StatsChartMWP: A Dataset for Evaluating Multimodal Mathematical Reasoning Abilities on Math Word Problems with Statistical Charts

Dan Zhu
TAL Education Group
Beijing, China
zhudan11@tal.com

Tianqiao Liu
TAL Education Group
Beijing, China
liutianqiao1@tal.com

Zitao Liu*
Jinan University
Guangzhou, China
liuzitao@jnu.edu.cn

Abstract

Recent advancements in Large Multimodal Models (LMMs) have showcased their impressive capabilities in mathematical reasoning tasks in visual contexts. As a step toward developing AI models to conduct rigorous multi-step multimodal reasoning, we introduce StatsChartMWP, a real-world educational dataset for evaluating visual mathematical reasoning abilities on math word problems (MWPs) with statistical charts. Our dataset contains 8,514 chart-based MWPs, meticulously curated by K-12 educators within real-world teaching scenarios. We provide detailed preprocessing steps and manual annotations to help evaluate state-of-the-art models on StatsChartMWP. Comparing baselines, we find that current models struggle in undertaking meticulous multi-step mathematical reasoning among technical languages, diagrams, tables, and equations. Towards alleviate this gap, we introduce CoTAR, a chain-of-thought (CoT) augmented reasoning solution that finetunes the LMMs with solution-oriented CoT-alike reasoning steps. The LMM trained with CoTAR is more effective than current open-source approaches. We conclude by shedding lights on challenges and opportunities in enhancement in LMMs and steer future research and development efforts in the realm of statistical chart comprehension and analysis. The code and data are available at https://github.com/ai4ed/StatsChartMWP.

1 Introduction

Recently developed Large Multimodal Models (LMMs), exemplified by the latest GPT-4o (OpenAI, 2024) and Qwen2-VL (Wang et al., 2024b), show promising capabilities of understanding both images and texts (Liu et al., 2024b). However, even the latest LMMs often falter when required to perform multi-step mathematical reasoning in visual contexts (Cobbe et al., 2021). Visual reasoning tasks require LMMs to interpret comprehensive mathematical implications from combined representations of visual entities, i.e., diagrams and tables, and natural languages, i.e., question content and execute stringent multi-step reasoning based on these learned representations simultaneously. This exposes a significant shortcoming in contemporary LMMs.

Various visual question answering (VQA) datasets, including FigureQA (Kahou et al., 2017), PlotQA (Methani et al., 2020), and ChartQA (Masry et al., 2022), have been developed to evaluate the mathematical reasoning capabilities of LMMs in visual contexts. As shown in Figure 1, which presents example images and question-answer pairs from these datasets, the average number of reasoning steps required (as analyzed by GPT-4) is surprisingly low - merely 0.09 for ChartQA and 0.45 for FigureQA. This reveals a critical limitation: most questions in these datasets only require direct visual extraction of information rather than sophisticated mathematical reasoning. This shallow level of cognitive demand fails to assess LMMs' true capabilities in complex visual-mathematical reasoning. Therefore, there is a pressing need for a more challenging benchmark that promotes the development of sophisticated mathematical reasoning in visual contexts and better evaluates the ability of LMMs to handle tasks requiring deep integration of visual understanding and mathematical reasoning.

In this paper, we choose to use math word problems (MWPs) with statistical charts to assess the multi-step mathematical reasoning abilities of LLMs and LMMs in visual contexts. A typical MWP with statistical charts is made up of mixed question contents of natural languages, notations and equations and a statistical chart that expresses mathematical quantities and relations. We provide a real-world example in Figure 1, where the dataset

*The corresponding author: Zitao Liu

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12944–12954
November 4-9, 2025

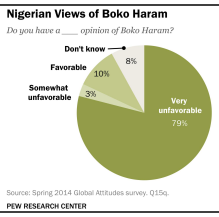
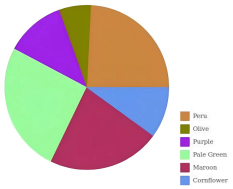
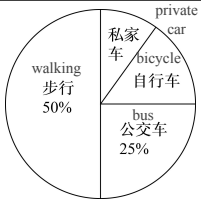
Dataset	ChartQA	FigureQA	StatsChartMWP
Image			
QA	Q: What percentage of people are favorable? A: 10 Q: What is the difference between the very unfavorable and the Favorable? A: 69	Q: Is Olive the minimum? A: Yes Q: Is Purple greater than Olive? A: Yes	学校为了解本校初三年级学生上学的交通方式，随机抽取了本校20名初三学生进行调查，其中有2名学生是乘私家车上学，图是收集数据后绘制的扇形图.如果该校初三年级有200名学生，那么骑自行车上学的学生大约有____人. Q: In order to understand the transportation methods of the students in the third year of junior high school, the school randomly surveyed 20 students, among which 2 students commute to school by private car. The pie chart was drawn after the data was collected. If there are 200 students in the third year of junior high school, then the number of students who commute to school by bicycle is approximately _____. A: 30
R-Steps	0.09	0.45	2.53

Figure 1: An example illustrating the enhanced mathematical reasoning capabilities of our dataset in comparison to other chart available datasets. R-Steps means the average reasoning steps of the dataset.

has an average reasoning step count of 2.53. Successfully solving the MWPs with statistical charts requires models to not only understand every mathematical quantity and relation, but conduct rigorous multi-step reasoning and computation.

We releasing StatsChartMWP, a dataset of 8.5K high-quality Chinese MWPs with statistical charts from elementary to high school levels. The StatsChartMWP dataset contains a rich variety of 11 chart types, including bar, line, line-function, pie, scatter, composite, radar, dual-axis, histograms, percentage-bar, tables, which provides opportunities to conduct fine-grained model performance analysis across different visual representations. Every corresponding statistical chart is meticulously scanned to guarantee optimal clarity and accuracy. Due to the space limit, we provide more StatsChartMWP questions samples from each category in Appendix A.1.

The StatsChartMWP dataset and its associated tasks introduce fresh avenues for research by presenting several technical hurdles: (1) the representation of innovative visual mediums of artificial figures such as diagrams, tables, and equations; (2) comprehension of technical language and equations; (3) the task of cross-modal alignment between figures and natural language; and (4) undertaking meticulous multi-step mathematical reasoning within visual contexts. Our quantitative studies reveal that current LMMs grapple with these challenges. To address the issue of weak multi-step multimodal mathematical reasoning abilities, we propose a **c**hain-**o**f-**t**hought (CoT) **a**ugmented **r**easoning approach, i.e., CoTAR, a solution of generating solution-oriented CoT-alike reasoning steps for LMMs training. While our approach yielding some advancement, the StatsChartMWP dataset continues to present unique challenges that promise to ignite future research in the fields of educational content modeling, multimodal reasoning, and question answering.

Overall this paper makes the following contributions:

- We present a comprehensive dataset of 8.5K K-12 school MWPs with 11 types of statistical charts and natural language solutions, useful for probing the informal reasoning ability in visual contexts of LLMs and LMMs.
- We conducted an in-depth analysis of error types and visual token lengths in current LMMs, providing insightful directions for future improvements in multi-step mathematical reasoning models within visual contexts.
- We introduce CoTAR, a data augmentation strategy that leverages Chain-of-Thought (Wei et al., 2022) reasoning to mitigate the cross-modal alignment issues between image and text representations. We can fine-tune any effective LMM base models with CoTAR-enhanced dataset to improve their multi-step mathematical reasoning abilities in visual contexts.

2 Related Work

Recently, many datasets have been curated to assess the mathematical reasoning abilities of AI systems within visual contexts. FigureQA (Kahou et al., 2017), PlotQA (Methani et al., 2020), ChartQA (Masry et al., 2022) and SimChart9K

Table 1: Comparison with existing datasets. OE: Open-ended, MC: Multiple-choice, FB: Fill-in-the-blank, J: Judgement. E: elementary school, M: middle school, H: high school, O: Olympiad-level, U: University-Level.

Dataset	Category	Math-skill	Task	Chart types	Multi-model	Question type	Grade-level
PlotQA	General VQA	Statistical	Chart Perception	3	×	OE	-
ChartQA	General VQA	Statistical	Chart Perception	3	×	OE	-
FigureQA	General VQA	Statistical	Chart Perception	5	×	OE	-
SimChart9K	General VQA	Statistical	Chart Perception	4	×	OE	-
TTC-QuAli	Math-Target	Statistical	Chart Perception	1	×	OE	-
TABMWP	Math-Target	Statistical	MWP	1	×	MC,OE	E,M
GeoQA	Math-Target	Geometry	MWP	-	×	MC	M,H
GeoQA+	Math-Target	Geometry	MWP	-	×	MC	M,H
UniGeo	Math-Target	Geometry	MWP	-	×	MC,OE	M,H
PGPS9K	Math-Target	Geometry	MWP	-	×	COM,MC	M,H
Geometry3K	Math-Target	Geometry	MWP	-	×	MC	M,H
GSM8K	Math-Target	Comprehensive	MWP	-	×	OE	E
U-MATH	Math-Target	Comprehensive	MWP	-	×	OE	U
MathVista	Math-Target	Comprehensive	MWP	5	×	MC,OE	-
MathVerse	Math-Target	Comprehensive	MWP	-	×	OE	H
CMM-Math	Math-Target	Comprehensive	MWP	-	×	MC,FB,J,OE	E,M,H
Math-Vision	Math-Target	Comprehensive	MWP	5	×	OE	O
OlympiadBench	Math-Target	Comprehensive	MWP	-	×	OE	O
StatsChartMWP	Math-Target	Statistical	Statistical WMP	11	×	MC,FB,J,OE	E,M,H

(Xia et al., 2023) are designed for chart comprehension and question-answering, serving as important tools for evaluating an AI model's capacity to interpret and analyze graphical data. However, these datasets generally focus more on the interpretation of charts. The generation of questions typically relies on predefined templates or manual annotations, and the answers are often directly retrievable from the charts themselves. TABMWP (Lu et al., 2023a) and TTC-QuAli (Dong et al., 2024) are datasets employed for resolving mathematical problems. However, this dataset offers solely tabular data, leading to a limited variety in the chart types offered. GeoQA (Chen et al., 2021), Geometry3K (Lu et al., 2021), GeoQA+ (Cao and Xiao, 2022), UniGeo (Chen et al., 2022) and PGPS9K (Zhang et al., 2023) have been developed to investigate multimodal mathematical reasoning capabilities. While these works provide a thorough evaluation for tackling mathematical geometry problems, all of these datasets and the corresponding approaches rely on dense annotation in formal language and require external geometric or symbolic programs or solvers to conduct multimodal mathematical reasoning. GSM8K (Cobbe et al., 2021) is commonly used to evaluate and enhance the capability of large language models in solving mathematical word problems. However, this dataset does not include visual images, making it unsuitable for assessing the mathematical reasoning abilities of multimodal models. OlympaidBench (He et al., 2024), MathVista (Lu et al., 2023b), Math-Vision (Wang et al., 2024a), MathVerse (Zhang et al., 2024b), CMM-Math (Liu et al., 2024c), U-MATH (Chernyshev et al., 2024) and We-Math (Qiao et al., 2024) provide comprehensive benchmarks for evaluating the LMMs' mathematical reasoning abilities. However, these datasets predominantly emphasize problems pertaining to geometry and functions, thereby overlooking the domain of statistics.

The StatsChartMWP dataset is the pioneer in providing question texts and visual diagrams encompassing a wide range of K-12 statistical knowledge, including various graphical types such as line charts, pie, and histograms, and is accessible to the research community. We summarize and compare recent work towards mathematical reasoning in visual contexts in Table 1.

3 StatsChartMWP

The dataset used with the CC0 1.0 license[1]. Besides visual reasoning benchmark task, the users

[1] https://creativecommons.org/publicdomain/zero/1.0

Table 2: Statistics of StatsChartMWP benchmark

Statistic	Number
Total Questions	8514
- multiple choice question	1091(12.81%)
- fill-in-the-blank questions	3415(40.11%)
- problem-solving questions	3974(46.68%)
- true or false question	34(0.4%)
Unique number of images	4801
Number of questions with split	2246
Number of questions without split	2555
Grade	
- primary school	3579(42.04%)
- middle school	2464(28.94%)
- high school	2471(29.02%)
Chart Type	11
maximum question length	1188
maximum answer length	266
average question length	151.25
average answer length	13.02

can use the dataset for other custom tasks such as building AI tutoring service under the license.

3.1 Data Statistics

The StatsChartMWP dataset comprises 8,514 unique MWPs with statistical charts and includes 11 different types of statistical chart representations. The distribution of these chart types is shown in Table 3, and we present the summarized data statistics in Table 2. For detailed definitions of each chart type and a statement on data provenance, please refer to Appendix A.1 and Appendix A.2, respectively.

3.2 Data Collection and Preprocessing

We mainly collect data from questions designed by teachers during lecture slides and homework assignments. Specifically, we collect Chinese MWPs with statistical charts from an online learning system that is developed by one of the largest educational technology companies in China. For each question, we collect both the question textual information and the corresponding visual charts. To construct a high-quality mathematical reasoning dataset containing explicit visual information, we conduct several simple yet critical data preprocessing and annotation steps to ensure the quality of the data, including decomposition, de-duplication and human annotation described as follows:

Decomposition. In the real-world educational scenarios, a MWP may contain multiple questions. To have a consistent and standardized evaluation, we manually decompose MWPs of multiple sub-problems into separate MWPs in our proposed dataset. A detailed decomposition of sub-questions, along with their explanations and illustrations, can be found in Appendix A.4.

De-duplication. We conduct a sample de-duplication process from both textual and visual perspectives to ensure there is no redundant MWPs in our dataset. In textual domain, we use a pre-trained BERT (Kenton and Toutanova, 2019) model to ascertain the extent of similarity across question texts and in visual domain, we apply a fine-tuned DePlot (Liu et al., 2023) model to filter out MWPs with statistical charts that contain repetitive mathematical information.

Human Annotation. We engage crowd workers with robust mathematical backgrounds to perform sample annotation and ensure data validity. Specifically, our annotators are initially tasked with eliminating charts that fail to meet our visual quality standards from the dataset. Following this, each question in StatsChartMWP is assigned a chart type from the specified 11 categories. At final, the annotators conduct a comprehensive review of all problem solutions in StatsChartMWP to confirm the accuracy of our ground-truth answers.

4 Experimental Setup

The StatsChartMWP dataset is designed to evaluate LMM's understanding and reasoning abilities of mathematical problems with statistical charts, as measured by its performance on correctly finding the solution. We conduct systematic evaluations on StatsChartMWP for state-of-the-art LLMs and LMMs, encompassing both closed- and open-source models and we are interested in understanding the performance gap between them. We also introduce CoTAR, a data augmentation method to enhance the multi-step reasoning capabilities of LMMs via mimicking the models' CoT process.

4.1 Baselines

LLMs: GPT-4 (Achiam et al., 2023) augmented with chart captions generated by GPT-4V and GPT-4o, respectively.

Closed-source LMMs: We select QwenVL-MAX (Bai et al., 2023b), GPT-4V (OpenAI, 2023) and GPT-4o (OpenAI, 2024), which have been repeatedly demonstrated to be among the most advanced

Table 3: Distribution of different chart types.

Chart	Hist	Bar	Line	Line-f	Scatter	D-axis	P-bar	Pie	Table	Comp	Radar
Distribution	21.55%	19.46%	13.81%	2.8%	1.88%	0.75%	0.26%	11.95%	16.48%	10.56%	0.28%

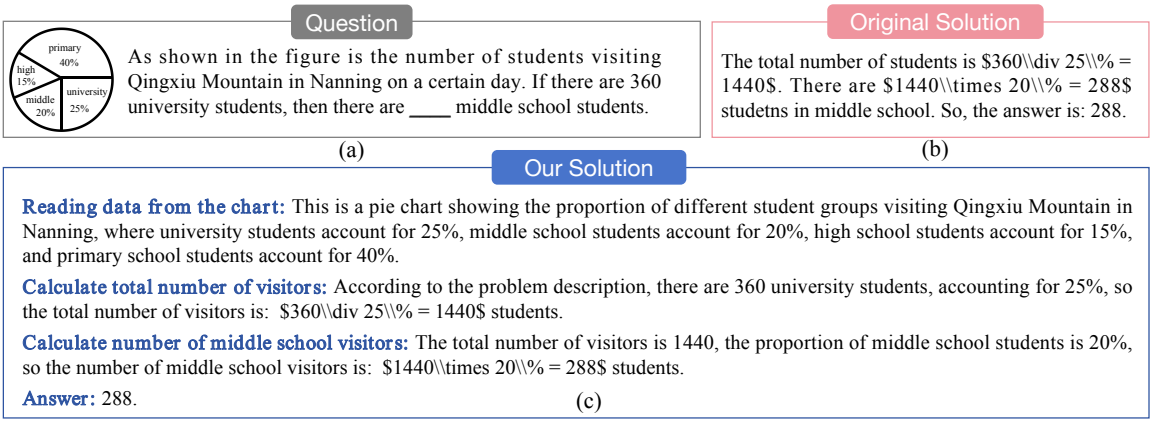


Figure 2: An illustration of CoTAR. (a) the original MWP with statistical chart. (b) the corresponding original solution. (c) the solution of CoTAR. The bold words are the step summaries and the following sentences are reasoning responses.

models available.

Open-source LMMs: HPT(HyperGAI, 2024): a dual network to learn both local and global features for vision-language alignment. DeepSeek-VL-7B (Lu et al., 2024): a hybrid vision encoder that efficiently processes high-resolution images within a fixed token budget, while maintaining a relatively low computational overhead. LLaVA-NeXT-34B (Liu et al., 2024a): a classic ViT-MLP-LLM architecture that exhibits superior visual reasoning and OCR capability by integrating visual instruction tuning data mixture. Cambrian-1 (Tong et al., 2024): a vision-centric multimodal large language model. This approach introduces the Spatial Visual Aggregator, a dynamic and spatially-aware connector that integrates high-resolution visual features with LLMs while reducing the number of tokens. InternLM-XC2d5 (Zhang et al., 2024a): a partial LoRA connector to interconnect vision encoder and an LLM. LLaVA-OneVision (Li et al., 2024) is a model capable of simultaneously advancing the performance boundaries of open-source LMMs in single-image, multi-image, and video scenarios. It allows for robust transfer learning across different modalities, thereby enabling new emergent capabilities. InternVL2 (OpenGVLab, 2024) and InternVL2.5 (Chen et al., 2024): enhancing the visual understanding capabilities of LMMs based on the large-scale visual foundation model InternVL-6B (Zhang et al., 2024a). Additionally, it can support 4K resolution input through adaptive resolution partitioning for images. Qwen2-VL (Wang et al., 2024b): introduces the naive dynamic resolution mechanism, enabling the model to dynamically process images of varying resolutions and convert them into different numbers of visual tokens. Additionally, it integrates multimodal rotary position embedding, facilitating the effective fusion of positional information across text, images, and videos.

Human performance: The human evaluation was conducted with approval from our institution's review board. We recruited 10 participants (5 male, 5 female; aged 22-25) after obtaining their informed consent. All participants held university degrees in relevant fields (mathematics, electronic information, computer science). Each participant individually undertook the evaluation on an identical set of test items, with a maximum duration of 5 minutes allocated per item. The protocol instructed them to attempt a correct answer but allowed for guessing when a solution was not apparent. Human accuracy was assessed using the same protocol as the computational models. In recognition of their contribution, all participants were provided with fair remuneration as outlined in a formal agreement.

4.2 Evaluation

Our evaluations are anchored in a zero-shot approach and CoT prompting to facilitate the models' engagement in comprehensive reasoning sequences (Wei et al., 2022) and the prompts are listed in Appendix A.6.

Table 4: Accuracy scores of models on our StatsChartMWP benchmark. Hist: histograms, Line-f:line-function chart, D-axis: dual-axis chart, P-bar: percentage-bar chart, Comp: composite chart. The **first** and second highest scores are bolded and underlined, respectively. The amalgamation of LLMs with image captioning tasks is denoted with the visual model utilized specified in parentheses.

Model	All	Bar	Hist	Line	Line-f	Scatter	D-axis	P-bar	Pie	Table	Comp	Radar
Closed-source LLMs (Image caption)												
GPT4 (GPT-4V)	31.47	38.11	8.61	39.12	22.18	20.62	35.94	4.55	34.71	52.46	24.36	20.83
GPT4 (GPT-4o)	46.95	59.98	13.30	52.72	35.98	27.50	45.31	27.27	59.19	71.85	38.82	20.83
Open-source LMMs												
HPT-1.0	10.10	9.91	5.07	17.77	9.62	10.62	26.56	9.09	7.18	10.62	11.56	29.17
DeepSeek-VL-7B	13.20	16.06	4.63	21.43	11.72	12.50	28.12	4.55	14.16	15.47	9.78	8.33
LLaVA-NeXT-34B	15.67	20.96	5.45	23.13	13.39	20.00	25.00	4.55	14.06	19.24	12.44	20.83
Cambrian-34B	18.15	22.03	8.77	27.89	14.23	18.75	46.88	22.73	16.52	20.24	14.02	41.67
IXC-2.5-7B	22.55	31.10	7.36	29.25	17.99	18.75	43.75	18.18	24.88	29.72	15.02	41.67
LLaVA-OV-72B	32.39	38.33	15.26	39.80	30.54	35.62	42.19	31.82	34.32	45.97	22.91	16.67
Qwen2-VL-7B	37.46	45.67	20.16	39.29	30.96	31.25	**65.62**	36.36	44.54	51.25	25.70	**62.50**
InternVL2-Llama3-76B	45.02	58.81	24.58	50.43	35.98	43.12	42.19	13.64	48.08	57.38	35.37	29.17
InternVL2_5-78B	55.25	70.93	29.26	56.12	40.59	48.75	57.81	54.55	57.01	74.27	51.84	37.04
Qwen2-VL-72B	59.33	69.91	39.29	60.03	46.44	43.75	62.50	59.09	65.78	77.12	50.39	**62.50**
Qwen2.5-VL-72B	**71.12**	**78.45**	**59.51**	**68.45**	**56.90**	**54.37**	**65.62**	**63.64**	**78.76**	**85.89**	**61.07**	41.67
Closed-source LMMs												
Qwen-VL-MAX	30.24	37.40	10.19	29.51	19.25	20.00	29.69	18.18	37.86	54.74	16.91	33.33
GPT-4V	34.28	38.57	12.10	40.48	28.87	30.00	39.06	18.18	38.25	55.67	27.89	33.33
GPT-4o	57.05	66.51	26.38	58.76	42.26	45.62	68.75	54.55	72.57	81.54	49.50	**45.83**
OpenAI o3	**82.75**	**81.73**	**77.71**	**76.96**	**71.97**	**83.12**	**82.81**	**90.91**	**93.23**	**88.10**	**83.98**	33.33
Human												
Human performance	93.88	96.51	92.21	95.76	88.81	84.76	98.21	85.71	91.36	93.77	93.85	92.31

Similar to MathVista (Lu et al., 2023b), we initially feed the model's generated sequence into GPT-4 for target value or option letter extraction. To augment the precision of our answer extraction, we devise complex rules for post-processing results in instances of GPT-4o's shortcomings. This strategy has facilitated an extraction accuracy exceeding 97%, mirroring the success rate documented in MathVista. The prompts' specifications and extraction protocols can be found in the Appendix A.6. The extracted outcomes are juxtaposed with the golden answers to establish the ultimate performance metric. Considering the model's aim to generate responses in diverse formats, either as the exact answer or as the corresponding option letter, we consider a prediction accurate if it aligns with either the golden answer or the golden option letter.

4.3 CoTAR

On top of these baselines, we further introduce CoTAR, a data augmentation strategy designed to enhance cross-modal alignment between visual representations and reasoning text. Specifically, instead of relying solely on the concise textual solutions provided in MWPs, we leverage state-of-the-art LMMs[2] to transform these solutions into detailed, step-by-step explanations in a CoT-alike format. Each step in CoTAR consists of two parts: (1) **Step Summary:** A brief directive outlining the purpose of the step and the corresponding image caption to strengthen the interaction between language and visual modalities; (2) **Concrete Reasoning Response:** A thorough explanation of the reasoning process, incorporating calculations, logical deductions, and references to summarized visual information and prior context;Recognizing that visual data from images is often inadequately explained in responses, CoTAR explicitly guides the model to identify **where** and **what** visual content should be utilized at each step. This design further enhances the model's ability to integrate and reason across language and visual modalities effectively. Once the CoTAR-enhanced dataset is generated, it can be used to fine-tune any LMMs base model,

[2]In this work, we utilize the GPT-4 API for this task.

Table 5: Fine-tuning evaluation results of Qwen2-VL-7B on different benchmark datasets.

Dataset	SFT	CoTAR	ΔAcc
StatsChartMWP	42.66	51.42	**8.76**↑
TABMWP	78.91	79.04	**0.13**↑
GEOQA+	66.09	70.4	**4.31**↑

significantly improving its multi-step mathematical reasoning capabilities in visually rich contexts. This approach not only bridges the gap between visual and textual representations but also fosters more accurate and interpretable reasoning in complex problem-solving scenarios.

5 Results

5.1 Model Performance

We present a comprehensive assessment of the performance accuracy exhibited by various models on the StatsChartMWP dataset. Table 4 summarizes the detailed performance results in terms of prediction accuracy on each statistical chart type. Furthermore, Figure 3 depicts the variation in models' accuracy according to different grade levels and question types. Our key observations can be summarized as the following:

Challenging of StatsChartMWP The dataset presenting considerable challenges to the current multimodal models. This is particularly evident in the struggles encountered by open-source models. Notably, Qwen2.5-VL-72B emerges as a standout among the evaluated models, securing a remarkable overall accuracy of 71.12%. Although this score represents a superior performance relative to other models, it still falls far short of human performance. We also conducted further comparisons between StatsChartMWP and other benchmarks, Figure 5 presents a comparison of accuracies across different models on StatsChartMWP, ChartQA, and MathVista. Results of their benchmarks are either from their official website.

Comparative across grades and question formats Figure 3 illustrates a consistent trend across all models, indicating that accuracy rates are generally higher for questions targeted at the primary level as opposed to those designed for middle and high school students. In terms of question types, there is a discernible pattern wherein models tend to exhibit elevated accuracy rates for true or false questions, with a descending order of accuracy observed for multiple-choice, fill-in-the-blank, and problem-solving questions. This also provides some insights for new inferential strategies, such as re-configuring open-ended questions into structured fill-in-the-blank tasks or applying transformations that involve evaluative judgment.

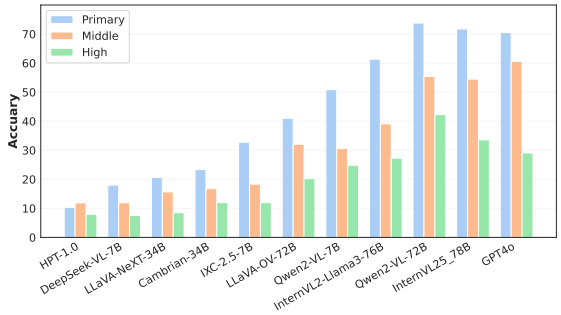
Table 6: Distribution of GPT-4o's errors.

Error Type	VE	RE	CE	KE
Distribution	46.44%	30.02%	16.81%	6.73%

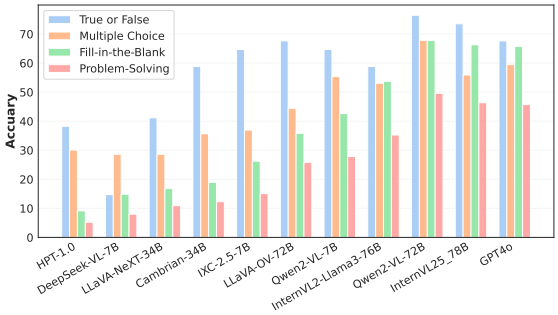
The effectiveness of CoTAR To validate the effectiveness of our method, we conducted further fine-tuning experiments on the Qwen2-VL-7B. As shown in Table 5, using the CoTAR data augmentation scheme resulted in an accuracy improvement of 8.76% on the StatsChartMWP dataset. Additionally, to further verify the generalizability of CoTAR, we performed comparative experiments on the TABMWP and GEOQA+. The results demonstrated improvements of 0.13% and 4.31% for TABMWP and GEOQA+ respectively. This indicates that our CoTAR prompt is effective, suggesting it can be used to fine-tune LMMs foundation model to enhance its multi-step mathematical reasoning capabilities in visual environments.

5.2 Error Analysis

Considering that the problem-solving process of LMMs may involve multiple aspects, it is common for an initial error to lead to a series of subsequent errors. Therefore, in analyzing the causes of model errors, we categorize all errors into four types: visual recognition error (VE), reasoning error (RE), calculation error (CE) and knowledge error (KE). Based on this classification, we conducted a detailed analysis of erroneous problems output by GPT-4o, as shown in Table 6. Visual recognition error account for 46.44%, indicating the difficulty multimodal models have in accurately interpreting visual information, suggesting that visual encoders remain a bottleneck in multimodal development. Reasoning error constitute 30.2%, highlighting significant challenges in the model's logical processing and reasoning capabilities. Calculation error represent 16.81% in our statistical graph tests, which also indicates that statistical questions often involve extensive calculations. In contrast, knowledge error account for 6.73%, demonstrating that knowledge errors do not significantly impede LMMs' mathematical reasoning abilities in our StatsChartMWP.

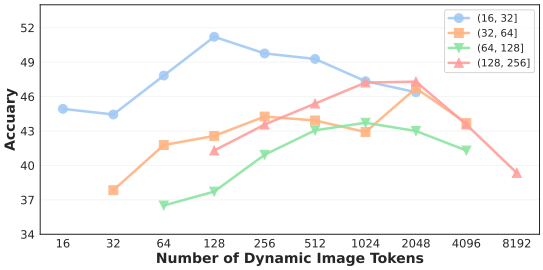


(a) Accuracy analysis for grade groups

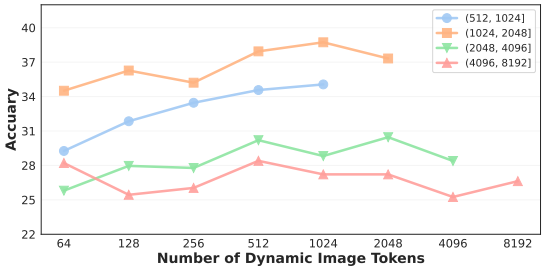


(b) Accuracy analysis for question types

Figure 3: Detailed accuracy results of different grade levels and question types.



(a) Accuracy analysis for token ranges (16, 256)



(b) Accuracy analysis for token ranges (512, 8192)

Figure 4: The accuary of Qwen2-VL-7B model with different dynamic image tokens.

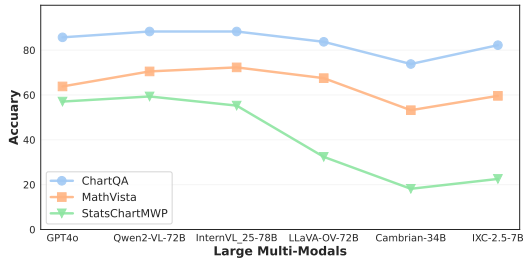


Figure 5: Comparative model performance on ChartQA, MathVista and StatsChartMWP.

Table 7: Accuracy scores of Qwen2-VL-7B with fixed resolution and dynamic resolution.

Res	Min	Max	Acc	ΔAcc
Fix	4	8192	37.46	-
Dyn	$token \times 4,\ if\ token < 128$ $512,\ others$	1280	38.96	**1.5**↑

5.3 Visual Encoder Ablation

We conducted a visual ablation study on Qwen2-VL-7B to explore the impact of visual token length on model performance. The dataset was divided based on the number of tokens corresponding to image resolution, with experiments performed by setting the minimum token for low-resolution images and the maximum token for high-resolution images. For image with fewer than 256 tokens, we set different minimum image tokens, as shown in Figure 4 (a). For image with more than 512 tokens, we set different maximum images tokens, and the results shown in Figure 4 (b). We observed when image resolution is low, appropriately increasing the number of tokens essentially performing super-resolution processing on the image significantly enhances accuracy. Our findings suggest that a four times increase in resolution yields optimal results. Conversely, for high-resolution images, the improvement in accuracy generally shows a positive correlation with the number of tokens. However, when the token length exceeds 4096, our experimental results indicate a deviation from this trend, implying that an excessive number of tokens does not necessarily lead to optimal model performance for high-resolution images. Based on these observations, we conclude that setting the maximum token count between 1024 and 2048 achieves a higher cost-performance ratio. Table 7 presents the accuracy of Qwen2-VL-7B using fixed and dynamic resolution tokens on StatsChartMWP, respectively. It is evident that employing the dynamic tokens scheme results in a 1.5% improvement in accuracy.

6 Conclusion

This study introduces StatsChartMWP, a dataset of 8,514 MWPs with statistical charts specifically developed to assess the proficiency of LMMs in performing multimodal mathematical reasoning in real-world K-12 educational contexts. We conduct a comprehensive evaluation of the state-of-the-art models and performed an in-depth error analysis, underscoring that enhancing multimodal performance requires attention to both visual and reasoning components. Additionally, we explored the relationship between image token length and model performance through experiments, providing recommended settings for image token lengths. Furthermore, we introduced CoTAR, a chain-of-thought data augmentation scheme that fine-tunes LMMs through chain-of-thought reasoning steps. Through benchmarking existing and newly proposed approach, we outline future research directions in tackling crossmodal alignment between figures and natural language and undertaking meticulous multi-step mathematical reasoning.

7 Limitation

While our StatsChartMWP takes a step forward in the field of visual multi-step MWPs with statistical charts for LMMs, it is important to recognize several limitations as follows: (1) We have categorized the problems in StatsChartMWP according to various standards, encompassing chart types, grade-levels, and question types. These categorization strategies evaluate the capabilities of LMMs from multiple dimensions. However, we are deficient in the necessary annotations pertinent to the fundamental knowledge points of the studied topic. The supplementation of these relevant knowledge points could facilitate a more comprehensive and robust assessment of the model's chart interpretation and inference abilities. (2) The StatsChartMWP dataset is predominantly in Chinese. Considering the development of multilingual benchmarks, the current evaluations may not completely uncover their potential when restricted to a single language. The incorporation of multilingual visual mathematical problems not only amplifies the dataset's global relevance but also elevates the evaluation of LMMs' linguistic diversity and understanding capabilities.

Acknowledgments

This work was supported in part by National Key R&D Program of China, under Grant No. 2023YFC3341200; in part by NFSC under Grant No. 62477025; in part by Key Laboratory of Smart Education of Guangdong Higher Education Institutes, Jinan University (2022LSYS003) and in part by Beijing Municipal Science and Technology Project under Grant No. Z241100001324011.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023a. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. QwenVL-MAX.

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.

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.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric 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*, pages 513–523.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Konstantin Chernyshev, Vitaliy Polshkov, Ekaterina Artemova, Alex Myasnikov, Vlad Stepanov, Alexei Miasnikov, and Sergei Tilga. 2024. U-MATH: A university-level benchmark for evaluating mathematical skills in llms. *arXiv preprint arXiv:2412.03205*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Haoyu Dong, Haochen Wang, Anda Zhou, and Yue Hu. 2024. TTC-QuAli: A text-table-chart dataset for multimodal quantity alignment. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 181–189.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. OlympiadBench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.

HyperGAI. 2024. Introducing HPT: A groundbreaking family of leading multimodal llms.

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*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. LlaVA-NeXT: Improved reasoning, ocr, and world knowledge.

Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong. 2024b. Multimodal pretraining, adaptation, and generation for recommendation: A survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6566–6576.

Wentao Liu, Qianjun Pan, Yi Zhang, Zhuo Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo Jiang, and Liang He. 2024c. CMM-Math: A chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. *arXiv preprint arXiv:2409.02834*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, and 1 others. 2024. DeepSeek-VL: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

P Lu, L Qiu, KW Chang, YN Wu, SC Zhu, T Rajpurohit, K Clark, and A Kalyan. 2023a. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. International Conference on Learning Representations (ICLR 2023).

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023b. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

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 (Volume 1: Long Papers)*, pages 6774–6786.

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.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. PlotQA: Reasoning over scientific plots. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1516–1525. IEEE.

OpenAI. 2023. GPT-4V(ision) system card.

OpenAI. 2024. GPT-4o contributions.

OpenGVLab. 2024. InternVL2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, and 1 others. 2024. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-Thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Renqiu Xia, Bo Zhang, Haoyang Peng, Ning Liao, Peng Ye, Botian Shi, Junchi Yan, and Yu Qiao. 2023. StructChart: Perception, structuring, reasoning for visual chart understanding. *arXiv preprint arXiv:2309.11268*.

Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. 2023. A multi-modal neural geometric solver with textual clauses parsed from diagram. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 3374–3382.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, and 1 others. 2024a. InternLM-XComposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024b. MathVerse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.