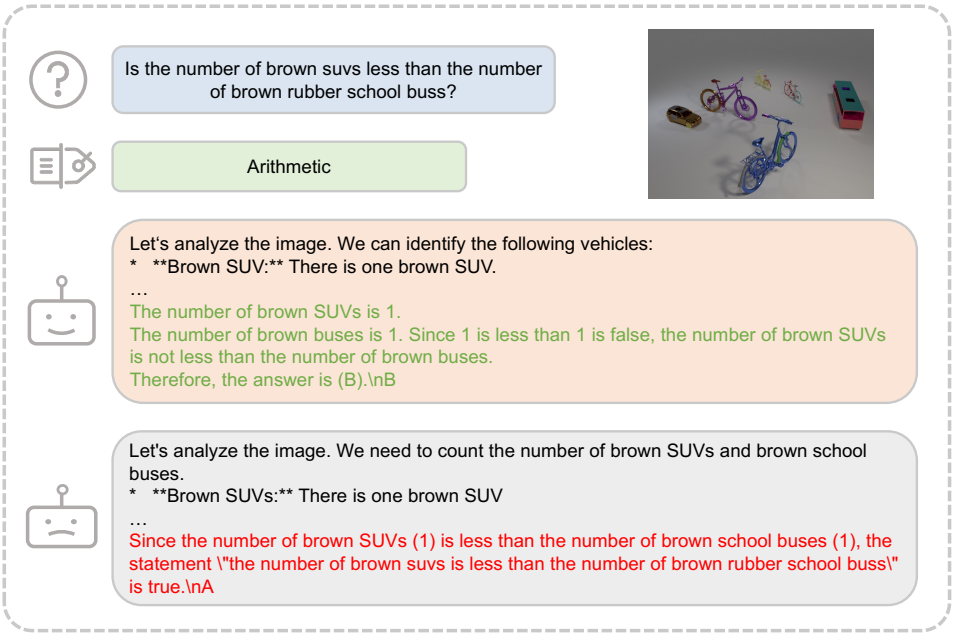

Figure 6: An arithmetic case.

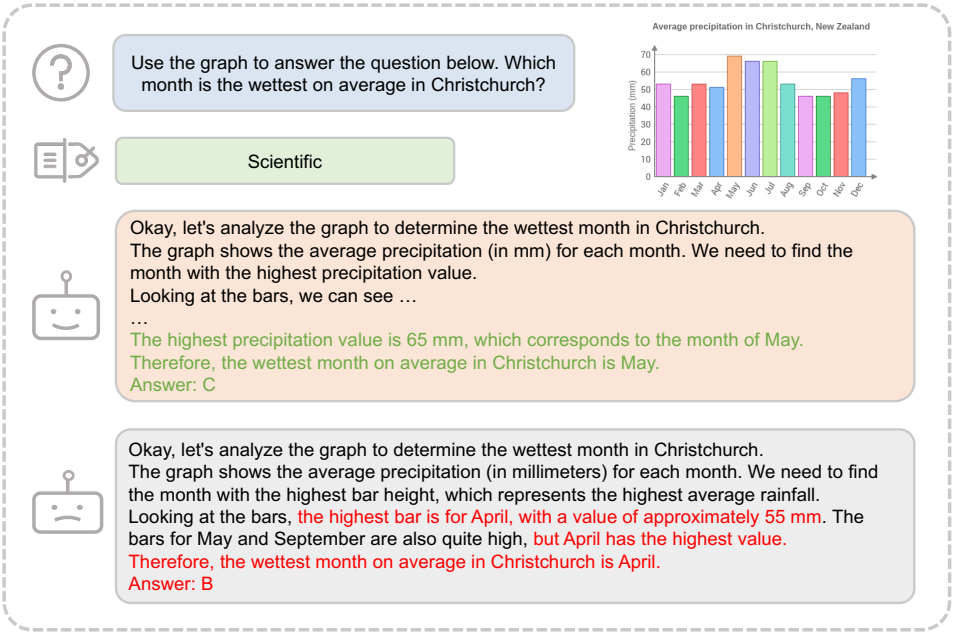

Figure 7: A numeric case.

D Broader Impacts

The provided dataset pipeline and the generated dataset contribute to enhancing the generalizable reasoning abilities of multimodal large language models (MLLMs). In narrow domains, they are particularly effective for improving the geometric problem-solving capabilities of MLLMs, while in broader domains, they support the development of mathematical reasoning skills applicable to everyday scenarios. As the dataset is limited to geometric mathematical problems, it is considered safe for release and is unlikely to pose direct negative social impacts.

Table 9: Accuracy of Gemma3-4B models trained on various sizes of AutoGeo, GeoPeP, and our dataset.

	MathVerse	**MathVista**
AutoGeo(1k)	23.9	47.6
AutoGeo(10k)	24.3	47.5
AutoGeo(100k)	24.7	46.1
GeoPeP(1k)	23.8	48.5
GeoPeP(10k)	24.0	48.2
GeoPeP(100k)	22.7	47.1
ours(1k)	25.0	47.9
ours(10k)	**26.3**	**48.9**