Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated impressive reasoning capabilities, particularly in textual mathematical problem-solving. However, existing open-source image instruction fine-tuning datasets, containing limited question-answer pairs per image, do not fully exploit visual information to enhance the multimodal mathematical reasoning capabilities of Multimodal LLMs (MLLMs). To bridge this gap, we address the lack of high-quality, diverse multimodal mathematical datasets by collecting 40K high-quality images with question-answer pairs from 24 existing datasets and synthesizing 320K new pairs, creating the MathV360K dataset, which enhances both the breadth and depth of multimodal mathematical questions. We introduce Math-LLaVA [1], a LLaVA-1.5-based model fine-tuned with MathV360K. This novel approach significantly improves the multimodal mathematical reasoning capabilities of LLaVA-1.5, achieving a 19-point increase and comparable performance to GPT-4V on MathVista's ministest split. Furthermore, Math-LLaVA demonstrates enhanced generalizability, showing substantial improvements on the MMMU benchmark. Our research highlights the importance of dataset diversity and synthesis in advancing MLLMs' mathematical reasoning abilities.

1 Introduction

Motivation. Large language models (LLMs) exhibit impressive reasoning capabilities, drawing significant research interest in mathematical problem-solving in textual form (Wei et al., 2022; Wang et al., 2023; Bin et al., 2023; Luo et al., 2023; Yue et al., 2023b; Gou et al., 2023; Zhou et al., 2023). However, the task of multimodal mathematical reasoning (Lu et al., 2023) requires models to interpret diverse images and apply advanced reasoning skills. While open-source multimodal large language models (MLLMs) like LLaVA (Liu et al., 2023) and Mini-GPT4 (Zhu et al., 2023) perform well on visual question answering tasks (Guo et al., 2023), they fall short of proprietary MLLMs (OpenAI; Google) in solving complex mathematical problems involving visual content.

Two common approaches to enhance MLLMs' mathematical reasoning skills are prompt methods and fine-tuning methods. Prompt methods (Lu et al., 2023; Wang et al., 2024b) leverage MLLMs' latent abilities through carefully designed prompts, while fine-tuning methods (Wang et al., 2024a; Hu et al., 2023a; Zheng et al., 2023) adjust model parameters using reasoning data collected from real-world or synthetic data from advanced LLMs (e.g., GPT-4). However, existing open-source image instruction fine-tuning datasets (Lu et al., 2022b; Li et al., 2023; Lu et al., 2022a), which contain limited question-answer pairs per image, do not fully exploit visual information to enhance MLLMs' multimodal mathematical reasoning capabilities.

Research Objectives. To bridge this gap, we select 40K high-quality images with corresponding question-answer pairs from 24 pre-existing datasets. These images and queries span various subjects, including algebra, arithmetic, geometry, logic, numeric commonsense, science, and visual question answering. The selection criteria were based on image clarity and comprehension complexity. Additionally, we propose a pipeline to synthesize 320K new pairs based on the 40K images and seed inquiries.

Constructing such a dataset presents significant challenges, including selecting diverse and high-quality multimodal question-answer data and enhancing question diversity. Selecting suitable data involves filtering for image clarity and comprehension complexity, ensuring the dataset covers a wide range of mathematical concepts and question types. Enhancing question diversity requires synthesizing

[1]We will make our data and model publicly available.

new questions that probe different aspects of the images and involve multiple reasoning steps. To further improve model robustness and comprehension, we focus on enhancing logical consistency (Tascon-Morales et al., 2023) and the ability to understand underspecified language (Pezzelle, 2023).

Contributions. Using the selected 40K data, the fine-tuned LLaVA-1.5 model, named Math-LLaVA-DS, achieved a significant improvement of 10.6% on MathVista (Lu et al., 2023). To further enhance multimodal mathematical reasoning capabilities, we synthesized an additional 320K question-answer pairs based on the 40K images and seed questions, resulting in the MathV360K dataset. This comprehensive dataset, containing around 40K images and 360K question-answer pairs, significantly expands the coverage of multimodal mathematical reasoning. Fine-tuning LLaVA-1.5 with MathV360K, we developed Math-LLaVA, which outperforms the original LLaVA-1.5 by 19% on MathVista's minitest split. We also evaluated Math-LLaVA on MMMU (Yue et al., 2023a), demonstrating its improved generalizability.

2 Related Works

2.1 Multimodal Large Language Models

The advancement of LLMs has spurred significant research interest in vision-language interaction, particularly in integrating visual knowledge into LLMs. The CLIP series (Radford et al., 2021; Li et al., 2022) aligned visual and language modalities using contrastive learning on extensive image-text pairs. Recent studies increasingly use pre-training alignment and visual instruction tuning on LLMs for complex tasks like visual question answering and multimodal reasoning. MiniGPT-4 (Zhu et al., 2023) engages in image-text dialogues by aligning visual features with text. Similarly, models like LLaVA (Liu et al., 2023) and Instruct-BLIP (Dai et al., 2024) use learnable projectors or query embeddings to interact with visual features. These approaches aimed to leverage high-quality pre-training and fine-tuning data to comprehend complex instructions. Models like mPLUG-Owl (Ye et al., 2023), SPHINX (Lin et al., 2023b), and MiniCPM-V2 (Hu et al., 2024) introduced new grounding data types and modularization training to minimize hallucinations and enhance grounding abilities. Despite these advancements, MLLMs face challenges in solving multimodal mathematical problems using diagrams. Further exploration of the quality and format of image instructions is needed to improve the reasoning capabilities of MLLMs.

2.2 Multimodal Reasoning

The rapid development of MLLMs has advanced research on multimodal reasoning (Chen et al., 2024a; You et al., 2023). Augmenting the original question and answer text data in restricted domains to further fine-tune MLLMs is a popular approach. For raw answers, rationales were either generated by humans (Zhang et al., 2023c) or gathered from prominent LLMs (Wang et al., 2024a; Lin et al., 2023a; Chen and Feng, 2023). Additionally, VPD (Hu et al., 2023b) proposed expanding answers by converting programming code formats to natural language formats. For raw questions, DDCoT (Zheng et al., 2023) used LLMs to decompose the original questions into sub-questions. These methods, however, only utilize LLMs to target text-only data within restricted domains, neglecting to fully exploit the visual information in raw images for further enhancement. To evaluate the multimodal reasoning abilities of MLLMs more comprehensively, MathVista (Lu et al., 2023), which involves various types of mathematical reasoning and skills, and MMMU (Yue et al., 2023a), which encompasses multidisciplinary tasks, have been proposed. There is still significant room for improvement in open-source MLLMs.

3 Data Synthesis

Existing open-source image instruction fine-tuning datasets (Lu et al., 2022b; Li et al., 2023; Lu et al., 2022a), containing limited question-answer pairs per image, do not fully exploit visual information to enhance the multimodal mathematical reasoning capabilities of MLLMs. To address this, we propose MathV360K, a robust dataset synthesized based on the 40K selected images and seed question-answer pairs from multiple sub-domains. As shown in the left side of Figure 1, we first select 40K high-quality data points based on the image clarity and comprehension complexity from 24 open-source multimodal question-answering datasets. In the second step, illustrated in the top right of Figure 1, we attempt to fully mine the visual information of the images to generate additional questions. The data generation pipeline includes creating diverse new questions to fully exploit the visual information, more complex questions to fur-

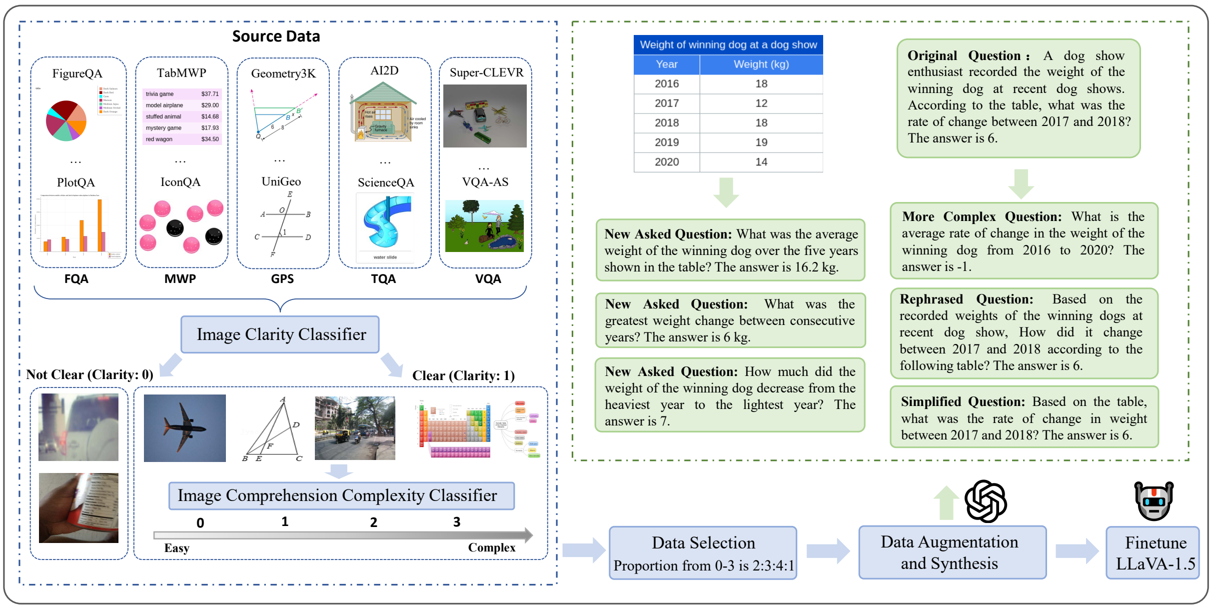


Figure 1: The overall flowchart of the proposed multimodal question-answer data selection and data augmentation. Our data selection depends on the fine-tuned ViT as image classifier. The data generation process depends on the vision-language model.

ther improve the reasoning capabilities, rephrased questions and underspecified questions to improve the robustness of the model. With the data generation pipeline, we collected 360K high-quality and diverse instruction-tuning data for the selected 40K data points to enhance the image understanding and mathematical reasoning capabilities of the LLaVA-1.5 open-source model.

3.1 Multimodal Reasoning Data Selection

3.1.1 Source Data

We collected 24 visual question answering and multimodal mathematical reasoning datasets, each targeting a specific task type and visual content. We focused on five problem task types requiring high-level reasoning to compile the source dataset: Figure Question Answering (FQA), Geometry Problem Solving (GPS), Math Word Problem (MWP), Textbook Question Answering (TQA), and Visual Question Answering (VQA). Table 5 in Appendix shows more details about the task type and visual context of each source dataset.

Each multimodal training sample consists of three components: an image I_i, a text question Q_i, and a ground-truth answer A_i. From this data format, the model aims to capture visual information and question semantics to reason the final answer.

3.1.2 Image Filtering and Proportioning

After acquiring the 24 source datasets, we intentionally selected data from the raw images based on the following criteria: (1) The clarity of the images, as poor-quality images introduce noise and interfere with learning image semantics; (2) The comprehension complexity of the images, which varies from easy to complex. By categorizing images into different levels of complexity and selecting proportionally, we can form a training set with an appropriate difficulty distribution; (3) The quality of the corresponding textual question data, ensuring that the difficulty aligns with the comprehension complexity of the images.

We fine-tuned two Vision Transformer (ViT) (Dosovitskiy et al., 2021) models to classify image clarity and image comprehension complexity, respectively. Due to the lack of annotated image data, we first sampled 10K images uniformly and randomly from the source datasets. These images were labeled for clarity and comprehension complexity using GPT-4V (OpenAI), with our designed prompt shown in Figure 2. For image clarity, label 0 indicates a blurred, poor-quality image, and label 1 indicates a clear, good-quality image. Image comprehension complexity is determined by the number of objects, their positional relationships, the need for mathematical calculations, detail level, texture, and material properties. Images are categorized into scores of 0, 1, 2, and 3, with lower values indicating easier visual context comprehension.

Based on the 10K annotated images, we trained two ViT models with initialized fully connected layers for classification using cross-entropy loss.

We first classified all source training dataset images using the fine-tuned image clarity classifier and filtered out images labeled as 0. Table 5 shows the number of images before (i.e., *Training Images*) and after (i.e., *Clear Images*) filtering.

Next, we used the image comprehension complexity classifier to score the filtered images. Table 5 shows that most images are classified as medium complexity, followed by easy, and finally the most complex. Considering that simple images are easier to learn from, while complex images are harder and require more reference samples, we sampled the first three complexity categories using a progressive scale from simple to complex. Since images with a score of 3 are the least abundant, we collected all of them. We selected 40K data points based on an overall ratio of complexity 2:3:4:1, ensuring samples from different complexities are uniformly selected from each source dataset. As a result, we obtained 40K high-quality (I, Q, A) real data points that are diverse in image information and questions are progressive in difficulty.

Prompt-Image Annotation:

[ROLE] You are an AI assistant to help me review the image.

[Task1] Your first task is to review the image and classify the clarity and quality of the given image into 0 or 1. **0 indicates that the image is not clear and of poor quality. 1 indicates that the image is clear enough and of high quality.** Your answer MUST be in the format: "The label is [0 or 1]".

[Task2] Your second task is to assess the complexity of the image. **Rate based on the number of objects in the image, their positional relationships, whether mathematical calculations are needed for understanding, detail level, texture and material properties.** The score ranges from 0 to 3, with higher scores indicating greater complexity. A score of 3 represents the highest complexity. Your answer MUST be in the format: "The ranking is [YOUR SCORE]".

Figure 2: The prompt template used in our GPT-4V API for image annotation. Image clarity is considered as binary classification and image comprehension complexity is viewd as multi-classification.

3.2 Data Augmentation

3.2.1 Mining Image for QA Generation

After selecting 40K multimodal reasoning data, we observed that each image typically corresponds to limited questions. As shown in the tabular image of Figure 1, the original question often focuses only on localized arithmetic differences. However, additional questions about overall averages, continuous variations, and more can also be asked, indicating that the visual information of an image is not fully exploited with just one question. Therefore, we can further augment the available real data by generating more question-answer pairs for each image.

We use GPT-4V to generate additional questions based on the input image and the original question. If questions are generated in a zero-shot manner, they often focus on one-sided visual scenes, lacking reasoning and mathematical skills. For images from specific tasks, such as geometric figures, more task-specific questions should be asked. Therefore, we adopt few-shot demonstrations for GPT-4V to generate new questions.

For an image belonging to one of the categories (FQA, GPS, MWP, TQA, VQA), we first internally cluster the questions into five classes for each source dataset within that task category. Specifically, features of text questions are obtained using TF-IDF and clustered using K-Means. As shown in Figure 4, we take IconQA as an example. After clustering the questions in the training set, each cluster internally represents a specific questioning format and pattern that can be referenced. Demonstrations are constructed by randomly sampling one question from each cluster of each source dataset belonging to a certain task type.

The prompt for generating new questions for an input image is shown in Figure 3. This method ensures that the newly generated questions are consistent with the distribution of the original reference questions while improving diversity. Using this approach, we generated 200K new question-answer pairs based on the selected 40K data points.

3.2.2 Augmentation of Original Question

We designed prompts to augment the original questions, as shown in Figure 5. Using GPT-4V, we generated 40K more complex questions, 40K simplified questions, and 40K rephrased questions. The augmentation focused on the following aspects:

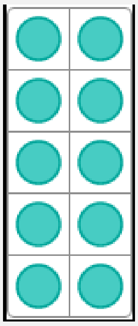
Complexity. More complex reasoning samples can enhance the reasoning capabilities of fine-tuned LLMs (Luo et al., 2023). Our first approach involves creating more complex questions based on the original image and corresponding inquiries.

Logical Consistency. Robust MLLMs should answer consistently about similar content in a given image (Tascon-Morales et al., 2023). We employed GPT-4V to ask the same question in different ways without changing the answer.

Underspecification. Robust MLLMs must deal with semantic underspecification, where the linguistic signal conveys only part of the necessary in-

Prompt-Ask New Questions:

Input Image



Q: How many dots are there?

[ROLE] You are an expert at understanding images and good at asking and answering questions based on the given images.

[TASK] You will be given some question examples. Please refer to the format of the examples to ask up to five high-quality questions on the given image. The original question of the image will also be given, please avoid asking the same question. Please provide the correct answer within ten words or answer with only an integer or float number.

[EXAMPLES]
From IconQA:
Question: There is 1 ball in the top row. How many balls are in the bottom row?
Question: What has been done to this letter? ...
From CLEVR-Math: ...
From TabMWP: ...
[ORIGINAL QUESTION] {Q}
[REQUIREMENT] Please follow and make full use of the image information. Please avoid asking questions for which you are not confident to give the definite correct answer. Please do not completely copy the content of the example questions. Ensure that provide final correct answer for each question.
[OUTPUT FORMAT] Your output MUST be "The question is [YOUR QUESTION]. The answer is [YOUR CORRECT ANSWER]."

Figure 3: The prompt template used in our GPT-4V API generates additional questions for each input image. Demonstrations are constructed by randomly sampling one question from each cluster of each source dataset belonging to a specific task type.

formation for successful communication (Pezzelle, 2023). Therefore, we simplified the original questions without affecting their semantic understanding when combined with the image.

4 Experiments

4.1 Model and Training

We employ the LLaVA-1.5 architecture as our base model, which primarily comprises the Vicuna-v1.5 language model (Team, 2023) and a pretrained Vision Transformer (ViT) as the image encoder. To preserve the foundational model's superior visual perception and descriptive abilities, we fine-tune LLaVA-1.5-13B using the proposed MathV360K instruction-tuning dataset. The diverse question patterns and rich visual content within this dataset enhance the model's multimodal mathematical reasoning capabilities while maintaining its general vision-language understanding skills.

4.2 Evaluation and Metrics

We evaluate our model using the minitest subset of MathVista (Lu et al., 2023) in a zero-shot manner. This minitest subset comprises 1,000 samples, including 540 multiple-choice questions and 460 questions that require free-form answers in the form of integers, floats, or lists. MathVista adequately assesses the MLLMs' multimodal mathematical skills, including algebraic reasoning (ALG), arithmetic reasoning (ARI), geometry reasoning (GEO), logical reasoning (LOG), numeric commonsense (NUM), scientific reasoning (SCI), and statistical reasoning (STA). Furthermore, MathVista questions can be categorized into the following subsets: FQA, GPS, MWP, TQA, and VQA. For evaluation, we first employ GPT-4 to extract the predicted choices or answers from responses, then report the answer accuracy, which entails determining whether the final answer matches the ground truth. Additionally, we evaluate our model's enhanced generalizability using the MMMU benchmark (Yue et al., 2023a). The MMMU benchmark, with 900 evaluation samples, encompasses six core disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Technology & Engineering, making it suitable for assessing the generalization of MLLMs' reasoning capabilities.

4.3 Implementation Details

We utilize GPT-4V (GPT-4 Vision Preview) for the data generation process. To classify image clarity and comprehension complexity, we fine-tune two ViT-Large-Patch16-224 models, each with a learning rate of 2e-4 and a training period of 5 epochs. For the LLaVA-1.5-13B model, the input image resolution is configured to 336 by 336 pixels. Both the projection linear layer and the language model

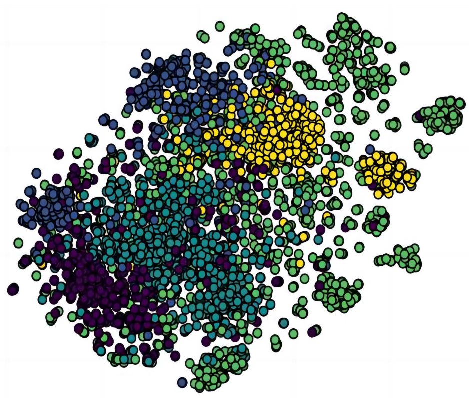


Figure 4: The visualization of the K-Means by T-SNE. We take IconQA as example. The questioning format of each cluster can be used as a reference to generate new questions for similar visual content.

are trainable. During the fine-tuning phase, we set a learning rate of 2e-5, employ a batch size of 16, and conduct fine-tuning over 2 epochs using A800 GPUs equipped with 80GB of memory.

5 Results and Analysis

5.1 Main Comparison on MathVista

We compare Math-LLaVA with other MLLMs on the minitest split of the MathVista benchmark in Table 1. As shown in the table, open-source MLLMs such as miniGPT4 (Zhu et al., 2023), instructBLIP (Dai et al., 2024), and LLaVA-1.5-13B have poor performance in multimodal mathematics, with overall accuracy lower than 30%. Compared to the base model, LLaVA-1.5-13B, with poor multimodal mathematical ability, Math-LLaVA achieves 46.6% overall accuracy with a significant improvement of 19%. More surprisingly, the proposed Math-LLaVA model outperforms close-source models Gemini 1.0 Pro (Team et al., 2023) and Claude 3 Haiku (Anthropic, 2024), even achieving comparable performance to GPT-4V (OpenAI), the most powerful close-source MLLMs. Interestingly, Math-LLaVA achieves 57.7% accuracy on GPS subset, outperforming G-LLaVA-13B (Gao et al., 2023), which has been trained on 170K high-quality geometric image-caption and question-answer pairs. The superior performance of Math-LLaVA indicates that the data selection and synthesis of high-quality, diverse multimodal question-answer pairs are effective in improving MLLM's multimodal mathematical reasoning capabilities.

Prompt-Complexity:
You will be given the question for the given image. Please ask a more complex question that requires more steps to answer than the given question.
Question: {Q}

Prompt-Logical Consistency:
You are an AI assistant to help me rephrase questions. Please ask the same question in a different way but have to make sure the answer won't be changed.
Question: {Q}
Rephrase the above question:

Prompt-Underspecification:
You are an AI assistant to help me rephrase question of the given image. Please simplify the question into a concise question, but does not affect the understanding and answering question with the image.
Question: {Q}
Simplify the above question:

Figure 5: The prompt templates used in our GPT-4V API to generate more complex, logically consistent and underspecified questions from original question text.

5.2 Generalizability of Math-LLaVA

The proposed Math-LLaVA model has demonstrated exceptional performance in multimodal mathematical reasoning tasks. To assess its generalization capability, we conduct evaluation experiments using the MMMU benchmark, which encompasses various disciplines and domains. The results are shown in Table 2. With only the selected data, Math-LLaVA has a performance drop on science subset. However, we can observe that the Math-LLaVA model fine-tuned on MathV360K can significantly outperforms the base model, LLaVA-1.5-13B, as well as several other open-source MLLMs on all six sub-domains. This superior performance underscores its capability to generalize to downstream multimodal understanding and reasoning tasks. Furthermore, the fine-tuning process using our synthetic data does not detract from the model's reasoning abilities in other domains; rather, it enhances its generalizability.

5.3 Overfitting to Synthesized Dataset

The proposed data synthesis pipeline generates additional question-answer pairs for each image to enhance the mathematical reasoning of MLLMs.

Model	MathVista												
	ALL	FQA	GPS	MWP	TQA	VQA	ALG	ARI	GEO	LOG	NUM	SCI	STA
Heuristics Baselines													
Random Chance	17.9	18.2	21.6	3.8	19.6	26.3	21.7	14.7	20.1	13.5	8.3	17.2	16.3
Frequent Guess (Lu et al., 2023)	26.3	22.7	34.1	20.4	31.0	24.6	33.1	18.7	31.4	24.3	19.4	32.0	20.9
Human	60.3	59.7	48.4	73.0	63.2	55.9	50.9	59.2	51.4	40.7	53.8	64.9	63.9
Close-Source Multimodal Large Langugae Models (MLLMs)													
Gemini 1.0 Nano 2 (Team et al., 2023)	30.6	28.6	23.6	30.6	41.8	31.8	27.1	29.8	26.8	10.8	20.8	40.2	33.5
Qwen-VL-Plus (Bai et al., 2023)	43.3	**54.6**	38.5	31.2	55.1	34.1	39.1	32.0	39.3	18.9	26.4	59.0	56.1
Gemini 1.0 Pro (Team et al., 2023)	45.2	47.6	40.4	39.2	61.4	**39.1**	45.2	38.8	41.0	10.8	**32.6**	54.9	**56.8**
Claude 3 Haiku (Anthropic, 2024)	46.4	-	-	-	-	-	-	-	-	-	-	-	-
GPT-4V (OpenAI)	**49.9**	43.1	**50.5**	**57.5**	**65.2**	38.0	**53.0**	**49.0**	**51.0**	**21.6**	20.1	**63.1**	55.8
Open-Source Multimodal Large Langugae Models (MLLMs)													
mPLUG-Owl-7B (Ye et al., 2023)	22.2	22.7	23.6	10.2	27.2	27.9	23.6	19.2	23.9	13.5	12.7	26.3	21.4
miniGPT4-7B (Zhu et al., 2023)	23.1	18.6	26.0	13.4	30.4	30.2	28.1	21.0	24.7	16.2	16.7	25.4	17.9
LLaVAR-13B (Zhang et al., 2023b)	25.2	21.9	25.0	16.7	34.8	30.7	24.2	22.1	23.0	13.5	15.3	42.6	21.9
InstructBLIP-7B (Dai et al., 2024)	25.3	23.1	20.7	18.3	32.3	35.2	21.8	27.1	20.7	18.9	20.4	33.0	23.1
LLaVA-13B (Liu et al., 2023)	26.1	26.8	29.3	16.1	32.3	26.3	27.3	20.1	28.8	24.3	18.3	37.3	25.1
SPHINX-V1-13B (Lin et al., 2023b)	27.5	23.4	23.1	21.5	39.9	34.1	25.6	28.1	23.4	16.2	17.4	40.2	23.6
LLaVA-1.5-13B (Liu et al., 2024)	27.6	-	-	-	-	-	-	-	-	-	-	-	-
LLaVA-1.5-13B† (Liu et al., 2024)	27.7	23.8	22.7	18.3	40.5	30.2	25.3	26.4	22.8	21.6	26.4	35.3	23.6
OmniLMM-12B (OpenBMB, 2024)	34.9	45.0	17.8	26.9	44.9	39.1	23.1	32.3	20.9	18.9	27.8	45.9	44.2
SPHINX-V2-13B (Lin et al., 2023b)	36.7	**54.6**	16.4	23.1	41.8	**43.0**	20.6	33.4	17.6	**24.3**	21.5	43.4	**51.5**
G-LLaVA-13B (Gao et al., 2023)	-	-	56.7	-	-	-	-	-	-	-	-	-	-
Math-LLaVA-DS	38.2	33.5	47.2	41.4	36.7	34.6	38.4	34.3	45.6	18.9	33.3	45.9	35.2
Math-LLaVA	**46.6**	37.2	**57.7**	**56.5**	**51.3**	33.5	**53**	**40.2**	**56.5**	16.2	**33.3**	**49.2**	43.9

Table 1: Comparison with baselines on the testmini set of MathVista benchmark. Baseline results are obtained from Lu et al. (2023). † represents our reproduced results of LLaVA-1.5-13B. The best results in both the close-source and open-source MLLMs are in bold. MathVista is divided in two ways: task type or mathematical skill, and we report the accuracy under each subset.

Intuitively, we should investigate whether the proposed model, Math-LLaVA, is overfitting on the generated question-answer pairs. If overfitting occurs, Math-LLaVA might memorize or retrieve image information without requiring any visual input. To examine this, we compare the performance of Math-LLaVA before and after data synthesis, referred to as Math-LLaVA-DS and Math-LLaVA, respectively, on MathVista using text inputs only. As shown in Table 3, Math-LLaVA exhibits similar performance, approximately 32.0%, as Math-LLaVA-DS on MathVista when inference is performed without any visual information. Furthermore, fine-tuning Math-LLaVA with only text data also yields similar observations. This indicates that the Math-LLaVA model is not overfitting on the synthesized question-answer pairs.

Interestingly, we also observe that with text-only input, LLaVA-1.5-13B achieves an accuracy of 23.3% on MathVista. Potential reasons for this, as explored in (Chen et al., 2024b), could be that visual content is unnecessary for many samples in MathVista and that unintentional data leakage may occur during the pre-training of LLMs and MLLMs.

5.4 Effectiveness of Synthesis

To verify the effectiveness of data selection and the proposed data augmentation strategies, we conduct experiments on various components of MathV360K independently. Initially, we fine-tune the LLaVA-1.5 model on 40K randomly sampled data points from the source dataset, without any selection, to demonstrate the efficacy of data filtering and proportioning. Subsequently, we separately combine the selected 40K data points with the generated data using four augmentation methods: mining images for QA generation (AskImg), posing complex questions (CompQ), rephrasing questions for logical consistency (RephQ), and simplifying questions for underspecification (SimpQ). Table 4 presents the accuracy achieved by different combinations of augmentations on MathVista. The results

Model	MMMU	Art & Design	Business	Sci.	Health & Med.	Human. & Social Sci.	Tech. & Eng.
Random Chance	22.1	29.2	24.7	18.0	20.7	20.0	21.4
Frequent Guess	26.8	23.3	**29.3**	27.3	30.0	25.8	24.8
miniGPT4-7B	26.8	29.2	21.3	28.7	30.7	29.2	23.8
mPLUG-Owl-7B	32.7	45.8	24.7	22.7	32.0	45.8	31.0
SPHINX-13B	32.9	48.3	24.7	26.7	30.7	50.0	26.2
InstructBLIP-7B	32.9	40.0	28.0	**32.7**	28.7	47.5	27.1
LLaVA-1.5-13B	36.4	51.7	22.7	29.3	38.7	53.3	31.4
Math-LLaVA-DS	36.9	**55.0**	24.7	23.3	38.7	56.7	32.4
Math-LLaVA	**38.3**	53.3	24.7	30.7	**38.7**	**58.3**	**33.3**

Table 2: Comparison with baselines on the MMMU benchmark.

Model	Training	Inference	MathVista
LLaVA-1.5-13B	Image-Text	Text	23.3
Math-LLaVA-DS	Image-Text	Text	32.2
Math-LLaVA	Image-Text	Text	32.4
Math-LLaVA-DS	Text	Text	32.1
Math-LLaVA	Text	Text	32.5

Table 3: Results of inference using only text of MathVista as input. Fine-tuning LLaVA-1.5 using image-text or text-only data.

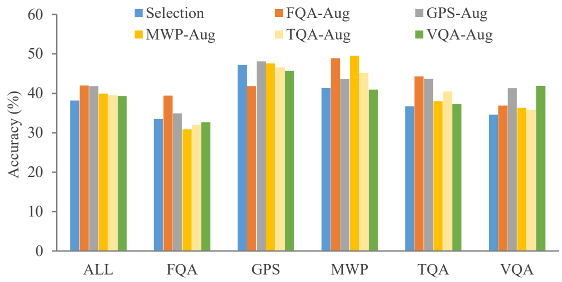


Figure 6: Accuracy on MathVista by augmentation for each task type.

indicate that our data synthesis approach, which incorporates data selection and each augmentation method, yields better performance. Collectively, these strategies result in a significant 11% improvement over randomly sampling 40K data points.

Select	AskImg	CompQ	RephQ	SimpQ	ALL
✗	✗	✗	✗	✗	35.6
✓	✗	✗	✗	✗	38.2
✓	✓	✗	✗	✗	42.2
✓	✗	✓	✗	✗	39.8
✓	✗	✗	✓	✗	40.9
✓	✗	✗	✗	✓	41.1
✓	✓	✓	✓	✓	**46.6**

Table 4: Effectiveness of data selection and different data augmentation strategies on MathVista.

5.5 Enhancements from Augmentation of Each Task Type

Given that we selected data from five different question-answering task types, our aim is to investigate which types or skills in multimodal mathematical reasoning could be enhanced by augmenting the source data from each individual task category. To this end, we conduct experiments with newly synthesized data for each task type, mixed with selected data. The results on MathVista are presented in Figure 6. We observe that augmentation of various types of source data can further improve the model's performance on the corresponding tasks. The enhancements are particularly pronounced for tasks involving FQA, MWP, and VQA. Interestingly, data augmentation for a single task type also shows improvements in effectiveness for other task types, likely due to the overlap in reasoning skills required across different tasks.

6 Conclusions

We addressed the shortage of high-quality and diverse multimodal mathematical training datasets by creating MathV360K, which consists of 40K high-quality multimodal questions and answers from 24 existing datasets, along with 320K newly synthesized question-answer pairs. This comprehensive dataset enhances both the breadth and depth of multimodal mathematical questions. Using MathV360K, we fine-tuned Math-LLaVA, significantly improving its capability in multimodal mathematical reasoning, outperforming LLaVA-1.5 by 19 points on the minitest split of MathVista. Additionally, Math-LLaVA was validated on the MMMU benchmark, demonstrating its generalizability. Our research underscores the importance of dataset diversity and synthesis in enhancing the mathematical reasoning abilities of MLLMs.

7 Limitations

The data we selected and synthesized are in the format of images, questions, and answers, lacking intermediate steps that could be further improved. In future work, we will introduce annotated intermediate steps and rationale to construct more comprehensive and high-quality datasets to further enhance the MLLMs's multimodal reasoning capability.

8 Ethics Statement

We do not envision that our work will result in any harm as defined in the ACL ethics policy. LLaVA-1.5 base model uses LLaMA 2 Community License and ViT-Large-Patch16-224 uses Apache License 2.0. For datasets, GEOS, A-OKVQA and MMMU use Apache License 2.0. Geometry3K, FigureQA and PMC-VQA use MIT License. Super-CLEVR uses BSD License. ChartQA uses GPL 3.0 License. GeoQA+, UniGeo and DocVQA are publicly available for research purposes. The rest of the dataset use permissive Creative Commons Licenses. The intended use of these source datasets and evaluation datasets is to train and test the model's multimodal reasoning capability, which is consistent with our utilization of all these data. Our proposed MathV360K can improve the multimodal mathematical reasoning ability of the open-source LLaVA-1.5 through training. We will make our data and model publicly available.

References

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Claude-3 Model Card.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. CoRR, abs/2308.12966.

Yi Bin, Mengqun Han, Wenhao Shi, Lei Wang, Yang Yang, See-Kiong Ng, and Heng Tao Shen. 2023. Non-autoregressive math word problem solver with unified tree structure. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 3290–3301.

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

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022. Mapqa: A dataset for question answering on choropleth maps. arXiv preprint arXiv:2211.08545.

Feng Chen and Yujian Feng. 2023. Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction. arXiv preprint arXiv:2306.14122.

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

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024b. Are we on the right way for evaluating large vision-language models? arXiv preprint arXiv:2403.20330.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems, 36.

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.

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

Google. Gemini. https://gemini.google.com.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Tora: A tool-integrated reasoning agent for mathematical problem solving. CoRR, abs/2309.17452.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6904–6913.

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.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3608–3617.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. arXiv preprint arXiv:2404.06395.

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

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

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. arXiv preprint arXiv:1710.07300.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, pages 235–251.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In Proceedings of the IEEE Conference on Computer Vision and Pattern recognition, pages 4999–5007.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. Scientific data, 5(1):1–10.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International conference on machine learning, pages 12888–12900.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. 2023. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14963–14973.

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

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. 2023b. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. arXiv preprint arXiv:2208.05358.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In Advances in Neural Information Processing Systems.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating 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 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 6774–6786.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems, 35:2507–2521.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. arXiv preprint arXiv:2209.14610.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony 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. arXiv preprint arXiv:2110.13214.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. CoRR, abs/2308.09583.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq 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, pages 2263–2279.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1697–1706.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1527–1536.

OpenAI. Gpt-4v(ision). https://openai.com/research/gpt-4v-system-card.

OpenBMB. 2024. Large multi-modal models for strong performance and efficient deployment. https://github.com/OpenBMB/OmniLMM.

Sandro Pezzelle. 2023. Dealing with semantic underspecification in multimodal nlp. arXiv preprint arXiv:2306.05240.

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.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In European Conference on Computer Vision, pages 146–162.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1466–1476.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8317–8326.

Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. 2023. Logical implications for visual question answering consistency. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6725–6735.

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

The Vicuna Team. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2024a. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 19162–19170.

Lei Wang, Wanyu Xu, Zhiqiang Hu, Yihuai Lan, Shan Dong, Hao Wang, Roy Ka-Wei Lee, and Ee-Peng Lim. 2024b. All in a single image: Large multimodal models are in-image learners. arXiv preprint arXiv:2402.17971.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In The Eleventh International Conference on Learning Representations.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023.

mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A. 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, pages 11289–11303.

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. 2023a. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. CoRR, abs/2311.16502.

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

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023b. Llavar: Enhanced visual instruction tuning for text-rich image understanding. arXiv preprint arXiv:2306.17107.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023c. Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv: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. Advances in Neural Information Processing Systems, 36:5168–5191.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In The Eleventh International Conference on Learning Representations.

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.

A Appendix

A.1 Source Data Statistics

We collected 24 visual question answering and multimodal mathematical reasoning datasets, each targeting a specific task type and visual content. We focused on five problem task types to compile the source dataset: Figure Question Answering (FQA), which involves analyzing charts and plots statistically; Geometry Problem Solving (GPS), which involves solving geometrical problems with diagrams and figures; Math Word Problem (MWP), which involves arithmetic calculations within the context of images; Textbook Question Answering (TQA), where reasoning is based on scientific knowledge and figures; and Visual Question Answering (VQA), which requires reasoning about objects, scenes, or relationships within images. These datasets from different domains can be combined to cover multiple tasks, incorporating diverse visual contexts and mathematical skills. Although TQA and VQA primarily involve questions about scenes and relationships, they also include questions requiring arithmetic or numeric skills. Such data enhances multimodal mathematical reasoning and generalizes to other question answering tasks.

The source data are summarized in Table 5 corresponding to Section 3.1.

A.2 Distribution Proportioning of Image Comprehension Complexity

We select images from the source data based on an overall complexity ratio of 2:3:4:1. Due to the limited number of the most complex images, all images with complexity level 3 are sampled. We employ a progressive distribution scale from easy to complex, as described in Section 3.1.2. In this section, we examine the impact of varying distribution proportions of the first three image comprehension complexity levels on model performance. We explore settings with different proportions of comprehension complexities 0, 1, and 2, including uniform distribution, decreasing proportions as complexity increases, and proportions that fluctuate with complexity. As demonstrated in Table 6, both uniform distribution of image complexity and decreasing proportion with increasing difficulty are less effective compared to a progressive proportional distribution aligned with complexity. These findings suggest that MLLMs require fewer simple images and question-answer pairs, but benefit from a larger proportion of complex training data to enhance multimodal mathematical reasoning.

Proportion	ALL	FQA	GPS	MWP	TQA	VQA
3:3:3:1	36.0	29.0	44.4	40.9	35.6	34.5
4:3:2:1	36.4	32.0	39.6	42.5	36.9	35.1
2:4:3:1	35.1	32.0	40.5	35.5	36.2	34.6
2:3:4:1	38.2	33.5	47.2	41.4	36.7	34.6

Table 6: Comparison with different distribution proportioning of image comprehension complexity on MathVista.

A.3 Cases Study

We present several examples of solutions generated by Math-LLaVA and LLaVA-1.5 for image-question pairs of high school or college-level difficulty in MathVista. As illustrated in Figure 7, the base model (LLaVA-1.5) often performs inadequately on numerical computations involving tables, geometric problems, and counting tasks at the high school level. In contrast, our Math-LLaVA model demonstrates superior proficiency in addressing these high school problems, thanks to its training on selected and synthesized data designed to tackle complex issues. Although LLaVA-1.5 faces challenges when dealing with more advanced functions and detailed tables, Math-LLaVA shows promise and capability in solving such intricate problems.

Additionally, we present several examples of newly generated questions, created by thoroughly mining images and questions from the selected dataset. As depicted in Figure 8, the existing dataset contains a limited number of image-question pairs. By fully utilizing the visual information from the images, we are able to generate a wider variety of questions from different perspectives, thereby enhancing the diversity of the problem set. The newly generated questions are created in a few-shot manner, referencing the format of existing question types. Consequently, these questions encompass more than just isolated visual content; they involve reasoning with the images. Moreover, the inclusion of complex questions, logically consistent rephrased questions, and simplified, underspecified questions increases the diversity and robustness of the dataset in terms of both question depth and format, compared to the original set of questions.

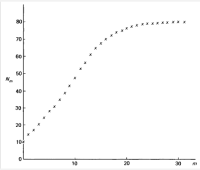
Interestingly, Multimodal Language Models (MLLMs) demonstrate biases when handling multimodal mathematical reasoning tasks, particularly

Dataset	Task	Visual Context	Training Images	Clear Images	Image Complexity			
					0	1	2	3
DocVQA (2022)	FQA	Document Image	8535	8227	2086	6007	125	9
FigureQA (2017)	FQA	Charts and Plots	18173	18173	687	16792	694	0
DVQA (2018)	FQA	Bar Chart	19092	19092	21	18021	1045	5
PlotQA (2020)	FQA	Bar, Line, Scatter	18782	18782	13	18759	10	0
ChartQA (2022)	FQA	Charts and Plots	3699	3699	0	3649	50	0
MapQA (2022)	FQA	Map Chart	10020	10016	1	10015	0	0
IconQA (2021b)	MWP	Abstract Scene	20000	19068	10991	8055	22	0
CLEVR-Math (2022)	MWP	Synthetic Scene	17552	17551	1	17550	0	0
TabMWP (2022b)	MWP	Table	20000	20000	14919	5081	0	0
GEOS (2015)	GPS	Geometry Diagram	66	64	2	57	5	0
Geometry3K (2021a)	GPS	Geometry Diagram	2101	2101	21	1508	568	4
GeoQA+ (2022)	GPS	Geometry Diagram	6027	5956	103	4399	1454	0
UniGeo (2022)	GPS	Geometry Diagram	3499	3432	72	2514	846	0
TQA (2017)	TQA	Scientific Figure	1499	1497	20	949	498	30
AI2D (2016)	TQA	Scientific Figure	3247	3235	32	2321	823	59
ScienceQA (2022a)	TQA	Scientific Figure	6218	6061	1533	4251	273	4
A-OKVQA (2022)	VQA	Natural Image	16540	14526	10	11724	2743	49
VQA2.0 (2017)	VQA	Natural Image	16912	14521	45	12783	1672	21
PMC-VQA (2023a)	VQA	Medical Image	19682	9846	62	2989	3501	3294
VizWiz (2018)	VQA	Natural Image	20,000	16400	790	14800	770	40
Super-CLEVR (2023)	VQA	Synthetic Scene	2000	1950	1	1568	381	0
VQA-AS (2015)	VQA	Abstract Scene	14065	14065	7	13996	62	0
VQA-RAD (2018)	VQA	Medical Image	259	248	0	91	95	62
TextVQA (2019)	VQA	Natural Image	15815	11350	179	9497	1598	76

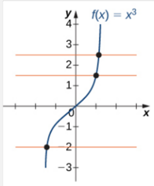
Table 5: Summary of the 24 different source traing datasets for collection. The table provides details on their task, visual context, distribution of image clarity and comprehension complexity according to fine-tuned ViT classification model. Among them, only the text data of GeoQA+ are in Chinese, the rest source data are in English.

with logically consistent rephrased or underspecified questions. As illustrated at the top of Figure 9, LLaVA-1.5 exhibits the ability to answer the original question correctly but tends to falter with simplified, underspecified questions. In contrast, Math-LLaVA proves to be more robust, consistently providing correct answers to underspecified questions. This trend is also observed with logically consistent rephrased questions. Therefore, the use of logically consistent and simplified underspecified questions for data augmentation can enhance the robustness of MLLMs in mathematical reasoning tasks.

Grade Level: College



Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.
Question: Based on the image, what is the most likely equilibrium population count?
Choices: (A) 40 (B) 60 (C) 80 (D) 100
LLaVA-1.5: The answer is (A) 40 ✗ **Math-LLaVA:** The answer is (C) 80 ✓



Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.
Question: Is the function (f: R to R) surjective?
Choices: (A) Yes (B) No
LLaVA-1.5: The answer is (B) No ✗ **Math-LLaVA:** The answer is (A) Yes ✓

Table 7: **Grounding segmentation performance.** The proposed FC-CLIP also demonstrates state-of-the-art performances on grounding segmentation

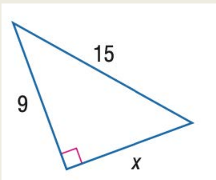
method	grounding PQ			grounding mIoU					
	ADE20K	Cityscapes	Mapillary Vistas	A-847	PC-459	A-150	PC-59	PAS-21	PAS-20
ALIGN [38, 28]	-	-	-	17.8	21.8	25.7	34.2	-	-
ALIGN w/ proposal [38, 28]	-	-	-	17.3	19.7	25.3	32.0	-	-
LSeg+ [46, 28]	-	-	-	10.5	17.1	30.8	56.7	-	-
OpenSeg [28]	-	-	-	21.8	32.1	41.0	57.2	-	-
OpenSeg [28] w/ L. Narr	-	-	-	25.4	39.0	45.5	61.5	-	-
FC-CLIP (ours)	38.4	48.1	21.5	33.4	41.2	54.1	74.9	88.7	98.5

Hint: Please answer the question requiring an integer answer and provide the final value, e.g., 1, 2, 3, at the end.
Question: How many methods in the table achieve an A-847 score higher than 20.0?
LLaVA-1.5: The answer is 1 ✗
Math-LLaVA: The answer is 3 ✓

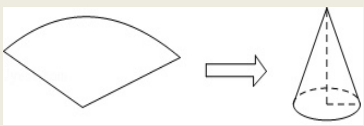
Grade Level: High School

Average waiting time at the DMV	
Month	Waiting time (minutes)
August	17
September	14
October	26
November	17
December	25

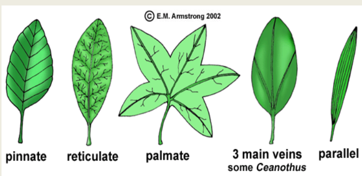
Hint: Please answer the question requiring an integer answer and provide the final value, e.g., 1, 2, 3, at the end.
Question: An administrator at the Department of Motor Vehicles (DMV) tracked the average wait time from month to month. According to the table, what was the rate of change between August and September? (Unit: minutes per month)
LLaVA-1.5: The answer is 11.5 ✗ **Math-LLaVA:** The answer is -3 ✓



Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.
Question: Find x.
Choices: (A) 10 (B) 11 (C) 12 (D) 13
LLaVA-1.5: The answer is (B) 11 ✗ **Math-LLaVA:** The answer is (C) 12 ✓



Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.
Question: Use a sector paper sheet with a central angle of 120.0 and a radius of 6.0 to roll into a conical bottomless paper cap (as shown in the picture), then the bottom perimeter of the paper cap is ()
Choices: (A) 2π cm (B) 3π cm (C) 4π cm (D) 5π cm
LLaVA-1.5: The answer is (B) 3π cm ✗
Math-LLaVA: The answer is (C) 4π cm ✓



Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.
Question: If you wanted the leaf with the least main veins, which would you choose?
Choices: (A) 3 main veins (B) pinnate (C) reticulate (D) palmate
LLaVA-1.5: The answer is (B) pinnate ✗
Math-LLaVA: The answer is (A) 3 main veins ✓

Figure 7: Cases on college and high school difficulty problems of MathVista, Math-LLaVA can solve better compared to LLaVA-1.5.

Cases of Questions Generation

purple T-shirt	$8.03
denim skirt	$8.67
purple dress	$7.67
pair of jeans	$3.81
gray sweatshirt	$9.74

Original Question:
Clara has $13.00. Does she have enough to buy a pair of jeans and a gray sweatshirt?

New Asked Question:
- How much would it cost to buy one of each clothing item listed?
- Can Clara afford to buy both a denim skirt and a purple dress with $13.00?
- If Clara buys two pairs of jeans, how much money will she have left?
- What is the average cost of the items listed?

More Complex Question:
Clara has $13.00. If she buys a pair of jeans and a gray sweatshirt, how much money will she have left?

Rephrased Question:
With $13.00, does Clara have sufficient funds to purchase both a pair of jeans and a gray sweatshirt?

Simplified Question:
Clara has $13.00. Does she have enough to buy some jeans and a sweatshirt?

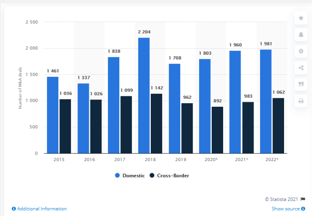
Original Question:
Tommy is writing a letter one evening. The clock shows the time. What time is it?

New Asked Question:
- How many minutes past 7 o'clock is it?
- If Tommy started writing his letter at 7:00, how many minutes has he been writing?
- If Tommy needs to finish his letter by 8:00, how many minutes does he have left to write?
- If the clock is 10 minutes fast, what is the actual time?

More Complex Question:
If Tommy started writing the letter at 7:30 P.M. and finished it 2 hours and 45 minutes later, what time did he finish writing the letter?

Rephrased Question:
What is the current time shown on the clock while Tommy is writing a letter in the evening?

Simplified Question:
The clock shows something one evening. What time is it?



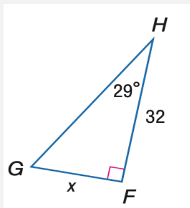
Original Question:
What is the difference between domestic and cross border in 2018?

New Asked Question:
- What is the total number of M&A deals in 2019?
- What is the percentage increase in domestic M&A deals from 2015 to 2016?
- What is the average number of cross-border M&A deals from 2015 to 2017?
- In which year did cross-border M&A deals surpass domestic deals?
- What is the difference in the number of domestic M&A deals between the years 2018 and 2020?

More Complex Question:
What is the total number of domestic and cross-border deals combined for the years 2018 and 2022?

Rephrased Question:
In 2018, what is the contrast between domestic and cross-border?

Simplified Question:
What is the difference between these two categories in 2018?



Original Question:
Find x. Round to the nearest tenth.

New Asked Question:
- What is the length of side FG in the triangle GFH?
- What is the measure of angle HGF in the triangle GFH?
- What is the length of side FH in the triangle GFH?
- Find the length of GH. Round to the nearest tenth.
- What is the perimeter of triangle GFH?

More Complex Question:
What is the area of triangle GFH?

Rephrased Question:
Determine the value of x and round it to the nearest tenth.

Simplified Question:
Determine x.

Figure 8: Examples of synthesizing new questions on source data.

Inference for Underspecified Question

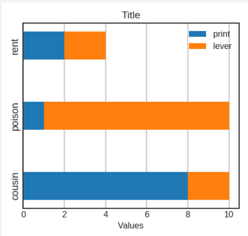
Original Question: What is the highest amount this class measures?
LLaVA-1.5: 400 ✓ **Math-LLaVA:** 400 ✓

Underspecified Question: ***What amount can this measure up to?***
LLaVA-1.5: 3000 ✗ **Math-LLaVA:** 400 ✓

Price	Quantity demanded	Quantity supplied
$270	20,200	10,600
$320	20,000	12,500
$370	19,800	14,400
$420	19,600	16,300
$470	19,400	18,200

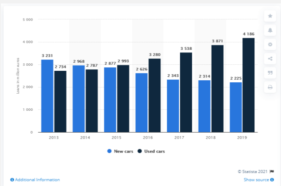
Original Question: Look at the table. Then answer the question. At a price of $320, is there a shortage or a surplus? Choices: (A) shortage (B) surplus
LLaVA-1.5: (A) shortage ✓ **Math-LLaVA:** (A) shortage ✓

Underspecified Question: ***At a price of $320, what is the market situation?***
LLaVA-1.5: (B) surplus ✗ **Math-LLaVA:** (A) shortage ✓



Original Question: What is the value of the smallest individual element in the whole chart?
LLaVA-1.5: 1 ✓ **Math-LLaVA:** 1 ✓

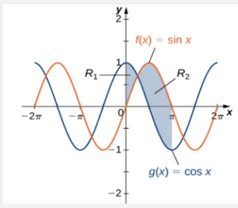
Underspecified Question: ***What is the smallest element value?***
LLaVA-1.5: 0 ✗ **Math-LLaVA:** 1 ✓



Original Question: Which year has the least difference between the used and new cars?
LLaVA-1.5: 2015 ✓ **Math-LLaVA:** 2015✓

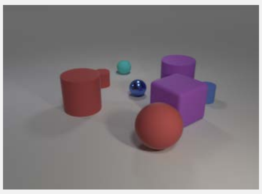
Underspecified Question: ***Which year has the least difference between these two types?***
LLaVA-1.5: 2014 ✗ **Math-LLaVA:** 2015 ✓

Inference for Rephrased Question



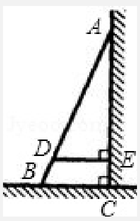
Original Question: Which region is larger? R1 or R2?
LLaVA-1.5: R2 ✓ **Math-LLaVA:** R2 ✓

Rephrased Question: ***Which region, R1 or R2, has a greater area?***
LLaVA-1.5: R1 ✗ **Math-LLaVA:** R2 ✓



Original Question: Subtract all blue metal things. Subtract all tiny objects. How many objects are left?
LLaVA-1.5: 4 ✓ **Math-LLaVA:** 4 ✓

Rephrased Question: ***Remove all blue metallic items. Remove all small things. What is the number of remaining things?***
LLaVA-1.5: 6 ✗ **Math-LLaVA:** 4 ✓



Original Question: As shown in the figure, AB is a long ladder leaning on the wall, the foot of the ladder B is away from the wall 1.6, the point D on the ladder is away from the wall 1.4, the length of BD is 0.55, then the length of the ladder is ()
LLaVA-1.5: 4.40 ✓ **Math-LLaVA:** 4.40 ✓

Rephrased Question: ***In the given figure, AB represents a ladder leaning against the wall, with the foot B of the ladder located 1.6 units away from the wall. Point D on the ladder is located 1.4 units away from the wall, and the length of BD is 0.55 units. What is the length of the ladder?***
LLaVA-1.5: 4.00 ✗ **Math-LLaVA:** 4.40 ✓

Figure 9: Examples of testing on underspecified and rephrased questions.