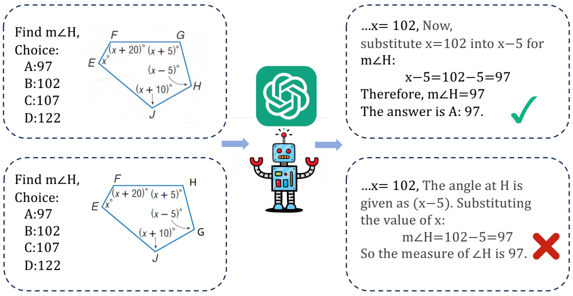
The Role of Visual Modality in Multimodal Mathematical Reasoning: Challenges and Insights

Yufang Liu[1*†], **Yao Du**[2*], **Tao Ji**[3,4], **Jianing Wang**[2],
Yang Liu[2], **Yuanbin Wu**[1], **Aimin Zhou**[1], **Mengdi Zhang**[2], **Xunliang Cai**[2]
[1]School of Computer Science and Technology, East China Normal University
[2] Meituan Inc. [3]Fudan University [4] Pazhou Laboratory
yfliu.antnlp@gmail.com taoji@fudan.edu.cn ybwu@cs.ecnu.edu.cn

Abstract

Recent research has increasingly focused on multimodal mathematical reasoning, particularly emphasizing the creation of relevant datasets and benchmarks. Despite this, the role of visual information in reasoning has been underexplored. Our findings show that existing multimodal mathematical models minimally leverage visual information, and model performance remains largely unaffected by changes to or removal of images in the dataset. We attribute this to the dominance of textual information and answer options that inadvertently guide the model to correct answers. To improve evaluation methods, we introduce the HC-M3D dataset, specifically designed to require image reliance for problem-solving and to challenge models with similar, yet distinct, images that change the correct answer. In testing leading models, their failure to detect these subtle visual differences suggests limitations in current visual perception capabilities. Additionally, we observe that the common approach of improving general VQA capabilities by combining various types of image encoders does not contribute to math reasoning performance. This finding also presents a challenge to enhancing visual reliance during math reasoning. Our benchmark and code would be available at https://github.com/Yufang-Liu/visual_modality_role.


Find m∠H,
Choice:
A:97
B:102
C:107
D:122

...x= 102, Now,
substitute x=102 into x−5 for m∠H:
x−5=102−5=97
Therefore, m∠H=97
The answer is A: 97. ✓

Find m∠H,
Choice:
A:97
B:102
C:107
D:122

...x= 102, The angle at H is given as (x−5). Substituting the value of x:
m∠H=102−5=97
So the measure of ∠H is 97. ✗

Figure 1: An illustrative example of GPT-4o on the HC-M3D. When modifying the image (by swapping the positions of points G and H) while keeping the original question unchanged, the model's output failed to correspondingly adjust, resulting in an incorrect response.

1 Introduction

Recent advancements in Large Vision-Language Models (LVLMs) (Zhang et al., 2023) have demonstrated remarkable potential in tasks that require the seamless integration of visual and linguistic understanding, such as image captioning (Lin et al., 2014), visual question answering (Antol et al., 2015), visual grounding (Yu et al., 2016), and autonomous agents (Xi et al., 2023; Durante et al., 2024). Among these, the domain of multimodal mathematical reasoning (Wang et al., 2024a; Lu et al., 2024; Zhang et al., 2024b) emerges as a particularly challenging and pivotal application. This task requires models to accurately interpret and abstractly represent mathematical concepts, such as geometry, through a tight collaboration between textual descriptions and visual elements in diagrams. The ability to align these modalities and resolve complex problems lies at the heart of LVLMs' capabilities, positioning multimodal mathematical reasoning as a critical field for advancing their effectiveness in abstract and logic-driven applications.

There are three common approaches to enhancing mathematical reasoning capabilities in LVLMs: prompting, fine-tuning, and chain-of-thought (CoT) reasoning. The prompting approach (Lu et al., 2024; Wang et al., 2024b) seeks to unlock the potential of LVLMs through carefully designed prompts. Fine-tuning (Shi et al., 2024; Gao et al., 2023), on the other hand, focuses on constructing higher-quality and more diverse datasets to enhance model performance. Chain-of-thought (Hu et al., 2024)reasoning mimics human reasoning by generating detailed step-by-step thought processes to

[*] Equal contribution.
[†]Work done during an internship at Meituan.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22596–22611
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

tackle complex mathematical tasks. These methods rely heavily on training data tailored for mathematical understanding. Existing research primarily emphasizes improving the diversity of problem types (Gao et al., 2023), refining the conciseness and clarity of questions (Zhang et al., 2024c), and providing more detailed answer explanations (Peng et al., 2024), with model performance evaluated on various benchmarks. *However, current datasets often fail to establish a strong connection between mathematical images and text.*

In this work, we explore whether images are genuinely learned and utilized by models for mathematical reasoning within current multimodal mathematical methods and examines the role they play in the process. We observe that in existing multimodal math models, the benefit brought by image patterns is minimal. Even if the images in the training set are shuffled or even removed, the impact on model performance is not significant (On some evaluation datasets, perturbing the visual modality even resulted in higher scores). By analyzing existing evaluation methods, we suspect that there are two main reasons why the role of visual patterns in multimodal math reasoning has been overestimated: first, the textual information is overly rich, and the model can make correct predictions based on text alone; second, the options leak the answers, and the model can guess the correct answer based on the options.

Hence, we propose the HC-M3D dataset, a **h**uman-**c**rafted **m**uli**m**odal **m**athematical **d**ataset, comprising 1,851 samples meticulously selected by humans, that ensures that the questions depend on the images and provides an additional image that looks similar but changes the correct answer for over 400 problems. We observe whether models can identify these subtle differences in the images and make correct predictions. In evaluating existing leading models, we find that they fail to accurately identify these differences, and in more than half of the cases, they stick to their original predictions. Figure 1 presents a representative example.

Furthermore, we attempt to delve into the causes of this phenomenon. Considering that the model's predictions remain unchanged after image alterations, we infer that the image encoder (mainly the CLIP encoder) fails to effectively recognize such subtle differences, a phenomenon also observed in other works (Liu et al., 2024b; Tong et al., 2024a). However, we observe that the common approach of improving general VQA capabilities by combining various types of image encoders does not contribute to math reasoning performance. This finding also presents a challenge to enhancing visual reliance during math reasoning.

Our contributions can be summarized as follows:

- We show that image shuffling or removal has minimal impact on performance, revealing an overestimation of these models' reliance on visual input.
- We introduce the HC-M3D dataset with 1,851 human-annotated samples, ensuring image-dependent questions. For 429 questions, similar but altered images are included, yet most mathematical LVLMs fail to adjust predictions, often remaining unchanged.
- We demonstrate the challenge of enhancing visual reliance in mathematical reasoning, as combining image encoders proves ineffective.

Based on these observations, we suggest several potential directions for future research: constructing higher-quality datasets with a stronger reliance on visual data (e.g., in the image-caption pretraining task, describing the differences between two images rather than generating captions for a single image.), improving image encoders to capture more fine-grained mathematical information, and designing better loss functions to enhance the model's dependence on the visual modality.

2 Related Work

Mathematical reasoning has garnered increasing attention from researchers and has become a critical benchmark for evaluating LLMs. Specialized LLMs such as MetaMath (Yu et al., 2024a), Math-Shepherd (Wang et al., 2024c), WizardMath (Luo et al., 2023), and DeepSeekMath (Shao et al., 2024) have been developed to tackle these tasks. One of the challenges in multimodal mathematical reasoning lies in visual-textual reasoning, involving geometric diagrams, scientific charts, and function graphs. Due to the scarcity of training data, existing methods often employ the LLaVA architecture for training, aiming to generate higher-quality and more diverse training data. For instance, G-LLaVA (Gao et al., 2023) obtains image caption data for alignment by rewriting the original QA data and paraphrases the remaining QA data to augment the SFT data. Math-LLaVA (Shi et al., 2024) synthesizes a large number of QA pairs based on seed questions, while MAVIS (Zhang et al., 2024c) automatically constructs QA pairs through templates.

With the rapid advancement of LVLMs, there is a growing need for high-quality benchmark tests to assess math problem-solving in visual settings. While previous attempts like GeoQA (Chen et al., 2021) and Geometry3K (Lu et al., 2021) primarily focused on geometric problems, the recently proposed MathVista (Lu et al., 2024) takes it a step further by encompassing a diverse range of multimodal tasks involving mathematical reasoning. Mathverse (Zhang et al., 2024b), on the other hand, emphasizes textual and visual richness to gauge whether models truly understand image content.

3 Does Visual Modality Matter for Mathematical LVLMs?

To validate the importance of visual modality, we replicated the training processes of mainstream mathematical LVLMs under a unified architecture, with the primary differences lying in the training sets. By perturbing the alignment between images and text—either by shuffling correct pairs or masking image information—we observed that the data utilized in mathematical LVLMs fails to enable the model to establish a strong correlation between images and text.

3.1 Setups

Mathematical LVLMs We study the recently proposed mathematical LVLMs, including: **G-LLaVA** (2023) which constructs an enhanced multimodal geometric dataset, Geo170K, based on the scalability of geometric problems and fine-tunes it based on LLaVA. Similarly, **MathLLaVA** (2024) also establishes a large-scale multi-type mathematics problem dataset, which includes filtered image clarity and question complexity for Chart Question Answering, Geometric Problem Solving, Math Word Problems, Textbook Question Answering, and Visual Question Answering. **MAVIS** (2024c) utilizes an automatic data engine to generate mathematical visual datasets covering areas such as plane geometry, analytic geometry, and functions, effectively bypassing the high cost of manual annotation. **MultiMath** (2024) collects textbooks, exercises, and exam questions from the K-12 curriculum and employs GPT to generate detailed CoT processes, featuring bilingual data.

Reproduction Details Since all LVLMs are based on the LLaVA architecture, which connects the language model core to the visual encoder, we standardize and choose `deepseek-math-rl-7B` as language model and `clip-vit-large-patch14-336` as image encoder for reproduction (same as Peng et al. (2024)). The training process is divided into three stages: the pre-training stage, where the connection module of LLaVA is fine-tuned on alignment data (including math alignment data, if available); the SFT stage, where the alignment module and the language model are fine-tuned on LLaVA's general data; and the Math SFT stage, where the alignment module and the language model are further fine-tuned on multimodal math QA questions. Detailed statistcs can be found in Table 11.

Mathematical Benchmarks $\mathcal{D}_{\mathbf{math}}$ We evaluate the following datasets: **MathVerse** (2024b) offers a collection of 3,940 high-quality mathematical problems across various disciplines, enriched with relevant charts, showcasing varied dependencies on visual and textual cues. Conversely, **MathVista** (2024) presents a comprehensive benchmark with 1,000 instances from 28 multimodal datasets, emphasizing their acute visual comprehension and intricate reasoning skills. **GeoQA** (2021) consists of 754 pieces of data, mainly multiple-choice questions about plane geometry. **MATHVision** (2024a) comprises 304 high-quality mathematical problems sourced from authentic mathematics competitions. The questions span 16 distinct mathematical disciplines and are categorized into 5 difficulty levels. **WeMath** (2024) consists of 1,740 visual math questions, covering 67 hierarchical knowledge concepts and 5 levels of knowledge granularity. [1]

3.2 Visual Modality Perturbation

Data w/ correct images 🖼 → 📄: The standard training process, the model can learn the correspondence between images and text as well as the templates for question-answering during the Math SFT stage.

Data w/ random images 🖼 🔀 📄: During the math SFT stage, by shuffling the correspondence between images and texts while keeping the picture distribution unchanged, we ensure that after the correspondence is disrupted, the Q&A for the original picture will be assigned to another picture. In this

[1] For Mathverse, MathVita, and MathVision, as they contain multiple test sets, we select the "testmini" subset for evaluation.

$\mathcal{D}_{\text{math}}$ ▸		Ins@Stage1/2/3	GeoQA	MathVerse	MathVista	MathVision	WeMath	Avg.
🖼→🗎	G-LLaVA	618K/665K/117K	68.0	24.7	34.9	8.9	39.3	35.2
	MathLLaVA	558K/665K/339K	57.4	25.6	**45.3**	**9.5**	42.6	36.1
	MAVIS	1.1M/665K/555K	70.6	**31.5**	37.0	6.9	39.4	37.1
	MultiMath	1.2M/665K/1.0M	**74.4**	29.6	44.5	7.2	**44.7**	**40.1**
🖼🔀🗎	G-LLaVA	*same as above*	71.0 $_{+3.0}$	22.4 $_{-2.3}$	**36.2** $_{+1.9}$	8.2 $_{-0.7}$	35.2 $_{-4.1}$	34.6 $_{-0.6}$
	MathLLaVA		56.6 $_{-0.8}$	23.3 $_{-2.3}$	34.4 $_{-10.9}$	9.5 $_{-0.0}$	38.5 $_{-4.1}$	32.5 $_{-3.6}$
	MAVIS		70.4 $_{-0.2}$	**26.0** $_{-5.5}$	32.8 $_{-4.2}$	**10.2** $_{+3.3}$	34.3 $_{-5.1}$	34.7 $_{-2.4}$
	MultiMath		**76.5** $_{+2.1}$	25.6 $_{-4.0}$	35.0 $_{-9.5}$	9.2 $_{+2.0}$	**39.5** $_{-5.2}$	**37.2** $_{-2.9}$
🗋→🗎	G-LLaVA	*same as above*	68.2 $_{+0.2}$	23.7 $_{-1.0}$	34.5 $_{-0.4}$	8.6 $_{-0.3}$	**41.8** $_{-2.5}$	35.4 $_{+0.2}$
	MathLLaVA		56.9 $_{-0.5}$	23.9 $_{-1.7}$	36.2 $_{-9.1}$	6.9 $_{-2.6}$	39.7 $_{-2.9}$	32.7 $_{-3.4}$
	MAVIS		66.1 $_{-4.5}$	**27.0** $_{-4.5}$	31.3 $_{-5.7}$	**10.2** $_{+3.3}$	36.3 $_{-3.1}$	34.2 $_{-2.9}$
	MultiMath		**70.3** $_{-4.1}$	24.7 $_{-4.9}$	**37.1** $_{-7.4}$	8.6 $_{+1.2}$	41.1 $_{-3.6}$	**36.4** $_{-3.7}$

$\mathcal{D}_{\text{general}}$ ▸		Ins@Stage1/2	VQAv2	MMBench	MM-Vet	MME_P†	MME_C†	Avg.
🖼→🗎	LLaVA-1.5	558K/665K	**79.2**	**66.8**	**32.4**	**1470.1**	**322.1**	**35.9**
🖼🔀🗎	LLaVA-1.5	*same as above*	46.2 $_{-33.0}$	26.6 $_{-40.2}$	12.2 $_{-20.2}$	708.9 $_{-761.2}$	276.8 $_{-45.3}$	17.1 $_{-18.8}$
🗋→🗎	LLaVA-1.5	*same as above*	60.8 $_{-18.4}$	54.3 $_{-12.5}$	20.9 $_{-11.5}$	706.0 $_{-764.1}$	271.8 $_{-50.3}$	27.3 $_{-8.6}$

Table 1: Mathematical reasoning performance of using correct images, shuffled images, and no images during the training phase. "**Ins@Stage1/2/3**" indicates the stages 1, 2, and 3 instances. The † means we normalize MME_P and MME_C for the average score.

process, the model can learn the statistical information of the picture distribution and the template for answering questions.

No images 🗋→🗎: For the Math SFT stage, we remove the images from the training set and retain the question and answer. This helps the model learn the template for question-answering and enhances the quality of the responses.

Observations The results are presented in the Table 1, we can observe that:

- Firstly, after standardizing the model structure and training methodologies, the discrepancies among different approaches diminish (compared to the contrasts presented in Zhang et al. (2024c); Peng et al. (2024)). Contrary to the direct fintuning from LLaVA for alignment as discussed in the G-LLaVA paper (Gao et al., 2023), we notice a significant improvement with a three-stage training approach, aligning with the findings of Peng et al. (2024). The MultiMath model exhibits superior performance across datasets, with the exception of MathVerse, attributable to its diverse dataset and detailed reasoning processes.

- Secondly, we observe that substituting the correct images with shuffled ones only slightly impacts model performance. This trend is consistent across different models and datasets. Specifically, average performance decreases by 0-4 percentage points. In contrast, the GeoQA and MathVisions datasets may even experience some degree of improvement. From a model perspective, high-performance models are more affected by image shuffling than lower-performance models. Compared to other models, the G-LLaVA model exhibits less performance fluctuation.

- Lastly, similar experimental outcomes are observed in the absence of images. These results suggest that during the mathematical SFT phase, image patterns do not significantly enhance reasoning performance, indicating that models primarily rely on text to learn question-and-answer capabilities in the mathematical domain.

Compared with general tasks $\mathcal{D}_{\text{general}}$ To observe if a similar phenomenon exists in the general VQA task, we follow the training process of LLaVA-1.5 with vicuna-7B (Chiang et al., 2023) and examine the impact of visual information on VQA models by comparing the performance differences between three scenarios: shuffled images, removed images, and provided correct images. We select the following commonly used VQA evaluation datasets: VQAv2 (2017), MMBench (2024a), MM-Vet (2024b), and MME (2023).

The results in Table 1 indicate a significant impact on model performance in general VQA tasks when images are either shuffled randomly or not provided at all. Not providing images leads to markedly better performance than offering random images: without images, model performance drops by approximately 23% on VQAv2 and 19% on

Model	Input	#Param.	GeoQA	M.Verse	M.Vista	M.Vison	WeM.	Avg.
Deepseek-Math-RL (2024)	📄	7B	39.8	20.7	26.1	11.5	33.3	26.3
QWen-2.5-Math-Instruct (2024)	📄	7B	62.2	23.7	31.4	9.2	40.2	33.3
Math-Shepherd-Mistral-RL (2024c)	📄	7B	30.6	12.4	28.6	4.9	24.0	20.1
OpenMath2-Llama3.1 (2024)	📄	8B	43.9	22.2	30.0	8.2	35.6	28.0
Dart-Math-dsmath-prop2diff (2024b)	📄	7B	43.6	19.2	26.6	11.5	38.1	27.8
Meta-Llama-3.1-Instruct (2024)	📄	70B	37.0	20.6	32.7	9.9	33.7	26.8
OpenMath2-Llama3.1 (2024)	📄	70B	48.3	24.6	31.6	12.2	**41.2**	31.6
Dart-Math-Llama3-prop2diff (2024b)	📄	70B	41.9	17.6	27.8	9.9	33.8	26.2
QWen2.5-Math-Instruct (2024)	📄	72B	**68.0**	**32.3**	34.6	**12.5**	32.9	**36.1**
QWen2-Instruct-Step-DPO (2024)	📄	72B	61.9	21.8	**35.0**	9.9	30.9	31.9
GPT-4o (2024)	📄	/	57.7	25.4	30.7	10.5	33.9	31.6
InternVL2 (2023)	🖼+📄	8B	56.8	31.5	56.9	8.6	45.4	39.8
QWen2-VL-Instruct (2024d)	🖼+📄	8B	50.3	32.2	60.9	8.6	33.9	37.2
InternVL2-Llama3 (2023)	🖼+📄	76B	59.2	38.4	62.8	10.5	55.8	45.3
QWen2-VL-Instruct (2024d)	🖼+📄	72B	**68.8**	46.7	**70.4**	13.8	61.5	**52.2**
GPT-4o (2024)	🖼+📄	/	58.9	**46.8**	53.1	**15.5**	**68.3**	48.5

Table 2: Results of common text and visual language models on mathematical datasets. "**#Param.**" indicates the number of parameters. "**M.**" is the short for "Math".

MMBench; with random images, the decline is more drastic at 42% and 61%, respectively. This suggests that models can detect inconsistencies between randomized image content and text, thereby disrupting learning. In contrast, the trend of performance decline in the mathematical domain is much weaker. We speculate this is because multimodal mathematical reasoning tasks rely more on generative aspects and less on image comprehension, a discrepancy that may not be as apparent in general VQA tasks.

4 Issues in Existing Benchmarks

We further explore the reasons for the overestimation of the image modality in multi-modal mathematical reasoning tasks. Through a careful examination of existing evaluation sets, we identify two main issues.

Firstly, a large proportion of the existing evaluation dataset can be answered correctly based on the text alone, which does not accurately reflect whether the image modality is effectively utilized. We conduct an evaluation on 7 common textual models and and 3 vision language models. The experimental results are shown in Table 2. As compared to the results in Table 1, the existing multimodal math methods show limited improvement based on the language model `DeepSeek-Math-7B-RL`. For example, MultiMath exhibits enhanced performance on the MathVerse and Wemath datasets, increasing from 20.7 to 29.6 and from 33.3 to 44.7, respectively. However, it even experiences a degree of decline on the MathVision dataset, dropping from 11.5 to 7.2.

We also observe that as the capability of the language model improves, the performance of the text model is further enhanced. In terms of average performance across all datasets, when the language model is switched from `DeepSeek-Math-7B-RL` to `QWen-2.5-Math-7B-Instruct`, the score increase from 26.3 to 33.3. Meanwhile, the average performance of the G-LLaVA model is only 35.2, indicating that the performance of language models is very close to that of multimodal models.

Secondly, the options in the question may leak the answer, and the model can guess the answer based on the option information. We shuffle the order of the options in the multiple-choice questions, observe the changes in the model's predictions before and after, and determine whether the model truly understands the question or is just making predictions based on option information. Table 3 shows the experimental results. We calculate the original accuracy (CR: Correct), the proportion of predictions are correct both before and after shuffled (BC: Both Correct), and the proportion of consistent predictions before and after (AR: Agreement). As can be seen from the table, in the case of multiple-choice questions, the BC indicator is significantly lower than the CR indicator. For example, the BC/CR indicators of the MultiMath model on the three datasets are 9.5%, 76.1%, and 80.7%, respectively. This indicates that the model's single-performance is not sufficient to reflect its overall performance on the dataset. Additionally, a very low BC/CR could indicate potential quality issues

Model	GeoQA			MathVerse			MathVista		
	CR↑	**BC↑**	**AG↑**	**CR↑**	**BC↑**	**AG↑**	**CR↑**	**BC↑**	**AG↑**
$\mathrm{G}_{\mathrm{LLaVA}}$	68.0	**16.0**	**20.8**	44.4	30.3	50.6	51.7	40.9	70.4
$\mathrm{Math}_{\mathrm{LLaVA}}$	57.4	7.6	13.3	40.0	26.2	53.4	**58.7**	**49.3**	**75.4**
MAVIS	70.6	8.2	10.1	41.9	29.0	49.1	54.6	44.8	72.0
MultiMath	**74.4**	7.1	8.6	**48.1**	**36.6**	**56.6**	55.0	44.4	72.4

Table 3: Results of shuffling the order of the multiple-choice options in the dataset. The table indicates the accuracy (CR: Correct), predictions are correct both before and after shuffled (BC: Both Correct), and predictions remained consistent across both attempts (AG: Agreement).

with the dataset.

Based on our observations, we believe that existing multimodal math datasets do not accurately measure the importance of the image modality. Since a large portion of the questions can be answered by textual models alone and the model can guess the answers, these issues lead to an overestimation of the capability of the visual modality, especially in multiple-choice questions. To improve the accuracy of the assessment, it is essential to mitigate the above two issues. Establishing a suitable benchmark will significantly advance the research on multimodal mathematic models.

5 The HC-M3D Benchmark

Based on the analysis above, we propose HC-M3D, a high-quality, vision-dependent multimodal benchmark. Our dataset construction adheres to the following three principles.

- First, correctness of the data, meaning that the collected data must be solvable based on the question and image asked and the answer is correct.
- Second, visual dependency, indicating that the data must rely on images for accurate answers.
- Lastly, high correlation between images and answers. Where possible, we aim to keep the question unchanged but alter the image to ensure that the answer changes. This methodology tests whether the model can detect changes in the images and make correct predictions.

Statistic	**Number**
Total questions	1,851
- Multiple-choice questions	1,851 (100.0%)
- Newly collected questions	**1,219 (65.9%)**
- Newly collected images	**1,084 (58.6%)**
Data Souce	
- GeoQA	896 (48.4%)
- MathVista	273 (14.7%)
- Jingyou Net	684 (37.0%)
Language	
- English	1,052 (56.8%)
- Chinese	799 (43.2%)
Number of unique images	1,851 (100.0%)
Number of unique questions	993 (53.6%)

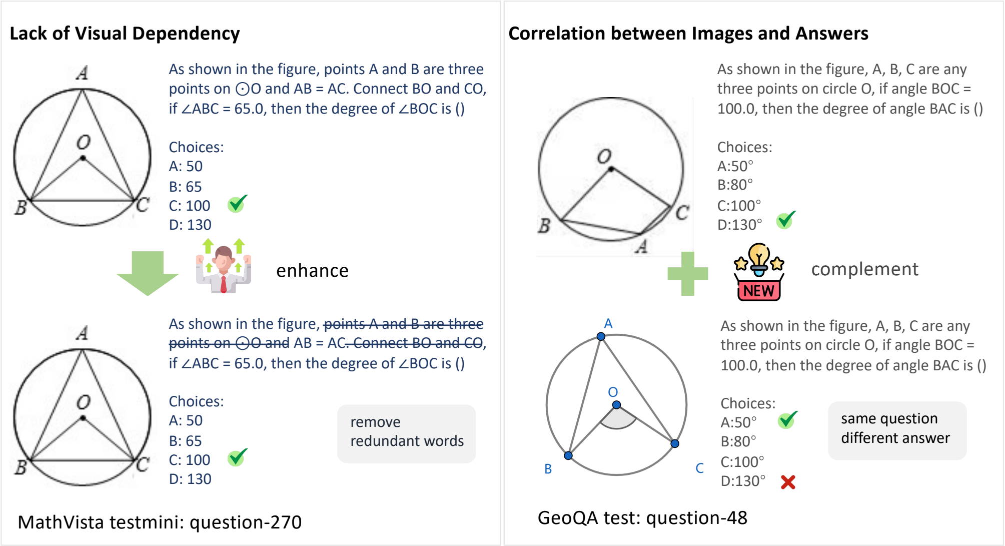
Table 4: Statistical Results on the HC-M3D Dataset.

Data Curation Our dataset is sourced from the GeoQA and MathVista GPS subsets, along with a selection of questions from the Chinese Jingyou Net [2], primarily focusing on plane geometry knowledge. We manually annotate each sample for correctness and visual dependency, modifying the original questions when these criteria are not met and attempting to supply an additional image for the question to change the correct answer. Instructions for human annotators of these two steps are shown in Figure 5 and Figure 6. Ultimately, we obtain 1,851 samples, including 429 questions for which an additional image was provided. Statistical results for the dataset are presented in Table 4.

Figure 2 illustrates the process of constructing the dataset. Upon manually identifying a lack of visual dependency in the data, the text (or image) is revised to ensure that the question requires reliance on the image for an answer. For instance, the image on the left can be drawn based on the question, demonstrating low dependency on the image; thus, a revision is necessary. Under conditions where dependency is met, as with the image on the right, changing the position of point A on the circle alters the correct answer from D to A. In this scenario, there is a high correlation between the image and the correct answer, necessitating that the model explicitly perceives the differences in the image to make accurate predictions.

Our HC-M3D stands out by modifying images and using a controlled-variable method to verify whether the model can truly and accurately understand the image content. The challenge with MathVerse lies in precisely distributing information, a process that is somewhat ambiguous. In contrast, H3-M3D changes the answer solely through image modifications, with the text remaining unchanged. Our control over the text information is thorough and successful (annotators simply need to determine whether the modified image corresponds to a new answer).

[2] https://www.jyeoo.com/



As shown in the figure, points A and B are three points on ⊙O and AB = AC. Connect BO and CO, if ∠ABC = 65.0, then the degree of ∠BOC is ()

Choices:
A: 50
B: 65
C: 100 ✓
D: 130

As shown in the figure, ~~points A and B are three points on ⊙O and~~ AB = AC~~. Connect BO and CO~~, if ∠ABC = 65.0, then the degree of ∠BOC is ()

Choices:
A: 50
B: 65
C: 100 ✓
D: 130

As shown in the figure, A, B, C are any three points on circle O, if angle BOC = 100.0, then the degree of angle BAC is ()

Choices:
A:50°
B:80°
C:100°
D:130° ✓

As shown in the figure, A, B, C are any three points on circle O, if angle BOC = 100.0, then the degree of angle BAC is ()

Choices:
A:50° ✓
B:80°
C:100°
D:130° ✗

Figure 2: Examples from our constructed HC-M3D benchmark. The left image demonstrates a scenario where the original data lacks visual dependency, allowing for the reconstruction of the question's corresponding image solely from text. This is addressed by rewriting or possibly changing the image to ensure the question requires image dependency for a correct answer. The right image shows an example where supplementing the original question with a similar image alters the correct answer. More examples can be found in Figure 3 and Figure 4.

Model	#Param.	ALL ↑	DI ↑	BC ↑	AG ↓
G-LLaVA	7B	45.4	41.5	15.2	52.2
MathLLaVA	7B	39.6	37.2	8.4	75.3
MAVIS	7B	42.8	37.7	9.8	58.0
MultiMath	7B	49.2	44.8	16.6	56.9
InternVL2	8B	41.9	38.3	16.6	**34.0**
QWen2-VL-Instruct	8B	40.9	40.7	13.8	58.7
InternVL2-Llama3	76B	47.4	45.5	19.4	40.8
QWen2-VL-Instruct	76B	**51.8**	**48.3**	**20.3**	51.5
GPT-4o	/	49.0	45.8	19.1	42.0

Table 5: Evaluation results on the HC-M3D benchmark. The ↑ and ↓ arrows respectively indicate that higher or lower values are preferred.

Metrics Performance on the dataset is evaluated based on the following metrics. Accuracy over all data (ALL); accuracy on a subset (DI: Diverse Image) consisting of 858 samples that share similar questions but have different images and answers; accuracy (BC: Both Correct) for cases within the DI subset where both samples sharing the same question are correctly predicted; and consistency rate (AG: Agreement) within the DI subset for samples that share the same question and are predicted consistently. Since these samples have differing answers, a lower AG metric is preferred. [3]

Evaluation Experimental results are shown in Table 5. It is observed that among existing multimodal mathematical models, MutiMath achieves the highest accuracy of 49.2, comparable to the performance of GPT-4o. The open-source model QWen2-VL-Instruct-76B, reaches an highest accuracy of 51.8. Furthermore, in the DI subset, which accounts for image and answer relevance, the proportion of question pairs correctly predicted/Performance on the DI subset (BC/DI) remains below 50%. This indicates that the models do not deeply understand questions for which half of the answers are correct; changing the image leads to incorrect predictions. At the same time, the AG metric for different models is generally above 50%, suggesting that models do not appropriately change their answers when the image is replaced. The GPT-4o and InternVL2 models perform best in this metric.

6 Challenges in Enhancing Visual Dependency

We observe that existing multimodal mathematical models exhibit weak reliance on visual information (Section 3), demonstrating insufficient sensitivity to changes in images (Section 5). Given recent findings indicating that the CLIP models struggle to capture object information (Liu et al., 2024b), spatial relationships (Kamath et al., 2023), and compositional understanding (Zhang et al., 2024a) in images during encoding, we hypothesize that these phenomena stem from the limited capabilities of

[3] The visual control experiment results on the proposed HC-M3D benchmark can be found in Table 8.

Image Encoder	General VQA Performance after SFT stage					Math Performance after Math SFT stage					
	VQA^{v2}	MMBench	MM-Vet	MME^{P}	MME^{C}	GeoQA	M.Verse	M.Vista	M.Vision	WeM.	HC-M3D
clip-B	72.6	60.7	25.7	1365.8	273.6	71.0	23.0	31.4	9.5	38.7	41.2
clip-L	77.0	64.8	30.4	1481.0	269.6	68.0	**24.7**	**34.9**	8.9	**39.3**	**45.4**
(a) combine features by concatenating the hidden feature											
clip-B+ siglip-B	74.6	62.0	25.7	1397.5	257.5	**72.3**	22.2	31.1	8.2	38.1	40.0
clip-L+dinov2-L	78.4	66.9	30.9	1500.1	253.2	70.3	23.2	30.0	**12.2**	38.6	40.3
clip-L+siglip-B+dino-B	74.4	62.5	25.5	1404.0	280.0	70.8	22.5	31.0	8.9	36.0	40.5
(b) combine features by concatenating image tokens											
clip-L+siglip-L	78.7	66.6	33.0	1445.2	266.4	69.9	22.6	30.8	8.9	37.2	39.2
clip-L+dinov2-L	78.6	66.7	32.3	1494.7	269.6	70.7	22.3	28.7	7.6	35.7	39.2
clip-L+siglip-L+dinov2-L	**80.0**	**69.8**	**35.8**	**1524.6**	**301.8**	69.9	22.6	28.5	8.9	38.1	39.4

Table 6: Performance on general VQA and mathematical tasks. clip-L, sigLip-L, and dino-L correspond to `clip-vit-large-patch14-336`, `siglip-so400m-patch14-384`, and `dinov2-large` models,respectively. siglip-B (siglip-base-patch16-224) and dino-B (`dino-vitb16`) are utilized alongside clip-B (`clip-vit-base-patch16`) for matching image token numbers in combinations. Detailed results can be found in Table 12 and Table 13.

Image Encoder	POPE	MathPOPE
clip-B	81.4	72.9
clip-L	**85.2**	73.7
(a) concatenating the hidden feature		
clip-B+siglip-B	82.2	72.6
clip-L+dinov2-L	84.4	70.1
clip-L+siglip-B+dino-B	80.1	76.2
(b) concatenating image tokens		
clip-L+siglip-L	83.9	67.6
clip-L+dinov2-L	84.2	71.9
clip-L+siglip-L+dinov2-L	84.1	**81.0**

Table 7: F1 performance on both POPE and MathPOPE datasets. POPE primarily focuses on questions about objects in images, whereas MathPOPE concentrates on points existing in geometric images.

the image encoders. Recent study (Al-Tahan et al., 2024) find that although scaling up the dataset and model parameters in the CLIP series proves effective for general tasks, such enhancements have minimal impact (even negative impact) on performance in complex tasks involving reasoning and relationship understanding. Therefore, we attempt to leverage the strengths of diverse types of encoders, which is also commonly used in general vqa tasks (Tong et al., 2024a; Yang et al., 2024), to enhance visual dependency and final performance.

SigLip (Zhai et al., 2023) and DINO (Caron et al., 2021; Oquab et al., 2024) are selected as the components of our combination, where SigLIP replaces the loss function used in CLIP with a simple pairwise sigmoid loss function, resulting in improved zero-shot classification performance; DINO, as a self-supervised visual model, is capable of capturing rich information from images. We consider two methods for feature concatenation:

- **Hidden Feature Concatenation** Images are input into multiple image encoders, and the dimensions of the hidden representation are concatenated, then mapped back to the original hidden layer output dimensions through the modality connection module. In this case, it is necessary for the image tokens across different representations to be consistent;

- **Image Token Concatenation** Images are input into multiple image encoders and mapped to the same hidden layer dimensions according to their respective modality connection module, with the image tokens then concatenated. This method does not require the dimensions across different representations to be the same, but it increases the overall number of image tokens. [4]

Results We experiment on G-LLaVA model [5], and the results are shown in Table 6. We can find that:

- First, we find that integrating multiple image encoders, specifically SigLip and DINO, either separately or combined, significantly improves performance in general Visual Question Answering (VQA) tasks. The optimal performance is achieved by concatenating image tokens from all three encoders without dimensionality constraints, which elevates MME perception scores from 1481 to 1524.

- Second, an in-depth analysis of various metrics within the MME demonstrates notable advancements in color perception and spatial relation-

[4] Experiment results of interleaving image tokens from different encoders as done in (Tong et al., 2024a) when combining multiple visual encoders can be found in Table 10.

[5] Experiment results on MathLLavA model can be found in Table 9.

ships (see Table 12). This suggests that utilizing diverse encoders allows the model to discern more detailed image features.

- Lastly, while enhancements are observed in general tasks, there is a consistent dip in performance on mathematical evaluation tasks, with only a slight increase seen in GeoQA. Moreover, the addition of multiple image encoders, such as combining SigLip and DINO, does not proportionally enhance performance beyond the use of either encoder on its own (i.e., the model's performance does not increase with the addition of more image encoders). For instance, in MathVista, the performance metrics for combining SigLip alone, DINO alone, and both combined are 30.8, 28.7, and 28.5, respectively.

Hallucination Evaluation We aim to further investigate the causes behind this decline in performance. Given that the general improvement across common tasks suggests an enhanced general capability of the model, we are curious to determine if a more severe presence of hallucinations has led to this performance decline. Considering the lack of assessments for hallucinations in the mathematical domain, inspired by POPE (Li et al., 2023), we construct MathPOPE (9,000 questions based on 500 images) based on annotations of points in images from the Geometry3K (Lu et al., 2021) test set, in order to assess model hallucinations. The format of the question in MathPOPE is: 'Is there a point X in the image?', where X refers to the point name in picture. The questions in the dataset are designed such that the objects are present and absent in equal measure, therefore the ideal 'yes' response rate should be around 50%.

Experimental results can be seen in Table 7. We observe that combining multiple image encoders indeed results in an increase in hallucinations to a certain extent. However, distinct from the general VQA tasks, in the domain of mathematical hallucination assessment, we note a reduction in hallucinations when three image encoders are combined (73.7 vs 81.0). Moreover, while there is a clear positive correlation between VQA performance in general tasks and hallucination performance, such correlation is absent in the domain of mathematical reasoning.

Analysis The results from the above experiments reveal a significant difference in the performance improvements within the mathematical domain compared to the findings in general VQA tasks. We believe there are two possible reasons for this. Firstly, images in the mathematical domain usually feature monochromatic colors and lower information density (see Figure 2), which markedly distinguishes them from general images. Secondly, while there is a close association between performance and hallucination assessment in general tasks, the linkage between reasoning performance and hallucinations appears to be weaker. Therefore, reasoning tasks in the mathematical domain present unique challenges, making it significantly difficult to improve modality dependence. We leave further exploration to future work.

7 Conclusion

We study the importance of the visual modality within multimodal mathematical models. Our findings indicate that the performance of existing multimodal mathematical models does not significantly deteriorate when images are shuffled or even removed. We propose HC-M3D benchmark to measure the model's sensitivity to changes in images. Experiments demonstrate that the model does not alter its answers corresponding to changes in the images when there is a high relevance between images and answers. We further discover that enhancing the dependency on the visual modality in the mathematical domain presents significant challenges, and we leave further exploration to subsequent work.

Limitations

Firstly, this paper focuses on the importance of the visual modality in mathematical LVLMs during pre-training and instruction fine-tuning, without addressing reasoning stages such as chain-of-thought, which will be part of our future work. Secondly, due to the high cost and time-consuming nature of manual annotation, especially generating new images for different options of the same question, the scale of our proposed HC-M3D dataset is relatively small, consisting of only 1,851 samples. Lastly, although the method of integrating multiple visual experts shows performance improvements in general VQA tasks, it offers little benefit (and in some cases, even negative impact) when applied to mathematical reasoning evaluation. This highlights the differences between mathematical reasoning tasks and general tasks, urging us to explore solutions tailored to mathematical reasoning tasks.

Ethics Statement

The HC-M3D dataset is sourced from publicly available academic datasets or manually collected math problems. The annotations were performed by a professional labeling team from the company, with the task of filtering math problems and generating images. Therefore, no ethical issues are involved.

Acknowledgement

The authors wish to thank all reviewers for their helpful comments and suggestions. The corresponding authors are Tao Ji, Yuanbin Wu, Aimin Zhou and Mengdi Zhang. The computations in this research were performed using the CFFF platform of Fudan University.

References

Haider Al-Tahan, Quentin Garrido, Randall Balestriero, Diane Bouchacourt, Caner Hazirbas, and Mark Ibrahim. 2024. Unibench: Visual reasoning requires rethinking vision-language beyond scaling. *CoRR*, abs/2408.04810.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 513–523. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. 2024. Agent AI: surveying the horizons of multimodal interaction. *CoRR*, abs/2401.03568.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. 2023. G-llava: Solving geometric problem with multi-modal large language model. *CoRR*, abs/2312.11370.

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 *2017 IEEE Conference* *on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9161–9175. Association for Computational Linguistics.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *CoRR*, abs/2406.18629.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024a. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer.

Yufang Liu, Tao Ji, Changzhi Sun, Yuanbin Wu, and Aimin Zhou. 2024b. Investigating and mitigating object hallucinations in pretrained vision-language (CLIP) models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 18288–18301. Association for Computational Linguistics.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6774–6786. Association for Computational Linguistics.

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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.

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

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma Gongque, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *CoRR*, abs/2407.01284.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

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

In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 4663–4680. Association for Computational Linguistics.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024a. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9568–9578. IEEE.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024b. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *CoRR*, abs/2407.13690.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. 2024. Openmathinstruct-2: Accelerating AI for math with massive open-source instruction data. *CoRR*, abs/2410.01560.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *CoRR*, abs/2402.14804.

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. *CoRR*, abs/2402.17971.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024c. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9426–9439. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024d. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *CoRR*, abs/2309.07864.

Shijia Yang, Bohan Zhai, Quanzeng You, Jianbo Yuan, Hongxia Yang, and Chenfeng Xu. 2024. Law of vision representation in mllms. *CoRR*, abs/2408.16357.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 69–85. Springer.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024a. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024b. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2023. Vision-language models for vision tasks: A survey. *CoRR*, abs/2304.00685.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. 2024a. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13774–13784. IEEE.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2024b. MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VIII*, volume 15066 of *Lecture Notes in Computer Science*, pages 169–186. Springer.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, and Hongsheng Li. 2024c. MAVIS: mathematical visual instruction tuning. *CoRR*, abs/2407.08739.

$\mathcal{D}_{\text{math}}$ ▸		ALL ↑	DI ↑	BC ↑	AG ↓
🖼 → 🗎	G-LLaVA	45.4	41.5	15.2	**52.2**
	MathLLaVA	**49.2**	**44.8**	**16.6**	56.9
🖼 ⤨ 🗎	G-LLaVA	41.2	**39.9**	**12.1**	**54.8**
	MultiMath	**42.3**	38.8	11.2	60.6
🗋 → 🗎	G-LLaVA	40.8	37.2	**10.8**	**57.3**
	MultiMath	**42.8**	**39.3**	10.3	67.4

Table 8: Performance on HC-M3D datset of using correct images, shuffled images, and no images during the training phase.

A Visual Control Experiment on HC-M3D Benchmark

We evaluate whether HC-M3D can effectively distinguish the utilization of visual modalities on GLLaVA and MultiMath datasets. The results (shown in Table 8) demonstrate that perturbing or removing the images leads to a significant decline in overall accuracy (ALL). Moreover, under the image removal setting, the prediction consistency (BC) is higher than in both the normal and perturbed image settings, suggesting that HC-M3D can partially identify the reliance on visual modalities.

B More Experiments on Exploring Different Visual Encoders

	GeoQA	MathVerse	MathVista	MathVision	WeMath
clip-L	57.4	**25.6**	**45.3**	9.5	**42.6**
clip-L+siglip-L	57.3	24.3	44.0	10.9	40.6
clip-L+dinov2-L	**58.6**	24.0	39.6	**12.2**	38.8
clip-L+siglip-L+dinov2-L	58.0	24.5	37.6	8.2	37.9

Table 9: Results of combining different visual encoders on mathematical tasks. Experiments are conducted on MathLLaVA model.

we additionally report results on the training set of MathLLaVA to provide a more comprehensive evaluation. (as shown in Table 1, MathLLaVA demonstrates a stronger dependency on image-based information.) Here we combine different encoders by concatenating image tokens. The experimental results on MathLLaVA (shown in Table 9) are generally consistent with those on G-LLaVA (Table 6). Combining multiple encoders does not enhance the performance of complex reasoning on images. Although Math-LLaVA's dataset exhibits a marginally higher degree of visual-textual dependency, the 3% disparity is not deemed to constitute a robust reliance (e.g., when compared with the LLaVA's training dataset). Consequently, it is not feasible to derive a definitive conclusion from such a dataset. The development of an additional general geometry dataset characterized by a pronounced visual-textual dependency is reserved for future research endeavors.

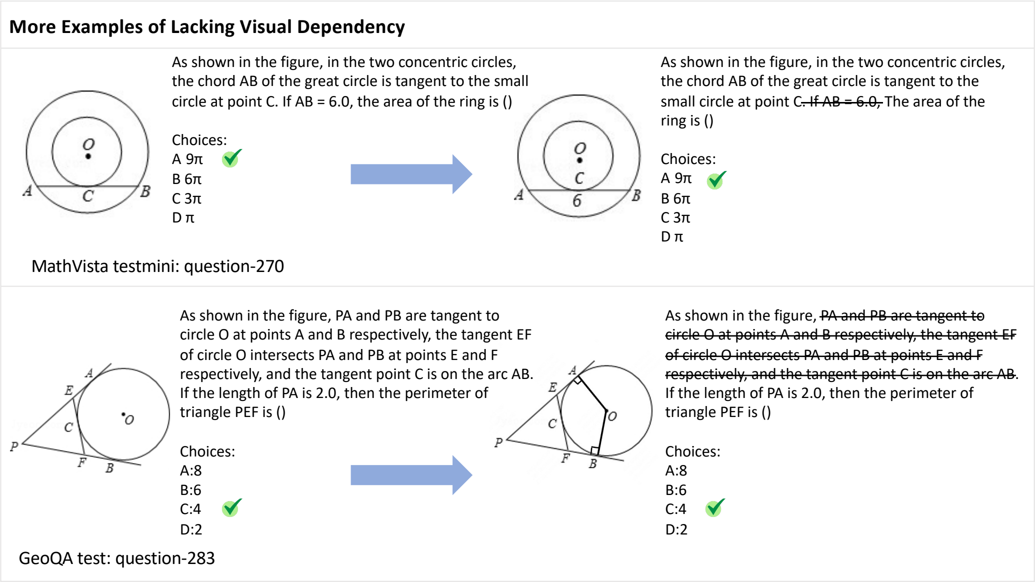
	GeoQA	MathVerse	MathVista	MathVision	WeMath	HC-M3D	Avg.
clip-L	68.0	**24.7**	**34.9**	**8.9**	**39.3**	**45.4**	**36.9**
clip-L+dinov2-L	**70.7**	22.3	28.7	7.6	35.7	39.2	34.0
clip-L+dinov2-L (2024a)	66.7	24.3	31.9	6.6	38.7	41.8	35.0

Table 10: Results of interleaving the images tokens from different encoders as done in (Tong et al., 2024a).

We also try interleaving the tokens from different encoders as done in (Tong et al., 2024a) when combining multiple visual encoders. Experimental results (shown in Table 10) indicate that the interleaving input strategy achieved slightly better performance (average score of 35.0) compared to directly concatenating the representations of CLIP and DINOv2 (average score of 34.0). However, the performance of this method remains significantly lower than that of using CLIP representations alone (average score of 36.9).

C Training Details

Detailed information on model training are shown in 11. Given that existing multimodal mathematical models are based on the same LLaVA architecture with the main differentiation being the training data, we standardize the training approach as three stage: Pretrain, SFT and Math SFT with language model `deepseek-math-rl-7b`. All the models are trained on 8 NVIDIA A100-80GB GPUs with the random seed 42.



More Examples of Lacking Visual Dependency

As shown in the figure, in the two concentric circles, the chord AB of the great circle is tangent to the small circle at point C. If AB = 6.0, the area of the ring is ()

Choices:
A 9π ✓
B 6π
C 3π
D π

As shown in the figure, in the two concentric circles, the chord AB of the great circle is tangent to the small circle at point C~~. If AB = 6.0,~~ The area of the ring is ()

Choices:
A 9π ✓
B 6π
C 3π
D π

MathVista testmini: question-270

As shown in the figure, PA and PB are tangent to circle O at points A and B respectively, the tangent EF of circle O intersects PA and PB at points E and F respectively, and the tangent point C is on the arc AB. If the length of PA is 2.0, then the perimeter of triangle PEF is ()

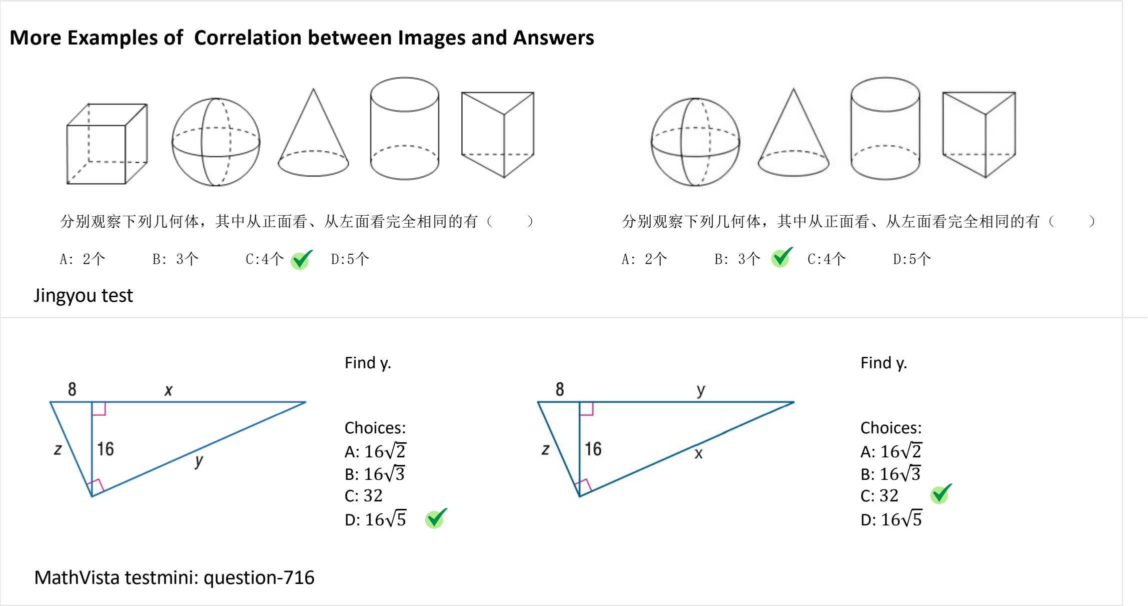
Choices:
A:8
B:6
C:4 ✓
D:2

As shown in the figure, ~~PA and PB are tangent to circle O at points A and B respectively, the tangent EF of circle O intersects PA and PB at points E and F respectively, and the tangent point C is on the arc AB.~~ If the length of PA is 2.0, then the perimeter of triangle PEF is ()

Choices:
A:8
B:6
C:4 ✓
D:2

GeoQA test: question-283

Figure 3: More examples from our constructed HC-M3D benchmark where original samples are lacking visual dependency. (Supplement to Figure 2)



More Examples of Correlation between Images and Answers

分别观察下列几何体，其中从正面看、从左面看完全相同的有（　　）

A: 2个　B: 3个　C:4个 ✓　D:5个

分别观察下列几何体，其中从正面看、从左面看完全相同的有（　　）

A: 2个　B: 3个 ✓　C:4个　D:5个

Jingyou test

Find y.

Choices:
A: $16\sqrt{2}$
B: $16\sqrt{3}$
C: 32
D: $16\sqrt{5}$ ✓

Find y.

Choices:
A: $16\sqrt{2}$
B: $16\sqrt{3}$
C: 32 ✓
D: $16\sqrt{5}$

MathVista testmini: question-716

Figure 4: More examples from our constructed HC-M3D benchmark where images highly related to answers. (Supplement to Figure 2)

Statistcs	**G-LLaVA**	**MathLLaVA**	**MAVIS**	**MultiMath**
		Pretrain stage		
Dataset	Geo170K-align +LLaVA-Pretrain	LLaVA-Pretrain	MAVIS-align +LLava-Pretrain	MultiMath-300K-align +Geo170K-align +LLaVA-Pretrain
#Samples	618K	558K	1.1M	1.2M
Training Module	mm adapter	mm adapter	mm adapter	mm adapter
Epoch	1	1	1	1
Batch_size * #gpu	32*8	32*8	32*8	32*8
Learning Rate	1e-3	1e-3	1e-3	1e-3
		SFT stage		
Dataset	LLaVA-Instruction	LLaVA-Instruction	LLaVA-Instruction	LLaVA-Instruction
#Samples	665K	665K	665K	665K
Training Module	mm adapter+llm	mm adapter+llm	mm adapter+llm	mm adapter+llm
Epoch	1	1	1	1
Batch_size * #gpu	8*8	8*8	8*8	8*8
Learning Rate	2e-5	2e-5	2e-5	2e-5
		Math SFT stage		
Dataset	Geo170K-qa	MathV360K	MAVIS-qa	MultiMath-300K-qa +Geo170K-qa +MathV360K
#Samples	117K	339K	555K	1.0M
Training Module	mm adapter+llm	mm adapter+llm	mm adapter+llm	mm adapter+llm
Epoch	2	2	2	2
Batch_size * #gpu	8*8	8*8	8*8	8*8
Learning Rate	2e-5	2e-5	2e-5	2e-5

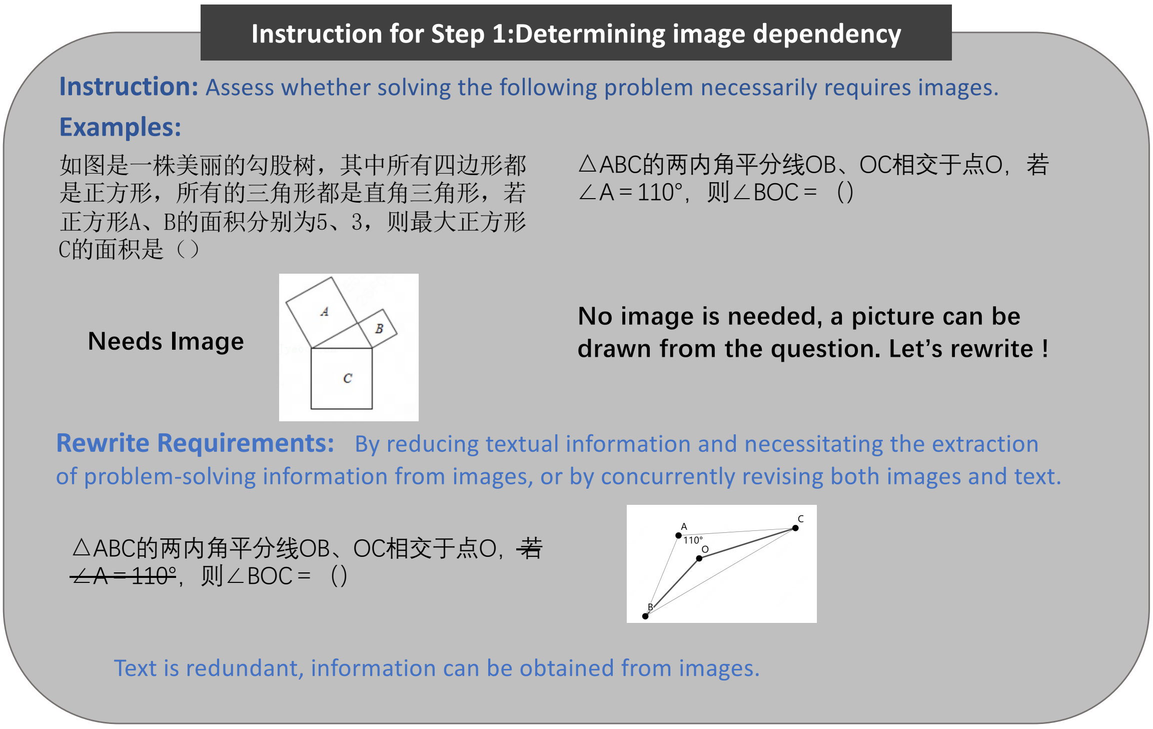
Table 11: Training detail information across different models.

Image Encoder	**existence**	**count**	**position**	**color**	**posters**	**celebrity**	**scene**	**landmark**	**artwork**	**OCR**	**CR**	**NC**	**TT**	**CR**
clip-B	185.0	141.7	141.7	150.0	120.7	105.0	**163.3**	138.5	102.5	117.5	108.6	62.5	50.0	52.5
clip-L	195.0	156.7	131.7	175.0	136.7	127.6	160.5	147.0	118.3	132.5	**112.1**	47.5	50.0	60.0
(a) combine features by concatenating the hidden feature														
clip-B+ siglip-B	**200.0**	151.7	140.0	160.0	114.6	113.2	158.5	140.3	111.8	107.5	110.0	47.5	50.0	50.0
clip-L+dinov2-L	195.0	148.3	**143.3**	**195.0**	131.3	122.6	158.0	147.3	119.3	140.0	105.7	47.5	50.0	50.0
clip-B+siglip-B, dino-B	195.0	151.7	136.7	175.0	107.8	108.8	158.3	140.5	105.3	125.0	110.0	65.0	50.0	55.0
(b) combine features by increasing image token number														
clip-L+siglip-L	190.0	145.0	120.0	185.0	117.3	120.6	159.3	158.0	117.5	132.5	101.4	42.5	**55.0**	**67.5**
clip-L+dinov2-L	**200.0**	158.3	141.7	180.0	136.1	117.6	160.5	155.3	**122.8**	122.5	107.1	57.5	50.0	55.0
clip-L+siglip-L, dinov2-L	185.0	**160.0**	130.0	178.3	**137.4**	**137.1**	161.3	**165.3**	**122.8**	**147.5**	109.3	**80.0**	50.0	62.5

Table 12: Detailed results on the MME dataset (Supplementary to Table 6). In the table, CR, NC, TT, and CR correspond to commonsense reasoning, numerical calculation, text translation, and code reasoning, respectively.

Image Encoder	**MathVista**					**MathVerse**			**HC-M3D**		
	TextQA	**VQA**	**geometry**	**Math Word**	**FigureQA**	**Plane Geometry**	**Functions**	**Solid Geometry**	**DI**	**BC**	**AG**
clip-B	33.5	27.4	62.0	17.2	19.0	29.7	17.1	23.0	39.6	8.6	62.0
clip-L	**38.6**	**35.2**	62.0	**21.0**	21.2	**31.4**	**19.1**	**24.7**	**40.8**	**14.5**	**53.4**
(a) combine features by concatenating the hidden feature											
clip-B+ siglip-B	37.3	24.0	**64.4**	12.9	19.0	27.8	18.1	22.2	37.5	11.0	56.6
clip-L+dinov2-L	36.1	25.7	59.1	11.3	19.7	29.8	18.1	23.2	38.8	12.1	60.8
clip-B+siglip-B, dino-B	38.6	26.8	60.6	12.4	19.3	28.8	17.5	22.5	38.3	9.8	58.3
(b) combine features by increasing image token number											
clip-L+siglip-L	31.7	31.8	54.8	14.0	**22.7**	29.4	16.9	22.6	37.7	11.0	59.4
clip-L+dinov2-L	32.3	24.0	53.9	15.1	19.7	29.4	15.7	22.3	37.3	9.8	55.0
clip-L+siglip-L, dinov2-L	29.1	24.6	58.7	11.8	19.0	29.4	16.9	22.6	36.8	9.6	57.1

Table 13: Detailed results on MathVista, MathVerse, and HC-M3D (Supplement to Table 6). Except for the AG (agreement) metric (lower is better), all other metrics indicate higher is better.



Instruction for Step 1:Determining image dependency

Instruction: Assess whether solving the following problem necessarily requires images.

Examples:

如图是一株美丽的勾股树，其中所有四边形都是正方形，所有的三角形都是直角三角形，若正方形A、B的面积分别为5、3，则最大正方形C的面积是（）

△ABC的两内角平分线OB、OC相交于点O，若∠A＝110°，则∠BOC＝（）

Needs Image

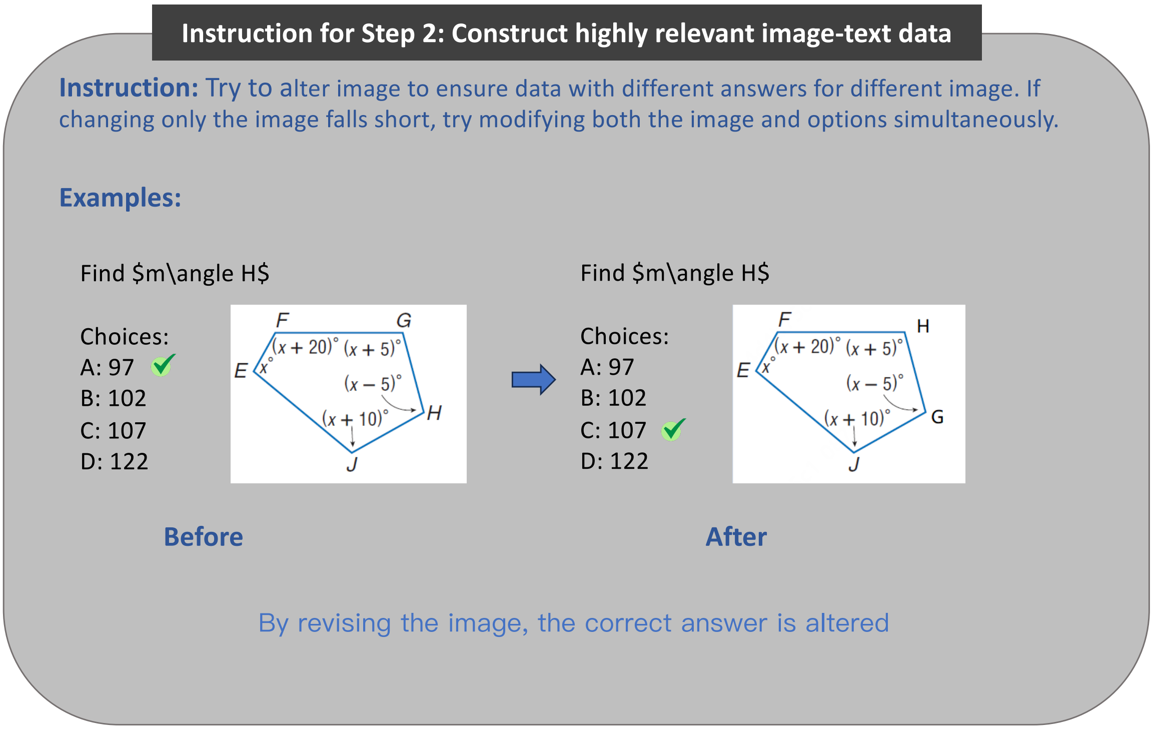
No image is needed, a picture can be drawn from the question. Let's rewrite !

Rewrite Requirements: By reducing textual information and necessitating the extraction of problem-solving information from images, or by concurrently revising both images and text.

△ABC的两内角平分线OB、OC相交于点O，~~若∠A＝110°~~，则∠BOC＝（）

Text is redundant, information can be obtained from images.

Figure 5: Instruction for step 1: determining image dependency.



Instruction for Step 2: Construct highly relevant image-text data

Instruction: Try to alter image to ensure data with different answers for different image. If changing only the image falls short, try modifying both the image and options simultaneously.

Examples:

Find $m\angle H$

Choices:
A: 97 ✓
B: 102
C: 107
D: 122

Find $m\angle H$

Choices:
A: 97
B: 102
C: 107 ✓
D: 122

Before

After

By revising the image, the correct answer is altered

Figure 6: Instruction for step 2: construct highly relevant image-text data.