

VisionMath: Vision-Form Mathematical Problem-Solving

Zongyang Ma^{1,2,4,5*}, Yuxin Chen^{2*†}, Ziqi Zhang^{1,5}✉, Zhongang Qi⁶, Chunfeng Yuan^{1,5},
 Shaojie Zhu³, Chengxiang Zhuo³, Bing Li^{1,5}, Ye Liu^{2,7}, Zang Li³, Ying Shan², Weiming Hu^{1,4,5,8}

¹State Key Laboratory of Multimodal Artificial Intelligence Systems,

Institute of Automation, Chinese Academy of Sciences ²ARC Lab, ³Tencent PCG

⁴School of Artificial Intelligence, University of Chinese Academy of Sciences

⁵Beijing Key Laboratory of Super Intelligent Security of Multi-Modal Information

⁶vivo Mobile Communication Co. ⁷The Hong Kong Polytechnic University

⁸School of Information Science and Technology, ShanghaiTech University

{mazongyang2020, zhangziqi2017}@ia.ac.cn, {cfyuan, bli, wmu}@nlpr.ia.ac.cn

{uasonchen, alisyzhu, felixzhuo, gavinzli, yingsshan}@tencent.com, zhongangqi@gmail.com,
 coco.ye.liu@connect.polyu.hk

Abstract

Mathematical problems in real-world scenarios are often presented in a purely vision-form, where textual problem statement and accompanying math figures, e.g., geometry figures and functional graphs, are integrated into a single image. This vision-form problem-solving task requires precise comprehension and reasoning on both textual and graphical elements in the images, posing significant challenge to current Multimodal Large Language Models (MLLMs), which process text and math figures in isolation. In this work, we propose VisionMath, the first exploration for vision-form mathematical problem-solving model, which employs a three-stage progressive multimodal reasoning alignment strategy to systematically enhance task-specific capabilities. Building upon a LLM proficient in unimodal mathematical reasoning, VisionMath first establishes foundational OCR capabilities through capturing rendered mathematical problem images. Subsequently, the model develops comprehensive understanding of figure structures and properties via learning from figure descriptions and mathematical educational videos. Finally, the model's reasoning capacity is activated using carefully constructed visual-form problem-solving datasets VisionMath-IT with chain-of-thought annotations. For comprehensive evaluation, we construct multilingual benchmarks covering diverse problem types, including geometry, algebra, function problems in both English and Chinese. Experimental results demonstrate that VisionMath significantly outperforms existing general-purpose

* Equal contribution. † Project Lead. ✉ Corresponding author.

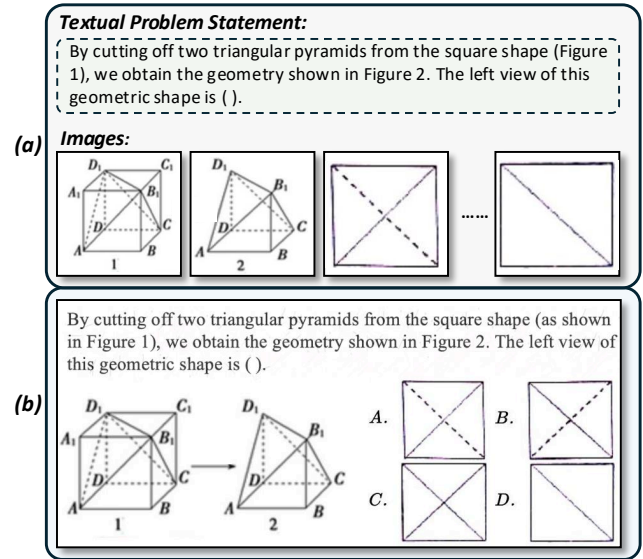


Figure 1. Demonstration of mathematical problem examples of image-text-separation-form (a) and vision-form (b).

and math-specialized models, showcasing superior multilingual and multi-type problem-solving capabilities. Our model weights, data and code will be made available at <https://github.com/mengqiDyange/VisionMath>.

1. Introduction

Recent advancements in reasoning capabilities have significantly enhanced the performance of both general-purpose [3, 7–9, 22, 27–29, 50, 52] and math-specialized MLLMs [10, 14, 32, 42], enabling them to demonstrate re-

markable proficiency in solving various types of mathematical problems. Current MLLMs primarily address “image-text-separation-form” mathematical problems, where the textual problem statement (including formulas) and math figures, such as geometric figures and functional graphs, are input separately into the model, as shown in Figure 1(a). However, in real-world scenarios, problem statement and math figures are often captured within a single image, as the demonstration in Figure 1(b). The lack of exploration has led to the inability of existing MLLMs to effectively process such “vision-form” problems. For instance, on the MathVista benchmark [32], when switching the input samples from image-text-separation-form to vision-form, the accuracies of GPT-4o mini for Geometry Problem Solving (GPS) and Math Word Problems (MWP) decreases from 58.1% to 53.4% and from 59.6% to 53.2%, respectively. Such performance gap limits the model’s real-world application, especially when the problem involves complex formulas and figures.

The above issue drives us to take the first step in exploring a MLLM tailored for vision-form math problem-solving. To enhance the model’s understanding and reasoning capabilities for both text and figures, we develop our MLLM, named VisionMath, focusing on three critical aspects: text recognition, figure comprehension, and integrated text-figure reasoning. Specifically, we design a three-stage progressive multimodal reasoning alignment strategy. In the first stage, we establish the basic *OCR capabilities*, enabling accurate perception of problem statements, especially formulas. The training combines document image OCR datasets [6, 16, 21, 37, 47, 51] with rendered math problem image datasets, requiring the model to transcribe the text present in the images. The second stage strengthens *figure comprehension* with a dual-path enhancement strategy. On one hand, math figure captioning datasets [14, 60] are employed to explicitly improve the structural visual parsing of geometric figures and functional graphs. Note that these captions generated by LLMs or rule-based systems often suffer from limited diversity and domain gap with real-world math figures. Thus, on the other hand, we foster a more in-depth math property understanding of real-world figures via learning from mathematical educational videos [61], which offer diverse and real-world math figures across complexity levels accompanied by teachers’ profound analysis. In the final stage, we employ vision-form problem-solving datasets to instruct the model to perform step-by-step reasoning based on the joint text-figures understanding, thus improving the *integrated text-figure reasoning capabilities*.

To deal with the lack of vision-form mathematical problem-solving datasets, we introduce VisionMath-IT, a large-scale, high-quality bilingual (English/Chinese) instruction tuning dataset. This dataset, developed through a

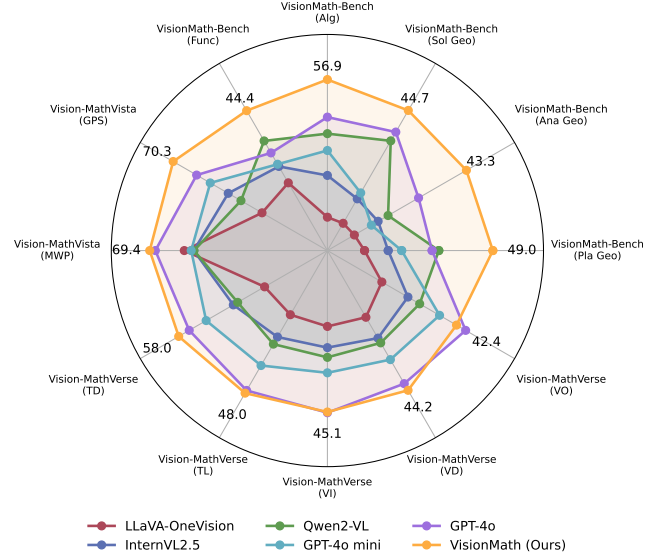


Figure 2. Performance comparisons between VisionMath and other representative MLLMs.

systematic pipeline involving data collection, filtering, rendering, and re-formatting, offers a rich collection of vision-form math problems. Furthermore, we develop specialized evaluation benchmarks to thoroughly assess the performance of MLLMs in solving various types of vision-form math problems. The first is VisionMath-Bench, which is a Chinese benchmark encompassing geometry, algebra, and function problems. Others are English benchmarks derived from commonly used image-text-separation-form datasets, including VisionMathVista [32] and VisionMathVerse [59]. Extensive experimental results demonstrate that our VisionMath significantly outperforms all open-source general-purpose and math-specialized MLLMs across these benchmarks. Simultaneously, VisionMath shows substantial improvements over GPT-4o on VisionMath-Bench and VisionMathVista, while achieving overall performance gains compared to GPT-4o on VisionMathVerse. Intuitive comparisons of representative models are visualized in Figure 2. Our contributions are summarized as follows:

- We present the first exploration of practical vision-form math problem-solving task, proposing a progressive multimodal alignment reasoning strategy to systematically enhance the model’s visual-form mathematical comprehension and reasoning capabilities.
- Through the design of a rigorous data construction pipeline, we meticulously curated a large-scale, high-quality dataset for vision-form mathematical problem-solving, named VisionMath-IT.
- We develop comprehensive benchmarks covering bilingual algebraic, geometric, and functional problems, and evaluate multiple state-of-the-art MLLMs to benchmark their performance on vision-form math problem-solving tasks. Extensive results highlights the superiority of our proposed approach.

		Samples	Dataset
VisionMath-OCR	Document image OCR and understanding	1.36M	ChartQA [35], Deepform [46], DocVQA [36], InfographicVQA [38], Kleister Charity [43], Tabfact [6], WikiTableQuestions [40], Visualmrc [47], Ccpdf [51], RVL-CDIP [15], synthdog [20], Vary [54], pubtabnet [62], Turl [11], OCRCC [5], Dvqa [18], FigureQA [19], PlotQA [39], MMC [26]
	Mathematical Problem OCR	1.34 M	VisionMath-IT-Chinese, MMathCoT-1M [33]
VisionMath-Figure	Figure Caption	0.55M	Geo170k-Caption [14], MAVIS-Caption [60]
	Mathematical Educational Video	0.60M	Multimodal-textbook [61]
VisionMath-IT	English Vision-Form Problem	1.96M	Geo170k [14], MathV360 [42], MultiMath-300k-English [41], MMathCoT-1M [33], Numbia [4]
	Chinese Vision-Form Problem	0.62M	VisionMath-IT-Chinese, MultiMath-300k-Chinese [41]

Table 1. Detailed subsets and data volumes of different training datasets.

2. Related Work

General-purpose Multimodal Large Language Model.

Recent years have witnessed remarkable advancements in general-purpose MLLMs. LLaVA series [22, 27–29] has pioneered systematic explorations of effective training strategies, including multi-stage alignment, multi-task data rationing, reinforcement learning, and synthetic data training, significantly improving the capabilities. Meanwhile, QwenVL series [3, 50, 52] demonstrate SOTA performance across diverse tasks such as high-resolution image understanding, document understanding, and long-form video understanding. InternVL series [7–9] use larger-scale visual encoder for multimodal pre-training to bridge the performance gap between commercial [2, 48, 49] and open-source models. EMU series [44, 45, 53] aims to unify multimodal understanding and generation. MiniCPM [57] substantially advances edge-device multimodal intelligence through architectural and efficiency innovations. While these advancements have collectively propelled MLLM capabilities across diverse vision tasks [13, 23–25, 30, 34–36, 58], their multimodal reasoning and mathematical problem-solving proficiency remains suboptimal due to the inherent trade-offs in general-purpose design.

Math-specialized Multi-Modal Language Model. To improve mathematical reasoning, several works have made multifaceted attempts. G-LLaVA [14] first collects geometry data to improve geometry problem solving. Math-LLaVA [42] and R-COT [10] employ rule-based synthetic data generation, significantly improving performance on geometry and algebra problems. MAVIS [60], MathGLM-Vision [56], and MultiMath [41] develop large-scale multi-source multimodal math datasets to strength mathematical reasoning through domain-adaptive fine-tuning. Math-PUMA [63] constructs paired samples conveying identical mathematical concepts to bridge the performance discrepancies between text-rich and image-rich problems. URSA [33] introduces error localization supervision to minimize logical errors and visual misinterpretations during reasoning. Despite these developments significantly expanding the capabilities frontier, previous works overlook fully vision-form mathematical problem-solving, leading to suboptimal performance in this practical task. This has inspired us to conduct an in-depth analysis on the characteristics of this task and make targeted improvements.

3. Method

3.1. Data Engine of VisionMath

High-quality data is crucial for effective model training. To meet the capability requirements for solving vision-based mathematical problems, ranging from basic perception to complex reasoning, we meticulously develop three specialized datasets: VisionMath-OCR, VisionMath-Figure, and VisionMath-IT. The detailed composition of each dataset is shown in Table 1.

3.1.1. VisionMath-OCR

VisionMath-OCR is designed to enable accurate recognition and parsing of general text, formulas, and symbolic notations within problem statements. Specifically, this dataset is partitioned into two components.

Document Image OCR. This component includes a collection of document image OCR and understanding datasets (as listed in Table 1), aimed at establishing general-purpose OCR capabilities with a focus on recognizing standard text, charts, and tables.

Mathematical Problem OCR. This component consists of rendered images with multilingual mathematical problems from the datasets in Table 1. It requires the model to accurately restate the complete problem statement, ensuring precise OCR of formulaic structures and symbolic notations.

3.1.2. VisionMath-Figure

VisionMath-Figure is introduced to enhance math figure perception and comprehension, particularly the properties and structural relationships of points, lines, and surfaces in geometry figures and function graphs. Based on the sources of data, the dataset can be divided into two parts:

Figure Captions. This portion of the data is derived from the Geo170k [14] and MAVIS [60] datasets, consisting of mathematical figures and their corresponding descriptions. These descriptions contain basic component information about the figures, which helps enhance the model’s figure structure perception and understanding. It is important to note that the figures in the Geo170k dataset are primarily planar geometric shapes, which limits their diversity. On the other hand, the figures in the MAVIS dataset are mainly constructed using rule-based methods, often resulting in a domain gap compared to real-world mathematical figures. To address this, we also utilize mathematical educational

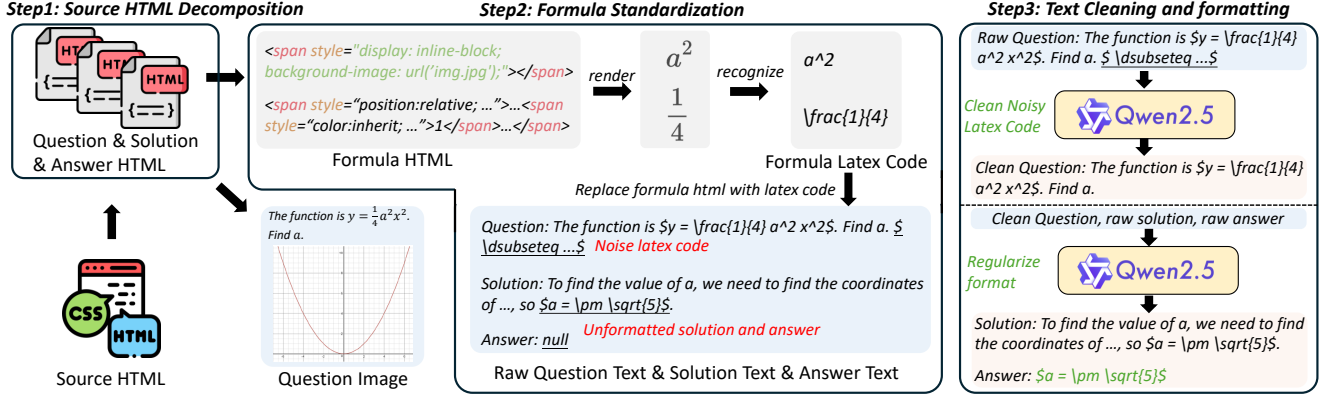


Figure 3. Pipeline of Chinese Vision-Form Problem Synthesis.

videos as an additional data source to further improve the understanding of mathematical figures.

Mathematical Educational Videos. We observe that mathematical educational videos contain a rich variety of real-world mathematical figures, along with in-depth analysis and extensive knowledge about these figures, such as the properties of geometric shapes. This information is beneficial for the understanding and reasoning of mathematical figures. Therefore, we collect mathematical videos from the Multimodal Textbook [61] and construct training samples with the following steps:

Step1: Subject Filtering. The videos in the Multimodal Textbook encompass multiple subjects. Therefore, we first utilize a Large Language Model (LLM) to filter the videos based on their titles and select those that are related to mathematics education.

Step2: ASR Text Filtering. Multimodal Textbook has already extracted frames from the videos and used ASR tools to transcribe the corresponding teacher explanations. However, some of the transcriptions contain non-informative filler phrases devoid of mathematical knowledge. Therefore, we employ an LLM to filter out text that does not include mathematical explanations.

3.1.3. VisionMath-IT

VisionMath-IT is a bilingual (English/Chinese) instruction-tuning dataset aimed at empowering end-to-end vision-form mathematical problem-solving, and English-Chinese examples of different problem types are shown in Sec. C. of the supplementary materials. Its construction adheres to two principles: (1) Diversity. Curating visually-presented mathematical problems spans algebra, geometry (including plane geometry, analytic geometry, solid geometry) and function problems with accompanying figures. (2) Structured Solution Formatting. The responses are unifying into (chain-of-thought, final answer) pairs. This not only covers a wide range of problem types to improve generalization but also eliminates style variations in solution formats, allowing to concentrate on learning the reasoning process.

To ensure VisionMath-IT meets rigorous diversity requirements and enables models to effectively learn heterogeneous reasoning patterns from various solution structures, we implement a multi-source data curation strategy.

English Vision-Form Problem Synthesis.

The publicly available English image-text-separation-form datasets already encompass diverse question types and substantial samples. Therefore, the focus is on selecting high-quality data for vision-form problem transformation and solution formalizing. The specific steps are as follows:

Step 1: Data Filtering. To filter out high-quality data from existing datasets, we first sample a portion of instances from different datasets and manually inspect the clarity of the mathematical images. We then retain the datasets with overall acceptable visual quality to serve as the source for training data. The retained dataset is shown in the “English Vision-Form Problem” section of Table 1.

Step 2: Image-form Problem Rendering. For each sample in the retained image-text-separation-form datasets, the text-based problem statement and the associated math figures are jointly rendered into a single image, resulting in vision-form mathematical datasets.

Step 3: Solution Formatting. For each problem, we design rules to split the original solution’s chain-of-thought process and final answer, and unify the response format using the template “{CoT process}. Therefore, the answer is \boxed{final answer}”.

Chinese Vision-Form Problem Synthesis.

Due to the lack of specialized high-quality and diversity in existing datasets for Chinese mathematical problems, we conduct extensive data collection on the Internet and perform automatic data filtering and curation strategies. As shown in Figure 3, the specific pipeline are as follows:

Step 1: Source HTML Decomposition. The HTML collected from the web often contains much irrelevant information. Therefore, we first extract the HTML parts corresponding to the question, solution, and answer based on predefined rules. And the question images can then be obtained by rendering the html code corresponding to question.

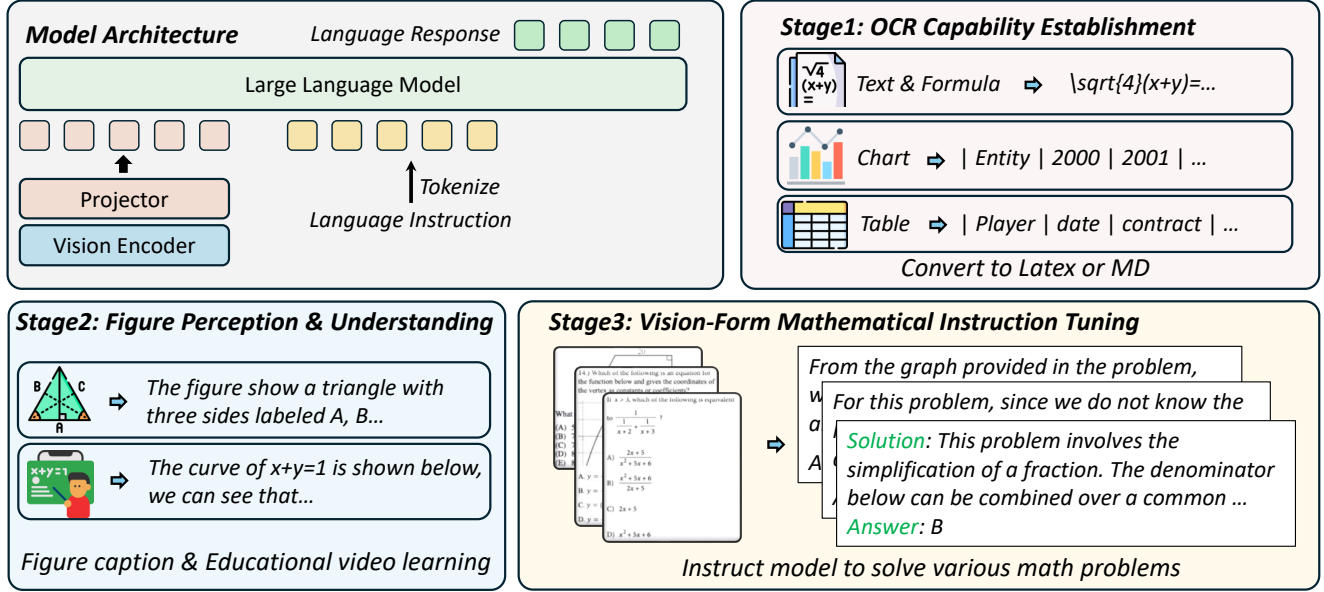


Figure 4. The architecture of VisionMath, and the accompanying three stage progressive multimodal reasoning alignment training strategy for enhancing Mathematical OCR, deepening math figure Understanding, and fine-tuning problem-solving capabilities.

Step 2: Formula Standardization. In the HTML code, mathematical formulas may appear as image paths or in style code formats. To obtain readable question and answer text, these image paths or style codes must be extracted, rendered into images, and subsequently converted into LaTeX code using OCR tools. By replacing the original mathematical formula codes with the LaTeX codes, we can ultimately obtain the raw question, solution, and answer text.

Step 3: Text Cleaning and Formatting. The raw texts obtained from stage 2 often encounter some quality issues, including incorrectly recognizing mathematical images as formulas and converting them into garbled LaTeX, along with improperly mixed presentation of solution steps and final answers. To clean and standardize the raw texts, we leverage the powerful text understanding capabilities of LLM to filter out any garbled LaTeX. Additionally, we also use LLMs to explicit separate the CoT process and answer text, fill in any missing answers, and organize them into the format “ $\{CoT\ process\}$. Therefore, the answer is $\boxed{final\ answer}$ ”.

The final dataset, VisionMath-IT-Chinese, consists of 332K math problems related to K-12 education. This corpus encompasses multiple formats including multiple-choice, fill-in-the-blank, and free-form problems, spanning core mathematical subjects such as geometry, algebra, function, and word problem-solving. Furthermore, we have also processed Multimath-300k-Chinese using the same pipeline as English Vision-Form Problem Synthesis.

3.2. Architecture of VisionMath

The architecture of the Vision-Math model is illustrated in Figure 4, which follows the structure that has been widely

verified to work well in general MLLMs. This comprises three core components: a pre-trained visual encoder InternViT [9], which is good at encoding semantically rich visual features; a lightweight projector, *i.e.*, a two-layer MLP [27], that transforms visual representations into the semantic space of LLMs; and an open-source LLM Qwen2.5-Math [55], which has advanced reasoning capabilities for text-form mathematical problems.

3.3. Progressive Multimodal Reasoning Alignment

To harness the synergistic integration of diverse datasets in Table 1 for building strong vision-form mathematical problem-solving model, we propose a three-stage multimodal reasoning alignment strategy that incrementally endows the model with capabilities ranging from foundational visual perception to advanced complex reasoning, as illustrated in Figure 4. The underlying design logic follows a hierarchical progression: first achieving accurate perception of problem statements, then advancing to comprehensive understanding of math figures, and finally enabling integrated text-figure reasoning.

3.3.1. Mathematical OCR Capability Establishment

During the first stage, we focus on equipping the model with the capabilities to accurately perceive fundamental problem-statement elements, including text, mathematical formulas, charts, and tables. To achieve this, we leverage the curated VisionMath-OCR dataset for training, updating all parameters in the visual encoder, projection layer, and language model. To prevent the LLM’s catastrophic forgetting of reasoning capabilities during multimodal training, we also strategically interleave a small proportion (25%) of

text-form mathematical dataset Numbia [4] into the training batches, ensuring balanced language skill retention.

3.3.2. Math Figure Perception & Understanding

In the second stage, we employ a dual-path enhancement strategy to improve the model’s comprehension of mathematical figures. On the first path, leveraging figure caption data, we instruct the model to generate descriptive captions for input figures, thereby enhancing its ability to accurately perceive and understand figure primitives (such as points, lines, planes) and their spatial relationships, *e.g.*, parallelism, intersections and containment. On the complementary path, utilizing the constructed educational video dataset, we task the model with predicting corresponding lecture content (*i.e.*, refined ASR text) based on input video frames. The diverse mathematical figures present in these videos further strengthen the model’s figure understanding capabilities. Additionally, the lecture content provided by educators contains rich knowledge, including in-depth figure analysis and geometric properties of figures, which further deepens the model’s understanding and reasoning capabilities regarding visual content. Similarly, a small amount (25%) of Numbia is added at this stage to prevent the language model from degrading.

3.3.3. Vision-Form Mathematical Instruction Tuning

After completing the above two-phase training curriculum, the model has demonstrated robust foundational comprehension of vision-form mathematical problems. Building upon the acquired capabilities, the final stage orchestrates a seamless transition toward advanced mathematical reasoning through end-to-end supervised instruction tuning using the complete VisionMath-IT dataset.

4. Evaluation Benchmarks

To comprehensively evaluate and compare model performance on mathematical problems across multiple languages and subjects, our evaluation benchmarks are designed to include the following parts.

4.1. Chinese Mathematical Proficiency Benchmark

VisionMath-Bench is a Chinese evaluation benchmark comprising 100 Plane Geometry (Pla Geo), 90 Analytic Geometry (Ana Geo), 103 Solid Geometry (Sol Geo), 72 Algebra (Alg) and 135 Function (Func) problems, with test samples systematically collected according to standardized mathematics curricula spanning primary to senior secondary education in China. Test accuracies for different problem types are reported separately during evaluation.

4.2. English Mathematical Proficiency Benchmarks

English mathematical evaluation benchmarks are derived through the conversion of commonly used image-text-separation-form evaluation datasets, *i.e.*, MathVista and

MathVerse. Detailed dataset processing and statistics are shown in the Sec. B. of the supplementary materials.

5. Experiment

5.1. Implementation Details

VisionMath employs InternViT-300M-448px as the visual encoder and Qwen2.5-Math-7B-Instruct as the base language model. The three-stage alignment training generally adopts the same training settings, except for using different datasets. The maximum number of image tiles for visual encoding and the maximum input sequence length for LLM are limited to 9 and 4096, respectively. The learning rate for the visual encoder is $2e-6$, while that for the projection layer and LLM is $1e-5$. Each stage undergoes 1 epoch of training with a batch size set at 384.

5.2. Comparisons with SOTAs

To make thoroughly performance comparisons of existing methods on vision-form mathematical problem-solving task, we extensively evaluate 8 open-source general-purpose MLLMs, 5 open-source math-specialized MLLMs, and 4 closed-source commercial general-purpose MLLMs, on the constructed evaluation benchmarks. General-purpose MLLMs are evaluated using VLMEvalKit [12], while math-specialized MLLMs are evaluated based on the original codebases. The detailed results are listed in Table 2, and the key findings are summarized below.

VisionMath substantially outperforms open-source MLLMs. As can be seen from Table 2, leading open-source general-purpose MLLMs (*e.g.*, Qwen2-VL) surpass math-specialized MLLMs (*e.g.*, URSA) across multiple benchmarks, despite URSA exhibiting superior performance than Qwen2-VL on image-text-separation-form problems. This divergence stems from math-specialized models’ excessive overfitting to figure-text separated problems, which induces catastrophic degradation of essential OCR for recognizing problem statements, making it difficult to generalize to the vision-form problem-solving tasks. However, both categories of current open-source models exhibit significant performance gaps compared to VisionMath across different languages and problem types. On the Chinese VisionMath-Bench, VisionMath outperforms the top general-purpose model Qwen2-VL by 16.0%, 24.4%, and 9.7% on plane geometry, analytic geometry, and solid geometry problems, while exceeds the dedicated math-specialized model MultiMath by 44.0%, 36.6%, and 34.3%. In Chinese algebra and function problem-solving, VisionMath achieves 18.0% and 9.6% advantages over Qwen2-VL and 48.6% and 31.8% leads against MultiMath. Similar superiority is also observed on English benchmarks Vision-MathVista and Vision-MathVerse, where VisionMath substantially outperforms

Model	Param	VisionMath-Bench					Vision-MathVista		Vision-MathVerse				
		Pla Geo	Ana Geo	Sol Geo	Alg	Func	GPS	MWP	TD	TL	VI	VD	VO
Closed-Source General-purpose MLLMs													
Gemini-1.5-flash [49]	–	23.0	12.6	20.4	26.4	31.1	62.3	64.4	53.9	44.3	44.2	40.0	44.2
Gemini-1.5-pro [49]	–	33.0	23.2	31.1	40.3	36.3	73.6	66.7	61.7	56.2	52.8	52.2	53.3
GPT-4o mini [17]	–	22.0	13.7	18.4	33.3	27.4	53.4	53.2	47.3	38.7	34.0	34.5	34.4
GPT-4o [17]	–	31.0	28.4	37.8	44.4	31.0	59.6	67.2	53.9	47.1	45.1	42.1	42.4
Open-Source General-purpose MLLMs													
Phi3V [1]	4B	4.0	3.2	3.9	4.2	5.9	18.8	21.0	16.4	17.0	15.5	15.4	14.0
Phi3.5V [1]	4B	0.0	0.0	4.9	4.2	1.5	19.7	37.6	18.8	15.9	14.3	14.1	17.5
DeepSeek-VL [31]	7B	2.0	1.1	2.9	1.4	4.4	18.3	26.9	8.5	9.8	10.4	10.2	12.8
MiniCPM-V2.6 [57]	8B	6.0	8.4	11.7	16.7	18.5	37.5	39.2	17.6	18.1	17.4	17.4	16.4
LLaVA-OneVision [22]	7B	11.0	8.4	8.7	11.1	21.5	29.8	55.9	24.5	21.6	21.1	21.1	16.7
InternVL2 [9]	8B	17.0	11.6	15.5	25.0	27.4	43.3	52.7	37.3	29.8	25.8	26.5	25.9
InternVL2.5 [7]	8B	18.0	15.8	16.5	25.0	26.7	45.2	52.2	36.7	29.1	27.0	27.7	24.7
Qwen2-VL [52]	7B	33.0	18.9	35.0	38.9	34.8	39.4	52.1	35.0	31.5	29.7	29.2	28.3
Open-Source Math-specialized MLLMs													
G-LLaVA [14]	7B	0.0	1.1	1.0	1.4	0.0	23.1	2.1	17.0	15.5	14.1	12.2	5.3
Math-LLaVA [42]	13B	1.0	2.2	0.0	2.8	4.4	31.3	27.4	17.4	17.5	16.4	16.0	12.6
Math-PUMA-Qwen2 [63]	7B	1.0	1.1	1.0	2.8	4.4	25.0	30.8	18.3	18.9	19.2	18.5	15.5
Math-PUMA-DeepSeek2 [63]	7B	1.0	2.2	1.1	4.2	5.9	19.7	29.3	19.7	19.5	20.0	17.9	12.2
MultiMath [41]	7B	5.0	6.7	10.4	8.3	12.6	31.8	26.3	17.1	18.2	17.6	18.1	11.7
URSA [33]	8B	2.0	0.0	6.8	4.2	10.4	22.1	29.0	15.0	21.3	21.7	20.5	13.3
VisionMath (Ours)	7B	49.0	43.3	44.7	56.9	44.4	70.3	69.4	58.0	48.0	45.3	44.2	39.6
OCR tool + Open-Source MLLMs													
OCR tool + InternVL2.5 [7]	8B	22.0	16.8	13.7	27.8	27.4	48.1	54.9	38.3	29.4	27.3	25.4	29.3
OCR tool + URSA [33]	8B	22.0	14.7	21.4	26.4	26.7	34.6	42.5	28.7	23.7	21.6	21.4	31.2

Table 2. Comparisons of representative MLLMs on vision-form mathematical problem-solving benchmarks. The best results of open-sourced and closed-sourced MLLMs are marked in orange and blue, respectively.

all open-source MLLMs across all problem types. This demonstrates that well-designed datasets combined with rationally progressive training strategies can methodically build the model’s capabilities—from foundational math element perception to complex reasoning, thereby achieving excellent vision-form problem-solving proficiency.

VisionMath is superior to the tool-integrated methods. A natural question is whether the end-to-end VisionMath model is better than pipelines that employ OCR tools to extract textual problem statements, and then feed vision-form problem image and extracted statements into MLLMs for solving. The results of InternVL2.5 and URSA integrated with GPT-4o as OCR tools show that our method still achieves all-around superiority across benchmarks. This demonstrates that current MLLMs’ performance bottleneck in solving vision-form math problems stems not merely from inaccurate mathematical OCR, but more from the need for enhancements in figure perception and reasoning.

VisionMath achieves competitive results compared to closed-source MLLMs. Compared to advanced small-scale closed-source MLLMs, such as GPT-4o mini and Gemini-1.5-flash, VisionMath maintains comprehensive performance superiority across all evaluation benchmarks. When benchmarked against large-scale closed-source MLLMs, *e.g.*, GPT-4o and Gemini-1.5-pro, VisionMath’s benchmark performance varies. It leads GPT-4o and Gemini-1.5-pro on VisionMath-Bench, surpasses GPT-

4o while matching performance with Gemini-1.5-pro on Vision-MathVista, and outperforms GPT-4o’s but trails behind Gemini-1.5-pro on Vision-MathVerse.

5.3. Ablations

In this section, we conduct ablation studies on dataset components and training strategies employed in our method, evaluating their contributions on Chinese VisionMath-Bench and English Vision-MathVerse (overall accuracy is reported due to space limitations) to validate the efficacy in enhancing problem-solving capabilities. Note that, Due to source constraints, 1M subsets of VisionMath-OCR and VisionMath-IT are sampled for training.

Is a dedicated OCR capability enhancement stage necessary? Yes. As demonstrated in Table 3, omitting the foundational mathematical OCR capability acquisition during initial training leads to significant performance degradation across both evaluation benchmarks. This verifies that precise OCR, specifically optimized for text, symbol, formulas of problem statement, serves as a critical enabler for subsequent correct reasoning execution.

Is the math figure comprehension strengthen stage beneficial? Yes. Table 4 presents results of different methods: (1) baseline without enhanced figure comprehension, (2) figure structural understanding enhancement using captions, (3) figure property understanding enhancement leveraging mathematical educational videos, and (4) a combined

	VisionMath-Benc					MathVerse
	Pla Geo	Ana Geo	Sol Geo	Alg	Func	
wo OCR-stage	37.2	31.1	34.3	31.3	35.2	37.5
Ours	39.0	33.2	36.9	34.3	37.8	41.1

Table 3. Effectiveness of Mathematical OCR enhancement.

Fig. Cap.	M.E. Vid.	VisionMath-Benc					MathVerse
		Pla Geo	Ana Geo	Sol Geo	Alg	Func	
		37.0	30.3	32.2	31.1	32.4	36.7
✓		38.0	31.3	33.0	33.6	33.3	38.2
	✓	38.5	33.8	34.6	31.9	34.0	39.8
✓	✓	39.0	33.2	36.9	34.3	37.8	41.1

Table 4. Contribution of Figure Captioning (“Fig. Cap.”) and Mathematical Education Video Learning (“M.E. Vid.”) to enhancing the mathematical figure comprehension.

Eng. Pro.	Chi. Pro.	VisionMath-Benc					MathVerse
		Pla Geo	Ana Geo	Sol Geo	Alg	Func	
✓		24.0	23.8	25.2	30.1	35.2	39.2
	✓	38.2	32.0	35.2	34.9	37.3	26.1
✓	✓	39.0	33.2	36.9	34.3	37.8	41.1

Table 5. Ablations on the effect of the collected vision-form English Problem (“Eng. Pro.”) and Chinese Problem (“Chi. Pro.”)-Solving datasets.

strategy integrating the above enhancement. Compared to the (1) baseline, both (2) figure structural enhancement and (3) property understanding enhancement strategies improve performance, and (4) combining them yields further gains. This underscores the importance of understanding and utilizing the component structures and properties of math figures for correct reasoning, making the specialized figure understanding strengthen stage essential.

Are collected English and Chinese vision-form problem-solving datasets effective and complementary? Yes. Table 5 demonstrates that instruction tuning with exclusively English (or Chinese) data achieves only acceptable performance on their language-specific benchmarks-MathVerse for English and VisionMath-Bench for Chinese. In contrast, simultaneous utilization of bilingual data yields performance improvements on both benchmarks compared to monolingual training. This demonstrates that mathematical reasoning capabilities transcend language barriers, thus aggregating multilingual datasets synergistically enhances reasoning to develop better problem-solving models.

5.4. Qualitative Analysis

Figure 5 presents visualization examples of vision-form problem-solving from advanced open-source model Qwen2-VL, closed-source model GPT-4o, and our VisionMath. Analysis of the geometry problem in (a) reveals that both Qwen2-VL and GPT-4o exhibit misinterpretations of geometric properties: Qwen2-VL incorrectly states “... AB is parallel to CD , angle A is equal to angle BOC ...”, while GPT-4o asserts “... AB is parallel to CD ... angle C (50°) and angle A are corresponding angles”. Erroneous geo-

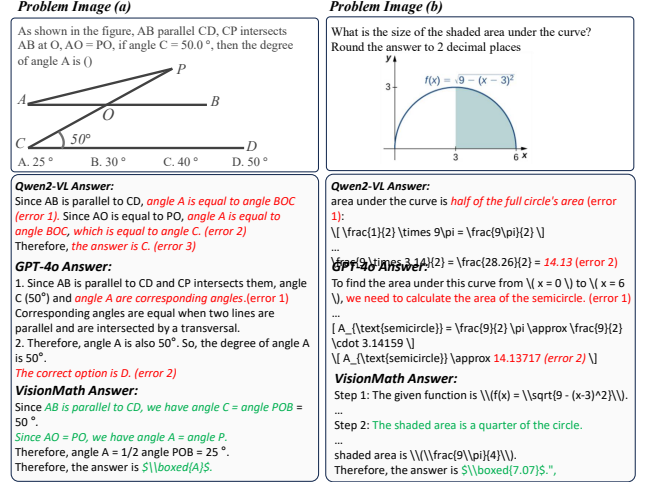


Figure 5. Quantitative comparisons of Qwen2-VL, GPT-4o, and VisionMath on the vision-form mathematical problem-solving task. Red and green text in answers mark incorrect and correct math understanding and reasoning, respectively.

metric propositions that propagate through their reasoning chains, resulting in incorrect answers. In contrast, our VisionMath precisely interprets the spatial relationships with the correct proposition “... AB is parallel to CD , we have angle $C =$ angle POB ...”, demonstrating rigorous geometric reasoning that ultimately arrives at the correct answer. This further substantiates the efficacy of the specialized enhanced figure comprehension training. In case (b), where the problem requires calculating the shaded area under the curve, both Qwen2-VL and GPT-4o fail to capture the constraint, erroneously computing the entire semicircle area beneath the curve. In contrast, VisionMath accurately parse the condition, correctly computing the quarter-circle area to obtain at the validated solution. More qualitative examples can be found in the Sec. D. of the supplementary materials.

6. Conclusion

In this work, we pioneer the exploration of a practical vision-form mathematical problem-solving task. We systematically deconstruct three core capabilities required for vision-form problem resolution, *i.e.*, precise problem statement OCR, advanced comprehension of math figure structures and properties, and excellent multimodal reasoning capacity. Correspondingly, we design a three-phase progressive multimodal reasoning alignment strategy to incrementally develop these competencies. To comprehensively evaluate model performance, we construct multilingual benchmarks spanning various problem types. Experimental results demonstrate that our method significantly outperforms open-source general-purpose, math-specialized and tool-integrated MLLMs, while achieve comparable performance to closed-source models with significantly larger parameter scales.

Acknowledgments This work is supported by the Natural Science Foundation of China (Grants No. 62192782, No. 62302501, No. 62403462), the Key Research and Development Program of Xinjiang Uyghur Autonomous Region, Grant No. 2023B03024.

References

- [1] Marah Abidin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 7
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1, 3
- [4] Edward Beeching, Shengyi Costa Huang, Albert Jiang, Jia Li, Benjamin Lipkin, Zihan Qina, Kashif Rasul, Ziju Shen, Roman Soletskyi, and Lewis Tunstall. NuminaMath-7B-CoT, 2024. 3, 6
- [5] Siow-Wee Chang, Sameem Abdul-Kareem, Amir Feisal Merican, and Rosnah Binti Zain. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC bioinformatics*, 14:1–15, 2013. 3
- [6] Wenhui Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyu Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019. 2, 3
- [7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 3, 7
- [8] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [9] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 3, 5, 7
- [10] Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, et al. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multi-modal models. *arXiv preprint arXiv:2410.17885*, 2024. 1, 3
- [11] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. Turl: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40, 2022. 3
- [12] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024. 6
- [13] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 3
- [14] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-llava: Solving geometric problem with multi-modal large language model, 2023. 1, 2, 3, 7
- [15] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*. 3
- [16] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE, 2015. 2
- [17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7
- [18] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018. 3
- [19] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 3
- [20] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 3
- [21] Sunjun Kweon, Yeonsu Kwon, Seonhee Cho, Yohan Jo, and Edward Choi. Open-wikitable: Dataset for open domain question answering with complex reasoning over table. *arXiv preprint arXiv:2305.07288*, 2023. 2
- [22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 3, 7

- [23] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 3
- [24] Zongzhao Li, Jiacheng Cen, Bing Su, Wenbing Huang, Tingyang Xu, Yu Rong, and Deli Zhao. Large language-geometry model: When llm meets equivariance. *arXiv preprint arXiv:2502.11149*, 2025.
- [25] Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. Star-rl: Spatial transformation reasoning by reinforcing multimodal llms. *arXiv preprint arXiv:2505.15804*, 2025. 3
- [26] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023. 3
- [27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 1, 3, 5
- [28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
- [29] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 3
- [30] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 3
- [31] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 7
- [32] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024. 1, 2
- [33] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*, 2025. 3, 7
- [34] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 3
- [35] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 3
- [36] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 3
- [37] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1697–1706, 2022. 2
- [38] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 3
- [39] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2020. 3
- [40] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015. 3
- [41] Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. Multimath: Bridging visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*, 2024. 3, 7
- [42] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Mathllava: Bootstrapping mathematical reasoning for multimodal large language models, 2024. 1, 3, 7
- [43] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer, 2021. 3
- [44] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yuez Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. 2023. 3
- [45] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yuez Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. 2023. 3
- [46] Stacey Svetlichnaya. Deepform: Understand structured documents at scale. 2020. 3
- [47] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13878–13888, 2021. 2, 3
- [48] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3
- [49] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 3, 7

- [50] Qwen Team. Qwen2.5-vl, 2025. [1](#), [3](#)
- [51] Michał Turski, Tomasz Stanisławek, Karol Kaczmarek, Paweł Dyda, and Filip Graliński. Ccpdf: Building a high quality corpus for visually rich documents from web crawl data, 2023. [2](#), [3](#)
- [52] 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. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. [1](#), [3](#), [7](#)
- [53] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. [3](#)
- [54] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*, 2023. [3](#)
- [55] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024. [5](#)
- [56] Zhen Yang, Jinhao Chen, Zhengxiao Du, Wenmeng Yu, Weihao Wang, Wenyi Hong, Zhihuan Jiang, Bin Xu, Yuxiao Dong, and Jie Tang. Mathglm-vision: Solving mathematical problems with multi-modal large language model. *arXiv preprint arXiv:2409.13729*, 2024. [3](#)
- [57] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. [3](#), [7](#)
- [58] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruochi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. [3](#)
- [59] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024. [2](#)
- [60] 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. Mavis: Mathematical visual instruction tuning, 2024. [2](#), [3](#)
- [61] Wenqi Zhang, Hang Zhang, Xin Li, Jiashuo Sun, Yongliang Shen, Weiming Lu, Deli Zhao, Yueting Zhuang, and Lidong Bing. 2.5 years in class: A multimodal textbook for vision-language pretraining. *arXiv preprint arXiv:2501.00958*, 2025. [2](#), [3](#), [4](#)
- [62] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer, 2020. [3](#)
- [63] Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. *arXiv preprint arXiv:2408.08640*, 2024. [3](#), [7](#)