InfiMM-WebMath-40B: Advancing Multimodal Pre-Training for Enhanced Mathematical Reasoning

Xiaotian Han[1] **Yiren Jian**[1] **Xuefeng Hu**[1] **Haogeng Liu**[1]
Yiqi Wang[2] Qihang Fan[2] Yuang Ai[1] Huaibo Huang[2]
Ran He[2] Zhenheng Yang[1] Quanzeng You[1]
[1]TikTok
[2]Institute of Automation, University of the Chinese Academy of Sciences

Abstract

Pre-training on large, high-quality datasets is essential for improving the reasoning abilities of Large Language Models (LLMs), particularly in specialized fields like mathematics. However, the field of Multimodal LLMs (MLLMs) lacks a comprehensive, open-source dataset for mathematical reasoning. To fill this gap, we present InfiMM-WebMath-40B[1], a high-quality dataset of interleaved image-text documents. It consists of 24 million web pages, 85 million image URLs, and 40 billion text tokens, all carefully extracted and filtered from CommonCrawl. We outline our data collection and processing pipeline in detail. Models trained on InfiMM-WebMath-40B demonstrate strong performance in both text-only and multimodal settings, setting a new state-of-the-art on multimodal math benchmarks such as MathVerse and We-Math.

1 Introduction

Recent advancements in Large Language Models (LLMs)(AI, 2024; Anthropic, 2024; Dubey et al., 2024) have improved their ability to handle complex reasoning and multi-step mathematical problems through techniques like Chain-of-Thought (CoT) prompting(Wei et al., 2022). These models excel from basic GSM8K word problems (Cobbe et al., 2021b) to high school-level MATH tasks (Hendrycks et al., 2021b). Specialized smaller LLMs like DeepSeekMath-7B (Shao et al., 2024) and InternLM-Math (Ying et al., 2024) have also made notable progress in mathematics, demonstrating strong performance in focused domains.

Although most mathematical knowledge is text-based, visual elements such as figures and diagrams are essential for understanding abstract concepts. To integrate these visual components, Multimodal LLMs (MLLMs) like G-LLaVA (Gao et al., 2023b), Math-LLaVA (Shi et al., 2024a), and MAVIS (Zhang et al., 2024d) have been developed. These models enhance reasoning by incorporating visual inputs through embeddings from pre-trained models like CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023), and use multimodal instruction datasets such as Geo170k (Cai et al., 2024), MathV360K (Shi et al., 2024b), and MAVIS-Instruct (Zhang et al., 2024c).

However, introducing new knowledge during instruction fine-tuning is challenging (Zhu and Li, 2023), often leading to hallucinations (Gekhman et al., 2024), particularly due to limitations in dataset scale and quality. While large corporations benefit from proprietary datasets, the open-source community lacks comprehensive pre-training datasets for mathematical reasoning that integrate text and visual data.

To address this gap, we introduce **InfiMM-WebMath-40B**, the first large-scale, publicly available multimodal mathematics pre-training dataset. Comprising 24 million web documents, 85 million image URLs, and 40 billion text tokens, it provides a valuable resource for training Multimodal LLMs (MLLMs). We validate the effectiveness of InfiMM-WebMath-40B through experiments on benchmarks like MathVerse (Zhang et al., 2024b) and WeMath (Qiao et al., 2024), showing improved performance in multimodal math reasoning.

Our contributions include: (1) We introduce InfiMM-WebMath-40B, the first large-scale, multimodal math dataset for pre-training, filling a critical gap in open-source research. (2) We provide a detailed preprocessing pipeline for filtering relevant content from CommonCrawl to ensure high-quality, relevant data. (3) We demonstrate the impact of InfiMM-WebMath-40B through experiments, where our models excel on multimodal mathematical benchmarks, showcasing the dataset's potential for advancing MLLM research.

[1] We have released our data at HuggingFace.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14221–14231
November 4-9, 2025 ©2025 Association for Computational Linguistics

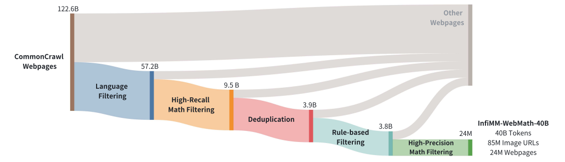


Figure 1: InfiMM-WebMath-40B data curation pipeline.

2 Related Work

LLMs have demonstrated potential in mathematical reasoning across various studies. To evaluate and enhance their capabilities, several math-specific benchmarks (Cobbe et al., 2021a; Hendrycks et al., 2021c,a; Azerbayev et al., 2024; Naeini et al., 2023; Liu et al., 2024; Zhou et al., 2024) and training datasets, both proprietary (Polu and Sutskever, 2020; Lightman et al., 2023; Lewkowycz et al., 2022) and open-source (Hendrycks et al., 2021b; Welleck et al., 2021; Paster et al., 2023a; Wang et al., 2023; Yue et al., 2023), have been introduced.

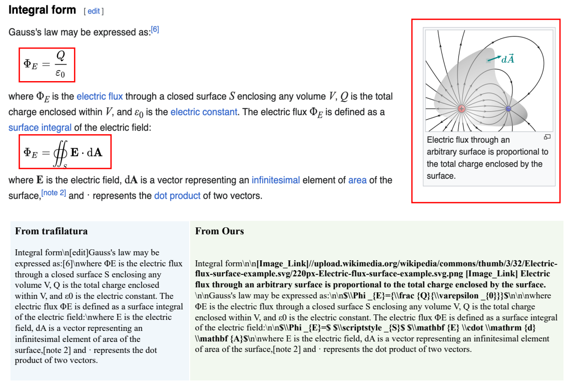
The rise of Multimodal LLMs (MLLMs) has sparked interest in enhancing their multimodal reasoning capabilities. To support this, various evaluation benchmarks (Zhang et al., 2024a; Lu et al., 2021; Kazemi et al., 2023; Xia et al., 2024; Masry et al., 2022; Xu et al., 2024; Lu et al., 2024; Zhang et al., 2024b; Qiao et al., 2024) and training datasets (Cai et al., 2024; Gao et al., 2023a; Shi et al., 2024b; Zhu et al., 2023; Laurençon et al., 2023; Awadalla et al., 2024; Li et al., 2024b) have been developed to assess and enhance MLLMs' mathematical reasoning skills.

3 Dataset Construction

In this section, we detail the methodology used to construct InfiMM-WebMath-40B from the CommonCrawl archives. InfiMM-WebMath-40B is a large-scale multimodal math dataset integrating interleaved text and image data, following approaches used in prior works (Penedo et al., 2023; Li et al., 2024a; Penedo et al., 2024). We enhance the methodology used in the OBELICS dataset (Laurençon et al., 2023) by incorporating both text and corresponding image URLs.

3.1 Text-only Data Curation Pipeline

Text Extraction and Language Filtering We chose Trafilatura (Barbaresi, 2021), a Python library widely used to extract text from web pages. While effective for text extraction, Trafilatura omits mathematical symbols and equations. Therefore, the subsequent section will outline our development of a specialized extraction tool tailored for math-related content.



Integral form [edit]

Gauss's law may be expressed as:[6]

$$\Phi_E = \frac{Q}{\varepsilon_0}$$

where Φ_E is the electric flux through a closed surface S enclosing any volume V, Q is the total charge enclosed within V, and ε_0 is the electric constant. The electric flux Φ_E is defined as a surface integral of the electric field:

$$\Phi_E = \oiint_S \mathbf{E} \cdot \mathrm{d}\mathbf{A}$$

where $\mathbf{E}$ is the electric field, $\mathrm{d}\mathbf{A}$ is a vector representing an infinitesimal element of area of the surface,[note 2] and · represents the dot product of two vectors.

Electric flux through an arbitrary surface is proportional to the total charge enclosed by the surface.

From trafilatura	**From Ours**
Integral form\n[edit]Gauss's law may be expressed as:[6]\nwhere ΦE is the electric flux through a closed surface S enclosing any volume V, Q is the total charge enclosed within V, and ε0 is the electric constant. The electric flux ΦE is defined as a surface integral of the electric field:\nwhere E is the electric field, dA is a vector representing an infinitesimal element of area of the surface,[note 2] and · represents the dot product of two vectors.	Integral form\n\n\n**[Image_Link]//upload.wikimedia.org/wikipedia/commons/thumb/3/32/Electric-flux-surface-example.svg/220px-Electric-flux-surface-example.svg.png [Image_Link] Electric flux through an arbitrary surface is proportional to the total charge enclosed by the surface.** \n\nGauss's law may be expressed as:\n\n**$\\Phi _{E}={\\frac {Q}{\\varepsilon _{0}}}$**\n\n\nwhere ΦE is the electric flux through a closed surface S enclosing any volume V, Q is the total charge enclosed within V, and ε0 is the electric constant. The electric flux ΦE is defined as a surface integral of the electric field:\n\n**$\\Phi _{E}=$ $\\scriptstyle _{S}$ $\\mathbf {E} \\cdot \\mathrm {d} \\mathbf {A}$**\n\nwhere E is the electric field, dA is a vector representing an infinitesimal element of area of the surface,[note 2] and · represents the dot product of two vectors.

Figure 2: A comparative illustration of extraction results from a Wikipedia webpage using Trafilatura and our enhanced version of Resiliparse, highlighting the successful retrieval of mathematical equations and image URLs.

Following DeepSeekMath (Shao et al., 2024), we focus on retaining only Chinese and English content when constructing our dataset. To achieve this, we apply language filtering to the CommonCrawl repositories with approximately 122 billion webpages, as shown in Figure 1. For language detection, we employ a fastText language identification model (Joulin et al., 2016). This language filtering process significantly reduces the dataset size, lowering the number of pages from 122 billion to 57.2 billion.

Mathematical Content Extraction Extracting mathematical content from HTML is challenging, as standard tools often fail to accurately capture LaTeX equations and image URLs. After evaluating multiple tools, we selected Resiliparse (Bevendorff et al., 2018) as the foundation for our approach. As shown in Figure 2, our enhanced version outperforms Trafilatura by preserving the original content order, ensuring the logical flow of mathematical arguments. Unlike conventional extractors, which often misinterpret equations or disrupt their structure, our method accurately retains LaTeX notation and its placement within the text. Additionally, it maintains image URLs and their positions, preserving crucial connections between text, equations, and visual elements.

High-Recall Filtering for Mathematical Content Inspired by DeepSeekMath (Shao et al., 2024), we

Model	Base LLM	All	Text Dom	Text Lite	Vision Intense	Vision Dom	Vision Only
Human	-	64.9	71.2	70.9	61.4	68.3	66.7
Proprietary Models							
GPT-4V	N/A	39.4	54.7	41.4	34.9	34.4	31.6
Gemini-Pro	N/A	23.5	26.3	23.5	23.0	22.3	22.2
Open-sourced Models							
SPHINX-Plus	LLaMA2-13B	14.0	16.3	12.8	12.9	14.7	13.2
G-LLaVA	LLaMA2-7B	15.7	22.2	20.4	16.5	12.7	6.6
InternLM-XC2	InternLM2-7B	16.5	22.3	17.0	15.7	16.4	11.0
Math-LLaVA	Vicuna-13B	19.0	21.2	19.8	20.2	17.6	16.4
ShareGPT4V	Vicuna-13B	17.4	21.8	20.6	18.6	16.2	9.7
LLaVA-NeXT	LLaMA3-8B	19.3	24.9	20.9	20.8	16.1	13.8
LLaVA-NeXT	Qwen-1.5-110B	24.5	31.7	24.1	24.0	22.1	20.7
MAVIS	Mammoth2-7B	27.5	41.4	29.1	27.4	24.9	14.6
Our Models							
InfiMM-Math	DS-Coder-1.3B	26.9	37.1	30.2	29.2	24.4	13.7
InfiMM-Math	DS-Coder-1.5-7B	34.5	46.7	32.4	38.1	32.4	15.8

Table 1: Evaluation of models on MathVerse. Further elaborations on performance of vision only tasks are discussed in Appendix H.

	CPT	IFT	Scores
DSC-1.3B		Mavis	20.2
DSC-1.3B	✓	Mavis	25.1
DSC-1.3B		Extended	22.3
DSC-1.3B	✓	Extended	26.9

Table 2: Datasets ablations (CPT and IFT) using Deepseek-coder-1.3B, evaluated on MathVerse w/o scores.

	CPT	IFT	Scores
DSC-1.5-7B		Mavis	22.8
DSC-1.5-7B	✓	Mavis	27.1
DSC-1.5-7B		Extended	23.8
DSC-1.5-7B	✓	Extended	29.1

Table 3: Datasets ablations (CPT and IFT) using Deepseek-coder-1.5-7B, evaluated on MathVerse w/o scores.

trained a fastText classifier to filter mathematical content, using half a million positive samples from OpenWebMath (Paster et al., 2023b) and negative samples from our earlier extracted content. This filtering reduced the dataset from 57.2 billion to 9.5 billion samples, prioritizing recall with a probability threshold set at 0.4.

Deduplication We applied MinHash (Broder, 1997) for content deduplication, following FineWeb's methodology (Penedo et al., 2024). Deduplication was performed within each snapshot and neighboring snapshot pairs, reducing the dataset by 43%, from 9.5 billion to 5.4 billion samples. URL deduplication further reduced the sample size to 3.9 billion.

Rule-based Filtering We applied a few essential filtering rules, such as removing "lorem ipsum" content, applying a punctuation ratio rule for English, filtering NSFW content, and excluding documents with Unicode errors. This step eliminated 3% of the samples, resulting in 3.8B samples.

High-Precision Filtering for Mathematical Content To enhance the accuracy of our labeling process, we utilized the LLaMA3-70B-Instruct model (Dubey et al., 2024), using prompt formats inspired by the FineWeb-Edu dataset (Lozhkov et al., 2024). This approach allowed us to score the mathematical quality of each sample on a scale from 0 to 10. The full prompt is displayed in Table 5 of Appendix.

From the data remaining after rule-based filtering, we randomly sampled approximately one million entries. We assigned math quality scores and applied a threshold of 6 to select 640,000 positive samples for training our updated fastText classifier, alongside an equivalent number of 640,000 randomly selected negative samples from prior filtering steps. These positive and negative samples were combined to train the new fastText classifier.[2]

During fastText training, we implement data cleaning rules to optimize the model's performance for mathematical content. Mathematical texts pose unique challenges due to specialized terminology, symbols, formulas, and numeric data, which differ from typical natural language and require more refined preprocessing techniques.

Our goal is to standardize and simplify the input training data while preserving essential mathematical information. Key considerations include maintaining consistency in token representation, minimizing noise from extraneous characters, and standardizing numeric values. The following steps reflect this approach:

- Utilizing the SpaCy English language model (`en_core_web_sm`), we preprocess the input text, tokenize it, and process each token by converting it to its lowercase and lemmatized form. Common placeholders are replaced, certain non-alphanumeric characters are removed, and patterns of special characters like dashes and underscores are normalized. We also strip any unnecessary whitespace, ensuring the text is well-prepared for downstream

[2] We also employ an LLM-based classifier for high-precision filtering, Appendix B shows the comparison.

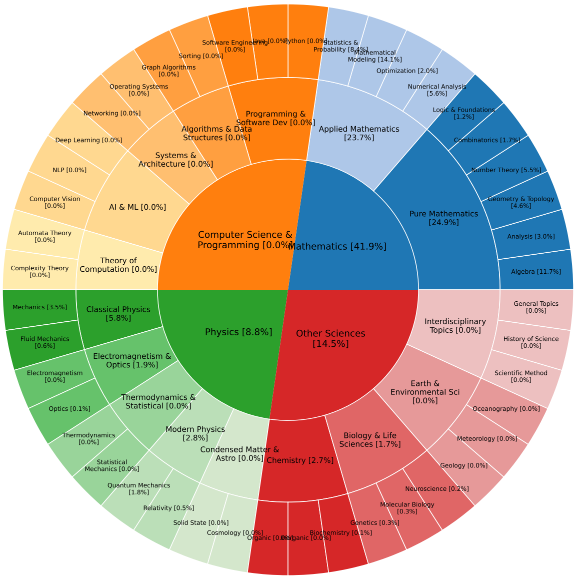


Figure 3: Topic Distribution of our sampled dataset.

processing.

- All numeric values are replaced with the <NUM> placeholder to standardize the representation, and line breaks along with carriage returns are removed. Tokens exceeding 100 characters in English are discarded.

For evaluation, we used all samples in the Geometry3K (Lu et al., 2021) benchmark as positive examples of mathematical content. With these refined preprocessing techniques, fastText's accuracy improved from 48.74% to 72.15%.

Text Topics Distribution We provide the data statistics of our sampled dataset. To assign mathematical topics, we use *LLaMA-3.1-70B-Instruct* (Dubey et al., 2024) to assign math topics to them. The distribution can be found in Figure 3. As shown in the figure, we can see that the *Infimm-WebMath-40B* covers a wide range of topics in the STEM domains, which possibly explains why it performs well in our experimental studies.

Text-Only Filtering Evaluation We pretrained a deepseek-coder-1.3b-base model on the filtered text dataset and evaluated its performance on GSM8K (Cobbe et al., 2021b) and the MMLU (STEM) (Hendrycks et al., 2021a). Our model outperformed both OpenWebMath and DeepSeek-Math, highlighting the quality of our dataset (results are shown in Appendix C).

3.2 Multimodal Data Construction

After filtering, 24 million documents with 85 million image URLs remained. Following the OBELICS format (Laurençon et al., 2023), all image URLs and extracted texts were preserved and organized into the interleaved image-text format, maintaining the same order as in the original document layout. We recognize that irrelevant images are often present in web documents. We present the detailed filtering process in Appendix F.

4 Experiments

Model Architectures We employ the SigLip model `siglip-so400m-patch14-384` to extract visual features, a 3-layer Perceiver Resampler (Jaegle et al., 2021) (see Appendix G for more details) with 64 latents to reduce the number of tokens/features per image to 64. These visual token/feature embeddings are then concatenated with text embeddings before being fed into the LLMs (DeepSeek-Coder (Guo et al., 2024): `deepseek-coder-1.3b-base` and `deepseek-coder-7b-v1.5`).

Training Details Our training data and processes involve a three-stage approach: modality alignment, continued pre-training using InfiMM-WebMath-40B, and instruction fine-tuning. Detailed training procedures are provided in the Appendix D. We refer to our resulting model as InfiMM-Math.

Evaluations on MathVerse In line with official MathVerse guidelines, we report the "w/o" score. The results in Table 1 show that our 7B model outperforms all open-source models, including the 110B LLaVA-NeXT, and surpasses Gemini-Pro and Qwen-VL-Max, trailing only GPT-4V. Our model demonstrates exceptional performance in the Text-Dominant, Text-Lite, Vision-Intense, and Vision-Dominant categories, highlighting its strong multimodal capabilities in processing both text and visual inputs. When comparing models of similar sizes, our model demonstrates competitive performance against state-of-the-art approaches on Vision-Only categories as well. Our 7B model achieved 15.8, outperforming LLaVA-Next-8B (13.8) and MAVIS-7B (14.6).

Evaluations on We-Math Here, we compare models on the We-Math benchmarks, consisting of 6.5K visual math questions. We report results

Model	Base LLM	AVG ↑	IK ↓	IG ↑	RM ↓
Proprietary Models					
Gemini-1.5-Pro	N/A	26.4	42.7	11.2	54.8
GPT-4V	N/A	31.1	39.8	14.5	47.9
Open-sourced Models					
LLaVA-1.6	Vicuna-7B	3.3	78.3	2.5	89.1
LLaVA-1.6	Vicuna-13B	5.2	69.1	3.2	86.9
DeepSeek-VL	DeepSeek-7B	6.3	69.1	4.6	84.8
G-LLaVA	Vicuna-13B	6.5	64.2	4.6	86.6
Math-LLaVA	Vicuna-13B	11.1	–	–	72.8
InternLM-XC2	InternLM2-7B	12.7	56.4	10.5	77.6
Our Models					
InfiMM-Math	DS-Code-1.3B	13.1	56.2	9.1	73.7
InfiMM-Math	DS-Base-7B	20.6	48.8	12.2	61.7

Table 4: Evaluations on the We-Math benchmark. AVG represents the primary metric of interest.

on the `testmini` set using four metrics: Insufficient Knowledge (IK), Inadequate Generalization (IG), Complete Mastery (CM), and Rote Memorization (RM). As shown in Table 4, our model, InfiMM-Math, surpasses all open-source models.

5 CPT and IFT Dataset Ablations on MathVerse

In this section, we compare models trained with and without our own mathematical multi-modal pre-training dataset, InfiMM-WebMath-40B. Additionally, we evaluate two IFT dataset configurations: (a) a combination of MAVIS-Caption-to-QA, MAVIS-Existing-Dataset-Augment, MAVIS-Caption, MAVIS-DataEngine-Geometry, and MAVIS-Meta-Question (referred to as the MAVIS dataset); and (b) a broader set consisting of the MAVIS datasets along with Vflan, VisualWebInstruct, AI2D, CHARTQA, DOCVQA, DVQA, GEOQA, DART-Math, and Numina-Math (referred to as the Extended dataset).

As shown in Table 2, in the 1.3B model, CPT improves the MathVerse scores by 4.9 and 4.6 points when IFT is performed with MAVIS and Extended datasets, respectively. Similarly, Table 3 shows that in the 7B model, CPT improves the MathVerse scores by 4.8 and 5.3 points with MAVIS and Extended datasets, respectively. In contrast, using broader IFT datasets typically enhances model performance by approximately 2 points. These results highlight the significant mathematical capabilities imparted to the models through our InfiMM-WebMath-40B for CPT.

6 Conclusions

In this work, we introduced InfiMM-WebMath-40B, the first large-scale multimodal pretraining dataset for mathematical reasoning, filling a crucial gap in open-source research. Our dataset significantly enhances models' performances on key benchmarks.

7 Potential Risk

Although we have applied rule-based filtering to exclude NSFW web pages, it could be possible that our dataset inevitably contains some harmful contents. Users are suggested to use with caution.

8 Limitations

Although our models excel compared to many open-sourced models due to the introduction of InifMM-WebMath-40B, the models lack enhanced vision capabilities to specifically read math content. For future work, we aim to develop enhanced higher-resolution vision encoders tailored to effectively process mathematical symbols, diagrams, and equations.

On the other hand, after being continual trained on a highly dense multimodal mathematical dataset, other capabilities (such as commonsense knowledge reasoning or domain-specific knowledge) may experience "catastrophic forgetting". This main come from underlying base LLM. Our multimodal CPT cannot compensate for potential shortcomings intrinsic to the base model itself.

References

Open AI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Kumar Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, Ran Xu, Yejin Choi, and Ludwig Schmidt. 2024. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *Preprint*, arXiv:2406.11271.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen Marcus McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. Llemma: An open language model for mathematics. In *The Twelfth International Conference on Learning Representations*.

Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131. Association for Computational Linguistics.

Edward Beeching, Shengyi Costa Huang, Albert Jiang, Jia Li, Benjamin Lipkin, Zihan Qina, Kashif Rasul, Ziju Shen, Roman Soletskyi, and Lewis Tunstall. 2024. Numinamath 7b tir. `https://huggingface.co/AI-MO/NuminaMath-7B-TIR`.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2018. Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In *Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

Shihao Cai, Keqin Bao, Hangyu Guo, Jizhi Zhang, Jun Song, and Bo Zheng. 2024. Geogpt4v: Towards geometric multi-modal large language models with geometric image generation. *Preprint*, arXiv:2406.11503.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*.

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021a. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. 2024. Data filtering networks. In *The Twelfth International Conference on Learning Representations*.

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

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

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. 2024. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021c. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

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

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2023. Geomverse: A systematic evaluation of large models for geometric reasoning. *Preprint*, arXiv:2312.12241.

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

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Preprint*, arXiv:2306.16527.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. *Preprint*, arXiv:2206.14858.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large scale language model society. *Preprint*, arXiv:2303.17760.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. 2024a. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *Preprint*, arXiv:2403.00231.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Naiming Liu, Shashank Sonkar, Myco Le, and Richard Baraniuk. 2024. Malalgoqa: A pedagogical approach for evaluating counterfactual reasoning abilities. *Preprint*, arXiv:2407.00938.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu. Software.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Preprint*, arXiv:2310.02255.

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. *Preprint*, arXiv:2105.04165.

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

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *The Eleventh International Conference on Learning Representations*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *Preprint*, arXiv:2203.10244.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Saeid Naeini, Raeid Saqur, Mozhgan Saeidi, John Giorgi, and Babak Taati. 2023. Large language models are fixated by red herrings: Exploring creative problem solving and einstellung effect using the only connect wall dataset. *Preprint*, arXiv:2306.11167.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2023a. Openwebmath: An open dataset of high-quality mathematical web text. *Preprint*, arXiv:2310.06786.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2023b. Openwebmath: An open dataset of high-quality mathematical web text. *arXiv preprint arXiv:2310.06786*.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2406.17557.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Stanislas Polu and Ilya Sutskever. 2020. Generative language modeling for automated theorem proving. *Preprint*, arXiv:2009.03393.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng

Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *Preprint*, arXiv:2407.01284.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Yu Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024a. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024b. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *Preprint*, arXiv:2406.17294.

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

Zengzhi Wang, Rui Xia, and Pengfei Liu. 2023. Generative ai for math: Part i – mathpile: A billion-token-scale pretraining corpus for math. *arXiv preprint arXiv:2312.17120*.

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

Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. 2021. Naturalproofs: Mathematical theorem proving in natural language. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *Preprint*, arXiv:2402.12185.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2024. Chartbench: A benchmark for complex visual reasoning in charts. *Preprint*, arXiv:2312.15915.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, et al. 2024. Internlm-math: Open math large language models toward verifiable reasoning. *arXiv preprint arXiv:2402.06332*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*.

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

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Jiaxin Zhang, Zhongzhi Li, Mingliang Zhang, Fei Yin, Chenglin Liu, and Yashar Moshfeghi. 2024a. Geoeval: Benchmark for evaluating llms and multi-modal models on geometry problem-solving. *Preprint*, arXiv:2402.10104.

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.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024b. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *Preprint*, arXiv:2403.14624.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, and Hongsheng Li. 2024c. Mavis: Mathematical visual instruction tuning. *Preprint*, arXiv:2407.08739.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. 2024d. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*.

Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek F. Wong, Xiaowei Huang, Qiufeng Wang, and Kaizhu Huang. 2024. Is your model really a good math reasoner? evaluating mathematical reasoning with checklist. *Preprint*, arXiv:2407.08733.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Preprint*, arXiv:2304.06939.

Zeyuan Allen Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*.

A Using Prompting with Llama-3-70B for Mathematical Annotation

We display the full prompt used in High-Precision Filtering for Mathematical Content in Table 5.

B Ablation Studies on High-Precision Mathematical Content Filtering

In this section, we examine the efficacy of two classifiers—LLM-based and fastText-based—focusing on high-precision mathematical content filtering. The comparison utilizes the DeepSeek-Coder 1.3B model, which we trained on a dataset previously introduced in Sec. High-Recall Filtering for Mathematical Content with a sequence length of 4096. This model was trained to score documents based on their relevance to mathematical content on a scale from 0 to 10.

We conduct the continue pretraining of the DeepSeekCoder 1.3B model using datasets filtered by both the LLM- and fastText-based classifiers. Table 6 shows the performance results. The results highlight a length bias in the LLM-based method, which tends to favor longer documents, averaging 2,500 tokens, compared to 1,700 tokens for the FastText filter. The length bias associated with the LLM-based classifier has adversely impacted the dataset's performance on the GSM8K dataset. As indicated in the table, the LLM-filtered dataset achieved lower accuracy (17.5%) on the GSM8K dataset compared to the fastText-filtered dataset (20.2%). This decrease in performance indicates that the LLM's preference for longer documents may not align well with the requirements of datasets like GSM8K, which demand concise and precise mathematical descriptions.

Given these insights, we have decided to continue utilizing the fastText classifier for high-precision filtering in our ongoing research. Nonetheless, the implications of the LLM-based classifier require further investigation to fully understand and address its biases.

C Text-Only Filtering Evaluation

To provide a preliminary evaluation of the quality of our filtered dataset, we continue pretraining a deepseek-coder-1.3b-base model for one epoch using the filtered mathematical content in Sec. High-Precision Filtering for Mathematical Content, excluding image URLs. We validate the effectiveness of our math-related filtering with a few-shot evaluation using the GSM8K (Cobbe et al., 2021b) and the STEM sections of the MMLU (Hendrycks et al., 2021a) benchmark.

As shown in Table 7, the model trained on our InfiMM-WebMath-40B text-only dataset demonstrates competitive performance compared to OpenWebMath and the DeepSeekMath Corpus, highlighting the high quality of our dataset and the effectiveness of our filtering procedures.

D Training Details

Modality Alignment Stage In this stage, we utilize general-purpose image-text pairs to align the visual encoder and the LLM via Perceiver Resampler. The primary objective is to minimize the domain gap between visual and linguistic modalities. To achieve this, we sample a 8 million image-text pair subset from the DFN-2B dataset (Fang et al., 2024) for the alignment training. During this stage, the vision encoder and LLM backbone are frozen, and training is focused on the Perceiver Resampler module. Training is conducted for one epoch using DeepSpeed Zero2, with the AdamW optimizer, configured with a cosine learning rate scheduler, a maximum learning rate of $1e^{-4}$, betas of $(0.9, 0.95)$, and a weight decay of 0.1.

Continue Pre-training Stage We further continue pre-training our models using the InfiMM-WebMath-40B dataset to enhance the model's mathematical knowledge acquisition in a multimodal setting. The training is conducted for one epoch using DeepSpeed Zero2, with the AdamW optimizer, configured with a cosine learning rate scheduler, a maximum learning rate of $5e^{-5}$, betas of $(0.9, 0.95)$, and a weight decay of 0.1. The context length for training examples is set to 4096, with a maximum of 32 images per example. During this stage, the visual encoder remains frozen, and training focuses on learning the Perceiver Resampler module (the visual-language connector) and the LLM.

Instruction Fine-tuning Stage In this stage of training, we fine-tune our models using instruction datasets, including PGPS9K (Zhang et al., 2023), Geo170k (Gao et al., 2023a), TABMWP

```
Below is an extract from a web page. Evaluate the mathematical value of the extract and its potential utility as a teaching resource in a mathematical
context using the additive 10-point scoring system described below. Points accumulate based on the satisfaction of each criterion, with special
attention to the presence and quality of mathematical equations:
- 0 points if the extract includes no mathematical content, such as only provides historical context, summarizes an article's abstract, or exclusively
features a person's resume.
- 1-2 points if the extract offers rudimentary information on mathematical subjects, even if interspersed with irrelevant material such as advertisements
or non-academic content.
- 2-4 points if the extract touches upon mathematical topics without rigorous adherence to academic standards and contains a mix of mathematical and
non-mathematical content, or if the presentation is haphazard and the writing lacks clarity.
- 4-6 points if the extract presents key concepts pertinent to educational curricula and includes mathematical equations, albeit potentially
non-comprehensive or alongside superfluous information. It should resemble a mathematical text, such as an introductory section of a textbook or a basic
tutorial.
- 6-8 points if the extract is highly relevant to mathematics, is well-structured, and offers a clear exposition, including a significant number of
mathematical equations and solutions. It should be akin to an in-depth textbook chapter or tutorial, with a strong focus on mathematical content and
minimal unrelated information.
- 8-10 points if the extract exhibits exceptional mathematical merit, characterized by detailed explanations, a comprehensive array of mathematical
equations, and a coherent, accessible writing style that provides profound insights into mathematical theories and applications.
The extract: <EXAMPLE>.
After examining the extract: - Briefly justify your total score. - Conclude with the score using the format: "mathematical score: <total points>"
```

Table 5: Prompt for evaluating mathematical content using Llama-3-70B following FineWeb-Edu (Lozhkov et al., 2024).

Method	MMLU (STEM)	GSM8K	Text Avg Len
LLM-Clf	32.8	17.5%	2500
FastText-Clf	31.1	20.2%	1700

Table 6: Ablations on high-precision filtering. "Text Avg Len" indicates the average document length after filtering.

Training Corpus	GSM8K	MMLU (STEM)
Baseline	4.8	25.6
OpenWebMath	11.0	29.6
DeepSeekMath Corpus	23.8	33.1
InfiMM-WebMath-40B (text)	26.1	35.6

Table 7: Evaluation of models on GSM8K and MMLU (STEM). The baseline is the deepseek-coder-1.3b-base without any training.

(Lu et al., 2023), ScienceQA (Lu et al., 2022), Vflan (Chen et al., 2024), VisualWebInstruct, AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), DVQA (Kafle et al., 2018), GeoQA (Chen et al., 2021), and MAVIS (Zhang et al., 2024c). We find that incorporating uni-modal text instruction datasets is crucial for enhancing the models' instruction-following capabilities. Therefore, we also include pure text instruction datasets such as Math(Li et al., 2023), MetaMathQA (Yu et al., 2024), DART-Math (Tong et al., 2024), and NuminaMath (Beeching et al., 2024). The objective of this stage is to acclimate the models to the common chat templates used in math VQA settings, thereby enabling them to better utilize the mathematical knowledge acquired in the previous stage.

We freeze the vision encoder and update the parameters of the Perceiver Resampler and LLMs. As in the previous stages, training is conducted using DeepSpeed Zero2 for one epoch, with the AdamW optimizer, configured with 2000 warmup steps, a maximum learning rate of $5e^{-6}$, betas of $(0.9, 0.95)$, a weight decay of 0.1, and cosine decay to $5e^{-7}$. The batch size is set to one per GPU, and the context length of the training examples is set to 4096. We utilize 32 A100-80G GPUs for the 1.3b models and 64 A100-80G GPUs for the 7b models.

E Use of AI Assistants

For this submission, we use ChatGPT to fix grammar, revise and polish the text at the sentence level.

F Elaborations on Images and Text matching

In our dataset, all image URLs and extracted texts were preserved and organized into the interleaved image-text format, maintaining the same order as in the original document layout, following the practices of the OBELICS dataset. However, we recognize that irrelevant images are often present in web documents. To address this, we implemented a two-step filtering process:

- URL-Based Filtering: As described in the manuscript, we filtered out irrelevant image URLs containing specific keywords (e.g., "logo", "banner", "avatar", "icon"), URLs that appeared in more than 10 documents, and documents with over 100 image URLs. This process removed over 1 million unique irrelevant or redundant images.

- Image-Text Similarity Filtering: After downloading images from the filtered URLs, we identified and removed unrelated or mismatched images by calculating their similarity to the corresponding document text using the SigLIP-so400m model. For this step, document texts were divided into 64-token chunks to fit the SigLIP text encoder, and images with a similarity score below 0.01 for all chunks were removed. This additional filtering step eliminated 12% of the remaining unique images.

Finally, we analyzed the filtered dataset and found that 52.4% of the remaining images had a similarity score of at least 0.5 with one of the text chunks from the same document, and 18.5% had a similarity score of at least 0.99. These results demonstrate the relevance of both our text and image data.

G Perceiver as the Vision-Language Connector

Due to the nature of interleaved image and text data, directly concatenating image and text tokens results in a significant computational cost during training. For example, a training instance containing a piece of text with four interleaved images—each consuming 500+ tokens (the direct output of a ViT without Resampler)—would exceed 2,000 tokens, quickly exhausting the context length of a Transformer-based model.

To address this, following the model architecture used in Idefics2, we adopt the Perceiver Resampler, which improves model performance while significantly reducing the number of visual tokens. With the Perceiver Resampler, the token count per image is restricted to 64, mitigating the computational burden. This strategy has been demonstrated to be effective for multi-image VLM.

H Performance on Vision Only Tasks

Vision-only category removes all textual input, rendering the textual content directly onto the diagram while reducing the text prompt to a negligible level.

Since we utilize publicly available SFT data with minimal modifications and have not integrated high-resolution support into our model, we expect the performance in the Vision-only category to be comparable to other models.

However, when comparing models of similar sizes, our model demonstrates competitive performance against state-of-the-art approaches. For example, our 7B model achieved a score of 15.8 in the Vision-only category, outperforming LLaVA-Next-8B (13.8) and MAVIS-7B (14.6).

Finally, the performance on vision-only tasks can be further improved by using high-resolution Vision Transformers (ViTs) as visual encoders, which we leave for future work.