Robo-DM: Efficient Robot Big Data Management

[†]**Kaiyuan Chen**[1], **Letian Fu**[1], **Siyuan Fu**[1*],
Lawrence Yunliang Chen[1], **Huang Huang**[1], **Kush Hari**[1], **Ashwin Balakrishna**[2],
Pannag R Sanketi[2], **John Kubiatowicz**[1], **Ken Goldberg**[1,4] ¶

Abstract: Recent work suggests that very large datasets of teleoperated robot demonstrations can train transformer-based models that have the potential to generalize to new scenes, robots, and tasks. However, curating, distributing, and loading large datasets of robot trajectories, which typically consist of video, textual, and numerical modalities - including streams from multiple cameras - remains challenging. We propose Robo-DM, an efficient cloud-based data management toolkit for collecting, sharing, and learning with robot data. With Robo-DM, robot datasets are stored in a self-contained format with Extensible Binary Meta Language (EBML). Robo-DM reduces the size of robot trajectory data, transfer costs, and data load time during training. In particular, compared to the RLDS format used in OXE datasets, Robo-DM's compression saves space by up to 70x (lossy) and 3.5x (lossless). Robo-DM also accelerates data retrieval by load-balancing video decoding with memory-mapped decoding caches. Compared to LeRobot, a framework that also uses lossy video compression, Robo-DM is up to 50x faster. In fine-tuning Octo, a transformer-based robot policy with 73k episodes with RT-1 data, Robo-DM does not incur any loss at training performance. We physically evaluate a model trained by Robo-DM with lossy compression, a pick-and-place task, and In-Context Robot Transformer. Robo-DM uses 75x compression of the original dataset and does not suffer any reduction in downstream task accuracy. Code and evaluation scripts can be found on website.[6].

1 Introduction

Recent work [1, 2, 3, 4, 5, 6, 7] suggests Vision-Language-Action models [4, 1, 5] can enhance robot capabilities and generalization in handling multiple settings in diverse environments. A key ingredient for large model training is large and well-curated datasets of teleoperated robot demonstration trajectories such as the Open-X Embodiment (OXE) dataset [2]. However, the curation of robot data is still inefficient [2]; each robot demonstration consists of sequences of actions and observations, making the learning samples much larger and richer in information compared to the images or text tokens in VLMs [8, 9, 10, 11, 12] and LLMs [13, 14]. The complexity and informational content per sample of robotics data are significantly higher, which presents unique challenges for model training. At large scale, which is sometimes characterized as Big Data [15], existing data storage methods can be inefficient. We propose Robo-DM, an efficient data format with a toolkit for robot data collection, management, and training.

Each robot dataset includes a number of episodes. An episode is a sequence of actions performed by an agent from a starting state to a terminal state. An episode contains multiple sensor data streams in addition to language instructions and other metadata such as robot, task, environment, and control

[*1]Department of Electrical Engineering and Computer Science
[†2]Google Deepmind
[‡3]Department of Industrial Engineering and Operations Research
[§1,3]University of California, Berkeley, CA, USA
[¶†]For correspondence and questions: kych@berkeley.edu
[6]https://github.com/BerkeleyAutomation/fog_x

8th Conference on Robot Learning (CoRL 2024), Munich, Germany.

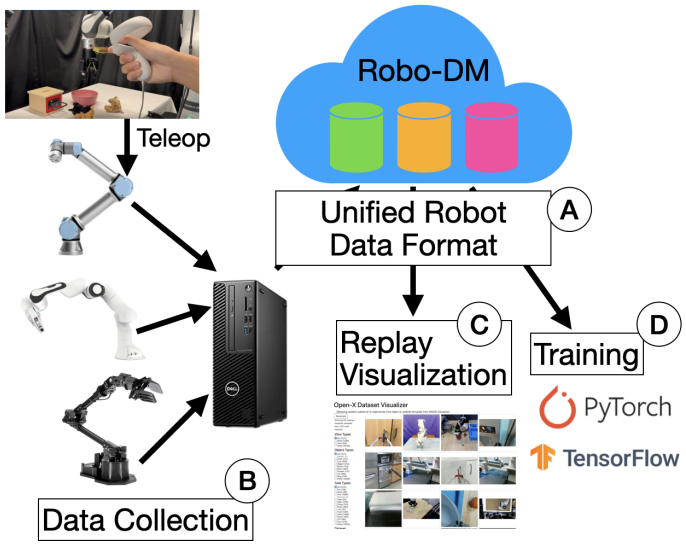


Figure 1: Robo-DM can streamline robot data collection, management, and learning. (a) Robo-DM uses a unified format for vision, language, and action that does not rely on assumptions about timestamps and data. (b) Robo-DM supports plug-and-play data collection to integrate with existing setups. (c) Robo-DM can facilitate replay and visualization. (d) Existing training frameworks can load from Robo-DM efficiently with minimal modification.

scheme specifications. The size of a typical episode ranges from 1 MB to 400 MB, depending on the episode length, compression level, number of cameras, and the camera resolution. Data streams may be recorded at different sampling rates. Episodes are typically stored as a sequence of matrices; for example, data collection with DROID [6] automates data storage with Hierarchical Data Format 5 (HDF5) [16], a format that supports hierarchical storage of matrices. OXE uses Reinforcement Learning Datasets (RLDS) [17], an extension of Tensorflow Datasets (TFDS) to store reinforcement learning demonstrations. Storing image and sensor data directly in matrices limits the capability of compression, and is thus not space efficient. One emergent framework, LeRobot [18], provides a platform to share robot models and datasets based on lossy video compression and HuggingFace datasets. However, its file structure is complex and loading is generally slower than storing matrices directly.

We observe the following challenges in robot data collection and usage:

(A) Transmission Efficiency: Distributing robotics datasets is costly. Cloud service providers, such as Google Cloud Platform (GCP) and Amazon Web Services (AWS), charge the *data host* for both data storage and outbound data transfers. Counterintuitively, the implicit cost of transferring data is more than the cost of storing it. For example, storing 8.9 TB of Open-X data on Google Cloud costs 172 US dollars per month, but *every full download* costs between 172 US Dollars and 1,540 US dollars. [7] Directly training with cloud storage requires repeated downloads if the local storage cannot store the full dataset, further increasing network traffic and cost to the *data host*. Thus, improving data compression and transmission efficiency can reduce costs and potentially encourage public sharing of datasets.

(B) Usability and Simplicity: Existing frameworks impose restrictions on file structure, data layout, semantics, and alignment. In particular, hybrid approaches rely on framework-specific assumptions to handle multiple formats simultaneously. Extending the current framework or migrating between

[7]The rate is calculated with the egress network traffic pricing in Google Cloud Platform (GCP), where the Open-X-Embodiment dataset is hosted. We use the size of Open-X v1.1 dataset with 8,964 GB in total. The rate differs by the downloading source and destination region. The rate does not consider retransmission of lost packets, so the actual cost is higher than the estimation.

frameworks can be challenging, resulting in complex structure and file organization. Figure 2 shows a comparison of LeRobot with other storage formats.

(C) Data Loading Performance: Large robot datasets are typically loaded into computationally training applications. In training, decoded frames are frequently reused and randomly accessed, and the decoded data is loaded on demand. Existing frameworks that use heavy compression sometimes lead to high computational resource utilization and interfere with the training performance. Thus, an efficient and perform ant data-loading framework should utilize available resources without contention.

We introduce Robo-DM, an efficient cloud-based toolkit for collecting, sharing, and learning with robot data. Robo-DM streamlines storage for vision, language, and action data via a unified container format with Extensible Binary Meta Language (EBML). Robo-DM efficiently orchestrates heterogeneous data streams, supporting flexible lossless compression and lossy compression for enhanced transmission efficiency. In prior work, LeRobot [18] empirically evaluates how lossy video compression parameters in FFmpeg affect robot policy accuracy. Octo is also pre-trained by compressing image frames in OXE to lossy images [1]. Robo-DM improves the data loading performance for training workloads, which requires repetitive data access by using memory-mapped caching for faster data retrieval and loading. Loading from cache and decoding are load-balanced to maximize the utilization of compute, memory and storage resources. Robo-DM requires minimal integration effort with existing frameworks. It supports plug-and-play data collection, training, replay, and visualization with mainstream frameworks, and can also be easily exported to other formats such as HDF5 and RLDS.

Experiments suggest that Robo-DM can reduce the size of data by up to 70 times with lossy compression compared to how Open-X-Embodiment currently shares the dataset, and up to 50x faster than LeRobot, a comparable framework that also uses lossy video compression to encode vision data. We fine-tune Octo, a transformer-based robot policy trained with an 800k Open-X-Embodiment dataset with 74k training episodes from RT-1. Robo-DM reduces the dataset size by 4.39 times, while being 3.0 times faster in small batch size (data loading intensive) and does not introduce any slowdown in the training pipeline with large batch size (compute intensive).

This paper makes the following contributions: (1) an extension of EBML to define a container format that unifies time-based robot data storage; (2) Robo-DM, a framework with 6 new features using this container format; (3) Experimental data that suggests Robo-DM can significantly reduce dataset size, improve loading speed, and incur marginal training performance degradation.

2 Related Work

Big Robot Data The robot learning community is actively building a number of open-source robot learning datasets [3, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 48, 49, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 41, 74, 75, 76, 77, 6, 78, 79, 80, 81, 82, 83, 84, 85, 86]. Recent work, such as Octo [1], Open-VLA [5], are trained on large datasets such as RT-1 [3], RT-2 [4], Open-X-Embodiment [2], Distributed Robot Interaction Dataset (DROID) [6]. Their initial results suggest training with large and diverse robotics datasets can enhance robot capabilities and generalization in handling multiple settings in diverse environments. In this work, we propose an efficient data pipeline for managing large and diverse robot datasets.

Robot Data Frameworks Existing frameworks for collecting, managing and storing robot data fall into the following three categories: (1) Serialized Log format that preserves timing information. This allows users to directly replay the data, e.g. with the official ROS2 tool, rosbag [87]. (2) Matrix format that can be directly integrated with training frameworks. For example, DROID [6] automates data storage with HDF5 Hierarchical Data Format (HDF5) [16] and existing OXE datasets use RLDS, an extension of Tensorflow Datasets (TFDS) [88] that store and retrieve the interaction between an agent and an environment with observation, action and reward. Storing image and sensor data directly

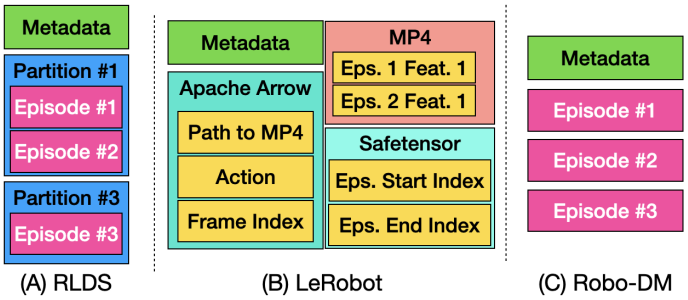


Figure 2: **A File Structure Comparison of Robo-DM with Alternative Storage Formats** All formats include metadata, storing descriptive information such as authors and dataset summary. (A) Reinforcement Learning Dataset (RLDS) stores episodes in partitions, where each partition is a Tensorflow Dataset Record file. All streams in episode data are compressed matrices that can be directly loaded and trained in Tensorflow. (B) LeRobot combines three formats for robot data. For vision data, it uses one MP4 per video stream in an episode, and uses HuggingFace Dataset (with Apache Arrow as backend[19]) to store language and action streams and the path to the MP4 files. It also uses safetensors [20] to store episode information. All the streams are scattered: to extract an episode, the framework needs to query safetensors for episode information - which is used to find the rest of the non-video streams in the HuggingFace Dataset - and finally use the frame information from the HuggingFace Dataset to find the corresponding MP4 files for vision streams. (C) In Robo-DM, robot data in all the episodes are stored and aligned in a self-contained format. To load an episode, one can simply read from Robo-DM files and load as trainable matrices.

in matrices limits the capability of compression, and is thus not space efficient. (3) Hybrid formats that store different features in separate files and require assumptions on how different features are aligned and synchronized, such as LeRobot [18], a platform to share robot models and datasets based on HuggingFace datasets.

Cloud and Fog Robotics Fog Robotics [89] utilizes cloud and edge resources for robotics applications. Existing Fog and Cloud robotics focus on deployment of robotics applications, such as grasp planning [90], motion planning [91], visual servoing [92], and human-robot interaction [93]. FogROS2 [94] automates cloud compute resources for robotics, addressing issues such as connectivity [95], latency [96], and cost [97]. We recognize the cost of the cloud required to distribute large robot datasets, and study how formats affect robotics learning in data collection, loading and management.

3 Robo-DM Features

We enumerate 6 new features to differentiate Robo-DM from existing data frameworks and alternative approaches.

(1) Self-Contained Robot Data Storage: Robo-DM uses a self-contained file format that integrates and stores heterogeneous robot data streams, ensuring all necessary data is consolidated within a single file.

(2) Vision, Language, Action Data Orchestration: The format of Robo-DM seamlessly unifies diverse binary robot data streams, including sensor data, environment specifications, language instructions, and kinematic controls.

(3) Data Flexibility: Robo-DM is extensible to different data streams, compression algorithms and video encoding formats. For example, Robo-DM enables users to flexibly choose from storing vision data as a sequence of serialized matrices, images, or encoding with lossy or lossless video codecs. With Robo-DM, one can record all the data with original timestamps without resorting to heuristics on data alignment.

(4) Efficient Dataset Size: Robo-DM efficiently encodes heterogeneous time-aligned streams. It uses video compression to significantly reduce the size of file transfer.

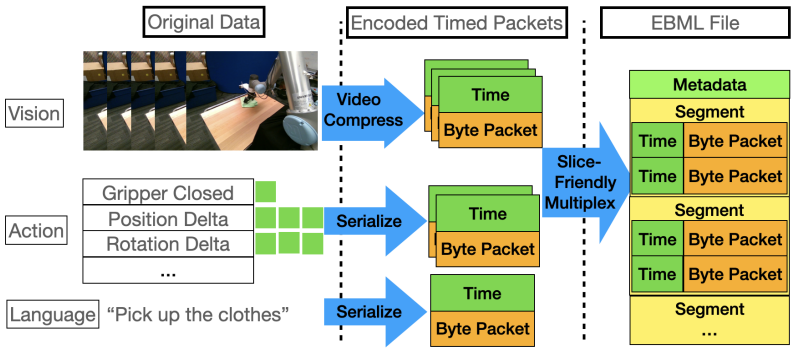


Figure 3: **How Robo-DM stores an episode with vision, language and action data** Robo-DM encodes vision, language and action data. For vision data, Robo-DM uses video or image compression; language and action data are serialized into bytes. All the bytes are encapsulated with an intake timestamp. Then Robo-DM multiplexes different streams of data into EBML file format similar to MKV video containers.

(5) Data Loading Efficiency: Robo-DM efficiently loads data by caching decoded frames and balancing resource utilization across available hardware.

(6) Simple Data Collection, Training and Visualization: Robo-DM adopts a concise interface for data collection that can fit into existing systems with minimal modification. It integrates seamlessly with TensorFlow and PyTorch interfaces, enabling easy adoption. It also allows for exporting of the collected data to existing state-of-the-art data storage frameworks, such as RLDS and HDF5. Robo-DM supports replaying messages through Robot Operating System (ROS) 2, the de-facto standard for developing robotics applications. One can use off-the-shelf ROS2 tools such as rviz [98] or Foxglove [99] to visualize the replayed streams.

4 Robo-DM Design

4.1 Unified and Self-Contained Robot Data Format

Robo-DM uses Extensible Binary Meta Language (EBML) [100] for data structuring. EBML is a versatile and extensible markup language that combines the flexibility of Extensible Meta Language (XML) with the efficiency of binary encoding. It organizes binary data elements in a hierarchical structure similar to XML, allowing for nested elements and coherent data management. This enables EBML to handle data streams from different sources within a single container, using self-describing elements that ensure compatibility and future extensibility. A notable application of EBML is in the MKV [101] video container format, which uses it to store multiple video and audio tracks, along with subtitles, in a time-aligned manner within a single container.

Figure 3 illustrates how Robo-DM encapsulates heterogeneous robot data streams. Robo-DM compresses vision streams and serializes robot data into byte packets. A byte packet encapsulates the raw bytes and descriptive information, such as timestamp and stream information. To efficiently replay the data and keep the relative timing information between data streams with different frequencies, all the data packets are stored with a relative timestamp to the beginning of the episode. Robo-DM extends MKV to store robot data to ensure the synchronization of multiple streams on vision, language, and action within the same container.

Data Collection and Post-Processing Compression can be computationally intensive. To prevent interference with the data collection process, Robo-DM uses its file format flexibility to first store all data in its serialized form. After the data collection is finished, Robo-DM iterates through the collected data, transcodes data that requires compression and re-arranges the collected data (remux) to arrange the data packets in favor of the access pattern. Because training applications sometimes access the episode at a given time frame, Robo-DM groups time-aligned data streams slices together. On querying a specific frame, metadata is used to identify the related segments and decode the video

starting from the latest keyframe before the start of the slice. All the decoded trajectories are cached to speed up future accesses.

4.2 Transmission-Efficient Storage, Retrieval and Loading

Transmission-Efficient Compression Robo-DM unifies heterogeneous data streams that require different mechanisms for compression and serialization. Because Robo-DM naturally supports byte streams, it is agnostic to mainstream byte compression algorithms and video encoders. For vision data, three channels (red, green, blue) can be compressed with off-the-shelf video compression algorithms, such as H.264 [102], H.265 [103], AV1 [104]. For large matrices that require full precision, such as stereo depth images, users can alternatively choose to compress them with lossless compression algorithms such as FFV1 [105, 106].

Efficient Decoding Cache For sequential access patterns, compression-based algorithms can reduce space usage by decoding all frames in order. In training, decoded frames are frequently reused and randomly accessed, and the decoded data is loaded on-demand. Robo-DM amortizes the random access patterns by memory-mapped files (mmap) [107, 16]. Mmap creates a new mapping in the virtual address space of a process to a cache file. If a slice of data is used, only the portions of the file that are actually used are brought into memory, conserving both I/O bandwidth and physical memory.

Load Balancing For Decoding and Decoding Cache Robo-DM automates the choice of computationally heavy decoding, loading directly cache in memory, and loading the decoded matrices from disk. To prevent overusing a single resource, Robo-DM estimates the potential latency of accessing the data and dynamically balancing the access. Specifically, if the memory resources are underutilized and a prior decoded matrix is available, this means the decoded data is likely in physical memory without being cached to the disk by mmap, and Robo-DM can directly use the decoded cache. In contrast, if the memory is full, cache miss is frequent and the data is not frequently accessed, Robo-DM does not load from cache, and directly decodes the video data instead.

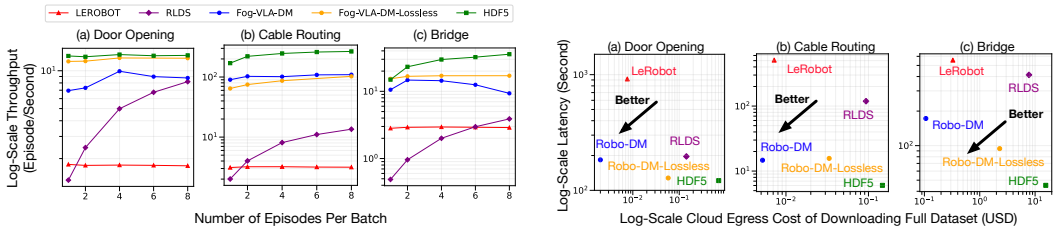
Dataset	Dataset Description			Total Dataset Size (GB)				
	# Image Streams	Resolution	Avg. Frames per Episode	Original RLDS	HDF5	Robo-DM-Lossless	LeRobot	Robo-DM
Cable Routing	3 RGB	(128, 128)	25	4.67 (18x)	7.38 (28x)	1.67 (6x)	0.36 (1.4x)	**0.26** (1x)
Door Opening	1 RGB	(720, 960)	42	7.12 (71x)	35.35 (354x)	2.89 (29x)	0.38 (4x)	**0.10** (1x)
AutoLab UR5	2 RGB, 1 Depth	(480, 640)	97	76.39 (23x)	258.33 (88x)	23.45 (7x)	(-)	**3.26** (1x)
Bridge	1 RGB	(480, 640)	34	387.49 (73x)	779.24 (147x)	114.63 (22x)	16.34 (3x)	**5.31** (1x)

Table 1: **Dataset information and Size Comparison with Different Formats in Gigabytes (GB)**. Compression ratios differ by the number of image streams and resolution. Robo-DM and LeRobot use lossy compression, while the rest are lossless. Both LeRobot and Robo-DM use AV1 codec with 30 Constant Rate Factor (CRF), a factor that balances compression and decoded video quality. These parameters are suggested by LeRobot video benchmark [108]. (-) LeRobot omits depth stream and some action streams at its conversion from RLDS [17].

5 Evaluation

Our experiments consider three questions: (1) How does Robo-DM's training data loader compare with state-of-the-art data loaders? (2) How does Robo-DM work with training workloads in terms of data loading speed, space saving, and training performance? (3) Does Robo-DM preserve the policy performance?

Setup We evaluate Robo-DM with a standard workstation setup: Intel i9-13900K Processor with 96GB RAM and NVidia 4070 Ti Super GPU. The workstation is equipped with 6TB NVMe M.2 SSD with the reading throughput up to 5000 MB/s and writing throughput up to 2500 MB/s. It connects Internet with a 1 Gbps Ethernet connection that can download from Open-X-Embodiment Google Cloud Bucket with 10 Mbps. We make sure the batch can fit in RAM without swap space. The video streams in Robo-DM are decoded with CPU without specialized GPU or additional hardware decoder.



(a) **Episode Per Second Throughput** (b) **Concurrent Loading Latency to Episode Size**

Figure 4: **Comparison of Robo-DM with baseline data loading methods in throughput and loading latency.** We compare Robo-DM with baseline data loading Methods RLDS, HDF5, and LeRobot. Complete episodes are loaded concurrently as a batch, and we record the average latency of 200 batches with a batch size of 8 episodes. We use the lowest GCP cost of 0.02 US Dollars (USD) per GB.

5.1 Data Loading Benchmarks with Open-X-Embodiment

We evaluate the data loading performance of Robo-DM with a number of exemplar datasets from Open-X-Embodiment (OXE). In the experiments, we concurrently load multiple entire episodes into memory, and we explicitly cast the data into in-memory numpy arrays. We measure the latency of issuing a number of concurrent reads (i.e. a batch) to the time that all the episodes are loaded. For each run, we measure the average latency over 200 data loads.

Datasets We use 1. *Bridge* [22]: two WidowX arms interact with household environments including kitchens, sinks, and tabletops. Skills include object rearrangement, sweeping, stacking, folding, and opening/closing doors and drawers. In the dataset,there are 4 RGB streams and 1 depth stream with 25,460 training episodes. 2. *UC Berkeley Cable Routing*: [109] one Franka robot arm routes a cable through a number of tight-fitting clips mounted on the table with 1,482 training episodes. 3. *NYU Door Opening*: [28] A Hello Stretch robot opens cabinet doors for a variety of cabinets with 435 training episodes. 4. *Berkeley AUTOLab UR5 [30]*: A UR5 robot arm pick-and-place of a stuffed animal between containers, sweeping a cloth, stacking cups with 896 training episodes.

Baselines We compare Robo-DM with the following baselines 1. *RLDS* [17] Open-X-Embodiment is stored and shared in RLDS format. In the evaluation, we directly download and load the datasets with official instructions. 2. *LeRobot* [18] We convert Open-X-Embodiment datasets in LeRobot datasets with the provided official script. Some features in Open-X-Embodiment are omitted in the conversion. We sequentially extract episodes suggested by the example instructions. 3. *HDF5* [16] We use Robo-DM to convert Open-X-Embodiment datasets to HDF5 formats. Since one HDF5 file per trajectory, we implement pre-fetch buffer and pytorch loader with the same setup as Robo-DM. We use a pre-fetch buffer of 50 episodes.

Episode Size Table 1 shows that Robo-DM significantly reduces file size (18x, 73x, 23x and 73x) per episode compared to the RLDS, a format in which these datasets are originally stored and shared. The episode size reduction leads to high accessibility to large robot datasets, transmission efficiency, and cost efficiency, shown in Figure 4b.

Loading Latency Figure 4 compares the throughput difference of Robo-DM compared against LeRobot, RLDS, and HDF5. The lossless version of Robo-DM has similar throughput as Robo-DM. It is faster than LeRobot by 33x, 20x and 5x. Robo-DM is slower than HDF5 because the HDF5 data is uncompressed and loaded in high disk throughput.

Limitation Because Robo-DM extensively uses RAM as a decoding cache to prevent repetitive decoding of the data, it leads to higher RAM usage and potentially degrades the performance when the per-episode data is large. For example, at bridge data, we see Robo-DM reduces the overall throughput when the batch size increases.

5.2 Case Study: Fine-tuning Octo with Robo-DM

Octo [1] is a transformer-based robot policy trained on 800k robot episodes from Open-X-Embodiment. We fine-tune the pre-trained Octo-small model with 25.6M trainable parameters. We fine-tune the entire model conditioned with both images and language instructions. For each configuration, we train with 50,000 iterations and measure the per-iteration average latency.

Dataset Compression We use RT-1 [3] dataset, a dataset containing 73,499 episodes. The dataset involves picking, placing, and moving 17 objects with Google Robot. The dataset contains 1 RGB video stream with resolution (320, 480). The original dataset is 111.06 GB. The final dataset size of Robo-DM is 36.50 GB with 4.39 times size reduction. The reason why the size reduction is smaller than other datasets from Open-X-Embodiment is that the per-trajectory size is small, with 1.51 MB on average per trajectory in RLDS. Robo-DM needs more space to store metadata for seeking and decoding.

Training Performance We run the training workload with batch size 64. Dataloader in Octo loads from Tensorflow dataloader and Robo-DM and lead to similar data loading latency (0.02 seconds) per iteration and overall training latency (0.10 seconds) per iteration. In validating the effect of lossy compression to the training outcome, we use lossless dataset for validation. The final image-conditioned Mean Squared Error of validation dataset is 1.86 with original lossless data and 1.91 with lossy data, leading to 2.6% increase in validation loss.

5.3 Case Study: Robo-DM with In-Context Robot Transformer Training

Task We evaluate the training performance of Robo-DM, hypothesizing the lossy compression of Robo-DM, despite a high compression rate, could reduce the accuracy of the trained model. Thus, we evaluate a model trained with 335 human-demonstrated trajectories with the lossy compression of Robo-DM. The trained model is tasked to pick up a stuffed toy tiger. Figure 5 shows the task setup with the Franka Emika robot.

Data We collect 335 human-demonstrated trajectories with one hand camera and one left-side-view camera. All video streams are recorded at resolution (320, 180). The trajectories were originally collected in HDF5 with gzip compression, with a total size of 5.8G. Stored in Robo-DM's format, the dataset with lossless codec leads to 1.7G (3.41x space reduction), and the size of lossy compression is 77MB (75.3x space reduction).

Model We use the ICRT [7], a transformer model that performs autoregressive prediction on sensorimotor trajectories. We train for 200 epochs with image brightness and contrast augmentation and a small proprioception noise ($\mathbf{N}(0, 0.01)$).

Results We randomize the position of the stuffed toy tiger at different places on the tabletop. We evaluate with consecutive 15 trials on the model trained with lossy data. The model is able to reliably identify the object, pick it up, and place it in a bowl with a 15 out of 15 success rate (100%).

6 Conclusion

In this paper we propose Robo-DM, which includes a new format for robot data, and a toolkit for data collection, management, and loading. Robo-DM significantly outperforms Open-X-Embodiment in terms of space saving. It also shows performant loading speed compared to LeRobot, a framework that also uses video compression. In the task of fine-tuning Octo and policy training, Robo-DM reduces dataset size and introduces marginal slowdown and accuracy degradation.

The file size reduction is mainly due to video compression. In future work, we will accelerate video compression and analyze the tradeoffs between parameters. In the evaluation, we used the off-the-shelf video processing library, pyav [110], without GPU acceleration. Recent works such as Decord [111] and GPU acceleration by Nvidia NVDEC [112] are demonstrated to be faster than

pyav. Also in future work, we will integrate and evaluate Robo-DM with larger-scale of existing and prospective Open-X-Embodiment datasets.

References

[1] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, D. Sadigh, C. Finn, and S. Levine. Octo: An open-source generalist robot policy. https://octo-models.github.io, 2023.

[2] O. X.-E. Collaboration, A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan, A. Raffin, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Ichter, C. Lu, C. Xu, C. Finn, C. Xu, C. Chi, C. Huang, C. Chan, C. Pan, C. Fu, C. Devin, D. Driess, D. Pathak, D. Shah, D. Büchler, D. Kalashnikov, D. Sadigh, E. Johns, F. Ceola, F. Xia, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Schiavi, H. Su, H.-S. Fang, H. Shi, H. B. Amor, H. I. Christensen, H. Furuta, H. Walke, H. Fang, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Kim, J. Schneider, J. Hsu, J. Bohg, J. Bingham, J. Wu, J. Wu, J. Luo, J. Gu, J. Tan, J. Oh, J. Malik, J. Tompson, J. Yang, J. J. Lim, J. Silvério, J. Han, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Zhang, K. Majd, K. Rana, K. Srinivasan, L. Y. Chen, L. Pinto, L. Tan, L. Ott, L. Lee, M. Tomizuka, M. Du, M. Ahn, M. Zhang, M. Ding, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, P. R. Sanketi, P. Wohlhart, P. Xu, P. Sermanet, P. Sundaresan, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Martín-Martín, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Moore, S. Bahl, S. Dass, S. Song, S. Xu, S. Haldar, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Dasari, S. Belkhale, T. Osa, T. Harada, T. Matsushima, T. Xiao, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, V. Jain, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Wang, X. Zhu, X. Li, Y. Lu, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Xu, Y. Wang, Y. Bisk, Y. Cho, Y. Lee, Y. Cui, Y. hua Wu, Y. Tang, Y. Zhu, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Xu, and Z. J. Cui. Open X-Embodiment: Robotic learning datasets and RT-X models. IEEE International Conference on Robotics and Automation, 2024.

[3] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. RT-1: Robotics transformer for real-world control at scale. *Robotics: Science and Systems (RSS)*, 2023.

[4] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[5] M. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[6] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Baijal, M. G. Castro, D. Chen, Q. Chen, T. Chung, J. Drake, E. P. Foster, J. Gao, D. A. Herrera, M. Heo, K. Hsu, J. Hu, D. Jackson, C. Le, Y. Li, K. Lin, R. Lin, Z. Ma, A. Maddukuri, S. Mirchandani, D. Morton, T. Nguyen, A. O'Neill, R. Scalise, D. Seale, V. Son, S. Tian, E. Tran, A. E. Wang, Y. Wu, A. Xie, J. Yang, P. Yin, Y. Zhang, O. Bastani, G. Berseth, J. Bohg, K. Goldberg, A. Gupta, A. Gupta, D. Jayaraman, J. J. Lim, J. Malik, R. Martín-Martín, S. Ramamoorthy, D. Sadigh, S. Song, J. Wu, M. C. Yip, Y. Zhu, T. Kollar, S. Levine, and C. Finn. Droid: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[7] L. Fu, H. Huang, G. Datta, L. Y. Chen, W. C.-H. Panitch, F. Liu, H. Li, and K. Goldberg. In-context imitation learning via next-token prediction. *arXiv preprint arXiv:2408.15980*, 2024.

[8] Gpt-4v(ision) system card. 2023. URL https://api.semanticscholar.org/CorpusID:263218031.

[9] Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[10] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[11] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[12] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, pages 8469–8488. PMLR, 2023.

[13] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[14] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[15] S. Sagiroglu and D. Sinanc. Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)*, pages 42–47. IEEE, 2013.

[16] The HDF Group. Hierarchical Data Format, version 5, 1997-2024. URL https://www.hdfgroup.org/HDF5/.

[17] L. Hussenot et al. Rlds: an ecosystem to generate, share and use datasets in reinforcement learning. *arXiv preprint arXiv:2111.02767*, 2021.

[18] R. Cadene, S. Alibert, A. Soare, Q. Gallouedec, and T. Wolf. Lerobot: Making ai for robotics more accessible with end-to-end learning. https://github.com/huggingface/lerobot, 2024.

[19] Apache Arrow. https://arrow.apache.org/. Accessed: 2024-09-14.

[20] HuggingFace SafeTensors. https://github.com/huggingface/safetensors. Accessed: 2024-09-14.

[21] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*, pages 651–673. PMLR, 2018.

[22] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[23] E. Rosete-Beas, O. Mees, G. Kalweit, J. Boedecker, and W. Burgard. Latent plans for task agnostic offline reinforcement learning. 2022.

[24] O. Mees, J. Borja-Diaz, and W. Burgard. Grounding language with visual affordances over unstructured data. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

[25] S. Dass, J. Yapeter, J. Zhang, J. Zhang, K. Pertsch, S. Nikolaidis, and J. J. Lim. Clvr jaco play dataset, 2023. URL https://github.com/clvrai/clvr_jaco_play_dataset.

[26] J. Luo, C. Xu, X. Geng, G. Feng, K. Fang, L. Tan, S. Schaal, and S. Levine. Multi-stage cable routing through hierarchical imitation learning. *arXiv pre-print*, 2023. URL https://arxiv.org/abs/2307.08927.

[27] A. Mandlekar, J. Booher, M. Spero, A. Tung, A. Gupta, Y. Zhu, A. Garg, S. Savarese, and L. Fei-Fei. Scaling robot supervision to hundreds of hours with roboturk: Robotic manipulation dataset through human reasoning and dexterity. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1048–1055. IEEE, 2019.

[28] J. Pari, N. M. Shafiullah, S. P. Arunachalam, and L. Pinto. The surprising effectiveness of representation learning for visual imitation, 2021.

[29] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. *6th Annual Conference on Robot Learning (CoRL)*, 2022.

[30] L. Y. Chen, S. Adebola, and K. Goldberg. Berkeley UR5 demonstration dataset. https://sites.google.com/view/berkeley-ur5/home.

[31] G. Zhou, V. Dean, M. K. Srirama, A. Rajeswaran, J. Pari, K. Hatch, A. Jain, T. Yu, P. Abbeel, L. Pinto, C. Finn, and A. Gupta. Train offline, test online: A real robot learning benchmark. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[32] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.

[33] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Learning agile robotic locomotion skills by imitating animals*, 2023.

[34] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 IEEE International Conference on Robotics and Automation (ICRA)*, 2019. URL https://arxiv.org/abs/1810.10191.

[35] S. Haldar, V. Mathur, D. Yarats, and L. Pinto. Watch and match: Supercharging imitation with regularized optimal transport. In *Conference on Robot Learning*, pages 32–43. PMLR, 2023.

[36] S. Belkhale, Y. Cui, and D. Sadigh. Hydra: Hybrid robot actions for imitation learning. *arxiv*, 2023.

[37] Y. Zhu, P. Stone, and Y. Zhu. Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation. *IEEE Robotics and Automation Letters*, 7(2):4126–4133, 2022.

[38] Z. J. Cui, Y. Wang, N. M. M. Shafiullah, and L. Pinto. From play to policy: Conditional behavior generation from uncurated robot data. *arXiv preprint arXiv:2210.10047*, 2022.

[39] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao, X. Yuan, P. Xie, Z. Huang, R. Chen, and H. Su. Maniskill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*, 2023.

[40] M. Heo, Y. Lee, D. Lee, and J. J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. In *Robotics: Science and Systems*, 2023.

[41] R. Mendonca, S. Bahl, and D. Pathak. Structured world models from human videos. *CoRL*, 2023.

[42] G. Yan, K. Wu, and X. Wang. ucsd kitchens Dataset. August 2023.

[43] Y. Feng, N. Hansen, and X. W. Ziyan Xiong abd Chandramouli Rajagopalan. Finetuning offline world models in the real world, 2023.

[44] S. Nasiriany, T. Gao, A. Mandlekar, and Y. Zhu. Learning and retrieval from prior data for skill-based imitation learning. In *Conference on Robot Learning (CoRL)*, 2022.

[45] H. Liu, S. Nasiriany, L. Zhang, Z. Bao, and Y. Zhu. Robot learning on the job: Human-in-the-loop autonomy and learning during deployment. In *Robotics: Science and Systems (RSS)*, 2023.

[46] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. BC-z: Zero-shot task generalization with robotic imitation learning. In *5th Annual Conference on Robot Learning*, 2021. URL https://openreview.net/forum?id=8kbp23tSGYv.

[47] G. Salhotra, I.-C. A. Liu, M. Dominguez-Kuhne, and G. S. Sukhatme. Learning deformable object manipulation from expert demonstrations. *IEEE Robotics and Automation Letters*, 7(4):8775–8782, 2022. doi:10.1109/LRA.2022.3187843.

[48] J. Oh, N. Kanazawa, and K. Kawaharazuka. X-embodiment u-tokyo pr2 datasets, 2023. URL https://github.com/ojh6404/rlds_dataset_builder.

[49] T. Matsushima, H. Furuta, Y. Iwasawa, and Y. Matsuo. Weblab xarm dataset, 2023.

[50] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn. Robonet: Large-scale multi-robot learning. In *Conference on Robot Learning*, pages 885–897. PMLR, 2020.

[51] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell. Real-world robot learning with masked visual pre-training. In *CoRL*, 2022.

[52] I. Radosavovic, B. Shi, L. Fu, K. Goldberg, T. Darrell, and J. Malik. Robot learning with sensorimotor pre-training. *arXiv:2306.10007*, 2023.

[53] M. Kim, J. Han, J. Kim, and B. Kim. Pre-and post-contact policy decomposition for non-prehensile manipulation with zero-shot sim-to-real transfer. 2023.

[54] K. Rana, B. B.-L. abd Jad Abou-Chakra, and N. S¨underhauf. Dgrasp: A large scale dataset for dynamic grasping of moving objects., 2023.

[55] A. Gupta, S. Tian, Y. Zhang, J. Wu, R. Martín-Martín, and L. Fei-Fei. Maskvit: Masked visual pre-training for video prediction. In *International Conference on Learning Representations*, 2022.

[56] T. Osa. Motion planning by learning the solution manifold in trajectory optimization. *The International Journal of Robotics Research*, 41(3):291–311, 2022.

[57] A. Padalkar, G. Quere, F. Steinmetz, A. Raffin, M. Nieuwenhuisen, J. Silvério, and F. Stulp. Guiding reinforcement learning with shared control templates. In *40th IEEE International Conference on Robotics and Automation, ICRA 2023*. IEEE, 2023.

[58] A. Padalkar, G. Quere, A. Raffin, J. Silvério, and F. Stulp. A guided reinforcement learning approach using shared control templates for learning manipulation skills in the real world. *Research square preprint rs-3289569/v1*, 2023.

[59] J. Vogel, A. Hagengruber, M. Iskandar, G. Quere, U. Leipscher, S. Bustamante, A. Dietrich, H. Hoeppner, D. Leidner, and A. Albu-Schäffer. Edan - an emg-controlled daily assistant to help people with physical disabilities. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[60] G. Quere, A. Hagengruber, M. Iskandar, S. Bustamante, D. Leidner, F. Stulp, and J. Vogel. Shared Control Templates for Assistive Robotics. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, page 7, Paris, France, 2020.

[61] Y. Zhou, S. Sonawani, M. Phielipp, S. Stepputtis, and H. Amor. Modularity through attention: Efficient training and transfer of language-conditioned policies for robot manipulation. In *Conference on Robot Learning*, pages 1684–1695. PMLR, 2023.

[62] Y. Zhou, S. Sonawani, M. Phielipp, H. Ben Amor, and S. Stepputtis. Learning modular language-conditioned robot policies through attention. *Autonomous Robots*, pages 1–21, 2023.

[63] H. Shi, H. Xu, S. Clarke, Y. Li, and J. Wu. Robocook: Long-horizon elasto-plastic object manipulation with diverse tools. *arXiv preprint arXiv:2306.14447*, 2023.

[64] G. Schiavi, P. Wulkop, G. Rizzi, L. Ott, R. Siegwart, and J. J. Chung. Learning agent-aware affordances for closed-loop interaction with articulated objects. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5916–5922. IEEE, 2023.

[65] S. Saxena, M. Sharma, and O. Kroemer. Multi-resolution sensing for real-time control with vision-language models. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=WuBv9-IGDUA.

[66] N. S. Federico Ceola, Krishan Rana. Lhmanip: A dataset for long horizon manipulation tasks., 2023.

[67] S. Guist, J. Schneider, H. Ma, V. Berenz, J. Martus, F. Gr¨uninger, M. M¨uhlebach, J. Fiene, B. Sch¨olkopf, and D. B¨uchler. A robust open-source tendon-driven robot arm for learning control of dynamic motions. *arXiv preprint arXiv:2307.02654*, 2023.

[68] Y. Wang, Z. Li, M. Zhang, K. Driggs-Campbell, J. Wu, L. Fei-Fei, and Y. Li. D3field: Dynamic 3d descriptor fields for generalizable robotic manipulation. *arXiv preprint arXiv:*, 2023.

[69] R. Shah, R. Martín-Martín, and Y. Zhu. MUTEX: Learning unified policies from multimodal task specifications. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=PwqiqaaEzJ.

[70] X. Zhu, R. Tian, C. Xu, M. Ding, W. Zhan, and M. Tomizuka. Fanuc manipulation: A dataset for learning-based manipulation with fanuc mate 200id robot. 2023.

[71] A. Sawhney, S. Lee, K. Zhang, M. Veloso, and O. Kroemer. Playing with food: Learning food item representations through interactive exploration. In *Experimental Robotics: The 17th International Symposium*, pages 309–322. Springer, 2021.

[72] L. Chen, S. Bahl, and D. Pathak. Playfusion: Skill acquisition via diffusion from language-annotated play. In *CoRL*, 2023.

[73] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023.

[74] D. Shah, B. Eysenbach, N. Rhinehart, and S. Levine. Rapid Exploration for Open-World Navigation with Latent Goal Models. In *5th Annual Conference on Robot Learning*, 2021. URL https://openreview.net/forum?id=d_SWJhyKfVw.

[75] G. Kahn, A. Villaflor, B. Ding, P. Abbeel, and S. Levine. Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5129–5136. IEEE, 2018.

[76] N. Hirose, D. Shah, A. Sridhar, and S. Levine. Sacson: Scalable autonomous data collection for social navigation. *arXiv preprint arXiv:2306.01874*, 2023.

[77] T. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *RSS*, abs/2304.13705, 2023. URL https://arxiv.org/abs/2304.13705.

[78] P. Mitrano and D. Berenson. Conq hose manipulation dataset, v1.15.0. https://sites.google.com/view/conq-hose-manipulation-dataset, 2024.

[79] N. M. M. Shafiullah, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto. On bringing robots home, 2023.

[80] J. Luo, C. Xu, F. Liu, L. Tan, Z. Lin, J. Wu, P. Abbeel, and S. Levine. Fmb: a functional manipulation benchmark for generalizable robotic learning. *arXiv preprint arXiv:2401.08553*, 2024.

[81] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023.

[82] Z. Fu, T. Z. Zhao, and C. Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *arXiv*, 2024.

[83] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4788–4795. IEEE, 2024.

[84] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 2023.

[85] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan. Vima: General robot manipulation with multimodal prompts. In *Fortieth International Conference on Machine Learning*, 2023.

[86] G. Thomas, C.-A. Cheng, R. Loynd, F. V. Frujeri, V. Vineet, M. Jalobeanu, and A. Kolobov. Plex: Making the most of the available data for robotic manipulation pretraining. In *CoRL*, 2023.

[87] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan, 2009.

[88] TensorFlow Datasets, a collection of ready-to-use datasets. https://www.tensorflow.org/datasets.

[89] S. C. Gudi et al. Fog robotics: An introduction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017.

[90] A. K. Tanwani, N. Mor, J. Kubiatowicz, J. E. Gonzalez, and K. Goldberg. A fog robotics approach to deep robot learning: Application to object recognition and grasp planning in surface decluttering. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pages 4559–4566. IEEE, 2019.

[91] J. Ichnowski, W. Lee, V. Murta, S. Paradis, R. Alterovitz, J. E. Gonzalez, I. Stoica, and K. Goldberg. Fog robotics algorithms for distributed motion planning using lambda serverless computing. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pages 4232–4238, 2020.

[92] N. Tian, A. K. Tanwani, J. Chen, M. Ma, R. Zhang, B. Huang, K. Goldberg, and S. Sojoudi. A fog robotic system for dynamic visual servoing. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1982–1988. IEEE, 2019.

[93] S. L. K. C. Gudi, S. Ojha, B. Johnston, J. Clark, and M.-A. Williams. Fog robotics for efficient, fluent and robust human-robot interaction. In *2018 IEEE 17th International Symposium on Network Computing and Applications (NCA)*, pages 1–5. IEEE, 2018.

[94] K. E. Chen, Y. Liang, N. Jha, J. Ichnowski, M. Danielczuk, J. Gonzalez, J. Kubiatowicz, and K. Goldberg. FogROS: An adaptive framework for automating fog robotics deployment. In *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, pages 2035–2042. IEEE, 2021.

[95] K. Chen, R. Hoque, K. Dharmarajan, E. Llontop, S. O. Adebola, J. Ichnowski, J. D. Kubiatowicz, and K. Goldberg. FogROS2-SGC: A ROS2 cloud robotics platform for secure global connectivity. *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8, 2023. URL https://api.semanticscholar.org/CorpusID:259287275.

[96] K. Chen, M. Wang, M. Gualtieri, N. Tian, C. Juette, L. Ren, J. Kubiatowicz, and K. Goldberg. FogROS2-LS: A location-independent fog robotics framework for latency sensitive ROS2 applications. *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024.

[97] K. Chen, K. Hari, R. Khare, C. Le, T. Chung, J. Drake, K. Dharmarajan, S. Adebola, J. Ichnowski, J. Kubiatowicz, and K. Goldberg. Fogros2-sky: Optimizing latency and cost for multi-cloud robot applications. 2024.

[98] H. R. Kam, S.-H. Lee, T. Park, and C.-H. Kim. Rviz: a toolkit for real domain data visualization. *Telecommunication Systems*, 60(2):337–345, 2015.

[99] Foxglove Technologies Inc. Foxglove. https://foxglove.dev/.

[100] S. Lhomme, D. Rice, and M. Bunkus. Extensible binary meta language. RFC 8794, RFC Editor, July 2020. URL https://www.rfc-editor.org/info/rfc8794.

[101] Matroska Video Container. https://www.matroska.org/index.html. Accessed: 2024-09-14.

[102] ITU-T. Advanced video coding for generic audiovisual services. Recommendation H.264, International Telecommunication Union, Geneva, Switzerland, 2003.

[103] ITU-T. High efficiency video coding. Recommendation H.265, International Telecommunication Union, Geneva, Switzerland, 2023. Version 9.

[104] Alliance for Open Media. Av1 bitstream & decoding process specification. https://aomediacodec.github.io/av1-spec/, 2019. Accessed: [Insert Date].

[105] Library of Congress. Ff video codec 1, version 0, 1 and 3. https://www.loc.gov/preservation/digital/formats/fdd/fdd000341.shtml, 2024. Accessed: [Insert Date].

[106] M. Niedermayer, D. Rice, and J. Martinez. Ffv1 video coding format versions 0, 1, and 3. RFC 9043, RFC Editor, August 2021. URL https://www.rfc-editor.org/info/rfc9043.

[107] Linux mmap(2) Manual. https://man7.org/linux/man-pages/man2/mmap.2.html. Accessed: 2024-09-14.

[108] LeRobot Video Benchmark. https://github.com/huggingface/lerobot/tree/main/benchmarks/video. Accessed: 2024-09-13.

[109] J. Luo, C. Xu, X. Geng, G. Feng, K. Fang, L. Tan, S. Schaal, and S. Levine. Multi-stage cable routing through hierarchical imitation learning. *IEEE Transactions on Robotics*, 2024.

[110] Pyav: Pythonic bindings for FFmpeg's libraries. https://github.com/PyAV-Org/PyAV. Accessed: 2024-09-14.

[111] Decord: An efficient video loader for deep learning with smart shuffling that's super easy to digest. https://github.com/dmlc/decord. Accessed: 2024-09-14.

[112] NVIDIA Video Codec SDK. https://developer.nvidia.com/video-codec-sdk. Accessed: 2024-09-14.

```python
import robo_dm

# Data Collection
trajectory = robo_dm.Episode("Franka-01-02-2024.vla")
trajectory.add(feature = "language_instruction",
    value = "pick up the tiger and place in the bowl")
trajectory.add(feature = "image", value = image)
trajectory.add(feature = "joint_state", value = state)

# Data Loading
trajectory = robo_dm.load(path = "Franka-01-02-2024.vla")
# [image_1, image_2...]
images = trajectory["image"]
# [joint_state_1, joint_state_2, ...]
joint_states = trajectory["joint_state"]

# Data Exporting
# Support HDF5, RLDS
dataset.export(format = "hdf5")
```

Listing 1: **Code Example of Robo-DM** Robo-DM adopts a minimalist data collection, loading and exporting interface that can be easily integrated with existing frameworks.

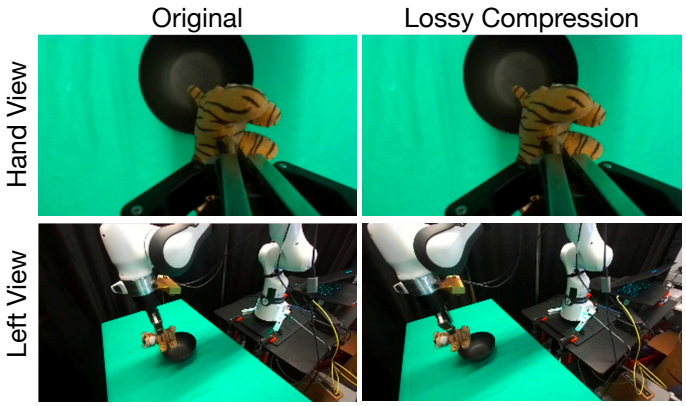


Figure 5: **ICRT Physical Experiment Setup with Robo-DM** We setup ICRT to pick up a stuffed toy tiger and place it into a black bowl with a Franka Emika robot arm. The Figure shows the view from the left camera and wrist camera used for training, for both the original dataset and reconstructed images from Robo-DM.

A Integration with existing Frameworks

Data Collection Interface In order to integrate with custom data collection software stacks, Robo-DM uses a concise programming interface for data collection. Listing 1 shows Robo-DM data collection library infers time and the data type from the input vision, action and language data. Due to the simplicity in Robo-DM's data storage format, the data collection library introduces minimal code complexity to the overall custom data collection software stack.

Plug-and-Play Data Collection and Visualization Robo-DM supports integration with ROS2-enabled setups to collect data in a plug and play manner. In ROS2, computational modules, *nodes*, can be deployed on different machines. ROS2 provides an off-the-shelf tool, *rosbag*, to capture data streams from sensors, logs, and various topics during robot operation. Robo-DM supports transcoding from and exporting robot data to rosbag, with all the timing information recorded. Rosbags also can be directly replayed in ROS2. The ROS2 community provides a number of frameworks, such as rviz [98], and Foxglove [99] from the open source community, a browser-based tool that enables visualization of ROS 2 topics. Besides replaying videos, these visualizers also support visualization in 3D, which is helpful for action data such as robot state and motions.

Data Loading Interface To support existing training frameworks with minimal modification, Robo-DM supports accessing robot data in the same way as accessing typical HDF5 files (shown in Listing

1). Robo-DM supports converting robot data to other state-of-the-art formats, such as HDF5 and Tensorflow dataset.