Supplementary Material: 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

1. Appendix

1.1. Prompt

We try to get best performance out of the selected models by guiding the generation through prompting. We notice that while models like GPT-4V [1] are great at following the instructions, ChartLLama [3] struggles to follow the instructions presented in the prompt.

<**instruction**> Directly answer after <answer> with a Python dictionary whose key is“answer” and value is a list containing elements of type string or float. Follow the examples below.

<**question**> Is the function integrable in the interval [-2, 2]?
<**answer**> {“answer”: [“Yes”]}

<**question**> What is the value of the function at x=5?
<**answer**> {“answer”: [2.1]}

<**question**> In this figure, in the subplot located in the top-right, is the function always increasing?
<**answer**> {“answer”: [“Yes”]}

<**question**> Find all points of discontinuity for the function X1.
<**answer**> {“answer”: [1.2, 3.3, 0.9]}

Table 1. Prompt for Chart-LLama [3] and LLava [4]

<**instruction**> Answer the following question and write **only** your final answer in the format: {“answer”: [<final answer>]}. If the answer is none, you can reply with {“answer”: [None]}.

Table 3. Instruction for GPT-4V (Copilot) [1]

<**instruction**> Always finish <answer> with“The answer is ___”

<**question**> Is the function integrable in the interval [-2, 2]?
<**answer**> Yes, the function is integrable in the interval [-2, 2]. The answer is yes.

<**question**> What is the value of the function at x=5?
<**answer**> From the figure, the value of the function seems to be 2.1. The answer is 2.1.

<**question**> In this figure, in the subplot located in the top-right, is the function always increasing?
<**answer**> In the subplot located in the top-right, we can see that the function is always increasing. The answer is yes.

<**question**> What is the period of the function X3325?
<**answer**> The period of the function is 2.54. The answer is 2.54.

Table 2. Prompt for Qwen-VL [2]

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

1.2. Dataset examples

We show some examples of our data in Table 4.

1.3. Metadata

Apart from the {image (Figure 1), question, answer} pairs, each plot in the training set of *MathSearch* also contains rich-metadata about each sub-plot as shown in Table 5. This includes information about the position of the subplot, title, legend, labels, ticks, etc., their positions in the form of bounding boxes and visual information such as linecolor, linewidth, markerstyle. This meta-data can be used for further expanding the dataset.

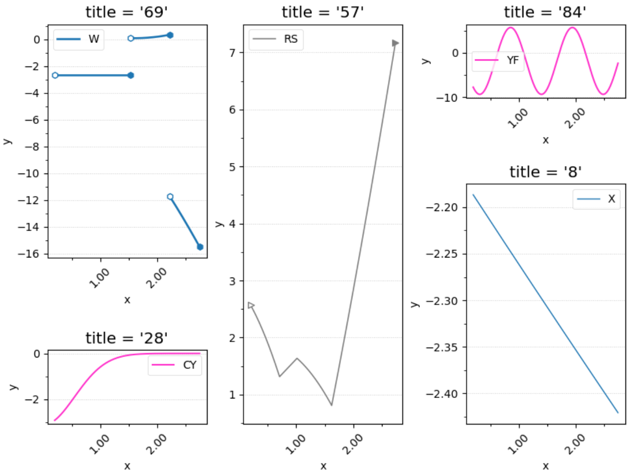


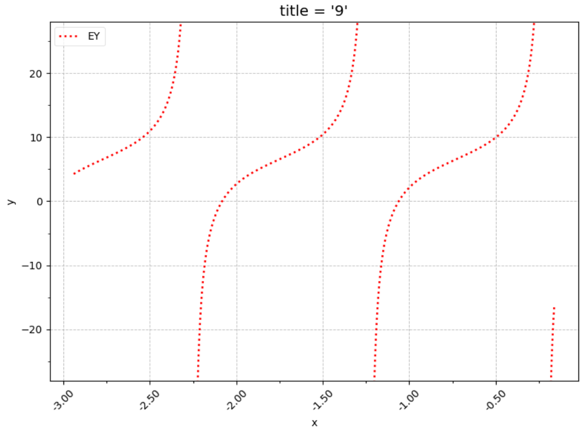
Figure 1. An example from the dataset.

Question: In the top-left subplot, find all visible points of discontinuity of the function.

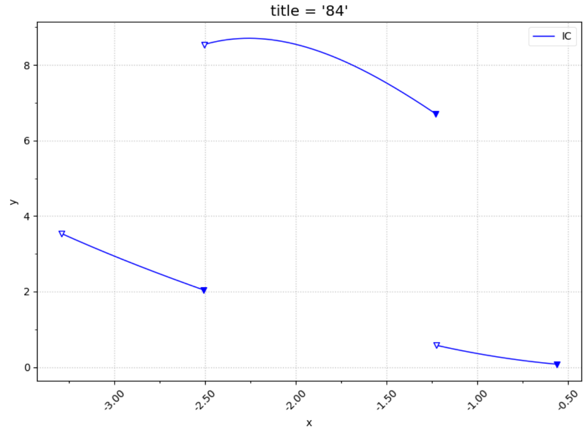
Answer: [1.53, 2.22]

title =“69”
xlabel =“x”
ylabel =“y”
line_styles = {“W”:“-”}
colors = {“W”:“#1f77b4”}
location =“top-left”
widths = {“W”: 1.86}
legends = [“W”]
xrange = [0.199, 2.742]
xticks = [0.0, 1.0, 2.0, 3.0]
yticks = [-18.0, -16.0, -14.0, -12.0, -10.0, -8.0, -6.0, -4.0, -2.0, 0.0, 2.0]
xticks_bbox = [“<bb> 0.05, 0.38, 0.09, 0.43 </bb>”, · · ·“<bb> 0.33, 0.38, 0.37, 0.43 </bb>”]
yticks_bbox = [“<bb> 0.03, 0.39, 0.07, 0.41 </bb>”, · · ·“<bb> 0.06, 0.96, 0.07, 0.99 </bb>”]
xaxis_bbox =“<bb> 0.08, 0.38, 0.33, 0.45 </bb>”
yaxis_bbox =“<bb> 0.03, 0.44, 0.08, 0.95 </bb>”
xlabel_bbox =“<bb> 0.20, 0.35, 0.21, 0.37 </bb>”
ylabel_bbox =“<bb> 0.01, 0.70, 0.02, 0.71 </bb>”
legend_bbox =“<bb> 0.09, 0.90, 0.17, 0.94 </bb>”
spine_bottom_bbox =“<bb> 0.08, 0.44, 0.33, 0.45 </bb>”
spine_left_bbox =“<bb> 0.07, 0.45, 0.08, 0.95 </bb>”
subplot_bbox =“<bb> 0.08, 0.45, 0.33, 0.95 </bb>”
curve_bboxes = {“W”:“<bb> 0.09, 0.85, 0.14, 0.84 </bb>”, · · · “<bb> 0.28, 0.47, 0.32, 0.58 </bb>”}
markers = {“W”:“h”}
title_bbox =“<bb> 0.14, 0.96, 0.27, 0.99 </bb>”

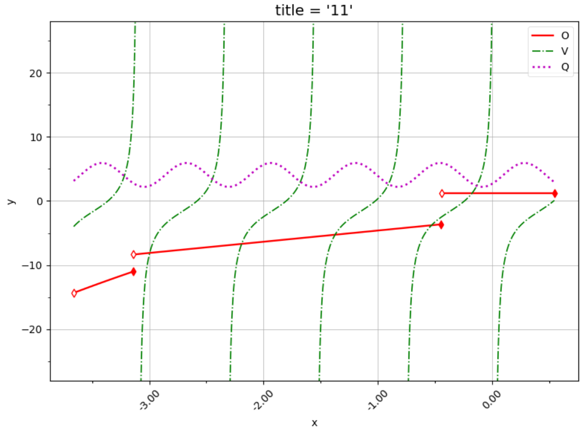
Table 5. **Top:** Question, answer corresponding to Fig. 1. **Bottom:** Rich meta-data containing plot attributes, labels, ticks, and their locations in the form of bounding boxes.



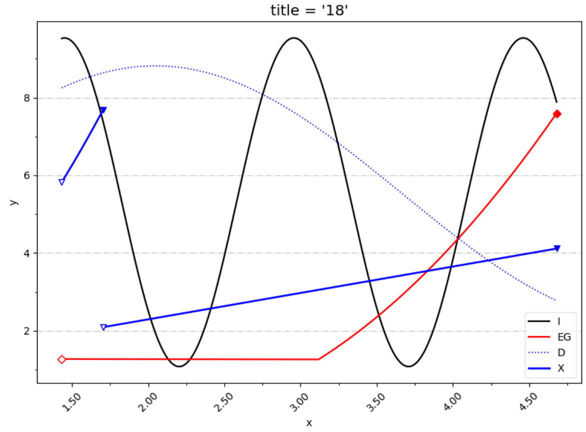
Question: Comparing the gradient of the function at -2.30 and -1.97, which one is higher?
Ground Truth: -2.30
GPT-4V: {“answer”: [-2.30]}



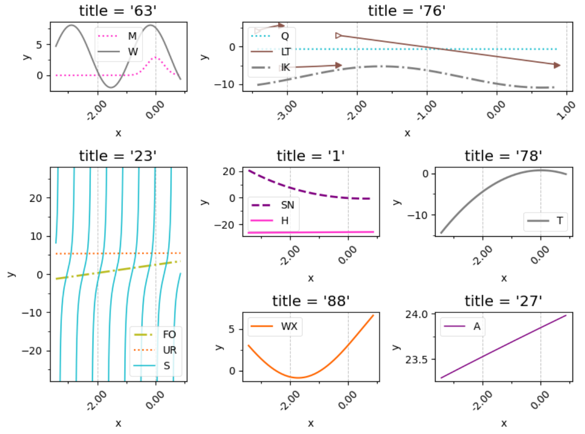
Question: Find all visible points of discontinuity of the function.
Ground Truth: [-2.5, -1.23]
GPT-4V: Refused to answer.



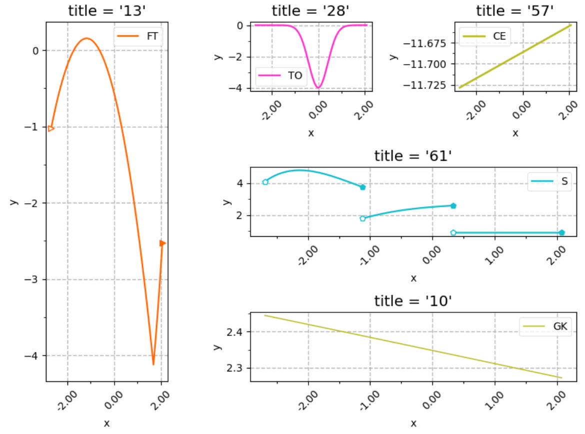
Question: At position -1.33, which function has the largest value?
Ground Truth: ”Q”
GPT-4V: {“answer”: [“O”]}



Question: What is the codomain of the function I (visible in the plot)?
Ground Truth: [1.07, 9.53]
GPT-4V: {“answer”: [0, 8]}



Question: In the 63 subplot, how many functions are increasing in the range [-2.36,-1.40]?
Ground Truth: 1
GPT-4V: Refused to answer



Question: Which subplots have at least one continuous function?
Ground Truth [“13”, “10”, “28”, “57”]
GPT-4V: [“13”, “28”, “58”]

Table 4. Examples from our dataset. **Top:** Single-function setting **Middle:** Multi-function setting **Bottom:** Multi-plot setting.

1.4. Finetuning on MathSearch

To study the impact of size of training data on our benchmark, we conduct an experiment where we scale the size of the training set from 10k examples to 300k examples and finetune LLaVa[4]. From Fig. 2 we notice that although the performance of the model improves with increasing the size of the training data, its performance on out-of-distribution subset i.e. on questions it has not seen during training improves only marginally. This showcases that the visually challenging tasks in our benchmark can't be solved with scaling the data alone and require further research.

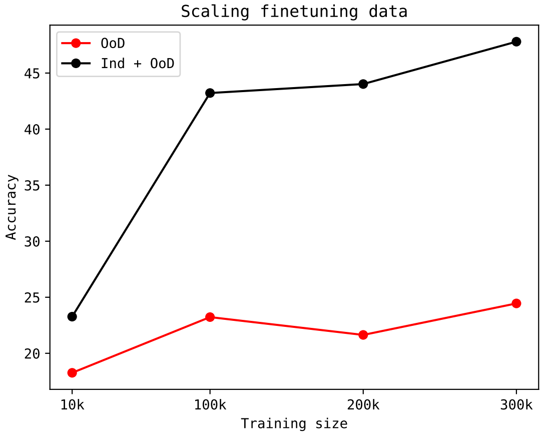


Figure 2. Finetuning LLaVa[4] on varying size of MathSearch problems.

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. 1

[2] 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. 1

[3] 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. 1

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