D Datasheets

D.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Plot2Code is a comprehensive and novel benchmark tailored for the specific multi-modal code tasks, enabling the assessment of advancements in multi-modal understanding and reasoning. We carefully collect 132 manually selected high-quality matplotlib plots across six plot types from publicly available matplotlib galleries. For each plot, we carefully offer its source code, and an descriptive instruction summarized by GPT-4. This approach enables Plot2Code to extensively evaluate MLLMs' code capabilities across various input modalities. We anticipate that Plot2Code will stimulate the research community to further explore and advance the realm of MLLMs, propelling us towards the realization of truly intelligent multi-modal systems.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

As released in 2024, the initial version of this dataset is created by The University of Hong Kong, ARC Lab from Tencent PCG, and Shanghai Jiao Tong University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ARC Lab from Tencent PCG and The University of Hong Kong funded the creation of the dataset.

D.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description. How many instances are there in total (of each type, if appropriate)?

This benchmark comprises a carefully curated dataset comprising 132 matplotlib plots across 6 plot types, incorporating a total of 293 subplots sourced from matplotlib galleries, as shown in Table 1. And each plot is paired with its corresponding code and a detailed description generated by GPT-4, as shown in Figure 2. In addition to the matlotlib, we also provide other plotting library data: 150 plots from Python's plotly and 86 plots R's plotly with the corresponding code and GPT-4 summarized descriptions.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of in- stances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains a small, representative sample of chart data from various plotting libraries, rather than being derived from any single larger dataset. To some extent, it can represent plots generated by programming languages (Python, R), as it comes from examples in these classic plotting libraries. However, it cannot represent all types of plots generated by programming languages.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Every instance contains the following components: 1. Images: Plots from from various plotting libraries examples. 2. Codes: Corresponding codes of the plots. 3. Instructions: GPT-4 summarized descriptions of the plots.

Is there a label or target associated with each instance? If so, please provide a description.

The task is for the MLLM to write the code that generates the corresponding plot based on the given instruction and plot.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

All instances contain the complete information (plot, code and instruction).

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Yes, the instances are explicitly grouped by the base type they were sampled from. The grouping is reflected in the directory structure.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

We split the dataset according to different libraries (matplotlib, Python's plotly, and R's plotly). Each chart corresponds to code and its instruction. All data is used for evaluating the MLLM, not for training. This split is designed to assess the MLLM's capabilities across different libraries. However, dataset users can freely design other splits according to their task requirements.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There are no errors, sources of noise, or redundancies in the dataset.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or re- lies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. In this study, we crawled every website link listed in the Matplotlib gallery and Plotly documentation to collect data for our analysis. Both Matplotlib and Plotly libraries are distributed under permissive open-source licenses. We have taken the following steps to ensure compliance with the respective license terms:

- **Acknowledgment of Licenses:** We acknowledge that the Matplotlib library and its gallery are distributed under the BSD 3-Clause License, and the Plotly library and its documentation are distributed under the MIT License.
- **Retention of Copyright Notices:** We have retained all copyright notices and license information from the original Matplotlib gallery content and Plotly documentation, as required by their respective licenses.
- **Usage and Distribution:** Our use of the Matplotlib gallery and Plotly documentation content is solely for academic and research purposes. We have not modified the original content from the Matplotlib gallery or Plotly documentation, and any distribution of our work will include proper attribution to the Matplotlib and Plotly projects.

By adhering to these guidelines, we ensure that our use of the Matplotlib and Plotly content is fully compliant with their respective licenses.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

The dataset does not contain data that might be considered confidential.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

The dataset does not contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

The dataset does not relate to any people.

D.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The instruction data for the plots was generated by GPT-4. The authors reviewed all the instructions to ensure their accuracy. The plots and codes were crawled from the corresponding plotting package webpages, and they are all validated by the authors.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset does not come from a larger dataset.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data collection process was automatic crawled and generated for the most part. All the programs for collecting and generating data are written by the authors.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The bulk of the data was collected and generated in March, 2024-May 2024. We crawled plots and codes from the latest plot package website.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No, there was no need for ethical review as the dataset is fully from open public code package.

D.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature ex- traction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Our data process for acquiring high-quality plot-code pairs to evaluate MLLM code generation capabilities involved three main steps: 1. Generation Filtering: Code was extracted from HTML files containing a single code block, resulting in 529 plot-code pairs. 2.Type Filtering: Only simple, static matplotlib figures were kept, excluding animations and interactive plots. 3.Manual Curation: Examples were manually selected based on criteria such as lack of external dependencies, diversity in plot characteristics, and varied difficulty levels.

D.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No.

What (other) tasks could the dataset be used for?

You can use this data for the ChartQA benchmark, as it includes plots and corresponding code. You can construct QA data based on the data in the code.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereo- typing, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

No.

Are there tasks for which the dataset should not be used? If so, please provide a description.

No.

D.6 Distribution

Will the dataset be distributed to third parties out- side of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset is available publicly for anyone interested to use.

How will the dataset will be distributed (e.g., tar- ball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset is distributed through Hugging Face, which will ensure the long term data availability, in `https://huggingface.co/datasets/TencentARC/Plot2Code`.

When will the dataset be distributed?

The dataset has been released now.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

This dataset is open-sourced under the Apache-2.0.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise re- produce, any relevant licensing terms, as well as any fees associated with these restrictions.

This dataset is open-sourced under the Apache-2.0. These evaluation code and datasets are fully open for academic research and can be used for commercial purposes with official written permission.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

D.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

Support and management will be provided by the dataset authors.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Main contact: Chengyue Wu (`hillwu@connect.hku.hk`)

Additional contact: Yixiao Ge (`yixiaoge@tencent.com`)

Is there an erratum? If so, please provide a link or other access point.

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes, the development of the dataset is planned to continue, and contributions from users are also welcomed.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

The dataset does not relate to people.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes, we plan to support versioning of the dataset so that all the versions are available to potential users. Hugging Face platform will maintain the history of version.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Everyone can contribute through the Hugging Face platform. We will carefully review the contributions to assess their value before merging them into our dataset.