Describe-then-Reason: Improving Multimodal Mathematical Reasoning through Visual Comprehension Training

Mengzhao Jia[1], **Zhihan Zhang**[1], **Wenhao Yu**[2], **Fangkai Jiao**[3], **Meng Jiang**[1]
[1]University of Notre Dame [2]Tencent AI Seattle Lab [3]Nanyang Technological University
[1]{mjia2, zzhang23, mjiang2}@nd.edu
[2]wenhaowyu@global.tencent.com [3]fangkai002@e.ntu.edu.sg

Abstract

Open-source multimodal large language models (MLLMs) excel in various tasks involving textual and visual inputs but still struggle with complex multimodal mathematical reasoning, lagging behind proprietary models like GPT-4V(ision) and Gemini-Pro. Although fine-tuning with intermediate steps (*i.e.,* rationales) elicits some mathematical reasoning skills, the resulting models still fall short in visual comprehension due to inadequate visual-centric supervision, which leads to inaccurate interpretation of math figures. To address this issue, we propose a two-step training pipeline VCAR, which emphasizes the **V**isual **C**omprehension training in **A**ddition to mathematical **R**easoning learning. It first improves the visual comprehension ability of MLLMs through the visual description generation task, followed by another training step on generating rationales with the assistance of descriptions. Experimental results on two popular benchmarks demonstrate that VCAR substantially outperforms baseline methods solely relying on rationale supervision, especially on problems with high visual demands.

1 Introduction

Open-source multimodal large language models (MLLMs), exemplified by models such as LLaVA (Liu et al., 2023b) and Mini-GPT4 (Zhu et al., 2023), has showcased impressive reasoning capabilities across tasks involving both textual and visual inputs such as visual question answering (Nam et al., 2017; Guo et al., 2023) and multimodal dialogue (Li and Tajbakhsh, 2023). Despite these advancements, when addressing a more complex task of multimodal mathematical reasoning (Lu et al., 2023a), these open-source models are far lagged behind proprietary MLLMs like GPT-4V(ision) (OpenAI) and Gemini-Pro (Reid et al., 2024). Such performance gap in solving math-related multimodal problems hinders their application potential in fields such as educational content generation (Kasneci et al., 2023) and statistical data analysis (Yang et al., 2023a). Notably, specific deficiencies have been identified in these open-source models, including suboptimal performance in step-wise reasoning processes (Yang et al., 2023b; Mitra et al., 2023) and a tendency to overly rely on textual inputs while falling short in processing visual information (Zhang et al., 2024).

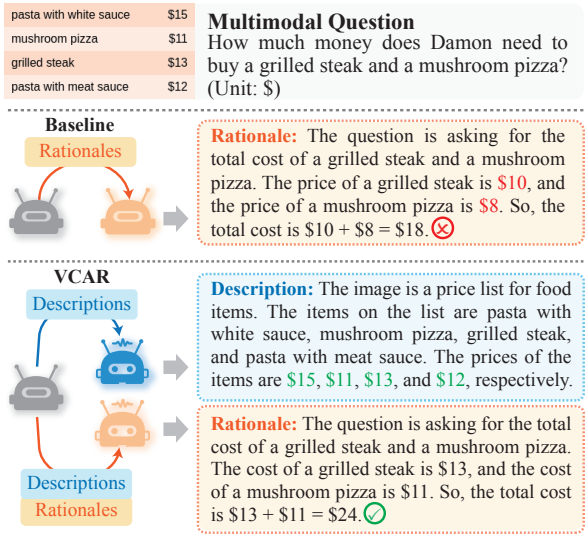

Figure 1: VCAR trains the model with an additional visual comprehension task in addition to the mathematical reasoning training, avoiding errors caused by inaccurate visual understanding. In contrast, the baseline method solely trained on rationales fails to correctly answer the question. It demonstrates the need for specified visual comprehension training.

Noticing this gap, great research interests have been dedicated to enhancing the multimodal reasoning ability of open-source MLLMs. A popular direction is step-wise supervised fine-tuning, which involves training with supervision signals of intermediate reasoning steps (*Rationales*) collected from language-based teacher models (*i.e.,* proprietary models such as GPT-4) (Wang et al., 2023; Zheng et al., 2023; Hu et al., 2023). Although the collected rationales help elicit some reasoning abil-

ity in the MLLM, they contain limited content related to the visual perception of math figures. Such inadequate supervision limits the model's ability to comprehend math figures effectively, and eventually leads to the performance being restricted by inaccurate visual comprehension such as the example in Figure 1. While certain visual information could be acquired through tools such as visual question answering (Zheng et al., 2023) and object detection (Hu et al., 2023), it is merely injected as supplementary to assist in composing coherent rationales, rather than being leveraged independently to teach models on how to recognize and understanding the visual content.

To address the aforementioned issue, we propose to improve the multimodal mathematical reasoning ability of MLLMs by emphasizing the importance of visual comprehension training. We introduce a novel two-step training pipeline that highlights **V**sual **C**omprehension training in **A**ddition to mathematical **R**easoning learning, dubbed as VCAR. Concretely, similar to the pre-training of MLLMs (Liu et al., 2023c) that utilizes image description for initial alignment, we employ an image description generation task to equip the MLLM with enhanced visual understanding capability, thereby generating high-quality descriptions to assist the subsequent development of mathematical reasoning abilities. Next, we train the MLLM to produce rationales based on the descriptions. With the text-form context provided by image descriptions, we isolate the training on mathematical reasoning skills from the demand of visual understanding. This two-step approach systematically facilitates both the visual comprehension and mathematical reasoning faculties of the MLLMs, ensuring a balanced development of multimodal mathematical reasoning capabilities.

To acquire supervision signals for the aforementioned two-step training, we utilize Gemini-Pro to collect: (1) descriptive contents, used for comprehending the image, and (2) rationales, focusing on reasoning the answer. Given the distinct objectives of each training step, we propose to optimize separate parameters tailored to each aspect to avoid imbalanced or disturbed learning. We adopt two Low-Rank Adaptation (LoRA) (Hu et al., 2022) modules to separately enhance visual understanding and mathematical reasoning abilities without the need for re-optimizing all model parameters.

To verify the effectiveness of our proposed VCAR training, we conduct experiments on two popular benchmarks MathVista (Lu et al., 2023a) and MathVerse (Zhang et al., 2024), where VCAR outperforms all baseline methods without explicit visual understanding training. Specifically, VCAR achieves significant improvements on math problems that demand high visual understanding ability, *i.e.*, 34.3% and 13.3% relative improvements on "visual-only" and "visual-dominant" categories on MathVerse. Such improvements are consistent across two base MLLMs, which demonstrates the universality of VCAR. Further ablation studies show that training in both visual comprehension and mathematical reasoning is crucial for performance improvement, and the two-step training pipeline is also indispensable in maximizing training efficacy. Our main contributions are as follows:

- We highlight the significance of supervising the visual comprehension capability in improving multimodal mathematical reasoning, yet often neglected by existing methods that concentrate on the training of reasoning capacity.
- We introduce a two-step training pipeline named VCAR, which separates the training for visual comprehension and mathematical reasoning skills.
- We conduct extensive experiments on two benchmarks, verifying the superiority of VCAR over various baselines. We also investigate the factors that contribute to the improvement. We released the codes and the collected supervision signals to facilitate other researchers in this community[1].

2 Related Work

2.1 Multimodal Large Language Models

In light of the recent flourishing of large language models (LLMs), integrating multimodal capabilities into LLMs to derive MLLMs has attracted intense attention (Zhang et al., 2023a; Alayrac et al., 2022). The architecture of contemporary MLLMs typically consists of a language foundation model (*e.g.,* LLaMA (Touvron et al., 2023) or Vicuna (Peng et al., 2023)), a visual encoder (*e.g.,* Vision Transformer (Dosovitskiy et al., 2021)), and a cross-modal alignment module. The development of MLLMs initiates with a cross-modal alignment training stage (Liu et al., 2023c), which utilizes paired image-description data to align the encoded

[1] https://anonymous.4open.science/r/Describe-then-Reason-0650.

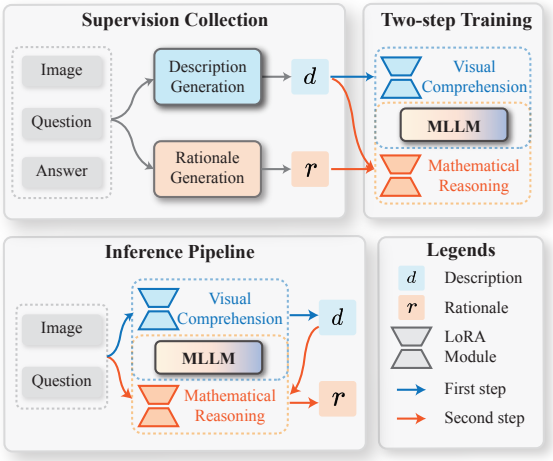

Figure 2: Illustration of the proposed method. VCAR consists of two components, namely supervision collection and two-step training. The inference pipeline of VCAR is also demonstrated.

visual features with the linguistic representation space of the LLM. This is followed by instruction tuning, where the model learns to interpret and execute instructions (Liu et al., 2023b). A prevalent method involves freezing the visual encoder's parameters during both training stages (McKinzie et al., 2024; Karamcheti et al., 2024). This practice, however, restricts the MLLMs' ability to fully grasp visual information (Guan et al., 2023). To mitigate this, our paper focuses on explicitly visual-centric supervision fine-tuning to specialize the training for visual comprehension capabilities.

2.2 Multimodal Reasoning

The boom of MLLMs has aroused researchers' interest in multimodal reasoning tasks (You et al., 2023; Chen et al., 2024; Surís et al., 2023). One popular direction is fine-tuning MLLMs with the supervision of intermediate reasoning steps (Rationales). Existing work in this area utilizes rationales either authored by humans (Zhang et al., 2023d) or collected from proprietary large language models (LLMs) (Lin et al., 2023; Chen and Feng, 2023; Lin et al., 2023), enhancing the model's reasoning capacity beyond direct answer learning. Notably, DD-CoT (Zheng et al., 2023) employs LLM to decompose the original question into sub-questions. Similarly, T-Sciq (Wang et al., 2023) and VPD (Hu et al., 2023) generate a question-resolving plan in natural language and programming code formats, respectively, thereby facilitate problem-solving. Despite these remarkable achievements, they primarily concentrate on supervising the training from a language-based reasoning aspect. Little effort is devoted to enhancing visual comprehension. This lack of visual comprehension supervision can lead to sub-optimal performance in overall multimodal reasoning capabilities.

3 Method

To tackle the issue of inadequate visual comprehension training in multimodal mathematical reasoning, we propose VCAR, a two-step training pipeline designed to enhance both visual comprehension and mathematical reasoning abilities. In the first step, MLLMs undergo training in the image description generation task to bolster visual comprehension. Subsequently, the second training step focuses on mathematical reasoning improvement through description-assisted rationale generation. In this section, we first introduce the process of collecting supervision signals for each learning objective, followed by the detailed two-step training pipeline.

3.1 Problem Formulation

Suppose that we have a set of N multimodal training samples $\mathcal{I} = \{x_i\}_{i=1}^{N}$. Each multimodal sample $x_i \in \mathcal{I}$ consists of an image v_i, a textual question q_i, and a gold answer y_i. We aim to train the model to learn a mapping function $\mathcal{F} : (v_i, q_i) \rightarrow y_i$. The index (*i.e.*, i) of each training sample is omitted for simplicity in the following sections.

3.2 Supervision Collection

The high cost of human annotation leads to an absence of fine-grained supervision signals such as rationales, which poses challenges to the training of reasoning ability. Noticing the advanced reasoning and sophisticated language generation capacity of proprietary LLMs or MLLMs, a prevalent practice is to collect the desired signals from them (Hsieh et al., 2023; Shridhar et al., 2023). Inspired by this, we gather supervision signals to carry out training from these proprietary MLLMs. Concretely, we employ the Gemini-Pro and GPT-4V(vision) models and design distinct prompts to acquire two types of supervision signals.

Rationale Generation. Utilizing rationales to assist the learning of mathematical reasoning has shown advantages over mere answer supervision in a series of studies (Liu et al., 2023a; Yue et al.,

2023). Motivated by this, we incorporate rationales into the training of MLLM's mathematical reasoning capability. A popular approach is to leverage prompts like "*Let's think step by step*" to elicit the generation of intermediate reasoning steps. However, the proprietary models sometimes fail to correctly answer certain questions, leading to rationales with erroneous and inaccurate expressions, thereby damaging the training (Lu et al., 2024). To alleviate this, we provide the ground truth answer when generating the rationale to enhance accuracy, similar to Li et al. (2022a). Concretely, we construct the rationale generation prompt for each sample as $p_r = (I_r, v, q, y)$, where I_r represents the rationale generation instruction. The detailed p_r is delineated below. By feeding p_r into the proprietary models, we get a rationale r for each sample.

> [v] Please analyze the question to generate a detailed rationale that leads to the correct answer.
> ### Question: [q]
> ### Answer: [y]

Description Generation. To impose explicit supervision on the MLLM's visual comprehension ability, we additionally construct a visual description d for each image v, besides the collection of the rationale r. One intuitive way to implement this is to directly prompt proprietary MLLMs to describe the given image, which, however, presents two primary drawbacks. Firstly, simply prompting MLLMs to exhaustively enumerate all details in the image often yields lengthy descriptions, thereby introducing unnecessary complexity for the description generation training and the following mathematical reasoning process. Besides, given the imperfect visual perception of MLLMs (Guan et al., 2023; Jing et al., 2023), generating the description without any references is likely to introduce erroneous information, which would also affect the subsequent learning process adversely.

To ameliorate these weaknesses, we include the corresponding question and the correct answer as guidance when acquiring the visual description, which not only emphasizes the description's relevance to the question but also avoids potential errors. In formulation, the visual description d is obtained using the prompt $p_d = (I_d, v, q, y)$, in which I_d is an instruction asking the model to describe the image regarding the current question and answer. Detailed prompt p_d is shown as follows:

> [v] Please describe the image only cover the information about the question and answer. Please only generate the description for the image.
> ### Question: [q]
> ### Answer: [y]

3.3 Two-step Training Pipeline

After acquiring the image descriptions and rationales, we can proceed with the training for both visual comprehension and mathematical reasoning abilities. Given the training sample (v, q, d, r, y), one straightforward approach is to train the model to generate both the description and rationale before reaching the final answer, leading to the loss function below:

$$\begin{aligned}\mathcal{L} &= -\log P(y, r, d|q, v) \\ &= -\log P(y|r, d, q, v)P(r|d, q, v)P(d|q, v).\end{aligned}$$

Benefiting from the auto-regressive decoding of LLMs, these conditional probabilities above can be calculated in a single forward pass. Nevertheless, we found that this joint optimization process leads to sub-optimal performance, which could be attributed to the imbalanced learning for visual comprehension and mathematical reasoning. Regarding this, we propose to separate the training process by modeling the two capabilities through independent parameters. In experiments of this work, we implement it by exploiting two sets of LoRA (Hu et al., 2022) modules.

Visual Comprehension Training. The first set of LoRA modules is trained to learn the visual comprehension capability through description generation, which, as mentioned above, is optimizing the following objective:

$$\mathcal{L}_d = -\log P\left(d|v, q; \theta_d\right),$$

where θ_d refers to the parameters of the LoRA module for visual comprehension.

Mathematical Reasoning Training. We utilize the rationale generation task to improve mathematical reasoning ability. The visual description d is integrated as a supplement, which provides clear text-form context, reducing dependence on visual comprehension in the mathematical reasoning training. We incorporate another set of LoRA modules with trainable parameters θ_r for optimization with the following objective:

$$\mathcal{L}_r = -\log P\left(r, y|v, q, d; \theta_r\right).$$

Inference. During the inference, we first enable θ_d to generate a targeted and precise visual description $\hat{d}$ for the question in the first step. Then we

Method	MathVista						MathVerse					
	ALL	FQA	GPS	MWP	TQA	VQA	ALL	TD	TL	VI	VD	VO
Proprietary MLLMs (Zero-shot)												
Gemini-Pro (Reid et al., 2024)	47.7	**48.3**	35.1	50.5	**65.8**	**42.5**	22.3	27.6	23.7	19.4	20.3	20.5
GPT-4V (OpenAI)	**49.9**	43.1	**50.5**	**57.5**	65.2	38.0	**38.3**	**52.1**	**40.9**	**34.9**	**33.6**	**29.8**
Open-source MLLMs (Zero-shot)												
LLaMA-Adapter-V2-7B (Gao et al., 2023)	23.9	21.2	**25.5**	11.3	32.3	31.8	5.7	6.2	5.9	6.1	4.2	6.1
TinyLLaVA-3B (Zhou et al., 2024)	26.7	20.4	19.7	**23.1**	44.9	31.8	**12.0**	**13.2**	**13.5**	**13.3**	**12.8**	7.0
LLaVA-1.5-7B (Liu et al., 2023b)	20.0	**22.7**	7.7	11.8	26.6	33.0	8.4	7.0	7.4	8.2	8.8	**10.8**
LLaVA-1.5-13B (Liu et al., 2023b)	**26.8**	19.7	20.2	18.8	**45.6**	**36.9**	7.6	8.8	7.6	7.4	7.4	6.9
Models Fine-tuned on LLaVA-1.5-7B												
Direct (question –> answer)	26.1	24.5	26.4	14.5	36.7	30.7	9.4	10.9	11.9	12.4	8.1	3.7
CoT (Hsieh et al., 2023)	25.8	21.9	23.1	21.5	**37.3**	29.1	7.3	7.6	6.5	7.1	7.1	8.4
CoT-T (filter out incorrect CoT)	27.5	26.0	26.9	21.0	36.1	29.6	13.7	15.7	13.7	14.0	14.5	10.5
CoT-GT (Li et al., 2022b)	28.8	27.1	31.2	22.6	31.6	**32.4**	14.0	14.6	15.5	15.5	15.0	9.3
T-SciQ (LM) (Wang et al., 2023)	24.6	26.8	14.4	15.1	36.7	**32.4**	6.8	7.6	4.9	5.3	5.6	10.5
T-SciQ (MM) (Wang et al., 2023)	27.0	28.6	14.9	26.3	36.1	31.3	4.9	6.3	5.6	4.8	4.6	3.2
VCAR (our method)	**33.7**	**30.9**	**34.6**	**38.7**	**37.3**	28.5	**16.2**	**17.1**	**16.8**	**16.0**	**17.0**	**14.1**
Models Fine-tuned on TinyLLaVA-3B												
Direct (question –> answer)	31.8	**30.1**	26.4	32.3	35.4	36.9	15.1	16.5	16.5	15.4	15.1	12.2
CoT (Hsieh et al., 2023)	30.3	20.8	29.3	32.3	38.6	36.3	6.2	5.5	4.9	4.7	5.1	10.7
CoT-T (filter out incorrect CoT)	30.0	26.4	23.1	30.6	40.5	33.5	14.2	15.6	14.5	15.4	15.2	10.2
CoT-GT (Li et al., 2022b)	32.7	26.0	32.2	36.0	37.3	35.8	16.6	**18.9**	16.4	17.5	17.0	13.0
T-SciQ (LM) (Wang et al., 2023)	29.6	29.0	30.3	19.9	36.7	33.5	13.1	15.4	12.4	11.8	12.2	**13.8**
T-SciQ (MM) (Wang et al., 2023)	31.0	26.0	25.5	31.2	**39.9**	36.9	15.1	17.5	15.5	15.4	14.1	12.8
VCAR (our method)	**34.6**	28.6	**35.6**	**34.4**	**39.9**	**38.0**	**16.7**	14.3	**17.9**	**18.5**	**19.7**	13.1

Table 1: Experimental results on MathVista and MathVerse. MathVista includes 5 types of math problems: figure question answering (FQA), geometry problem solving (GPS), math word problem (MWP), textbook question answering (TQA), and visual question answering (VQA). MathVerse problems are categorized according to the degree of information content in multi-modality, including text-dominant (TD), text-lite (TL), visual-intensive (VI), visual-dominant (VD), and visual-only (VO). Highest scores within each category are marked as **Bold**.

switch to θ_c for subsequent mathematical reasoning based on the question, the image, and the generated description.

4 Experiments

To verify the effectiveness of the proposed method, we conduct extensive experiments on two benchmarks. In this section, we first introduce the experimental settings, followed by the performance comparison and analysis of our method.

4.1 Setup

Data. We assess the effectiveness of our pipeline on two benchmarks, namely MathVista (Lu et al., 2023a) and MathVerse (Zhang et al., 2024). Both of them are multimodal benchmarks that focus on mathematical reasoning. We use the *minitest* split of the two benchmarks, which are composed of 1,000 and 3,940 testing samples, respectively. Since these benchmarks do not provide a corresponding training set, we construct a training set from five existing datasets, namely DVQA (Kafle et al., 2018), ChartQA (Masry et al., 2022), Geometry3K (Lu et al., 2021a), TabMWP (Lu et al., 2023b), and IconQA (Lu et al., 2021b). These datasets cover a wide range of domains (*e.g.,* chart plot, geometric graph, abstract diagram, and tabular image) in multimodal mathematical reasoning. We randomly select 1,000 samples from the training split of each dataset, resulting in a total of 5,000 training samples. The image description and rationales used in training (§3.2) are collected with Gemini-Pro (Reid et al., 2024) unless otherwise specified.

Implementation Details. We experiment with two foundation MLLMs: LLaVA-7B (Liu et al., 2023b) and TinyLLaVA-3.1B (Zhou et al., 2024). Due to constraints of computational resources, we train all models with LoRA (Hu et al., 2022) whose rank is set to 16 and 8 for LLaVA and TinyLLaVA, respectively. The learning rate is set to 2e-4 with a warmup during the first 3% steps. The batch size is set to 4 and 8 for LLaVA-7B and TinyLLaVA-3.1B, respectively. All models, if not specified, are trained for 1 epoch on the training set under the same hyperparameters.

Model	Method	MathVista						MathVerse					
		ALL	FQA	GPS	MWP	TQA	VQA	ALL	TD	TL	VI	VD	VO
LLaVA	**VCAR**	**33.7**	**30.9**	**34.6**	**38.7**	37.3	28.5	**16.2**	17.1	**16.8**	16.0	**17.0**	**14.1**
	w/o Rationale	30.3	29.4	29.3	28.0	41.8	25.1	14.8	16.4	16.1	**17.0**	15.4	9.3
	w/o Description	28.8	27.1	31.2	22.6	31.6	**32.4**	14.0	14.6	15.5	15.5	15.0	9.3
	Concatenation Training	32.5	29.7	31.2	28.5	**44.3**	31.8	15.4	**17.4**	16.5	15.2	16.9	10.8
	Multi-task Training	30.8	29.4	31.2	26.9	35.4	**32.4**	15.5	**17.4**	16.6	16.0	16.5	10.9
TinyLLaVA	**VCAR**	**34.6**	28.6	35.6	34.4	39.9	**38.0**	**16.7**	14.3	**17.9**	**18.5**	**19.7**	13.1
	w/o Rationale	34.1	**29.0**	**36.5**	**37.6**	38.0	31.8	**16.7**	16.9	16.9	18.0	16.1	**15.5**
	w/o Description	32.7	26.0	32.2	36.0	37.3	35.8	16.6	18.9	16.4	17.5	17.0	13.0
	Concatenation Training	33.9	28.3	34.6	33.9	39.9	36.3	15.9	17.1	17.8	16.9	17.0	10.7
	Multi-task Training	33.2	27.5	31.7	37.1	**41.1**	32.4	16.0	**19.2**	16.1	17.1	15.7	12.1

Table 2: Performance of the proposed method compared to different variants described in §4.3.

Evaluation. We use accuracy as the evaluation metric. All models are evaluated under a zero-shot generative setting: Given the question and the image, the model is prompted to generate the image description and then the rationale. We use greedy decoding to ensure deterministic results. For answer extraction, after the model generates the rationale, we additionally prompt it with "*So the final answer is*"[2] and use regular expressions to extract the answer from this final response.

Baselines. We compare VCAR with other common methods to construct the fine-tuning data.

- **Direct**: Training models to directly predict the answer without generating any rationales.
- **CoT** (Chain-of-Thought): Similar to Hsieh et al. (2023), given the image v and question q, we ask Gemini to produce a rationale r for model training. It is worth noting that Gemini-generated rationales may contain inaccuracies.
- **CoT-T**: Building on the CoT approach, this variant filters out instances where the final answer predicted in r does not match the gold answer.
- **CoT-GT**: To improve the accuracy of the constructed rationale, this variant includes the gold answer y when prompting Gemini to produce r. This variant employs the identical prompt as described in §3.2.
- **T-SciQ(LM)** and **T-SciQ(MM)** (Wang et al., 2023): These methods deconstruct the training sequence into three components: *skills* necessary to answer the question, a *plan* outlining the reasoning steps, and a *rationale* to derive the final answer. We test both the language model version[3] (LM) and the multimodal version (MM) of Gemini-Pro to gather data in the T-SciQ format.

[2]This additional turn is also included in model training.

[3]The original paper only used language models for data collection.

4.2 Main Results

Results on MathVista and MathVerse are presented in Table 1. We highlight the following findings:

VCAR significantly outperforms all baseline methods across both benchmarks. This underscores the effectiveness of VCAR in boosting MLLMs' abilities in multimodal mathematical reasoning. Moreover, the consistent enhancement observed across two base models proves the broad applicability of the VCAR approach.

VCAR shows marked improvements in categories that demand stronger visual understanding abilities. In MathVerse, examples in "visual-dominant" and "visual-only" categories require models predominantly leverage visual interpretation skills for solving problems. Here, VCAR shows an average of 14.6% relative improvements over the best baseline, including a 34.4% relative increase on LLaVA across the "visual-only" examples. This is in contrast to an average 5.3% improvement observed in text-dominant instances. These results indicate that VCAR enhances the visual comprehension abilities of MLLMs, leading to remarkable performance gains in scenarios requiring intensive visual perception.

Visual comprehension is a greater bottleneck than mathematical reasoning in multimodal math tasks. While the baseline method T-SciQ aims to improve model training by intricately designing a three-step format for the math reasoning rationale, VCAR outstrips T-SciQ across both benchmarks. This suggests that deficiencies in visual understanding play a critical role in limiting MLLMs' performance, which are not adequately addressed by solely refining the design of the rea-

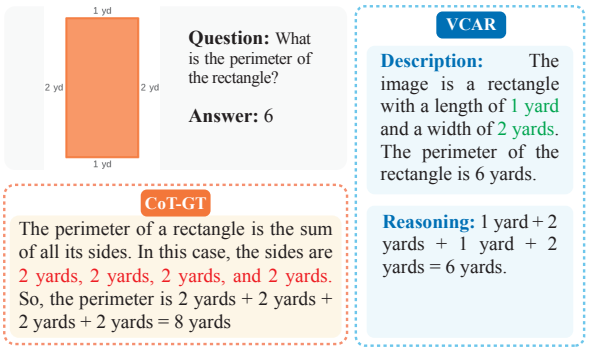

Figure 3: Inference results of VCAR and CoT-GT baseline on an example from MathVista. Correct and incorrect image descriptive expressions are highlighted in green and red, respectively.

soning rationale.

4.3 Ablation Study

To verify the effectiveness of each component in the proposed framework, we compare VCAR with the following variants.

- **w/o Rationale** and **w/o Description**: In these two variants, we ablate the rationale or the visual description in training and rely solely on the remaining element to derive the answer.
- **Concatenation Training**: Instead of the proposed two-step training, we concatenate the visual description and the reasoning rationale into one response $[d; r]$ to train the model.
- **Multi-task Training**: We train visual description generation and rationale generation simultaneously, using task-specific prompts.

As indicated in Table 2, experiments reveal that: 1) The exclusion of any component in VCAR results in reduced performance, underscoring that each component, including the description generation, the rationale generation, and the two-step training pipeline, is indispensable in terms of VCAR's significant improvement. 2) The omission of description notably impairs performance more than the absence of rationales, indicating that inadequate visual understanding is a critical bottleneck in multimodal mathematical reasoning. VCAR effectively addresses this issue through explicit visual description training.

4.4 Case Study

To get an intuition of VCAR's effect, we present an example from MathVista in Figure 3. The baseline method CoT-GT fails to accurately recognize the values depicted in the input image, leading to an incorrect calculation and answer (*2 yards + 2 yards + 2 yards + 2 yards = 8 yards*). In contrast, VCAR enhances the model's visual understanding, enabling it to describe key information in the image precisely. Consequently, the model accurately computes the correct answer (*1 yard + 2 yards + 1 yard + 2 yards = 6 yards*) to the posed question. This instance underscores the importance of accurate visual information in solving multimodal math problems. A similar pattern is observed in the example in Figure 1.

Data Source	Method	ALL	FQA	GPS	MWP	TQA	VQA
Gemini-Pro	CoT-GT	28.8	27.1	31.2	22.6	31.6	**32.4**
	VCAR	**33.7**	30.9	**34.6**	**38.7**	37.3	28.5
GPT-4V	CoT-GT	30.5	29.4	28.4	31.2	36.7	28.5
	VCAR	**33.1**	29.7	**29.3**	**39.8**	39.2	**30.2**

Table 3: VCAR and CoT-GT with different data sources on MathVista.

5 Analysis

5.1 Using GPT-4V to Collect Data

To demonstrate the universality of our method, we employ GPT-4V (*gpt-4-vision-preview*) to generate training data in addition to Gemini-Pro. We compare VCAR with the strongest baseline, CoT-GT. Table 3 presents the comparison results using two data sources. Our method consistently outperforms CoT-GT, irrespective of the data source—GPT-4V or Gemini-Pro. However, the model performance slightly declines compared to training with Gemini-generated data. One possible reason is that GPT-generated responses are longer and more complex, which poses challenges for smaller MLLMs (e.g., LLaVA-7B) to learn. Detailed data collection methods with GPT-4V and statistics comparisons of two data sources are provided in Appendix B.

5.2 The Format of Visual Information

Constructing image descriptions is an intuitive and direct way of converting visual content into textual modality (Radford et al., 2021; Yarom et al., 2024). To explore alternative visual information formats, we substitute the image description in VCAR crafting *visual-augmented questions*. Specifically, we prompt Gemini to rephrase the question by including pertinent visual details observed in the image. The prompt used for question augmentation is demonstrated in the Appendix A.1. As depicted in the "Question Augmentation" category in Figure 4, incorporating visual information into questions im-

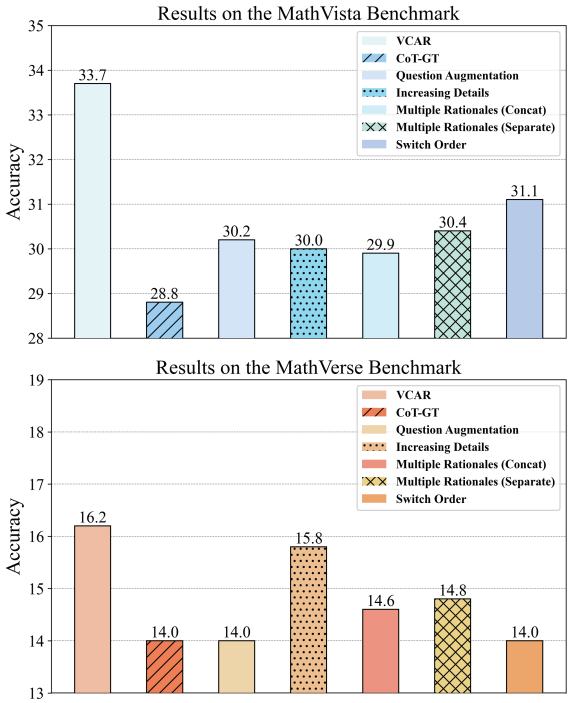

Figure 4: Performance comparisons of various variants on two benchmarks. We present the averaged accuracy. Detailed category-specific results are at Appendix C.

proves over the CoT-GT baseline. This once again validates that relying solely on rationales leads to insufficient visual understanding supervision. However, such question augmentation cannot match the improvement brought by VCAR. This discrepancy is likely due to the image description task being more consistent with the alignment objective during MLLM pre-training.

5.3 Impact of Training Sequence Length

To rule out the influence of increased sequence lengths due to the inclusion of visual descriptions, we develop variants that directly extend the length of reasoning rationales.

- **Increasing Details in Rationales**. We deliberately prompt Gemini-Pro to incorporate more details based on the previously derived rationale, resulting in a longer version. The specific prompt used is detailed in the Appendix A.2.

- **Multiple Rationales**. Previous research suggests that aggregating multiple rationales improves performance (Shridhar et al., 2023). Accordingly, we consider using two distinct rationales as another way to expand the number of training tokens. In practice, we prompt Gemini-Pro twice with a sampling temperature of 0.7. We explored two training configurations: (1) concatenating the rationales into a single training instance, and (2) pairing each rationale with the input question to create multiple training instances.

Figure 4 presents the comparative results of the aforementioned three variants, namely *Increasing Details*, *Multiple Rationales (Concat)*, and *Multiple Rationales (Separate)*. Notably, all three methods fail to yield the same performance as VCAR, which suggests that the improvements brought by VCAR are not merely due to the increased number of training tokens.

5.4 Order of the Training Pipeline

VCAR incorporates a sequential two-step training pipeline. Initially, the pipeline generates image descriptions, followed by the generation of rationales that derive the answer. To assess the necessity of this order, we introduce a variant where the order of the pipeline is reversed: the model first constructs rationales, subsequently generating image descriptions before deriving the final answer.

The "*Switch Order*" category results in Figure 4 illustrate that modifying the sequence order adversely impacts performance. This outcome is intuitive as the accurate visual details provided by the image descriptions can assist the rationale generation process. Altering the order obstructs the use of these descriptions in generating rationales.

6 Conclusion

In this work, we introduce VCAR, a two-step training pipeline designed to enhance multimodal mathematical reasoning via targeted visual comprehension training. VCAR enhances MLLMs with visual comprehension through image description generation training in the first step. Thereafter, descriptions are incorporated in the second rational generating step to assist in executing mathematical reasoning. We conduct comprehensive experiments across two widely used benchmarks, demonstrating that VCAR outperforms approaches without specified visual comprehension training. Notably, VCAR achieves significant performance improvements on problems requiring a high demand of visual understanding, underscoring the benefits of specialized visual comprehension training in multimodal mathematical reasoning.

Limitations

To our knowledge, this work has the following limitations:

- Due to our limited computational resources, we train all models, including the baselines, with LoRA to maintain a fair comparison. As a result, our two-step training pipeline is implemented with training separate LoRA modules. In future work, we plan to extend our approach to include experiments on full model fine-tuning.
- As mentioned in §4.1, our training data originate from the *training* splits of 5 published datasets. Meanwhile, a part of the *test* splits of these 5 datasets are included in the MathVista benchmark. We adopt the training splits of these datasets due to the scarcity of available training data in the field of multimodal math reasoning, while utilizing stronger models like GPT-4V to generate new large-scale data is beyond our budget. However, this setting remains equitable since all models are trained on the same training instances and there is no instance overlap between the training and test splits of these datasets. We leave the exploration of augmenting high-quality multimodal math data as future work.

Acknowledgements

This work was supported by NSF IIS-2119531, IIS-2137396, IIS-2142827, IIS-2234058, CCF-1901059, and ONR N00014-22-1-2507.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Feng Chen and Yujian Feng. 2023. Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction. *CoRR*, abs/2306.14122.

Jiaxing Chen, Yuxuan Liu, Dehu Li, Xiang An, Ziyong Feng, Yongle Zhao, and Yin Xie. 2024. Plug-and-play grounding of reasoning in multimodal large language models. *arXiv preprint arXiv:2403.19322*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*.

Yanyang Guo, Fangkai Jiao, Zhiqi Shen, Liqiang Nie, and Mohan S. Kankanhalli. 2023. UNK-VQA: A dataset and A probe into multi-modal large models' abstention ability. *CoRR*, abs/2310.10942.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8003–8017. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2023. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. *CoRR*, abs/2312.03052.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.

Kushal Kafle, Brian L. Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: understanding data visualizations via question answering. In *2018 IEEE*

Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 5648–5656. Computer Vision Foundation / IEEE Computer Society.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. *CoRR*, abs/2402.07865.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Shengzhi Li and Nima Tajbakhsh. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *CoRR*, abs/2308.03349.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022a. Explanations from large language models make small reasoners better. *CoRR*, abs/2210.06726.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022b. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 9114–9128. Association for Computational Linguistics.

Bingbin Liu, Sébastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. 2023a. Tinygsm: achieving >80% on gsm8k with small language models. *CoRR*, abs/2312.09241.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023a. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *CoRR*, abs/2310.02255.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021a. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 6774–6786. Association for Computational Linguistics.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Pan Lu, Liang Qiu, Jiaqi Chen, Tanglin Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021b. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Zimu Lu, Aojun Zhou, Houxing Ren, Ke Wang, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024. Mathgenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of llms. *CoRR*, abs/2402.16352.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2263–2279. Association for Computational Linguistics.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. 2024. MM1: methods, analysis & insights from multimodal LLM pre-training. *CoRR*, abs/2403.09611.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2023. Compositional chain-of-thought prompting for large multimodal models. *CoRR*, abs/2311.17076.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2156–2164. IEEE Computer Society.

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

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *CoRR*, abs/2304.03277.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7059–7073. Association for Computational Linguistics.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11854–11864. IEEE.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2023. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. *CoRR*, abs/2305.03453.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023a. Fingpt: Open-source financial large language models. *CoRR*, abs/2306.06031.

Kaiwen Yang, Tao Shen, Xinmei Tian, Xiubo Geng, Chongyang Tao, Dacheng Tao, and Tianyi Zhou. 2023b. Good questions help zero-shot image reasoning. *CoRR*, abs/2312.01598.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. 2024. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2023. IdealGPT: Iteratively decomposing vision and language reasoning via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11289–11303, Singapore. Association for Computational Linguistics.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *CoRR*, abs/2309.05653.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2023a. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *CoRR*, abs/2309.15112.

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

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2023b. PLUG: leveraging pivot language in cross-lingual instruction tuning. *Arxiv preprint 2311.08711*.

Zhihan Zhang, Shuohang Wang, Wenhao Yu, Yichong Xu, Dan Iter, Qingkai Zeng, Yang Liu, Chenguang Zhu, and Meng Jiang. 2023c. Auto-instruct: Automatic instruction generation and ranking for black-box language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023d. Multimodal chain-of-thought reasoning in language models. *CoRR*, abs/2302.00923.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Annual Conference on Neural Information Processing Systems*.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *CoRR*, abs/2402.14289.

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

Source	Type	Averatge #Token
Gemini-Pro	Description	48.92
	Rationale	55.39
GPT-4V(ision)	Description	138.85
	Rationale	101.57
GPT-4V(ision) (Disentangle)	Description	60.20
	Rationale	81.43

Table 4: Statistics of data from two sources. GPT-4V tends to produce relatively lengthy responses compared to Gemini-Pro, which places challenges for training. Using the disentangle prompt alleviates this phenomenon.

A Detailed Prompts

A.1 Question Augmentation Prompt

We provide the prompt used to augment the question (detailed in Section 5.2) as follows.

> [v]
> ### Question: [q]
> ### Answer: [y]
> Please rephrase the question by adding visual content of the image. The added visual information should be helpful for answering the question.

A.2 Increasing Detail Prompt

Based on a previously obtained rationale s, we use the following prompt to deliberately incorporate more details (detailed in Section 5.3) as follows.

> [v]
> ### Question: [q]
> ### Answer: [y]
> ### Rationale: [r]
> Please extend the rationale by adding more details.

B Statistics Comparison across Data Sources

We detail the supervision collection process with GPT-4V in Section 5.1. Notably, GPT-4V tends to produce lengthy responses that intermingle image descriptions with rationales. To collect targeted responses that solely focus on description or step-wise reasoning that is suitable for VCAR, we tailor the prompt by asking GPT-4V to disentangle these two parts into separate paragraphs. We illustrate the prompt as follows.

Method	MathVista						MathVerse					
	ALL	FQA	GPS	MWP	TQA	VQA	ALL	TD	TL	VI	VD	VO
VCAR	**33.7**	**30.9**	**34.6**	**38.7**	37.3	28.5	**16.2**	17.1	**16.8**	16.0	17.0	**14.1**
Variants												
Question Augmentation	30.2	29.7	31.7	24.2	37.3	29.1	14.0	14.3	14.9	14.1	16.4	10.2
Increase Details in Ratinales	30.0	29.7	25.5	22.6	39.9	**34.6**	15.8	15.9	16.0	**16.9**	**17.8**	12.6
Rationales Ensemble (Concat)	29.9	27.5	25.5	29.0	40.5	30.2	14.6	**17.3**	14.6	14.8	14.7	11.5
Rationales Ensemble (Separate)	30.4	29.7	28.4	31.7	36.1	27.4	14.8	15.7	15.2	14.0	16.9	12.1
Switch order	31.1	26.8	30.8	22.6	**45.6**	34.1	14.0	14.5	14.3	14.5	15.6	11.0

Table 5: Comparison among several variant methods. We highlight the best performances in **Bold**.

[v]
Question: [q]
Answer: [y]
Please first generate an image description that contains only the information that will help answer the question.
Then provide a rationale for this question and answer.
Your response should consist of two parts: <Image Description> and <Rationale>.

The statistics comparison between two data sources, namely Gemini-Pro and GPT-4V, is provided in Table 4

C Detailed Results of Variants

We present the category-specific results of the variants in Section 5 in Table 5.