LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark

Zhenfei Yin[1, 3]*, Jiong Wang[1, 4]*, Jianjian Cao[1, 4]*, Zhelun Shi[1, 2]*, Dingning Liu[1, 5], Mukai Li[1]
Xiaoshui Huang[1], Zhiyong Wang[3], Lu Sheng[2], Lei Bai[1]†, Jing Shao[1]†, Wanli Ouyang[1]
[1]Shanghai Artificial Intelligence Laboratory
[2]Beihang University [3]The University of Sydney [4]Fudan University [5]Dalian University of Technology
`{yinzhenfei,bailei,shaojing}@pjlab.org.cn`

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.

- This dataset is created to facilitate research on building large language model. For example, given an image, human can talk with large language model in natural language.

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

- This dataset is presented by Shanghai Artificial Intelligence Laboratory, Shanghai AI Lab in short. Contributors include Zhenfei Yin (Shanghai AI Lab), Jiong Wang (Fudan University), Jianjian Cao (Fudan University), Zhelun Shi (Beihang University), Dingning Li (Dalian University of Technology), Mukai Li (Shanghai AI Lab), Lu Sheng (Beihang University), Xiaoshui Huang (Shanghai AI Lab), Lei Bai (Shanghai AI Lab), Zhiyong Wang (The University of Sydney), Jing Shao(Shanghai AI Lab), Wanli Ouyang (Shanghai AI Lab).

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.

- This work was sponsored by Shanghai Artificial Intelligence Laboratory.

Any other comments?

- No.

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.

- There are 186 thousand image-text pairs and 10 thousand point cloud-text pairs for training. Each pair consists of a piece of image/point cloud and corresponding texts. For benchmarking, there are 70 thousand image/point cloud-text pairs covering 14 datasets and 12 tasks.

---

*Equal Contribution

†Corresponding Authors: Jing Shao (shaojing@pjlab.org.cn) and Lei Bai (bailei@pjlab.org.cn)

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

How many instances are there in total (of each type, if approciate)?

- For image-instruction pairs, there are 49K samples of n-round daily dialogue, 49k 1-round detailed description, 42K samples of n-round factual knowledge dialogue and 40K samples of 1-round visual task dialogue. Point cloud-instruction pairs consists of 3.3K 1-round visual task dialogue, 2K 1-round detailed description and 4.9K n-round daily dialogue instances.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances 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).

- For GPT generated data, we provide all instances obtained. For visual task dialogue data, we select samples from existing dataset and reorganize as instructions by templates we provide. No tests were run to determine representations.

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

- Each instance consists of the instructions in format of conversations and corresponding image or point cloud.

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

- The label is response of assistant in instructions.

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.

- Everything is included.

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.

- Each instance is labeled by corresponding task type.

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.

- The instances are divided into training set and test set. No dataset sharing between training and testing set for zero-shot setting in our benchmark.

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

- For instructions generated by GPT-API, instructions may not cover all information in image/point cloud accurately because the GPT-API cannot accept visual content directly.

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 relies 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 dataset consumer? 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 entirely self-contained.

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

- No. All data leveraged in dataset construction are publicly available or from public tools (e.g. GPT-API).

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

- Instructions generated by GPT-API may contain some inappropriate messages due to its bias. We briefly filtered data in order to minimize such occurrences as much as possible.

Does the dataset identify any subpopulations (e.g. by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

- No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

- Some human images are included in this dataset but they should be anonymous.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

- No.

Any other comments?

- No.

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 the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

- Data was observable as raw text, images or point cloud. Images and point cloud are selected from existing public vision dataset, while instructions are either fully generated by ChatGPT or reorganized by template from ChatGPT.

What mechanisms or procedures were used to collect data (e.g. hardware apparatuses or sensors, manual human curation, software programs, software APIS)? How were these mechanisms or procedures validated?

- Instructions and templates are obtained by ChatGPT API. Part of data are manual confirmed.

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

- For images from large-scale datasets such COCO **?** and Bamboo **?**, instances are randomly selected. For items from small datasets, all instances are adopted.

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)?

- Data collection process is fully conducted by authors.

Over what timeframes 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.

- Our dataset is collected in May 2023. No instance is associated with timeframe.

Any other comments?

- No.

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 extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

- Texts generated by GPT-API of incorrect format are removed in preprocessing.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

- Yes, source dataset and original file names are saved.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

- Yes, GPT-API for instruction and template generation is available on `https://openai.com/chatgpt`.

Any other comments?

- No.

5 Uses

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

- At the time of publication, the data has been used for multimodal large language model instruction tuning and relative performance evaluation.

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.

- There is a repository at `https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models` containing some papers used the dataset.

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

- The dataset could be used for image generation or image captioning.

Any other comments?

- No.

6 Distribution

Will the dataset be distributed to third parties outside 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 publicly available.

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

- The dataset is distributed on Huggingface (`https://huggingface.co/datasets/openlamm/LAMM_Dataset/tree/main`) and OpenDataLab (`https://opendatalab.com/LAMM/LAMM`). No DOI for the dataset.

When will the dataset be distributed?

- The dataset was released in June 2023.

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.

- The dataset is publicly available under license of Creative Commons licenses (CC-BY) if no special state for source datasets.

Any other comments?

- No.

7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

- The dataset is supporting/maintaining by Jiong Wang and Zhelun Shi.

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

- The curators, Zhenfei Yin, Lei Bai and Jing Shao, can be contacted at `{yinzhenfei, bailei, shaojing}@pjlab.org.cn`.

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 dataset consumers (e.g., mailing list, GitHub)?

- Updates of the dataset will be posted at Github page (`https://github.com/OpenLAMM/LAMM`) and project page (`https://openlamm.github.io`).

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 dataset consumers? If so, please provide a description.

- Yes, contributions to our dataset are welcomed. These contributions can be raised in huggingface community and commited after manual review.

Any other comments?

- No.