2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

RH20T: A Comprehensive Robotic Dataset for Learning Diverse Skills in One-Shot

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu and Cewu Lu
Shanghai Jiao Tong University
fhaoshu@gmail.com, {galaxies, tang_zhenyu, jirong, wcx1997, sjtuwjb3589635689, zhuhaoyi, lucewu}@sjtu.edu.cn

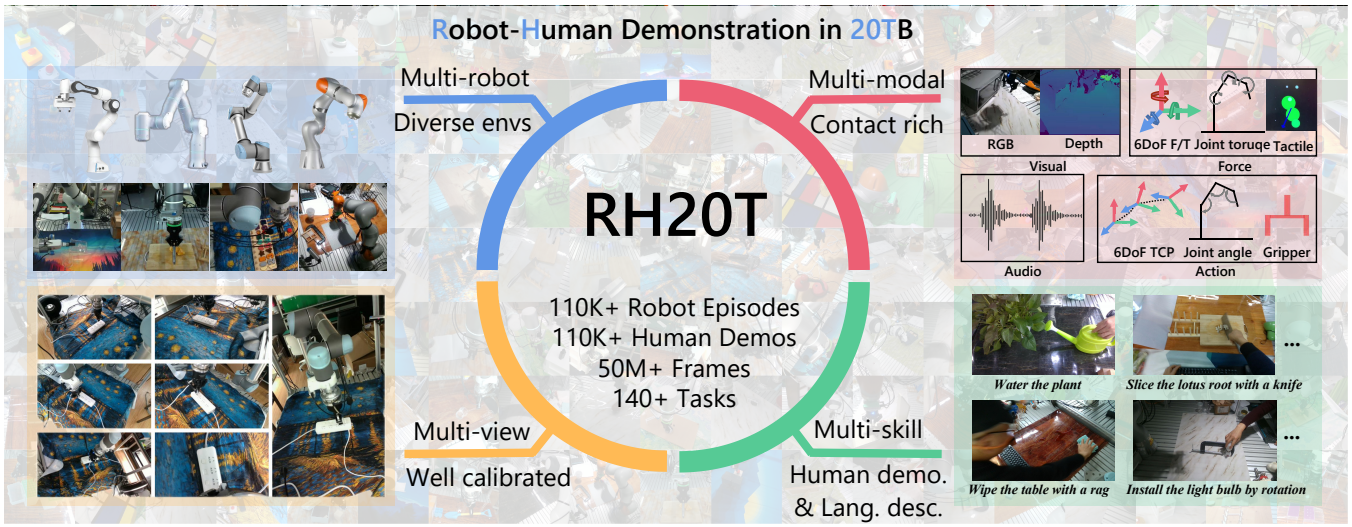

Fig. 1: Overview of our RH20T dataset. We adopt multiple robots and set up diverse environments for the data collection. The robot manipulation episodes include multi-modal visual, force, audio, and action data. For each episode, we collect the manipulation process with well-calibrated multi-view cameras. Our dataset contains diverse robotic manipulation skills and each episode has a corresponding human demonstration and language description. In total, we provide over 110K robot episodes and 110K corresponding human demonstrations. The dataset contains over 50 million frames and over 140 tasks.

***Abstract*— A key challenge for robotic manipulation in open domains is how to acquire diverse and generalizable skills for robots. Recent progress in one-shot imitation learning and robotic foundation models have shown promise in transferring trained policies to new tasks based on demonstrations. This feature is attractive for enabling robots to acquire new skills and improve their manipulative ability. However, due to limitations in the training dataset, the current focus of the community has mainly been on simple cases, such as push or pick-place tasks, relying solely on visual guidance. In reality, there are many complex skills, some of which may even require both visual and tactile perception to solve. This paper aims to unlock the potential for an agent to generalize to hundreds of real-world skills with multi-modal perception. To achieve this, we have collected a dataset comprising over 110,000 *contact-rich* robot manipulation sequences across diverse skills, contexts, robots, and camera viewpoints, all collected *in the real world*. Each sequence in the dataset includes visual, force, audio, and action information. Moreover, we also provide a corresponding human demonstration video and a language description for each robot sequence. We have invested significant efforts in calibrating all the sensors and ensuring a high-quality dataset. The dataset is made publicly available on our website: rh20t.github.io.**

I. Introduction

Robotic manipulation requires the robot to control its actuator and change the environment following a task specification. Enabling robots to learn new skills with minimal effort is one of the ultimate goals of the robot learning community. Recent research in one-shot imitation learning [10, 15] and emerging foundation models [3, 5] draw an exciting picture of transferring trained policies to a new task given a demonstration. This paper shares the same aspiration.

While the future is promising, most research in robotics only demonstrates the effectiveness of their algorithms on simple cases, such as pushing, picking, and placing objects in the real world. Two main factors hinder the exploration of more complex tasks in this direction. Firstly, there is a lack of large and diverse robotic manipulation datasets in this field [3], despite the community's long-standing eagerness for such datasets. The fundamental problem stems from the huge barriers associated with data acquisition. These challenges include the arduous task of configuring diverse robot platforms, creating varied environments, and gathering manipulation trajectories, which require significant effort and resources. Secondly, most methods focus solely on visual guidance control, yet it has been observed in physiology that humans with impaired digital sensibility struggle to accomplish many daily manipulations with visual guidance alone [22]. This indicates that more sensory information should be considered in order to learn various manipulations in open environments.

To address these problems, we revisit the data collection process for robotic manipulation. In most imitation learning literature, expert robot trajectories are manually collected using simplified user interfaces like 3D mice, keyboards,

or VR remotes. However, these control methods are inefficient and pose safety risks when the robot engages in rich-contact interactions with the environment. The main reasons are the unintuitive nature of controlling with a 3D mouse or keyboard, and the inaccuracies resulting from motion drifting when using a VR remote. Additionally, tele-operation without force feedback degrades manipulation efficiency for humans. In this paper, we equipped the robot with a force-torque sensor and employed a haptic device with force rendering for precise and efficient data collection. With the goal that the dataset should be representative, generalized, diverse and close to reality, we collect around 150 skills with complicated actions other than simple pick-place. These skills were either selected from RLBench [20] and MetaWorld [42] (but collected in real-world), or proposed by ourselves. Many skills require the robot to engage in contact-rich interactions with the environment, such as cutting, plugging, slicing, pouring, folding, rotating, etc. We have used multiple different robot arms commonly found in labs worldwide to collect our dataset. The diversity in robot configurations can also aid algorithms in generalizing to other robots.

So far, we have spent 8 months and collected around 110,000 robot episodes and 110,000 corresponding human demonstration videos for the same skills. This amounts to over 40 million frames of images for the robot episodes and over 10 million frames for the human demonstrations. Each robot sequence contains abundant visual, tactile, audio, and action information from multiple sensors. The dataset is carefully organized, and *we believe that a dataset with such diversity and scale is crucial for the future emergence of foundation models in general skill learning,* as promising progress has been witnessed in the NLP and CV communities [6, 34, 24].

II. Related Works

We briefly review related works in robotic manipulation datasets, zero/one-shot imitation learning, and vision-force learning methods.

a) Dataset: Our community has been striving to create a large-scale and representative dataset for a significant period of time. Previous research in one-shot imitation learning has either collected robot manipulation data in the real world [15] or in simulation [28]. However, their datasets are usually small and the tasks are simple. Some attempts have been made to create large-scale real robot manipulation datasets [9, 13, 16, 21, 23, 29, 36]. For example, RoboTurk [29] developed a crowd-sourcing platform and collected data on three tasks using mobile phone-based tele-operation. MIME [36] collected 20 types of manipulations using Baxter with kinesthetic teaching, but they were limited to a single robot and simple environments. RoboNet [9] gathered a significant amount of robot trajectories with various robots, grippers, and environments. However, it mainly consists of random walking episodes. BC-Z [21] presents a manipulation collection of 100 "tasks", but as pointed out in [28], they are combinations of 9 verbs and 6-15 objects. Concurrently to our work, BridgeData V2 [38] collects a dataset with 13 skills across 24 environments, RT-X [30] combines multiple existing datasets and provides a unified interface, but the data is dominated with pick-place tasks. In this paper, we present a larger dataset with a wider range of skills and environments, with more comprehensive information. More importantly, all previous datasets put less emphasis on contact-rich manipulation. Our dataset focuses more on this case and includes the crucial force modality during manipulation.

b) Zero/One-shot imitation learning: The objective of training policies that can transfer to new tasks based on robot/human demonstrations is not new. Early works [35, 31, 16] focused on imitation learning using high-level states such as trajectories. Recently, researchers [15, 10, 44, 19, 43, 33, 32, 46, 17, 37, 4, 41, 8, 27, 21, 28] have started exploring raw-pixel inputs with the advancement of deep neural networks. Additionally, the requirement of demonstrations has been reduced by eliminating the need for actions. Recent approaches have explored various one-shot task descriptors, including images [19, 4], language [37, 27, 5, 2], robot video [15, 8, 28], or human video [43, 21]. These methods can be broadly classified into three categories: model-agnostic meta-learning [15, 43, 19, 4, 46], conditional behavior cloning [10, 8, 21, 5, 28], and task graph construction [17, 18]. While significant progress has been made in this direction, these approaches only consider visual observations and primarily focus on simple robotic manipulations such as reach, pick, push, or place. Our dataset offers the opportunity to take a step further by enabling the learning of *hundreds* of skills that require *multi-modal perception* within a single imitation learning model.

c) Multi-Modal Learning of Vision and Force: Force perception plays a crucial role in manipulation tasks, providing valuable and complementary information when visual perception is occluded. The joint modeling of vision and force in robotic manipulation has recently garnered interest within the research community [12, 26, 14, 25, 1, 7, 39]. However, most of these studies overlook the asynchronous nature of different modalities and simply concatenate the signals before or after the neural network. Moreover, the existing research primarily focuses on designing multi-modal learning algorithms for specific tasks, such as grasping [7], insertion [25], twisting [12], or playing Jenga [14]. A recent attempt [40] explores jointly imitating the action and wrench on 6 tasks respectively. Overall, the question of how to effectively handle multi-modal perception at different frequencies for various skills in a coherent manner remains open in robotics. Our dataset presents an opportunity for exploring multi-sensory learning across diverse real-world skills.

III. RH20T Dataset

We introduce our robotic manipulation dataset, Robot-Human demonstration in 20TB (RH20T), to the community. Fig. 1 shows an overview of our dataset.

Dataset	# Traj.	# Skills	# Robots	Human Demo	Contact Rich	Depth Sensing	Camera Calib.	Force Sensing
MIME [36]	8.30k	12	1	✔	✗	✔	✗	✗
RoboTurk [29]	2.10k	2	1	✗	✗	✗	✗	✗
RoboNet [9]	162k	N/A	7	✗	✗	✗	✗	✗
BridgeData [11]	7.20k	4	1	✗	✗	✔*	✗	✗
BC-Z [21]	26.0k	3	1	✔	✗	✗	✗	✗
RoboSet [2]	98.5k	12	1	✗	✔	✔	✗	✗
BridgeData V2 [38]	60.1k	13	1	✗	✔	✔*	✗	✗
RH20T	110k	42	4	✔	✔	✔	✔	✔

TABLE I: Comparison with previous public datasets: "Camera Calib." indicates extrinsic calibration of all cameras and the robot. "✔*" indicates that only a portion of the images are paired with depth sensing. This comparison highlights the comprehensiveness of our dataset, which is the most extensive dataset for robotic manipulation to date.

Conf.	Robot	Gripper	6DoF F/T Sensor	Tactile
Cfg 1	Flexiv	Dahuan AG95	OptoForce	N/A
Cfg 2	Flexiv	Dahuan AG95	ATI Axia80-M20	N/A
Cfg 3	UR5	WSG50	ATI Axia80-M20	N/A
Cfg 4	UR5	Robotiq-85	ATI Axia80-M20	N/A
Cfg 5	Franka	Franka	Franka	N/A
Cfg 6	Kuka	Robotiq-85	ATI Axia80-M20	N/A
Cfg 7	Kuka	Robotiq-85	ATI Axia80-M20	uSkin

TABLE II: Hardware specification of different configurations.

Conf.	Modal	Size	Frequency
Cfg 1-7	RGB image	1280×720×3	10 Hz
	Depth image	1280×720	10 Hz
	Binocular IR image	1280×720	10 Hz
	Robot joint angle	6 / 7	10 Hz
	Robot joint torque	6 / 7	10 Hz
	Gripper Cartesion pose	6 / 7	100 Hz
	Gripper width	1	10 Hz
	6DoF F/T	6	100 Hz
	Audio	N/A	30 Hz
Cfg 7	Tactile	2×16×3	200 Hz

TABLE III: Data information of different configurations. The first 9 data modality are the same for all robot configurations. The last data modality of fingertip tactile sensing is only available in Cfg 7.

A. Properties of RH20T

RH20T is designed with the objective of enabling general robotic manipulation, which means that the robot can perform various skills based on a task description, like a video or language, while minimizing the notion of rigid tasks. The following properties are emphasized to fulfill this objective, and Tab. I provides a comparison between our dataset and previous representative publicly available datasets.

a) Diversity: The diversity of RH20T encompasses multiple aspects. To ensure task diversity, we selected 48 tasks from RLBench [20], 29 tasks from MetaWorld [42], and introduced 70 self-proposed tasks that are frequently encountered and achievable by robots. In total, it contains 147 tasks, consisting of 42 skills (*i.e.,* verbs). Hundreds of objects were collected to accomplish these tasks. To ensure applicability across different robot configurations, we used 4 popular robot arms, 4 different robotic grippers, and 3 types of force-torque sensors, resulting in 7 robot configurations. Details about the robot configurations are provided in Tab. II.

To enhance environment diversity, we frequently replaced over 50 table covers with different textures and materials, and introduced irrelevant objects to create distractions. Manipulations were performed by 19 volunteers, ensuring diverse trajectories. To increase state diversity, for each skill, volunteers were asked to change the environmental conditions and repeat the manipulation 10 times, including variations in object instances, locations, and more. Additionally, we conducted robotic manipulation experiments involving human interference, both in adversarial and cooperative settings. Further details about each task are provided in the appendix.

b) Multi-Modal: We believe that the future of robotic manipulation lies in multi-modal approaches, particularly in open environments, where data from different sensors will become increasingly accessible with advancements in technology. In the current version of RH20T, we provide visual, tactile, audio, and action information. Visual perception includes RGB, depth, and binocular IR images from three types of cameras. Tactile perception includes 6 DoF force-torque measurements at the robot's wrist, and some sequences also include fingertip tactile information. Audio data includes recordings from both in-hand and global sources. Action encompasses joint angles/torques, end-effector Cartesian pose and gripper states. We include both proprioception of robots and commands from user input for our action. All information is collected at the highest frequency supported by our workstation and saved with corresponding timestamps, and the details are given in Tab. III.

c) Scale: Our dataset consists of over 110,000 robot episodes and an equal number of human sequences, with more than 50 million images collected in total. On average, each task contains approximately 750 robot manipulations. Fig. 3 provides a detailed breakdown of the number of manipulations across different tasks in the dataset, showing a relatively uniform distribution. Fig. 4 presents statistics on the manipulation time for each sequence in our dataset. Most sequences have durations ranging from 10 to 100 seconds. With its substantial volume of data, our dataset stands as the largest in our community at present.

d) Data Hierarchy: Humans can accurately understand the semantics of a task based on visual observations, regardless of the viewpoint, background, manipulation subject, or object. We aim to provide a dataset that offers dense <human demonstration, robot manipulation> pairs, enabling models to learn this property. To achieve this, we organize the dataset in a tree hierarchy based on intra-task similarity.

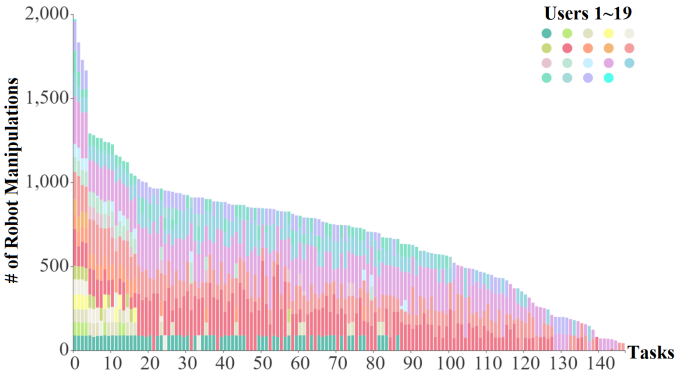

Fig. 3: Statistics on the amount of robotic manipulation for different tasks.

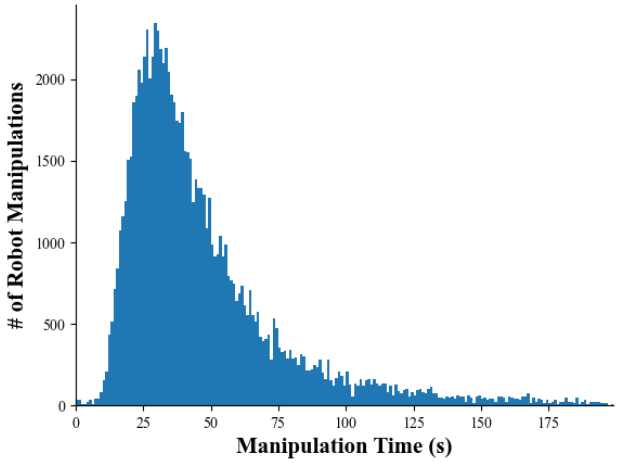

Fig. 4: Statistics on the execution time of different robot episodes in our dataset.

Fig. 5 illustrates an example tree structure and the criteria at different levels. Leaf nodes with a more recent common ancestor are more closely related. For each task, millions of <human demonstration, robot manipulation> pairs can be constructed by pairing leaf nodes with a common ancestor at different levels.

e) Compositionality: RH20T includes not only short sequences that perform single manipulations but also long manipulation sequences that combine multiple short tasks. For example, a sequence of actions such as grabbing the plug, plugging it into the socket, turning on the socket switch, and turning on the lamp can be considered as a single task, with each step also being a task. This task composition allows us to investigate whether mastering short sequences improves the acquisition of long sequence tasks.

B. Data Collection and Processing

Unlike previous methods that simplify the tele-operation interface using 3D mice, VR remotes, or mobile phones, we place emphasis on the importance of intuitive and accurate tele-operation in collecting contact-rich robot manipulation data. Without proper tele-operation, the robot could easily collide with the environment and generate significant forces, triggering emergency stops. Consequently, previous works either avoid contact [21] or operate at reduced speeds to mitigate these risks.

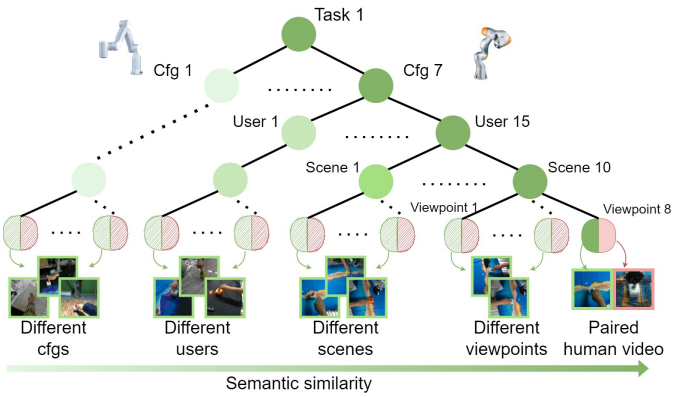

Fig. 5: Example of data hierarchy: The leaf nodes in the hierarchy consist of human demonstrations (highlighted in green) and robot manipulations (highlighted in red, only the right-est example is shown in the figure). We can pair a robot manipulation sequence with human demonstration videos captured from different viewpoints, scenes, human subjects, and environments. Zoom in to explore the details of various human demonstrations.

a) Collection: Fig. 6 shows an example of our data collection platform. Each platform contains a robot arm with force-torque sensor, gripper and 1-2 inhand cameras, 8-10 global cameras, 2 microphones, a haptic device, a pedal and a data collection workstation. For tele-operation, our haptic device can control the robot in real-time with a precision at 1 mm and render the end-link force-torque from the robot. All the cameras are extrinsically calibrated before conducting the manipulation. The human demonstration video is collected on the same platform by human with an extra ego-centric camera. 19 volunteers conducted the robotic manipulation according to our task lists and text description. The time for teaching the operator how to perform the tasks before data collection is less than 1 hour. The volunteers are also required to specify ending time of the task and give a rating from 0 to 9 after finishing each manipulation. 0 denotes the robot enters the emergency state (e.g., hard collision), 1 denotes the task fails and 2-9 denotes their evaluation of the manipulation quality. The success and failure cases have a ratio of around 10:1 in our dataset.

b) Processing: We preprocess the dataset to provide a coherent data interface. The coordinate frame of all robots and force-torque sensors are aligned. Different force-torque sensors are tared carefully. The end-effector Cartesian pose and the force-torque data are transformed into the coordination system of each camera. Manual validation is performed for each scene to ensure the camera calibration quality. Fig. 7 shows an illustration of rendering different component of the data in a unified coordinate frame and demonstrates the high-quality of our dataset. The detailed data format and data access APIs are provided on our website.

IV. Experiments

We introduce the RH20T dataset in pursuit of enabling robots to acquire novel skills within unfamiliar environments using minimal data. While the ultimate objective is to train a large model capable of performing such tasks in a one-shot learning fashion, we acknowledge the significant com-

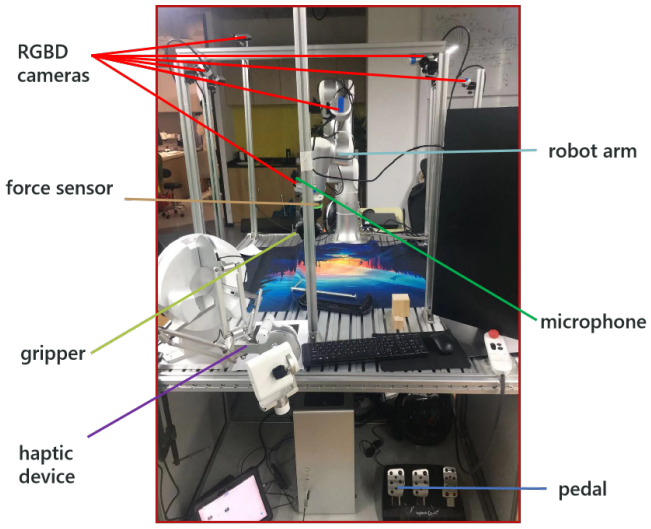

Fig. 6: Illustration of our data collection platform

Fig. 7: We display the point cloud generated by fusing the RGBD data from the multi-view cameras mounted in our data collection platform. The red pyramids indicate the camera poses. Additionally, the robot model is rendered in the scene based on the joint angles recorded in our dataset. It is evident that all the cameras are calibrated with respect to the robot's base frame, and all the recorded data are synchronized in the temporal domain.

putational resources required for this endeavor, which are presently beyond our reach. Consequently, this paper primarily focuses on demonstrating the dataset's effectiveness in enhancing the transferability of a baseline model within a few-shot learning framework.

To assess the efficacy of our dataset, we adopt the Action Chunking with Transformers (ACT) model as our baseline network. ACT, as proposed by in a recent work [45], has demonstrated remarkable capabilities in handling complex robot manipulation tasks. It leverages the power of transformers to learn intricate action sequences from hundreds of demonstrations.

A. Experimental Setup

a) Platform: In our experiments, we utilize a Flexiv robot arm equipped with an Intel RealSense RGB-D camera in front of the robot for perceiving the environment and a Dahuan-95 gripper for interacting with objects. We set up a new environment where the camera pose and table cover are different from those in our RH20T dataset. Fig. 8 (a) illustrates our robot platform.

b) Procedure: We setup a task involves grasping a block and placing it on a weight. In the new environment, we collect 75 robot episodes, including RGB images and actions, through teleoperation. From our dataset, we select 335 robotic manipulation from the same task and 195 manipulation from 3 different but similar tasks (pick up a block; pick up a block and place it at the designated location; pick up a block and move it from left to right). All the manipulation sequences from our dataset have different camera views, table covers, objects and robot embodiments from the robotic environment in our current experiment.

We initiate the training process by pretraining the ACT model on different subsets of the data selected from our dataset. By exposing the model to a range of robotic manipulation scenarios, we aim to enhance its ability to generalize across various tasks and environmental conditions. Following pretraining, we finetune the ACT model on specific portions of the newly collected data, focusing on the task involving grasping and weight placement. This stage aims to refine the model's performance on the target task.

We evaluate the performance of the ACT model both with and without pretraining on our dataset. The experiments are carried out on the real robot platform and repeated for 20 times for each configuration. We divide the task into 3 stages, namely whether the robot can reach the block, grasp it, and place it on the weight, and measure the success rate at each stage. Additionally, we examine how well the model generalizes to variations in object properties, as well as how the task and embodiment settings in pretraining affect the performances. The evaluation time limit is set as 60 seconds.

c) Implementation Details: For ACT model, we set the hidden channel and the feedforward channel in the network to 512 and 3200 respectively. The model is trained with a learning rate of 2×10^{-5} for 10 epochs in pretraining, and 10^{-5} for 750 epochs during fine-tuning. Although fewer than the original implementation [45], we boost sample density per epoch by incorporating all valid sub-trajectories from demonstrations. Hence, 750 epochs ensure the model converges effectively. The chunk size is set to 20, corresponding to 2 seconds with a frequency of 10Hz. The images are scaled to 640×360 during training and testing. We apply temporal ensembling and set its coefficient $k = 0.01$ following [45].

B. Experimental Results

We present the model's success rates under different training configurations in Tab. IV. When training the network with 75 demonstrations, we observe that pretraining the model with selected data from our dataset, despite differences in camera viewpoints, robot embodiments, and backgrounds, enhances the final success rate. Additionally, the inclusion

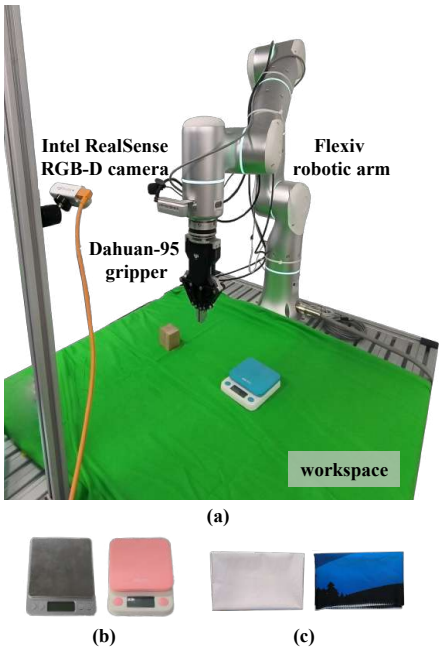

Fig. 8: (a) The experimental robot platform. (b) Varied weights (metal, pink) assessing the model's generalization ability. (c) Distinct table covers (white, blue) evaluating the model's generalization ability.

# Demos	Pretrain Task		Training Epochs	Success Rate (%) ↑		
	Same	Multi.		Reach	Pick	Place
75			500	35	10	0
	✓		500	70	15	15
	✓	✓	500	65	20	15
			750	55	5	0
	✓		750	**80**	20	15
	✓	✓	750	**80**	**25**	**25**
40			750	45	10	0
	✓		750	65	**25**	5
	✓	✓	750	70	**25**	**15**
10			750	15	0	0
	✓		750	30	**15**	**5**
	✓	✓	750	**50**	10	**5**

TABLE IV: Experimental results of ACT trained in different settings and tested in the original environment (20 trials).

	Type	w/wo pretrain	Success Rate (%) ↑		
			Reach	Pick	Place
weight	metal		20	0	0
		✓	**80**	**30**	**10**
	pink		40	10	0
		✓	**70**	**20**	**10**
table cover	white		20	0	0
		✓	**50**	0	0
	blue		30	0	0
		✓	**80**	**20**	**10**

TABLE V: Experimental results of ACT trained in different settings and tested in different environments (10 trials).

Pretrain Setting			Success Rate (%) ↑		
#Demo	Task	Embo.	Reach	Pick	Place
196	same	same	55	25	15
	diff.	same	40	10	10
	same	diff.	40	15	5
	diff.	diff	35	10	5
139	same	same	65	15	15
	diff.	same	35	15	10
	same	diff.	20	5	0
	diff.	diff.	35	10	5

TABLE VI: Experimental results of ACT pretrained in different settings and tested in the original environments (20 trials).

of data from different tasks during pretraining further improves the overall success rate. Comparing the results of training for 500 epochs with pretraining to training for 750 epochs without pretraining, we find that pretraining on our dataset also accelerates model convergence. These results demonstrate that leveraging the diverse training data from our dataset enhances the adaptability and robustness of the robotic manipulation model.

We then reduce the number of the robot demonstrations to simulate a few-shot learning scenario. With 40 robot demonstrations, the results of pretraining on our dataset outperform the counterpart trained with 75 demonstrations without pretraining. Further reducing the demonstrations to 10, the results of pretraining on multiple tasks from our dataset still surpass the one trained with 75 demonstrations without pretraining. This demonstrates the beneficial impact of our dataset on few-shot learning in robotic manipulation.

After that, we test the models' generalization ability in new environments by replacing the object and table cover with novel ones, as depicted in Fig. 8(b) and (c). Two models are compared in this experiment: both trained with the original 75 demonstrations for 750 epochs, one with pretraining on multiple similar tasks from our dataset and one without. Results in Tab. V consistently show that the model pretrained on our dataset outperforms its counterpart without pretraining, highlighting the enhanced generalization ability provided by our dataset.

Finally, we explore various pretraining settings to study the impact of pretraining tasks and embodiments. For fairness, we randomly chose 139 and 196 demonstrations for pretraining in these scenarios. The results in Tab. VI show that, within the same embodiment configurations, pretraining effectiveness surpasses that of differing embodiment configurations. This is due to significant differences in end-effector workspaces among different embodiments, making transfer more challenging. Additionally, within the same task configurations, pretraining efficacy is superior to that with different task setups. This is because different task pretraining may introduce action patterns unrelated to the current task, thus impacting performance.

V. Discussion and Conclusion

In this paper, we present the RH20T dataset for diverse robotic skill learning. We believe it can facilitate many areas in robotics, especially for robotic manipulation in novel environments. The current limitations of this paper are that (i) the cost of data collection is expensive and (ii) the potential of robotic foundation models is not evaluated on our dataseet. We have tried to duplicate the results of some recent robotic foundation models but haven't succeeded yet due the limit of computing resources. Thus, we decide to open source the dataset at this stage and hope to promote the development of this area together with our community. In the future, we hope to extend our dataset to broader robotic manipulation, including dual-arm and multi-finger dexterous manipulation.

Acknowledgement: This work was supported by the National Key Research and Development Project of China (No. 2022ZD0160102), National Key Research and Development Project of China (No. 2021ZD0110704), Shanghai Artificial Intelligence Laboratory, XPLORER PRIZE grants.

REFERENCES

[1] Michal Bednarek, Piotr Kicki, and Krzysztof Walas. "On robustness of multi-modal fusion—Robotics perspective". In: *Electronics* 9.7 (2020), p. 1152.

[2] Homanga Bharadhwaj et al. "RoboAgent: Generalization and Efficiency in Robot Manipulation via Semantic Augmentations and Action Chunking". In: *arXiv preprint arXiv:2309.01918* (2023).

[3] Rishi Bommasani et al. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021).

[4] Alessandro Bonardi, Stephen James, and Andrew J Davison. "Learning one-shot imitation from humans without humans". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 3533–3539.

[5] Anthony Brohan et al. "RT-1: Robotics Transformer for Real-World Control at Scale". In: *Robotics: Science and Systems (RSS)*. 2023.

[6] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), pp. 1877–1901.

[7] Shaowei Cui et al. "Self-Attention Based Visual-Tactile Fusion Learning for Predicting Grasp Outcomes". In: *IEEE Robotics and Automation Letters* 5.4 (2020), pp. 5827–5834.

[8] Sudeep Dasari and Abhinav Gupta. "Transformers for one-shot imitation learning". In: *Conference on Robot Learning (CoRL)*. PMLR. 2020, pp. 2071–2084.

[9] Sudeep Dasari et al. "RoboNet: Large-Scale Multi-Robot Learning". In: *Conference on Robot Learning (CoRL)*. Vol. 100. PMLR. 2019, pp. 885–897.

[10] Yan Duan et al. "One-shot imitation learning". In: *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).

[11] Frederik Ebert et al. "Bridge Data: Boosting Generalization of Robotic Skills with Cross-Domain Datasets". In: *Robotics: Science and Systems (RSS)*. 2022.

[12] Mark Edmonds et al. "Feeling the force: Integrating force and pose for fluent discovery through imitation learning to open medicine bottles". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, pp. 3530–3537.

[13] Hao-Shu Fang et al. "Robust grasping across diverse sensor qualities: The GraspNet-1Billion dataset". In: *The International Journal of Robotics Research* 42.12 (2023), pp. 1094–1103.

[14] Nima Fazeli et al. "See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion". In: *Science Robotics* 4.26 (2019), eaav3123.

[15] Chelsea Finn et al. "One-shot visual imitation learning via meta-learning". In: *Conference on Robot Learning (CoRL)*. PMLR. 2017, pp. 357–368.

[16] Maxwell Forbes et al. "Robot programming by demonstration with crowdsourced action fixes". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 2. 2014, pp. 67–76.

[17] De-An Huang et al. "Neural task graphs: Generalizing to unseen tasks from a single video demonstration". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 8565–8574.

[18] Tiancheng Huang, Feng Zhao, and Donglin Wang. "One-Shot Imitation Learning on Heterogeneous Associated Tasks via Conjugate Task Graph". In: *International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–8.

[19] Stephen James, Michael Bloesch, and Andrew J Davison. "Task-embedded control networks for few-shot imitation learning". In: *Conference on Robot Learning (CoRL)*. PMLR. 2018, pp. 783–795.

[20] Stephen James et al. "Rlbench: The robot learning benchmark & learning environment". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 3019–3026.

[21] Eric Jang et al. "BC-z: Zero-shot task generalization with robotic imitation learning". In: *Conference on Robot Learning (CoRL)*. PMLR. 2021, pp. 991–1002.

[22] Roland S Johansson, J Randall Flanagan, and Roland S Johansson. "Sensory control of object manipulation". In: *Sensorimotor control of grasping: Physiology and pathophysiology* (2009), pp. 141–160.

[23] Dmitry Kalashnikov et al. "Mt-opt: Continuous multi-task robotic reinforcement learning at scale". In: *arXiv preprint arXiv:2104.08212* (2021).

[24] Alexander Kirillov et al. "Segment Anything". In: *arXiv:2304.02643* (2023).

[25] Michelle A Lee et al. "Making sense of vision and touch: Learning multimodal representations for contact-rich tasks". In: *IEEE Transactions on Robotics* 36.3 (2020), pp. 582–596.

[26] Fengming Li et al. "Manipulation skill acquisition for robotic assembly based on multi-modal information description". In: *IEEE Access* 8 (2019), pp. 6282–6294.

[27] Corey Lynch and Pierre Sermanet. "Language conditioned imitation learning over unstructured data". In: *Robotics: Science and Systems (RSS)*. 2021.

[28] Zhao Mandi et al. "Towards More Generalizable One-shot Visual Imitation Learning". In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2022.

[29] Ajay Mandlekar et al. "Roboturk: A crowdsourcing platform for robotic skill learning through imitation". In: *Conference on Robot Learning (CoRL)*. PMLR. 2018, pp. 879–893.

[30] Abhishek Padalkar et al. "Open x-embodiment: Robotic learning datasets and rt-x models". In: *arXiv preprint arXiv:2310.08864* (2023).

[31] Peter Pastor et al. "Learning and generalization of motor skills by learning from demonstration". In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2009, pp. 763–768.

[32] Deepak Pathak et al. "Zero-shot visual imitation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 2050–2053.

[33] Rouhollah Rahmatizadeh et al. "Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration". In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 3758–3765.

[34] Aditya Ramesh et al. "Zero-Shot Text-to-Image Generation". In: *International Conference on Machine Learning (ICML)*. PMLR. 2021, pp. 8821–8831.

[35] Nathan Ratliff, J Andrew Bagnell, and Siddhartha S Srinivasa. "Imitation learning for locomotion and manipulation". In: *2007 7th IEEE-RAS International Conference on Humanoid Robots*. IEEE. 2007, pp. 392–397.

[36] Pratyusha Sharma et al. "Multiple interactions made easy (mime): Large scale demonstrations data for imitation". In: *Conference on Robot Learning (CoRL)*. PMLR. 2018, pp. 906–915.

[37] Simon Stepputtis et al. "Language-conditioned imitation learning for robot manipulation tasks". In: *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), pp. 13139–13150.

[38] Homer Walke et al. "BridgeData V2: A Dataset for Robot Learning at Scale". In: *arXiv preprint arXiv:2308.12952* (2023).

[39] Zheng Wu et al. "Learning dense rewards for contact-rich manipulation tasks". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 6214–6221.

[40] Taozheng Yang et al. “MOMA-Force: Visual-Force Imitation for Real-World Mobile Manipulation”. In: *arXiv preprint arXiv:2308.03624* (2023).

[41] Sarah Young et al. “Visual Imitation Made Easy”. In: *Conference on Robot Learning (CoRL)*. Vol. 155. PMLR. 2020, pp. 1992–2005.

[42] Tianhe Yu et al. “Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning”. In: *Conference on Robot Learning*. PMLR. 2019, pp. 1094–1100.

[43] Tianhe Yu et al. “One-Shot Imitation from Observing Humans via Domain-Adaptive Meta-Learning”. In: *Robotics: Science and Systems (RSS)*. 2018.

[44] Tianhao Zhang et al. “Deep imitation learning for complex manipulation tasks from virtual reality teleoperation”. In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 5628–5635.

[45] Tony Z Zhao et al. “Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware”. In: *Robotics: Science and Systems (RSS)*. 2023.

[46] Allan Zhou et al. “Watch, Try, Learn: Meta-Learning from Demonstrations and Rewards”. In: *International Conference on Learning Representations (ICLR)*. 2019.