MINEDOJO Datasheet

Linxi Fan[1], **Guanzhi Wang**[2*], **Yunfan Jiang**[3*], **Ajay Mandlekar**[1], **Yuncong Yang**[4], **Haoyi Zhu**[5], **Andrew Tang**[4], **De-An Huang**[1], **Yuke Zhu**[1 6†], **Anima Anandkumar**[1 2†]
[1]NVIDIA, [2]Caltech, [3]Stanford, [4]Columbia, [5]SJTU, [6]UT Austin
[*]Equal contribution [†]Equal advising
https://minedojo.org

1 Motivation

For what purpose was the dataset created? We create this internet-scale multimodal knowledge base to facilitate research towards open-ended, generally capable embodied agents.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? This knowledge base was created by Linxi Fan (Nvidia), Guanzhi Wang (Caltech), Yunfan Jiang (Stanford), Ajay Mandlekar (Nvidia), Yuncong Yang (Columbia), Haoyi Zhu (SJTU), Andrew Tang (Columbia), De-An Huang (Nvidia), Yuke Zhu (Nvidia and UT Austin), and Anima Anandkumar (Nvidia and Caltech).

Who funded the creation of the dataset? NVIDIA Corporation.

2 Author Statement

We bear all responsibilities for the licensing, distribution, and maintenance of our datasets. This document follows the Datasheet format [1] whenever applicable.

3 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? Yes, the dataset is publicly available on the internet.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? All datasets can be downloaded from `https://zenodo.org/`. Please refer to this table of URL, DOI, and licensing:

Database	DOI	License
YouTube	`10.5281/zenodo.6641142`	Creative Commons Attribution 4.0 International (CC BY 4.0)
Wiki	`10.5281/zenodo.6640448`	Creative Commons Attribution Non Commercial Share Alike 3.0 Unported
Reddit	`10.5281/zenodo.6641114`	Creative Commons Attribution 4.0 International (CC BY 4.0)

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

Submitted to the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks. Do not distribute.

4 Maintenance

Who will be supporting/hosting/maintaining the dataset? The authors will be supporting, hosting, and maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Please contact Linxi Fan (`linxif@nvidia.com`), Guanzhi Wang (`guanzhi@caltech.edu`), and Yunfan Jiang (`yunfanj@cs.stanford.edu`).

Is there an erratum? No. We will make announcements if there is any.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? Yes. New updates will be posted on `https://minedojo.org`.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? Yes, old versions will be permanently accessible on zenodo.org.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Yes, please refer to `https://minedojo.org`.

5 Composition

What do the instances that comprise the dataset represent? For YouTube videos, our data is in JSON format with video URLs and metadata. We do not provide the raw MP4 files for legal concerns. For Wiki, we provide the text, images, tables, and diagrams embedded on the web pages. For Reddit, our data is in JSON format with post IDs and metadata, similar to YouTube. Users can reconstruct the Reddit dataset by running our script after obtaining an official Reddit API license key.

How many instances are there in total (of each type, if appropriate)? There are more than 740K YouTube videos with 2.2B words of transcripts, 6,735 Wiki pages with 5.8M bounding boxes of visual elements, and more than 350K Reddit posts with 6.8M comments.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? We provide all instances in our Zenodo data repositories.

Is there a label or target associated with each instance? No.

Is any information missing from individual instances? No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? We provide metadata for each YouTube video link and Reddit post ID.

Are there recommended data splits (e.g., training, development/validation, testing)? No. The entire database is intended for pre-training.

Are there any errors, sources of noise, or redundancies in the dataset? Please refer to Sec. 4 in `supplementary.pdf`

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? We follow prior works [2] and only release the video URLs of YouTube videos due to legal concerns. Researchers need to acquire the MP4 and transcript files separately. Similarly, we only release the post IDs for the Minecraft Reddit database, but we also provide a script that can reconstruct the full Reddit dataset given a free official license key.

Does the dataset contain data that might be considered confidential? No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? We have made our best efforts to detoxify the contents via an automated procedure. Please refer to Sec. 4 in `supplementary.pdf`.

6 Collection Process

The collection procedure, preprocessing, and cleaning are explained in details in Sec. 4 of `supplementary.pdf`

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)? All data collection, curation, and filtering are done by MINEDOJO coauthors.

Over what timeframe was the data collected? The data was collected between Dec. 2021 to May 2022.

7 Uses

Has the dataset been used for any tasks already? Yes, we have used the MINEDOJO YouTube database for agent pre-training. Please refer to Sec. 5 in our main paper.

Is there a repository that links to any or all papers or systems that use the dataset? N/A.

What (other) tasks could the dataset be used for? Our knowledge base is primarily intended to facilitate research in open-ended, generally capable embodied agents. However, it can also be broadly applicable to research in video understanding, document understanding, language modeling, multimodal learning, and so on.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? Nothing we are aware of.

Are there tasks for which the dataset should not be used? We do not condone any research that intentionally generates harmful or toxic contents using our YouTube, Wiki, and Reddit data.

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12): 86–92, 2021. doi: 10.1145/3458723. URL `https://doi.org/10.1145/3458723`.

[2] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv: Arxiv-1705.06950*, 2017.