CMM-Math: A Chinese Multimodal Math Dataset To Evaluate and Enhance the Mathematics Reasoning of Large Multimodal Models

Wentao Liu*
Qianjun Pan*
wtliu@stu.ecnu.edu.cn
East China Normal University
Shanghai, China

Yi Zhang
East China Normal University
Shanghai, China

Zhou Liu
East China Normal University
Shanghai, China

Ji Wu
East China Normal University
Shanghai, China

Jie Zhou
East China Normal University
Shanghai, China
jzhou@cs.ecnu.edu.cn

Aimin Zhou
East China Normal University
Shanghai, China
amzhou@cs.ecnu.edu.cn

Qin Chen
East China Normal University
Shanghai, China

Bo jiang
East China Normal University
Shanghai, China

Liang He
East China Normal University
Shanghai, China

Abstract

Large language models (LLMs) have obtained promising results in mathematical reasoning, a foundational human intelligence skill. Most previous studies focus on improving or measuring the performance of LLMs via textual math datasets (e.g., MATH, GSM8K). Recently, a few researchers released English multimodal math datasets (e.g., MATHVISTA and MATH-V) to evaluate the effectiveness of large multimodal models (LMMs). In this paper, we release a Chinese multimodal math (CMM-Math) dataset, including benchmark and training parts, to evaluate and enhance the mathematical reasoning of LMMs. CMM-Math contains over 28,000 high-quality samples, featuring a variety of problem types (e.g., choice, fill-in-the-blank, analysis) with detailed solutions across 12 grade levels from elementary to high school in China. The problem may contain multiple images, and the visual context may be present in the questions or opinions, which makes this dataset more challenging. Our comprehensive analysis reveals that state-of-the-art LMMs on the CMM-Math dataset face challenges, emphasizing the necessity for further improvements in LMM development. We also propose a Multimodal Mathematical LMM (`Math-LMM`) to handle the problems with mixed input of multiple images and text segments. The `Math-LMM` is trained using three stages: foundational pre-training, foundational fine-tuning, and mathematical fine-tuning. The extensive experiments indicate that our model effectively improves math reasoning performance by comparing it with the SOTA LMMs over three multimodal mathematical datasets. We will release the datasets, codes, and weights on github[1] and huggingface [2].

Keywords

Mathematical Reasoning, Large Multimodal Models, Benchmark, Chinese

ACM Reference Format:
Wentao Liu, Qianjun Pan, Yi Zhang, Zhou Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo jiang, and Liang He. 2025. CMM-Math: A Chinese Multimodal Math Dataset To Evaluate and Enhance the Mathematics Reasoning of Large Multimodal Models. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Both authors contributed equally to this research.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1]https://github.com/ECNU-ICALK/EduChat-Math
[2]https://huggingface.co/datasets/ecnu-icalk/cmm-math

Supplementary Materials

This supplementary materials are organized as follows.

- In Section A, we provide more details about the dataset, covering manual inspection and correction, data labeling standards, classification of question types, and subjects descriptions.
- In Section B, we present additional content for the experiment, including evaluation prompt design and the performance outcomes of LLMs across different mathematical subjects.

A Dataset Details

Table 1 presents the detailed statistical results of the dataset.

Table 1: Detailed statistics of CMM-Math datasets.

Statistic	#Evaluation	#Training	#Total
Total problems	5,821	22,248	28,069
Total images	3,794	11,419	15,213
Total detailed solutions	5,204	18,621	23,825
Images in questions	2,144(56.51%)	7,346(64.33%)	9,490(62.38%)
Images in answers	1,650(43.49%)	4,073(35.67%)	5,723(37.62%)
Type	4	4	4
- Choice	2,222(38.17%)	8,618(38.74%)	10,840 (38.62%)
- fill-in-the-blank	1,668(28.65%)	6,382(28.69%)	8,050(28.68%)
- Yes-no	18(0.31%)	88 (0.40%)	106(0.38%)
- Analysis	1,913(32.86%)	7,170(32.23%)	9,083 (32.36%)
Level	12	12	12
- Level-1	319(5.48%)	1,180(5.30%)	1,499(5.34%)
- Level-2	439(7.54%)	1,648(7.41%)	20,87(7.44%)
- Level-3	444(7.63%)	1,680(7.55%)	2,124(7.57%)
- Level-4	574(9.86%)	2,210(9.93%)	2,784(9.92%)
- Level-5	534(9.17%)	1,939(8.72%)	2,473(8.81%)
- Level-6	463(7.95%)	1,783(8.01%)	2,246(8.00%)
- Level-7	458(7.87%)	1,751(7.87%)	2,209(7.87%)
- Level-8	361(6.20%)	1,372(6.17%)	1,733(6.17%)
- Level-9	493(8.47%)	1,900(8.54%)	2,393(8.53%)
- Level-10	587(10.08%)	2,284(10.27%)	2,871(10.23%)
- Level-11	646(11.10%)	2,512(11.29%)	3,158(11.25%)
- Level-12	503(8.64%)	1,989(8.94%)	2,492(8.88%)
Subjects	13	13	13
- Analytic Geometry	707(12.15%)	2,756(12.39%)	3,463(12.34%)
- Metric Geometry	738(12.68%)	2,876(12.93%)	3,614(12.88%)
- Solid Geometry	546(9.38%)	2,092(9.40%)	2,638(9.40%)
- Arithmetic	1,999(34.34%)	7,855(35.31%)	9,854(35.11%)
- Algebra	676(11.61%)	2,640(11.87%)	3,316(11.81%)
- Counting	407(6.99%)	1,546(6.95%)	1,953(6.96%)
- Transformation Geometry	85(1.46%)	274(1.23%)	359(1.28%)
- Graph Theory	26(0.45%)	44(0.20%)	70(0.25%)
- Combinatorial Geometry	140(2.41%)	495(2.22%)	635(2.26%)
- Combinatorics	217(3.73%)	747(3.36%)	964(3.43%)
- Logic	127(2.18%)	416(1.87%)	551(1.96%)
- Descriptive Geometry	135(2.32%)	465(2.09%)	603(2.15%)
- Statistics	18(0.31%)	42(0.19%)	60(0.21%)

A.1 Data Labeling Standards

We hired 12 human annotators to inspect and correct the questions. We paid each annotator ¥0.5 per question for inspection and ¥3 per question for correction. To ensure the accuracy of the question information, we conducted two rounds of inspections and one round of review (after correction) for each question.

A.2 Manual Inspection and Correction

During the data cleaning phase, we check and correct issues related to text and image recognition. Figure 1 provides examples of errors in text and image recognition when converting PDFs to Markdown format. In the examples of text incorrectly identified as images, typical errors involve text containing special symbols or formulas, leading to these elements being misidentified as images. Additionally, some images were incorrectly recognized as text, resulting in garbled characters in the text. We asked human annotators to check and manually correct these errors.

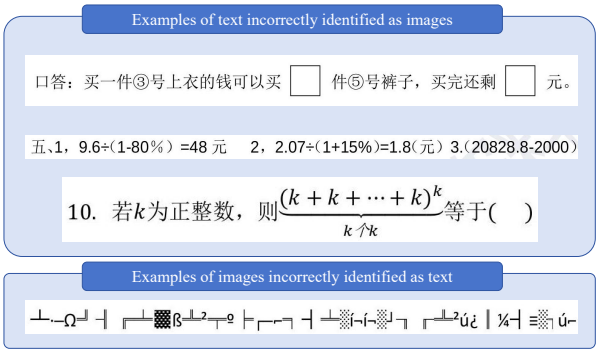

Figure 1: Examples of text and image recognition errors.

A.3 Classification of Question Types

Since the original materials come from real-world scenarios, the questions cover a variety of types, including single-choice, multiple-choice, fill-in-the-blank, true/false, analysis, calculation, problem-solving, and application. These questions were mapped to our defined question types to facilitate subsequent evaluation and testing. Specifically, single-choice and multiple-choice questions were mapped to multiple-choice questions to allow for comparing model performance using accuracy. Analysis, calculation, and problem-solving questions were mapped to Analysis questions to evaluate model output using the GPT-4 Score. Meanwhile, the types of fill-in-the-blank and true/false questions were retained. Therefore, our CMM-Math actually features a more diverse range of question types.

A.4 Subjects Descriptions

We designed 13 subjects, referencing MATH-V [9], including logic, algebra, counting, arithmetic, combinatorics, graph theory, topology, statistics, solid geometry, metric geometry, analytic geometry, descriptive geometry, combinatorial geometry, and transformation geometry. Detailed descriptions of these subjects can be found in the appendix of MATH-V [9]. Unlike MATH-V [9], we did not subdivide metric geometry because we found it difficult to finely categorize metric geometry into metric geometry - angle, metric geometry - area, and metric geometry - length subjects in CMM-Math.

B Experiment Details

B.1 Evaluation Prompt Design

In our experiments, we designed three prompt formats: zero-shot and 3-shot prompt templates for obtaining model outputs, and a scoring prompt template for calculating the GPT-4 score metric to evaluate fill-in-the-blank and analysis problems.

The zero-shot and 3-shot prompts are designed as shown in Figure 2. Since the questions of CMM-Math are in Chinese, we

Table 2: Comparison of model performances in accuracy across 13 mathematical subjects. Alg: algebra, AnaG: analytic geometry, Ari: arithmetic, CombG: combinatorial geometry, Comb: combinatorics, Cnt: counting, Desc: deseriptive geometry, GrphT: graph theory, Log: logic, MetG: metric geometry, SolG: solid geometry, Stat: statistics, TransG: transformation geometry.

Models	Overall	Alg	AnaG	Ari	CombG	Comb	Cnt	Desc	GrphT	Log	MetG	SolG	Stat	TransG
Open-source LMMs														
CogVLM2	25.85	21.94	25.22	24.76	28.85	24.68	26.54	29.79	44.44	19.05	33.79	23.36	45.45	25.58
InternLM-VL	19.82	21.94	26.38	20.68	21.15	12.99	16.67	19.15	0.00	9.52	20.48	12.15	27.27	13.95
Qwen2-VL-Instruct	43.04	36.45	42.03	50.65	36.54	40.26	45.68	27.66	44.44	39.68	43.34	37.85	45.45	37.21
LLaVA-v1.5	18.08	15.16	13.91	21.82	15.38	18.18	12.96	21.28	22.22	12.70	19.45	19.16	36.36	25.58
LLaVA-v1.6-mistral	16.83	17.10	16.52	22.31	7.69	11.69	10.49	8.51	0.00	19.05	13.99	15.42	27.27	16.28
CogVLM2 (3-Shot)	31.21	30.97	28.12	32.57	32.69	24.68	30.25	29.79	22.22	25.40	36.18	30.84	45.45	27.91
InternLM-VL (3-Shot)	25.09	29.03	30.14	24.76	15.38	27.27	25.93	12.77	0.00	26.98	24.57	17.76	9.09	25.58
Qwen2-VL-Instruct (3-Shot)	46.29	37.42	40.58	53.42	46.15	40.26	56.17	36.17	33.33	46.03	49.15	40.19	54.55	51.16
LLaVA-v1.5 (3-Shot)	19.69	19.03	14.20	28.66	3.85	23.38	21.60	14.89	11.11	15.87	14.68	16.36	45.45	2.33
LLaVA-v1.6-mistral (3-Shot)	21.88	22.26	17.97	32.25	5.77	20.78	22.22	4.26	11.11	28.57	15.70	13.55	27.27	16.28
Math-LMM (Ours 7B)	32.10	26.13	29.86	36.64	28.85	32.47	35.19	27.66	33.33	30.16	35.15	28.04	36.36	25.58
Math-LMM (Ours 72B)	48.57	48.06	44.35	60.59	32.69	50.65	53.70	17.02	22.22	57.14	44.03	35.98	63.64	27.91
Closed-source LMMs														
Qwen-VL-Max	49.91	49.68	48.12	60.59	44.23	50.65	54.94	21.28	11.11	50.79	42.66	36.45	63.64	51.16
Gemini	41.88	36.77	37.97	55.37	26.92	37.66	48.15	31.91	33.33	26.98	37.54	32.71	54.55	25.58
GPT-4o	29.02	26.13	28.41	35.67	26.92	20.78	32.10	25.53	33.33	28.57	24.57	21.96	18.18	37.21
Qwen-VL-Max (3-Shot)	64.91	70.00	63.19	74.43	50.00	66.23	69.75	21.28	44.44	73.02	54.61	57.01	63.64	53.49
Gemini (3-Shot)	41.65	38.71	35.07	54.07	19.23	31.17	51.85	27.66	44.44	44.44	36.18	30.84	54.55	44.19
GPT-4o (3-Shot)	65.98	62.90	59.71	82.57	46.15	66.23	72.84	40.43	77.78	68.25	59.04	47.66	81.82	55.81
Mean accuracy of LMMs	35.66	33.87	33.43	42.88	27.14	33.33	38.17	23.17	27.16	34.57	33.62	28.74	44.44	31.27

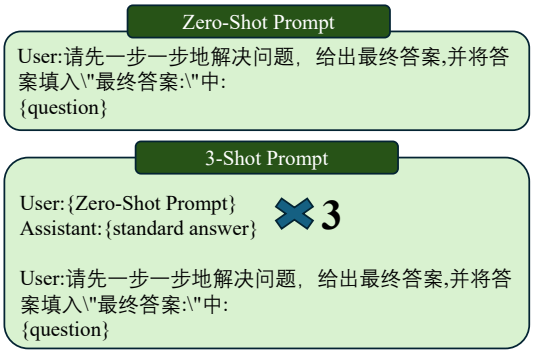

Figure 2: Prompts for evaluation.

designed the prompts to be in Chinese as well. The zero-shot prompt instructs the model to: Please solve the problem step by step, provide the final answer, and fill it in the "Final Answer:" field. The 3-shot prompt is similar to the zero-shot prompt but includes three rounds of questions (from the user) and standard answers (from the assistant) in the model's input history to guide the model towards better performance.

GPT-4o [7] is used to calculate the GPT-4o score metric of generated answers by providing the Questions, Standard answers, and Outputs of Models. The prompt format we designed is shown in Figure 3. Here, {question}, {options}, {answer}, and {analysis} represents the standard information provided by the test data. The {options}, will be ignored if the test question is not a multiple-choice question. {ansA} represents the generated response by the test model, which is used to compare with the standard information to obtain the score. We will subsequently open-source the related evaluation code along with the model and data.

Prompt:
Please, as a rigorous and impartial mathematical referee, conduct a comprehensive evaluation of the Chinese mathematical problem solutions provided by this language model (Model A). The evaluation criteria are the accuracy, completeness, logicality, and depth of understanding of the answer. You will receive standard answers and explanations, as well as answers to Model A. Please note: objectivity and fairness: avoid subjective factors such as model name, answer length, or other factors affecting the rating. Comprehensive evaluation: Compare the model answer with the standard answer and analyze the reasons. Avoid bias: It is not advisable to set preferences for the model in advance. Format specification: Please strictly output the final rating in the following format: Model A rating: [1-10 points]

[Question]
{question} {options}
[Standard answer]
{answer} {analysis}

[Output of Model A]: {ansA}

Figure 3: The prompt used for scoring with GPT-4o.

B.2 Detailed Experimental Results of SOTA LLMs on CMM-Math

Tables 2 and 3 present the accuracy and GPT-4o scores of LMMs across 13 mathematical subjects, respectively. These results are used to compare the performance of LMMs on different mathematical topics. In addition, Tables 3 and 4 show the GPT-4o scores of LMMs

Table 3: Comparison of model performances in GPT-4o score across 13 mathematical subjects.

Models	Overall	Alg	AnaG	Ari	CombG	Comb	Cnt	Desc	GrphT	Log	MetG	SolG	Stat	TransG
Open-source LMMs														
CogVLM2	2.82	2.60	2.44	3.21	2.81	2.44	2.29	2.05	2.41	2.89	2.91	2.41	2.71	3.36
InternLM-VL	4.48	4.51	4.02	5.24	2.69	4.28	4.08	2.11	3.82	4.75	4.16	3.78	5.14	3.07
Qwen2-VL-Instruct	5.26	5.73	4.77	6.07	3.81	4.30	4.87	3.69	2.94	4.92	4.71	4.38	7.43	4.00
LLaVA-1.5	2.56	2.46	2.38	2.72	2.45	2.30	2.56	2.51	2.35	2.31	2.37	2.63	3.43	2.90
LLaVA-v1.6-mistral	2.81	2.47	3.35	2.68	2.73	2.60	3.09	3.57	3.71	2.52	2.93	2.61	3.86	3.74
CogVLM2 (3-Shot)	2.72	2.53	2.62	2.94	2.42	2.59	2.65	1.90	2.12	2.84	2.68	2.51	2.57	3.14
InternLM-VL (3-Shot)	4.35	4.58	4.14	5.08	2.91	4.25	3.73	2.06	3.06	4.44	3.91	3.65	4.71	2.93
Qwen2-VL-Instruct (3-Shot)	4.09	3.83	3.67	4.91	3.61	3.34	3.60	3.26	2.82	3.53	3.75	3.23	5.43	2.95
LLaVA-v1.5 (3-Shot)	3.34	2.91	3.92	2.74	5.36	2.76	3.23	5.27	3.82	2.95	4.04	3.86	3.43	4.90
LLaVA-v1.6-mistral (3-Shot)	3.78	3.27	4.07	2.94	5.90	3.08	3.51	6.24	4.82	3.27	5.05	5.02	2.43	4.83
Math-LMM (Ours 7B)	2.46	2.19	2.23	2.76	2.40	1.90	1.84	2.41	2.06	2.62	2.40	2.53	1.43	2.43
Math-LMM (Ours 72B)	4.04	4.25	3.90	4.52	2.80	3.59	3.56	3.26	2.65	4.44	3.74	3.46	7.00	3.24
Closed-source LMMs														
Qwen-VL-Max	6.50	6.66	5.80	7.41	4.30	6.07	5.84	4.65	3.82	6.16	6.12	5.86	6.29	5.07
Gemini	6.02	5.65	5.29	6.72	5.72	5.15	5.32	5.17	5.47	5.30	6.06	5.71	6.00	5.05
GPT-4o	7.94	8.15	7.28	8.65	6.23	7.89	7.70	6.22	7.41	7.56	7.65	7.20	8.00	7.14
Qwen-VL-Max (3-Shot)	6.21	6.52	6.14	6.73	4.74	5.44	5.45	5.74	5.59	6.06	5.96	5.81	4.86	4.81
Gemini (3-Shot)	5.89	5.72	5.41	6.45	5.62	5.49	5.36	5.34	6.18	5.33	5.78	5.41	6.14	5.02
GPT-4o (3-Shot)	7.85	7.99	7.07	8.61	6.12	7.74	7.51	6.31	8.24	7.52	7.46	7.22	8.57	6.90
Mean accuracy of LMMs	4.62	4.56	4.36	5.02	4.03	4.18	4.23	3.99	4.07	4.41	4.54	4.29	4.97	4.19

Table 4: Comparison of model performances in GPT-4o score across different levels. The levels from 1 to 12 correspond to primary to high school grades. The maximum score is 10.

Models	Overall	LV1	LV2	LV3	LV4	LV5	LV6	LV7	LV8	LV9	LV10	LV11	LV12
Open-source LMMs													
CogVLM2	2.82	2.96	3.66	3.52	2.99	2.29	2.90	2.46	3.05	2.37	2.43	2.45	2.43
InternLM-VL	4.48	4.05	5.54	5.27	4.90	4.20	5.12	4.00	4.03	2.60	4.13	4.41	4.35
Qwen2-VL-Instruct	5.26	4.23	5.54	6.07	5.40	5.18	6.05	5.51	5.53	4.21	4.99	4.94	4.64
LLaVA-v1.5	2.56	2.51	2.80	2.96	2.76	2.50	2.37	2.66	2.66	2.37	2.43	2.27	2.34
LLaVA-v1.6-mistral	2.81	2.55	2.88	2.94	2.57	2.43	2.42	2.91	3.36	3.85	3.04	2.68	2.59
CogVLM2 (3-Shot)	2.72	2.73	3.53	3.18	2.78	2.30	2.66	2.35	2.97	2.31	2.46	2.48	2.74
InternLM-VL (3-Shot)	4.35	3.95	5.46	5.06	4.62	3.77	4.78	3.96	4.05	3.08	4.11	4.31	4.36
Qwen2-VL-Instruct (3-Shot)	4.09	3.94	4.99	4.94	4.28	3.82	4.73	3.75	4.06	3.28	3.61	3.49	3.47
LLaVA-v1.5 (3-Shot)	3.34	2.75	2.75	3.10	2.70	3.14	2.49	3.77	4.89	5.79	3.75	2.88	3.42
LLaVA-v1.6-mistral (3-Shot)	3.78	3.27	3.28	3.50	3.30	3.86	2.62	3.78	5.67	6.23	4.31	3.05	3.73
Math-LMM (Ours 7B)	2.46	2.27	3.17	2.95	2.97	2.40	1.83	2.16	2.67	2.16	2.13	2.26	2.11
Math-LMM (Ours 72B)	4.04	3.61	4.66	4.52	3.71	3.97	4.49	3.81	3.84	3.64	4.00	3.93	3.86
Closed-source LMMs													
Qwen-VL-Max	6.50	5.91	7.30	7.54	7.11	6.61	7.02	6.1	5.91	4.73	6.01	6.18	6.17
Gemini	6.02	6.30	6.85	6.83	6.46	5.75	6.40	5.67	6.46	5.01	5.42	4.96	5.62
GPT-4o	7.94	7.44	8.70	8.76	8.34	8.04	8.30	7.83	7.65	6.44	7.67	7.48	7.74
Qwen-VL-Max (3-Shot)	6.21	5.66	6.70	6.43	6.18	6.12	6.91	6.00	5.72	5.50	5.67	6.57	6.51
Gemini (3-Shot)	5.89	5.83	6.54	6.47	6.06	5.44	6.26	5.51	6.59	5.36	5.77	5.21	5.44
GPT-4o (3-Shot)	7.85	7.27	8.53	8.63	8.45	8.09	8.36	7.87	7.65	6.06	7.48	7.22	7.46
Mean accuracy of LMMs	4.62	4.29	5.16	5.15	4.75	4.44	4.76	4.45	4.82	4.17	4.41	4.26	4.39

after evaluation by GPT-4o. These scores are used to further analyze the models' performance on short-answer questions.

Comparisons Across Different Subjects. Most LMMs excel in arithmetic and statistics but show limited proficiency in geometry (see Appendix). We can observe that the mean accuracy of all LMMs across all subjects is 35.66, with LMMs achieving an accuracy exceeding 40 in arithmetic and statistics. In contrast, in geometry, including analytical, combinatorial, descriptive, metric, and transformation geometry, the accuracy of LMMs is below 35.66, with the worst performance at 23.17 for descriptive geometry.

Disparity Between Accuracy and GPT-4o Score. When comparing the performance of LMMs in both accuracy and GPT-4o score, we observe that closed-source models maintain consistent performance while open-source models show inconsistencies. Specifically, GPT-4o consistently ranks highest, while Gemini ranks lowest. CogVLM2 has higher accuracy than InternLM-VL but is lower on the GPT-4o score. This discrepancy may arise from differences in the training data distribution, where the datasets might focus more on choice and yes-no questions while lacking materials for problem-solving and analytical tasks. Moreover, this inconsistency could also result from insufficient training data, leading to models that obtain correct answers but struggle to provide high-quality analytical processes. This highlights that our CMM-Math, unlike other math test sets that offer only choice questions, provides a more comprehensive evaluation of models' analytical and problem-solving abilities.

Conversely, regarding the GPT-4o score, most models perform better with zero-shot prompting than with few-shot prompting, except for the LLaVA-v1.5 and LLaVA-v1.6-mistral. This suggests that few-shot prompting does not necessarily lead to higher-quality outputs. Our view is that the content of few-shot prompts may influence how models conduct their own analysis and reasoning, potentially leading to poorer performance.

Furthermore, regarding the GPT-4o score, we notice that `Math-LMM` achieves only ordinary results, while Qwen2-VL-Instruct obtains

Table 5: Comparison of model performances on MATH-V across various mathematical subjects. Alg: algebra, AnaG: analytic geometry, Ari: arithmetic, CombG: combinatorial geometry, Comb: combinatorics, Cnt: counting, DescG: descriptive geometry, GrphT: graph theory, Log: logic, Angle: metric geometry - angle, Area: metric geometry - area, Len: metric geometry-length, SolG: solid geometry, Stat: statistics, Topo: topology, TransG: transformation geometry.

Model	Overall	Alg	AnaG	Ari	CombG	Comb	Cnt	DescG	GrphT	Log	Angle	Area	Len	SolG	Stat	Topo	TransG
Open-source LMMs																	
LLaVA-v1.5-7B	8.52	7.00	7.10	10.70	7.10	4.80	10.50	7.70	10.00	9.20	15.60	10.20	9.80	5.30	8.60	4.40	4.80
SPHINX (V2)	9.70	6.70	7.10	12.90	7.50	7.70	6.00	9.60	16.70	10.10	11.00	11.80	12.50	8.20	8.60	8.70	6.00
ShareGPT4V-7B	10.53	5.50	3.60	12.90	10.10	4.80	7.50	11.50	14.40	10.90	16.20	11.80	12.30	9.80	15.50	17.40	11.30
LLaVA-v1.5-13B	11.12	7.00	14.30	14.30	9.10	6.60	6.00	13.50	5.60	13.50	10.40	12.60	14.70	11.50	13.80	13.00	10.70
ShareGPT4V-13B	11.88	7.50	15.50	16.40	10.70	8.90	9.00	11.50	8.90	7.60	11.60	13.00	17.40	10.30	8.60	8.70	12.50
SPHINX-MoE	14.18	7.80	17.90	14.30	15.60	9.50	11.90	12.50	15.6	12.60	16.20	15.60	17.80	13.50	12.10	8.70	16.10
InternLM-VL	14.54	9.30	15.50	12.10	15.30	11.30	10.50	14.40	22.20	19.30	19.70	15.6	15.00	11.90	15.50	26.10	15.50
Math-LMM (Ours 7B)	11.58	7.30	8.30	10.70	14.00	7.10	7.40	16.40	12.20	9.20	14.50	10.60	14.90	9.00	8.60	26.10	16.70
Math-LMM (Ours 72B)	17.53	10.70	28.60	15.00	20.10	11.30	11.90	15.40	16.70	21.00	22.50	18.40	20.00	15.60	20.70	8.70	19.60
Closed-source LMMs																	
Qwen-VL-Plus	10.72	11.30	17.90	14.30	12.70	4.80	10.50	15.40	8.90	14.30	11.60	6.40	10.00	14.30	6.90	8.70	11.31
Qwen-VL-Max	15.59	10.70	19.10	20.00	16.90	12.50	17.90	16.40	12.20	21.00	13.30	14.20	19.80	11.50	20.70	13.00	17.30
Gemini Pro	17.66	15.10	10.70	20.70	20.10	11.90	7.50	20.20	21.10	16.80	19.10	19.00	20.00	14.30	13.80	17.40	20.80
GPT-4V	22.76	27.30	32.10	35.70	21.10	16.70	13.40	22.10	14.40	16.80	22.00	22.20	20.90	23.80	24.10	21.70	25.60
Human Performance																	
Human (testmini)	75.66	57.90	79.00	100.00	100.00	47.40	94.70	89.50	63.20	63.20	36.80	52.60	73.70	89.50	89.50	100.00	73.70

the best result of 5.26. This could be because the training data for Math-LMM is still limited, leading to weaker language expression and analytical reasoning capabilities.

B.3 Details of the experimental results of Math-LMM on MATH-V.

We compare our Math-LMM with existing LLMs over MATH-V [9], including Qwen-VL-Plus [1], Qwen-VL-Max [1], Gemini Pro [8], GPT-4V [7], LLaVa-v1.5-7B [5], SPHINX [4], ShareGPT-4V-7B/13B [2], LLaVa-v1.5-13B [5], InternLM-XComposer2-VL [3], and SPHINX-MoE [4].

In Table 5, our Math-LMM (72B) achieves the best performance among in open-source models. Although the Math-LMM (7B) does not achieve the second-best performance, compared to other open-source models of the same parameter scale, such as LLaVa-v1.5-7B [5], SPHINX [4], and ShareGPT-4V-7B [2], Math-LMM (7B) still achieves the best performance. This may be because MATH-V [9] is a more challenging test set than MATHVISTA [6], where model parameter scale has a greater impact on performance. Meanwhile, among open-source models, Math-LMM almost achieves the best performance across all 16 subjects, except for the subjects of arithmetic and graph theory. Notably, compared to closed-source models, Math-LMM also achieves the best performance among all models on the subjects of logic, metric geometry-angle, and topology. These results indicate that Math-LMM also has a certain level of competitiveness on harder English multimodal mathematical problems.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966

[2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. arXiv:2311.12793

[3] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model. *arXiv preprint arXiv:2401.16420* (2024).

[4] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. 2023. SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models. arXiv:2311.07575

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *NeurIPS* 36 (2024).

[6] 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 *ICLR*.

[7] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[8] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, and Michael Isard. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805

[9] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. *arXiv preprint arXiv:2402.14804* (2024).