DATASHEET:
MFE-ETP

Xian Fu
College of Intelligence and Computing,
Tianjin University
fuxian1224@gmail.com

Min Zhang
College of Intelligence and Computing,
Tianjin University
min_zhang@tju.edu.cn

This document is based on *Datasheets for Datasets* by Gebru *et al.* [1]. Please see the most updated version here.

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.
The creation of this dataset is to facilitate the research of multimodal foundation models for embodied task planning, that is, to output executable action sequences, i.e. plans, given visual observations and text instructions. This dataset was intentionally created with consideration for this task, focusing on four supporting capabilities related to embodied task planning. Our website is https://mfe-etp.github.io/.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was created by Min Zhang, Xian Fu, Peilong Han, Hao Zhang, and Lei Shi at Tianjin University.

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
The creation of this dataset was not funded or supported by any grants, companies, or government agencies. It was independently developed without external financial assistance.

Any other comments?
None.

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.
These instances are robot egocentric visual observations extracted from the selected simulation environment (BEHAVIOR-100, VirtualHome), as well as manually annotated corresponding text questions and answers.

How many instances are there in total (of each type, if appropriate)?
There are 1,184 instances in total, including 330 Object Understanding instances, 340 Spatiotemporal Perception instances, 244 Task Understanding instances, and 270 Embodied Reasoning 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).
This dataset does not contain all possible samples. It is intended to test the capabilities of multimodal foundation models in embodied task planning scenarios. Due to the current lack of similar datasets, each instance was artificially constructed, which may introduce biases from human cognition. No tests were run to determine representativeness.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
For each capability dimension, there is an image folder called "images" and a JSON file called "data.json" that contains instances. Each instance contains the instance number "sample_id", which specifies the corresponding capability dimension prefix prompt with the label "task_instruction_id". This corresponds to a list of all image names called "images.path", the task name (explicitly given as "task_name" or included in "images.path"), the specific text question prompt for that instance, and the manually annotated answer "response". There is a slight difference in the dimension of embodied reasoning. The embodied reasoning dimension does not have "task_instruction_id" and "response", but there is an unmentioned "task_description"

used to describe the task goal, similar to the "context" above.

Is there a label or target associated with each instance? If so, please provide a description.
As mentioned above, this label is the answer given by humans for each specific question.

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. No data is missing.

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.
None explicitly, but the original instance includes specific problems and task names, so if necessary, some information can be extracted (such as the objects involved, instances of the same task or dimension).

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.
This benchmark focuses on testing multimodal basic models and has not been used for training, so data segmentation has not been considered yet; as described in the Section. 3.3 in the paper, results are measured by GPT-3.5 or human.

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

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 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 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 confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
No, 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.
No, there is no such data in the dataset.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
No.

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.
Skipped.

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.
Skipped.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial 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.
Skipped.

Any other comments?
None.

COLLECTION

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 data is annotated by experienced AI researchers, so it is more like an indirect inference/inference from other data. The data has been validated, and specifically, we manually checked and screened each data to ensure the quality.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
The data is manually collected, first selecting appropriate visual observation from simulation environment, and

then providing questions and corresponding answers. The introduction of experienced AI researchers ensures the effectiveness of this process.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell *et al.*[2] for approaches in this area.)
Computing resources only need to meet the requirements of the simulation environment, such as that AMD Ryzen 7 5800H CPU and NVIDIA GeForce RTX 3070 Laptop GPU is fine for VirtualHome. Due to the open-source simulation environment used, the related financial expenses are the compensation of human annotators, which meets the local wage standards For energy consumption, we resulted in approximately 20lbs of CO_2e.

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 is not a sample from a larger set.

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)?

Students participated in the data collection process. Remuneration is provided by the university and mentor respectively, meeting the standards

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.
Our dataset does not involve ethical issues and no ethical review processes have been conducted.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
Skipped.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
Skipped.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
Skipped.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)
Skipped.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
Skipped.

Any other comments?
None.

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 remainder of the questions in this section.
Instances that cannot construct a suitable QA have been discarded. The specific questions and answers are manually annotated. In the subsequent human verification, the following fixes were made: (1) some inappropriate and incorrectly labeled instances. (2) Some instances have incorrect image names and have been corrected.

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.
No, but the current dataset completely contains "raw" data.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
Yes. We use Python scripts to implement these processes. Pycharm is a choice: https://www.jetbrains.com/pycharm/download.

Any other comments?
None.

USES

Has the dataset been used for any tasks already? If so, please provide a description.
At the time of submission, only the original paper.

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.
Yes. Although there are currently no other papers using our dataset, further relevant information will be published in our website: https://mfe-etp.github.io/.

What (other) tasks could the dataset be used for?
This dataset can be used for any content related to embodied intelligence and task planning. For example, this dataset can be used for fine-tuning to enhance the corresponding capabilities of the model.

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., stereotyping, 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?
The risks and harms are minimal. Users do not need to pay special attention to anything.

Are there tasks for which the dataset should not be used? If so, please provide a description.
These data are only collected in the field of embodied task planning, so the system trained on them may or may not be extended to tasks outside of it. Therefore, caution should be exercised when selecting this dataset for other tasks without additional validation.

Any other comments?
None.

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 on the internet.

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 our Github repository: https://github.com/TJURLLAB-EAI/MFE-ETP. The dataset does not have a DOI and there is no redundant archive.

When will the dataset be distributed?
The dataset is already released.

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.
YOUR ANSWER HERE

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 reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
The crawled data copyright belongs to the authors of the reviews unless otherwise stated. There is no license, but there is a request to cite the corresponding paper if the dataset is used.

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.

Any other comments?
None.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
Min Zhang and Xian Fu are supporting/maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The curators of the dataset, Min Zhang and Xian Fu, can be contacted at Min Zhang's email and Xian Fu's email, respectively.

Is there an erratum? If so, please provide a link or other access point.
There is not an explicit erratum, but updates and known errors will be specified for higher versions in our repository.

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)?
This will be posted on our repository website.

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.
Older versions will be kept around for consistency.

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.
Others may do so and should contact the original authors about incorporating fixes/extensions.

Any other comments?
None.

REFERENCES

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, January 2020.

[2] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]*, June 2019.