LHManip: A Dataset for Long-Horizon Language-Grounded Manipulation Tasks in Cluttered Tabletop Environments

Federico Ceola[1], Lorenzo Natale[1], Niko Sünderhauf[2] and Krishan Rana[2]

[1]Humanoid Sensing and Perception (HSP), Istituto Italiano di Tecnologia (IIT), Genoa, Italy.

[2]Queensland University of Technology Centre for Robotics (QCR), Brisbane, Australia.

Abstract—Instructing a robot to complete an everyday task within our homes has been a long-standing challenge for robotics. While recent progress in language-conditioned imitation learning and offline reinforcement learning has demonstrated impressive performance across a wide range of tasks, they are typically limited to short-horizon tasks – not reflective of those a home robot would be expected to complete. While existing architectures have the potential to learn these desired behaviours, the lack of the necessary long-horizon, multi-step datasets for real robotic systems poses a significant challenge. To this end, we present the *Long-Horizon Manipulation* (*LHManip*) dataset comprising 200 episodes, demonstrating 20 different manipulation tasks via real robot teleoperation. The tasks entail multiple sub-tasks, including grasping, pushing, stacking and throwing objects in highly cluttered environments. Each task is paired with a natural language instruction and multi-camera viewpoints for point-cloud or NeRF reconstruction. In total, the dataset comprises 176,278 observation-action pairs which form part of the Open X-Embodiment dataset. The full *LHManip* dataset is made publicly available here.

Fig. 1: Robot and environment setup used for data collection.

I. INTRODUCTION

Solving long-horizon manipulation tasks is crucial for addressing real-world applicability of robotic problems. Many practical tasks and activities, such as meal preparation, room cleaning, or workspace organization, involve a sequence of actions performed over an extended period. These tasks are inherently more complex than the short-term ones, as they require robots not only to manipulate objects but also to plan and execute actions across multiple steps. Long-horizon manipulation datasets have potential to allow the development of algorithms capable of generalizing across diverse scenarios, adapting to new settings, and addressing the challenges posed by tasks that require several steps to be executed.

Long-horizon manipulation tasks bring forth the importance of perceptual skills in robotic systems. While requiring to execute low-level control policies to solve the sub-tasks in which they are decomposed, these tasks also require high-level planning and reasoning capabilities. Furthermore, as robots become increasingly integrated into human-centered settings, they will be required to understand and follow natural language instructions for such tasks. Understanding natural language instructions offers the opportunity to develop robotic systems that are capable of autonomously learning, generalizing and adapting to novel settings and environments, rather than being explicitly programmed for each task [16, 22].

The wide availability of datasets for short-horizon manipulation tasks fostered the development of effective learning-based approaches for such tasks [9, 7]. In contrast, the robotics literature lacks real-world datasets for long-horizon tasks. With *LHManip*, our objective is to address this existing gap in the robotics literature, providing data to develop novel approaches to solve real-world long-horizon manipulation tasks, benchmark existing methods for such tasks evaluated only in simulation [18] on real data, and evaluate generalization properties of state-of-the-art approaches [2]. Our dataset consists of 20 tabletop manipulation tasks involving 33 everyday objects. For each of these tasks, we provide a natural language description and 10 different demonstrations collected via teleoperation. Different demonstrations of the same task either involve manipulation of different object instances (e.g. objects with different texture or size) or consider different environment conditions (e.g. different distractors on the table or different initial configurations of the objects involved in the tasks). For each demonstration, we provide visual *RGB* and *depth* observations from a wrist-mounted and two static cameras, and robot proprioceptive information. Furthermore, for each timestep of the episode, we provide the cartesian displacement of the end-effector of the robot and the position offset applied to the gripper. Fig. 1 shows a visual observation in *LHManip* from one of the two external static cameras.
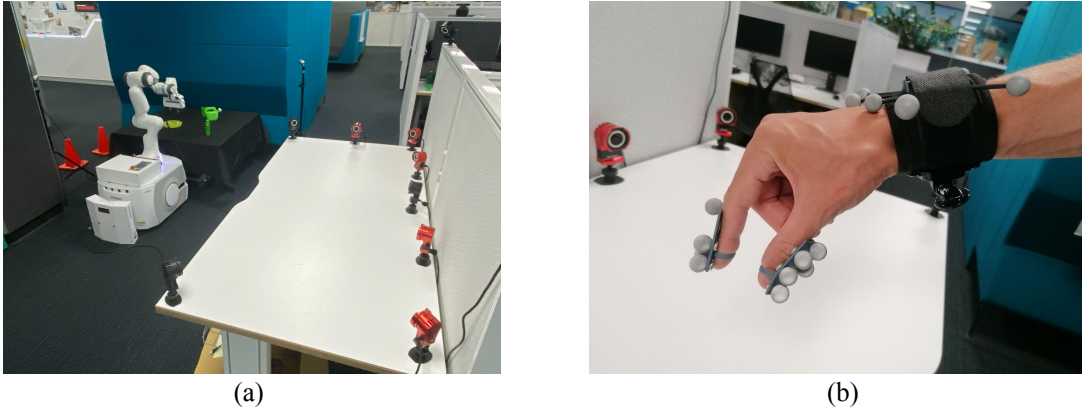
Fig. 2: (a) Motion capture and robot setup. (b) The robot was teleoperated by a human operator equipped with a motion capture system for hand gestures and movements detection in the 3D space.

This dataset forms part of the larger effort by the Open X-Embodiment collaboration project [13], and this dataset paper serves to provide full details of our contribution.

II. RELATED WORK

Long-horizon manipulation tasks are challenging robotics problems because they require both high-level reasoning capabilities to decompose the tasks in sequences of sub-tasks and low-level reasoning to solve the short-horizon subtasks. Existing approaches based on *Hierarchical Reinforcement Learning* [12, 18] have been extensively studied to solve long-horizon tasks, but their application in real robotic tasks is hampered by the need of huge amounts of real training data. Other approaches solve long-horizon tasks with methods based on *Imitation Learning* [10], combinations of *Task and Motion Planning* (*TAMP*) and *Reinforcement Learning* (*RL*) [3], or skills learning with *Large Language Models* (*LLM*) [19], and would benefit from the availability of real-world data for training.

While it has been shown in the literature the benefit of using real robotic datasets to learn short-horizon manipulation tasks such as grasping [9], pick and place [1], pouring and scooping [21], object pushing [4], or combinations of different tasks [17], datasets for long-horizon tasks are mainly provided in simulated environments. For example, the IKEA furniture assembly environment [8] provides simulated environments for assembly tasks that require long-horizon manipulation skills. CALVIN [11], instead, provides a dataset and a benchmark for language-guided long-horizon tasks with the aim of evaluating robot capabilities of learning new skills. This differs from the ALFRED [15] benchmark, that combines seven predefined skills described through language instructions for navigation and manipulation tasks. LoHoRavens [20], instead, is a simulated benchmark composed of ten long-horizon language-conditioned tasks that require at least five pick-and-place steps to be achieved.

To overcome the limitations of learning robotic tasks in simulated environments, FurnitureBench [6] proposes a benchmark for long-horizon assembly tasks, such as lamp screwing and assembling table legs. While providing a huge amount of real teleoperated data (5100 demonstrations), the considered environments are constrained to tabletop settings with 3D-printed objects with multiple markers attached on them. The dataset used to benchmark results in [14], instead, composes short-horizon language-guided tasks demonstrations into long-horizon tasks in constrained tabletop environments. With our dataset, we aim at overcoming these limitations and to provide a language-grounded dataset to perform cluttered tabletop manipulation tasks on everyday objects, relying only on a single language description for task description, multiple RGB-D observations of the environment and the proprioceptive state of the robot.

III. LHMANIP

A. Experimental Set-Up and Data Collection

We collect data via teleoperation, tracking the movements of a human operator via 10 OptiTrack *Motion Capture* (*MoCap*) cameras. Fig. 2 (a) shows the set-up. We equip the operator with three different sets of markers to track their movement in the cartesian space of the wrist, used to move the robot end-effector, and to measure the distance between the human thumb and index fingers, used to measure the robot gripper aperture. Fig. 2 (b) shows the sets of markers used to capture the motion of the operator.

We perform the tasks on a Franka Panda 7-DoF arm [5] mounted on a LD-60 Omron mobile base[1], keeping the mobile base fixed throughout data collection.

We teleoperate the end-effector of the Panda arm relying on the *servo* implementation of the *armer*[2] library. Specifically, we move the robot to mimic the human movement in the *x*, *y*, *z* position coordinates and the *yaw* orientation of the end-effector, i.e. the rotation around the axis perpendicular to the floor. Furthermore, we perform position control of the gripper joints, to mimic the fingers aperture performed by the operator.

[1]https://www.ia.omron.com/products/family/3664/dimension.html
[2]https://github.com/qcr/armer

Fig. 3: **Sub-tasks decomposition** of a *Place the bowl on the plate and the cup in the bowl matching the color* sequence.
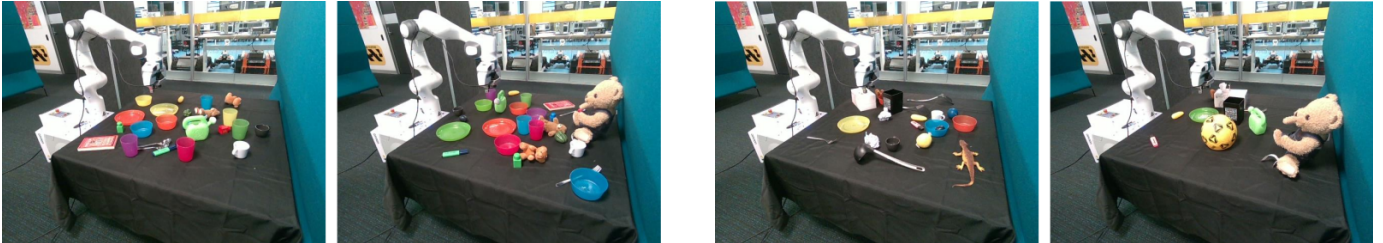

Fig. 4: **Tasks variations**: we consider different plate-bowl colors for the *Place the bowls on the appropriate plates* task (left) and different plates for the *Dry the plate* task (right).

We perform all the tasks in the dataset in the cluttered tabletop setting shown in Figs. 1, 3, and 4. We record proprioceptive information from the robot, and acquire visual and depth information from the environment. Specifically, we acquire RGB and depth information from an *Intel(R) RealSense D435* mounted on the wrist of the robot and from two external *Intel(R) RealSense D455*. We refer the reader to the *Observation and Action Space* section for a detailed description of the information provided with the dataset.

B. Dataset

The proposed dataset consists of 20 long-horizon tabletop manipulation tasks. In the following we describe the tasks and the information that we make available to the users.

1) Tasks: We report in Tab. I an overview of the 20 tasks considered in our dataset. The tasks are performed in a highly-cluttered tabletop environment and require the robot to manipulate everyday objects. We provide the list of objects used in the dataset in Tab. II. The considered tasks are composed by at least two sub-tasks that the robot must complete to successfully achieve the task. We report in Fig. 3 an exemplar sequence from the *Place the bowl on the plate and the cup in the bowl matching the color* task. As it can be noticed, to achieve the task, the robot needs to perform several sub-tasks: it firstly grasps the orange bowl and places it on the orange plate. Then it picks the orange cup and places it in the orange bowl. Furthermore, with our dataset we aim at providing data to solve tasks that require high-level reasoning capabilities. In that we specify the task at hand via a natural language instruction, without providing any other low-level environment descriptions (i.e. object details like shape, spatial information, or color). We believe that the robot must be able to infer this information autonomously from the environment. For these reasons, for each task, we provide different variations of either the initial placement of the objects

in the environment or variations of the objects involved to achieve the task. Fig. 4 shows two exemplar task variations for two different tasks, where the robot is required to solve the same task with different objects.

2) Observation and Action Space: LHManip provides a set of visual and proprioceptive information that captures the robot movement and the action that the robot is required to perform via teleoperation. In the dataset, we provide for each timestep observations and actions as reported in Tab. III. We control the robot in a non-blocking mode at 30 *Hz*.

While we control the cartesian end-effector of the robot and the aperture of the gripper, in the dataset, we also provide joint-level information such as joint positions and velocities which can be used for *keypoints* extraction and enable task training with keypoint-based methods as PerAct [16].

Please note also that, while we provide the full quaternion displacements at each timestep, we control only the rotation of the end-effector around the axis perpendicular to the floor (z quaternion value). Values different from zero in the other coordinates, therefore, aim at correct the error between the real orientation of the end-effector measured via *Forward Kinematics* and the desired null orientation around the x and y axes.

3) Dataset Access: The *LHManip* dataset is publicly available[3]. We provide data in *.png* format for *RGB* and *Depth* images, and in *.pkl* files for numerical and textual information. We provide a *Python* snippet code and the instructions to parse the dataset on *GitHub*[4]. Furthermore, *LHManip* is included as part of the larger dataset released in the *Open X-Embodiment* project [13]. We release additional depth and unprocessed sensory data together with this white paper, as well as the

---

[3]https://huggingface.co/datasets/fedeceola/LHManip
[4]https://github.com/fedeceola/LHManip

Task	Success Condition
Clean the pan.	The robot picks the sponge and performs a movement on the pan.
Cook the capsicum and place it on a plate.	The robot picks the capsicum, puts it in the pan, and then places it on a plate.
Cook the vegetables.	All the vegetables are in the pan.
Dry the plate.	The robot grasps a tissue and performs a movement on the plate.
Hide the teddy bear in the red bowl.	The teddy bear is in the red bowl, and there is an object on it.
Match the cups with the appropriate bowls.	The cups are in the bowls of the same color.
Place the bowl on the plate and the cup in the bowl matching the color.	The bowl is on the plate of the same color, and the cup is in the bowl of the same color.
Place the bowls on the appropriate plates.	The bowls are on the plates of the same color.
Prepare two cups of tea.	The two tea bags are placed in two different cups.
Put a highlighter on each book.	There is a highlighter on each book.
Put the ball in the red pot.	The ball is in the red pot.
Roll the dices in the bowl.	The dices are thrown in the bowl.
Serve the vegetables in different plates.	The vegetables in the pan are moved to different plates.
Set the table.	The fork and the spoon are to the right of the plate, and the cup is beyond the plate from the robot's point of view.
Sort the balls from left to right in order of size.	The balls are sorted in decreasing order from left to right in the robot's point of view.
Stack green blocks.	The three green blocks are stacked.
Stack the bowls.	The three bowls on the table are stacked.
Stack the cups.	The three cups on the table are stacked.
Throw away the rubbish paper.	The rubbish paper is in the trash bin.
Water the potted plant and put the can on the plate.	The robot performs a movement with the can close to the potted plant and then puts the can on the plate.

TABLE I: **Tasks**. We provide the list of tasks in the dataset, specifying the conditions that must be met to achieve them.

LHManip Items		
Ball	Block	Book
Bowl	Capsicum	Cob
Cup	Dice	Eggplant
Fennel	Fork	Highlighter
Ladle	Lizard	Marker
Mug	Pan	Plate
Plug Adapter	Plush Dog	Pot
Potted Plant	Rubbish Paper	Sippy Cup
Spatula	Sponge	Spoon
Tea Bag	Teddy Bear	Tissue Box
Trash Bin	Watering Can	Zucchini

TABLE II: **Items** considered in the *LHManip* dataset.

code to preprocess this dataset[5] back into the desired *RLDS*[6] data format as required by the *Open X-Embodiment* project.

IV. CONCLUSION

The resolution of long-horizon tasks is crucial for integrating robots performing everyday tasks in our homes. Motivated by the success of learning-based approaches from short-term manipulation datasets, we presented *LHManip*, a dataset for long-horizon robotic manipulation tasks, with the aim of addressing the current gap in the literature where such datasets are lacking.

In the perspective of developing robots that can interact with humans in everyday environments, we believe that they must possess the ability to address long-horizon tasks based on a single high-level natural instruction, relying solely on visual and proprioceptive feedback. Existing real datasets are either constrained to simplified environments or have long-horizon instructions composed of short-horizon task descriptions. *LHManip*, instead, presents several challenges, such as natural language instruction understanding, visual perception of the environment in the presence of changing challenging conditions, and learning of low-level control policies or *keypoint*-based methods for sub-tasks execution.

We hope that our work will motivate the need for more datasets, benchmarks and methodologies to learn long-horizon manipulation tasks, thereby taking a significant stride toward the integration of robots into human-centered environments.

[5]https://github.com/fedeceola/rlds_dataset_builder
[6]https://github.com/google-research/rlds

Observation	Description	Details
Cameras	Main static *RGB* camera	$640 \times 480 \times 3$
	Main static *Depth* camera	$848 \times 480 \times 1$
	Secondary static *RGB* camera	$640 \times 480 \times 3$
	Secondary static *Depth* camera	$848 \times 480 \times 1$
	Wrist-mounted *RGB* camera	$640 \times 480 \times 3$
	Wrist-mounted *Depth* camera	$848 \times 480 \times 1$
Proprioceptive	End-effector position	(x, y, z) w.r.t. root_frame
	End-effector orientation	(x, y, z, w) quaternion w.r.t. root_frame
	Robot joint angles	7 values in *rad*
	Gripper position	2 values in [0, 0.0404]
	Robot joint velocities	7 values in *rad/s*
Instruction	Natural language instruction	*String*
Action	**Description**	**Details**
Robot Action	End-effector position displacement	(x, y, z) w.r.t. root_frame
	End-effector orientation displacement	(x, y, z, w) quaternion w.r.t. root_frame
	Gripper opening displacement	1 value in [-0.0808, 0.0808]

TABLE III: **Observations and Actions** provided in *LHManip*.

REFERENCES

[1] Anthony Brohan, Noah Brown, and Justice Carbajal et al. Rt-1: Robotics transformer for real-world control at scale. In *arXiv preprint arXiv:2212.06817*, 2022.

[2] Lili Chen, Shikhar Bahl, and Deepak Pathak. Playfusion: Skill acquisition via diffusion from language-annotated play. In *Conference on Robot Learning*, pages 2012–2029. PMLR, 2023.

[3] Shuo Cheng and Danfei Xu. League: Guided skill learning and abstraction for long-horizon manipulation. *IEEE Robotics and Automation Letters*, 8(10):6451–6458, 2023.

[4] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[5] Sami Haddadin, Sven Parusel, Lars Johannsmeier, Saskia Golz, Simon Gabl, Florian Walch, Mohamadreza Sabaghian, Christoph Jähne, Lukas Hausperger, and Simon Haddadin. The franka emika robot: A reference platform for robotics research and education. *IEEE Robotics & Automation Magazine*, 29(2):46–64, 2022.

[6] Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. *arXiv preprint arXiv:2305.12821*, 2023.

[7] Dmitry Kalashnikov, Alex Irpan, and Peter Pastor et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning (CoRL)*, pages 651–673. PMLR, 2018.

[8] Youngwoon Lee, Edward S. Hu, and Joseph J. Lim. Ikea furniture assembly environment for long-horizon complex manipulation tasks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6343–6349, 2021.

[9] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5):421–436, 2018.

[10] Ajay Mandlekar, Danfei Xu, Roberto Martín-Martín, Silvio Savarese, and Li Fei-Fei. GTI: Learning to Generalize across Long-Horizon Tasks from Human Demonstrations. In *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, July 2020.

[11] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

[12] Ofir Nachum, Shixiang (Shane) Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[13] Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

[14] Erick Rosete-Beas, Oier Mees, Gabriel Kalweit, Joschka Boedecker, and Wolfram Burgard. Latent plans for task agnostic offline reinforcement learning. 2022.

[15] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[16] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[17] Homer Walke, Kevin Black, and Abraham Lee et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

[18] Xintong Yang, Ze Ji, Jing Wu, Yu-Kun Lai, Changyun Wei, Guoliang Liu, and Rossitza Setchi. Hierarchical reinforcement learning with universal policies for multi-step robotic manipulation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4727–4741, 2022.

[19] Jesse Zhang, Jiahui Zhang, Karl Pertsch, Ziyi Liu, Xiang Ren, Minsuk Chang, Shao-Hua Sun, and Joseph J Lim. Bootstrap your own skills: Learning to solve new tasks with large language model guidance. In *7th Annual Conference on Robot Learning*, 2023.

[20] Shengqiang Zhang, Philipp Wicke, Lütfi Kerem Şenel, Luis Figueredo, Abdeldjallil Naceri, Sami Haddadin, Barbara Plank, and Hinrich Schütze. Lohoravens: A long-horizon language-conditioned benchmark for robotic tabletop manipulation. *arXiv preprint arXiv:2310.12020*, 2023.

[21] Gaoyue Zhou, Victoria Dean, Mohan Kumar Srirama, Aravind Rajeswaran, Jyothish Pari, Kyle Hatch, Aryan Jain, Tianhe Yu, Pieter Abbeel, Lerrel Pinto, Chelsea Finn, and Abhinav Gupta. Train offline, test online: A real robot learning benchmark. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[22] Brianna Zitkovich, Tianhe Yu, and Sichun Xu et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 06–09 Nov 2023.