Math-Search: A Benchmark for Multi-hop Visual Reasoning over Plots

Pulkit Madan Sanjay Haresh Apratim Bhattacharyya Litian Liu
Reza Pourreza Sunny Panchal Roland Memisevic
Qualcomm AI Research
{pmadan, sanjayh, aprabhat, litiliu, pourreza, sunnpanc, rmemisevic}@qti.qualcomm.com

Abstract

We present a benchmark for visual understanding in multi-modal language models (MLLMs). A common failure mode of state-of-the-art MLLMs is that they focus on global, texture based features and fail to fully utilize local information in an image. As a result, they perform poorly on tasks that require fine-grained visual information. There has been a revived interest in models that utilize local information, specifically in tasks such as visual search. However, these models utilize only local information and do not incorporate global image context. In this work, we propose to use mathematical plot analysis to test the ability of MLLMs to utilize both local and global information for visual understanding. Our benchmark is designed to be straightforward for humans to solve, while requiring a combination of local and global understanding that is challenging for existing MLLMs. We also present an evaluation showing that state of the art MLLMs fall behind human performance by a very large margin.

1. Introduction

Our ability to solve difficult computer vision tasks has progressed rapidly in recent years thanks to the advent of foundation models, that combine vision with language and are pre-trained on a large amount of labeled data [10]. However, despite the impressive recent progress, existing vision-language models have surprisingly many failure modes that let them still remain far behind human visual capabilities on a wide variety of visual inference tasks [5, 16, 19, 20].

A particular failure mode is their inability to properly utilize fine-grained, local information during inference. This can can be traced back to common pre-training objectives (such as captioning), which instills in a model the ability to aggregate global texture information at the cost of more local information. Equipping vision models with local information extraction and search capabilities has been subject of revived recent interest [22] but it remains a largely open problem.

The most impressive showcase of the human ability to flexibly search for and then combine local visual information is the task of understanding diagrams and plots. These are detail-rich human-made artifacts, conveying complex, and often nuanced, numerical or symbolic information, directly via the visual system. The task of understanding plots often relies, by design, on a sophisticated interplay between detection of local cues on the one hand, such as axes, legends, intersection points, etc., and the methodological aggregation of those cues into a coherent inference.

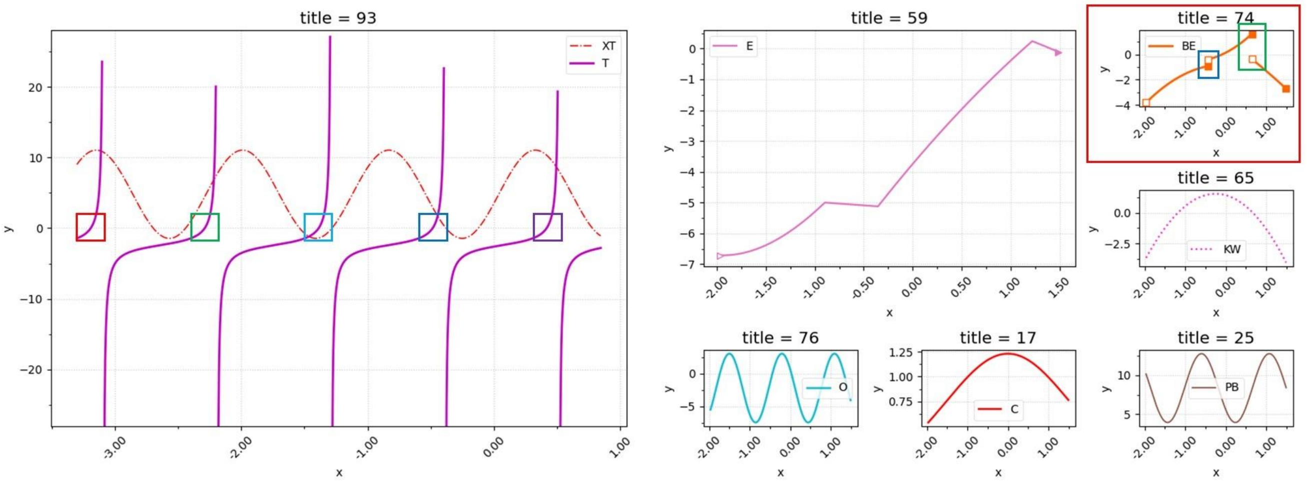
To evaluate the ability of MLLMs to reason on mosaics of mathematical imagery, we introduce a new benchmark: *MathSearch*, containing over 2200 image-question pairs. While the benchmark is highly challenging for current vision-language models, the skills it tests for are foundational capabilities in understanding diagrams, as evident by much stronger human performance on the benchmark. Our tasks on plot understanding require step-by-step aggregation of local visual information, and it is reminiscent of the aggregation of symbolic information in textual reasoning tasks, commonly referred to as “rationales”, or “chains of thought” [21].

2. Related Work

2.1. Vision-Language Benchmarks

There has been a barrage of vision benchmarks proposed for evaluating large scale multi-modal language models [4, 8, 11, 23] which evaluate the models’ ability to reason and carry out general visual tasks in the real world. However, most of these tasks only require

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.



Question: How many zeros does the function T have?

Question: List all the points of discontinuity in the top-right sub plot.

Figure 1. Our benchmark tests both local and global reasoning capabilities of multi-modal language models. **Left:** To answer the question the model needs to follow the visual chain of **1.** focusing on the legend box and identifying the function T **2.** finding $y = 0$ on the y-axis and **3.** then follow the imaginary $y = 0$ line to find the intersecting points on the corresponding curve, and finally, **4.** find the corresponding x-values. **Right:** Similarly, the model needs to **1.** find the correct subplot, **2.** follow the curve closely and recognize "breaks", **3.** find the corresponding values on the x-axis.

a coarse or global understanding of the image. Some of the works, such as [22], show that SOTA models on these benchmarks are unable to locate simple objects in the scene. At the same time, [5, 19] also show that existing models struggle with fine-grained features in the image. In this work, we present a benchmark that evaluates the models along both these axes and argue that mathematical plot understanding is well-suited for this type of evaluation. There has been some work on understanding plots [13, 15], which mostly requires OCR abilities for reading values from the figures and diagrams whereas [7, 12] require a deep mathematical understanding. In contrast, our benchmark focuses on visual understanding and only requires grade-school level mathematical understanding. [6] are the closest to our work however, they mostly focus on chart understanding, whereas we are more interested in visual understanding in MLLMs. We also show that even after being trained extensively for chart understanding tasks, ChartLlama [6] fails to perform on our benchmark.

2.2. Multi-Modal Language Models

Multi-modal language models [1, 3, 10] are usually built by combining pre-trained language models and large scale vision encoders like CLIP [17, 18] using some form of adapter layers [2, 10]. [10] have previously shown the importance of quality of data for fine-tuning these models. These models have been used to enable many applications, including chart and plot understanding [6, 9, 14]. However, despite training on very relevant data, we show that these models still fail on our benchmark which requires *extra* visual capabilities that do not emerge from just training on larger or cleaner data.

3. Benchmark

We generate plots by plotting randomly sampled functions from a set of 10 standard parameterized functions. We randomly sample the parameters of the function from a set range for each function. The function is then plotted using a standard plotting library. We categorize our plots into *single plots* (one plot and one function), *multi-function* plots (one plot and multiple functions), and *multi-plots* (multiple plots and multiple functions). Example plots are shown in Fig. 2. We also define 84 different types of question associated with the plots, ranging from simple function value lookups to counting roots of a function.

3.1. Dataset

Here we describe our dataset construction process. We first construct a set of 10 standard mathematical functions and generated variety of plots as shown in Figure 2. Our dataset is arranged across 3 different settings: *single-function, multi-function, multi-plot*. For *single-function*, we select one function at random which is then drawn on a plot. For *multi-function* plots, we

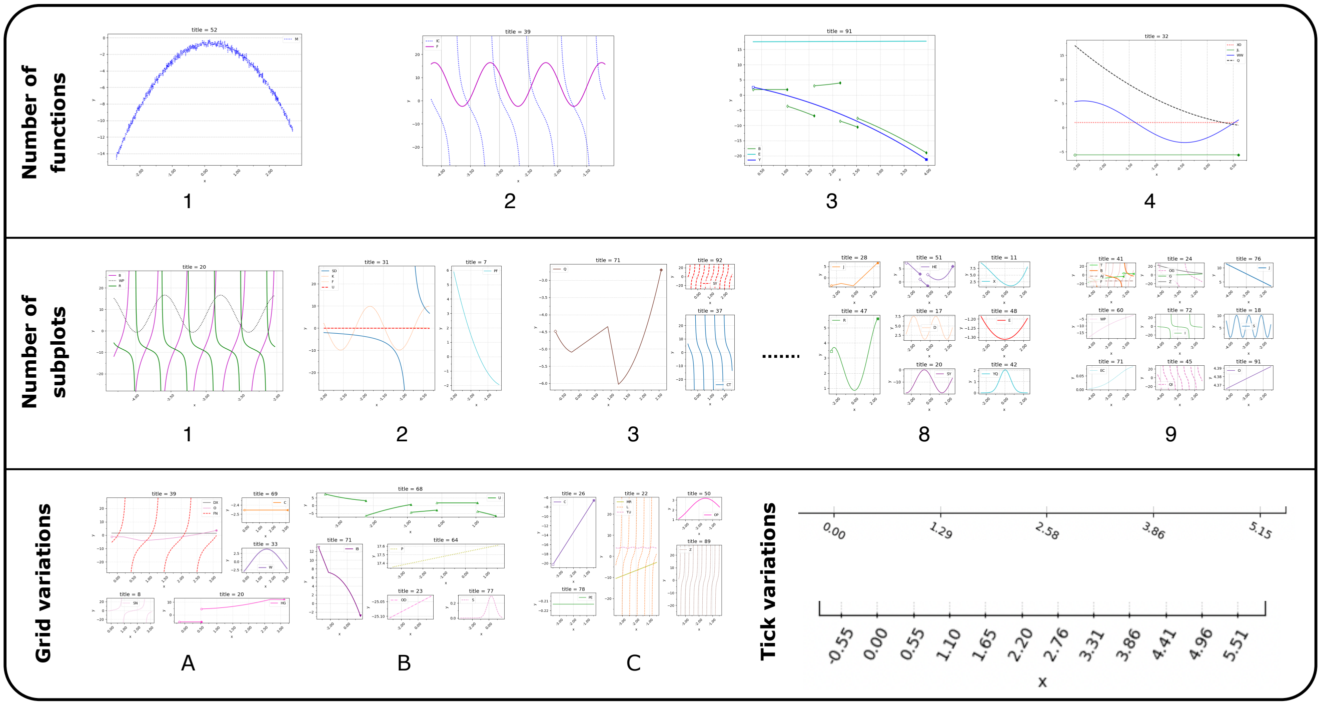


Figure 2. Diversity of plots in our data. **Top:** Each (sub)plot can contain up to 4 functions. **Middle:** The functions can be plotted in up to 9 subplots. **Bottom left:** Plots containing same number of sub-plots but diverse layouts. **Bottom right:** Variations in size, count, rotation of ticks.

Table 1. Data statistics. *q-len* denotes the length of question, *#func* represent the number of functions on a single plot, *#plots* denotes the number of subplots on a single figure and *#ex* denotes the number of examples in the split.

Stat	Sin-Fn	Multi-Fn	Multi-Pl	Avg.
Avg. q-len	47.51	51.03	70.83	56.45
Avg. #func	1.00	3.03	7.71	3.91
Avg. #plots	1.00	1.00	5.52	2.50
Train #ex	39.9k	80k	80k	67k

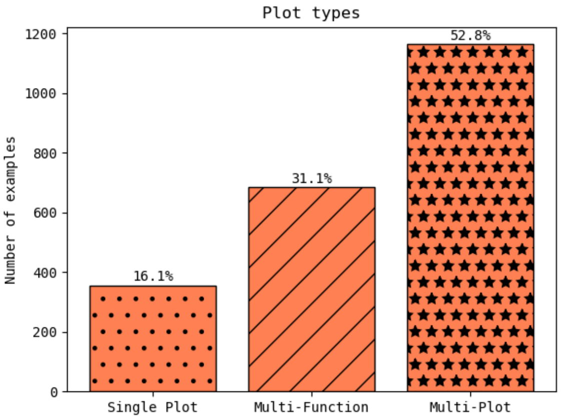


Figure 3. Breakdown of number of questions in each plot-type in the benchmark.

sample between 2-4 functions and plot them on the same plot. For the *multi-plot* setting, we first randomly sample between 2-9 subplots and for each subplot, we either treat it as a single plot or as a multi-function plot. We systematically vary all the aspects of plotting, including the layouts, line-styles, colors, margins, etc., and retain the meta-data for question construction. All plots have a resolution of 800 x 600. Some statistics of the dataset are shown in Table 1.

For question construction, we first hand-design a set of 84 different templated questions which range from simple function value lookup to more complicated referential questions like gradient comparison of functions at different points across different subplots. We also show the total number of questions in each of the splits in the benchmark in Figure 3.

In contrast to the prevailing setup of using multiple choice as answers, we keep our answers in free-form text. This constrains the language model to predict precise answer values which requires precise lookup of information in the plots. In Figure 4 we show the dif-

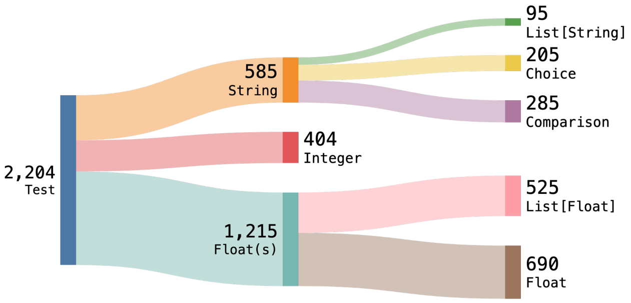


Figure 4. Breakdown of answer types in our benchmark.

ferent forms that answer values can take. As noted earlier, we design the data and questions such that it is not possible to solve them using just OCR. This allows us to truly test the visual reasoning capabilities of the model.

We generate a test-set containing $2,204$ plot-question pairs which forms our benchmark. We then also generate a training set of 200k plot-question pairs for fine-tuning purposes. The datasets will be available at https://qualcomm.com/developer/software/ai-datasets.

4. Experiments

4.1. Metrics

Our benchmark tasks MLLMs to collect information present in a small regions of the figure in the form of subplots, its axes and labels, etc,. As inferring the final answer based on imprecise information can be challenging, for comparing floats and integers, we use a relaxed evaluation criterion, by regarding the final answer as a success if it falls within 20% of the ground truth answer, similar to [15]. For answers containing strings, we consider semantically similar words, for example, “False” or “0” for ground truth: “No” to be correct.

4.2. User study

To establish a benchmark of how difficult it is for humans to perform on our dataset, we conduct a user study. We sub-sample a smaller test set of 310 questions called *test-mini* and get 10 different users to solve the questions. We report the relaxed accuracy similar to other baselines in Table 2. We can see the humans perform reasonably well by achieving an aggregate of 71.11%. We also see a clear decline in human performance with increasing difficulty of plots as it is much difficult to look up information in multi-plot and multi-function settings compared to single-plot setting.

Table 2. Baseline results. As we can see, existing MLLMs struggle to perform well on our dataset. *Sin-Fn* denotes the Single-function category of plots, *Multi-Fn* represents the Multi-function, and *Multi-Pl represents the Multi-plot category.* † denotes the baseline was evaluated on *test-mini.*

Model	Sin-Fn	Multi-Fn	Multi-Pl	Avg.
Human†	**75.45**	**74.78**	**67.49**	**71.11**
LLava [10]	12.89	8.46	9.36	9.64
Qwen-VL [3]	13.44	10.70	12.05	11.85
ChartLlama [6]	19.33	13.96	8.61	11.99
GPT-4V [1]†	**34.00**	**13.00**	**32.73**	**26.57**

4.3. Results

We evaluate GPT-4V [1][*], Qwen-VL [3], Llava [10] and ChartLlama [6] on our benchmark. We use pre-trained versions of the all these baselines. We use the prompts shown in Appendix 1.1 for each of the model. For benchmarking human and GPT-4v performance, we utilize *test-mini* containing 310 examples and for rest of the models, we use the full *test* set containing 2204 examples. As we can see from Table 2, all the models struggle on our dataset, including ChartLlama [6] which is trained on highly relevant large-scale dataset of mathematical charts. This might be due to the fact that ChartLlama [6] dataset mostly consists of questions that can solved by using OCR capabilities to convert the image into a table and looking up answers in it. Since, our data requires more complicated visual reasoning, ChartLlama [6] struggles to perform well. GPT-4V unsurprisingly performs the best among the baselines. Even though GPT-4V is the leading MLLM on a range of visual reasoning tasks [1] it only achieves 26.57% (compared to a reasonable 71.11% achieved by the humans) on our benchmark showing that complex visual reasoning required by our benchmark does not emerge from training on large amounts of internet data and captioning-like tasks. We also provide fine-tuning and scaling training data results in Appendix 1.4.

5. Conclusion

We present a new benchmark to evaluate the ability of MLLMs to find and aggregate the right set of visual cues for reasoning over mathematical plots. We show that traditional approaches of scaling training data or aggregating cleaner data are not enough for achieving strong performance on the benchmark, although the tasks are not very challenging for humans.

[*] GPT-4V was accessed via Copilot and evaluated under *Precise* mode.

References

[1] 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. 2, 4

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 4

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[5] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models, 2023. 1, 2

[6] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023. 2, 4

[7] Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023. 2

[8] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 1

[9] Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table translation. *arXiv preprint arXiv:2212.10505*, 2022. 2

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2, 4

[11] 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? *arXiv preprint arXiv:2307.06281*, 2023. 1

[12] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021. 2

[13] 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. 2

[14] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023. 2

[15] 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. 2, 4

[16] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens Continente, Larisa Markeeva, Dylan Sunil Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[18] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2

[19] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. 1, 2

[20] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences, 2024. 1

[21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 1

[22] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 17, 2023. 1, 2

[23] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi 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. *arXiv preprint arXiv:2311.16502*, 2023. 1