Low-Cost Exoskeletons for Learning Whole-Arm Manipulation in the Wild

Hongjie Fang*, **Hao-Shu Fang***, **Yiming Wang***,
Jieji Ren, Jingjing Chen, Ruo Zhang, Weiming Wang, Cewu Lu†
Shanghai Jiao Tong University
galaxies@sjtu.edu.cn, fhaoshu@gmail.com, sommerfeld@sjtu.edu.cn,
{jiejiren, jjchen20}@sjtu.edu.cn, ruozhang0608@gmail.com,
{wangweiming, lucewu}@sjtu.edu.cn

Abstract

While humans can use parts of their arms other than the hands for manipulations like gathering and supporting, whether robots can effectively learn and perform the same type of operations remains relatively unexplored. As these manipulations require joint-level control to regulate the complete poses of the robots, we develop *AirExo*, a low-cost, adaptable, and portable dual-arm exoskeleton, for teleoperation and demonstration collection. As collecting teleoperated data is expensive and time-consuming, we further leverage *AirExo* to collect cheap in-the-wild demonstrations at scale. Under our in-the-wild learning framework, we show that with only 3 minutes of the teleoperated demonstrations, augmented by diverse and extensive in-the-wild data collected by *AirExo*, robots can learn a policy that is comparable to or even better than one learned from teleoperated demonstrations lasting over 20 minutes. Experiments demonstrate that our approach enables the model to learn a more general and robust policy across the various stages of the task, enhancing the success rates in task completion even with the presence of disturbances.

1 Introduction

Robotic manipulation has emerged as a crucial field within the robot learning community and attracted significant attention from researchers. With the advancement of technologies such as deep learning, robotic manipulation has evolved beyond conventional grasping [9, 11, 33] and pick-and-place tasks [32, 43], encompassing a diverse array of complex and intricate operations [2, 3, 6, 10].

Most of the current robotic manipulation research focuses on interacting with the environment solely with the end-effectors of the robots, which correspond to the hands of human beings. However, as humans, we can also use other parts of our arms to accomplish or assist with various tasks in daily life. For example, holding objects with lower arms, closing fridge door with elbow, *etc*. In this paper, we aim to investigate

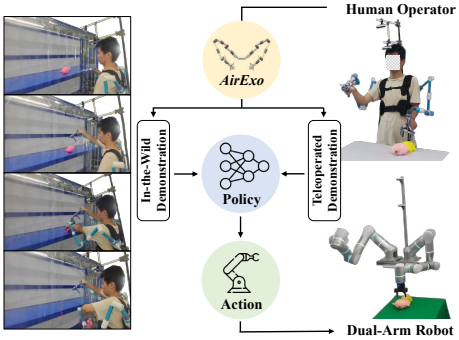


Figure 1: The methodology of our in-the-wild learning framework with low-cost exoskeletons *AirExo*. Our learning framework leverages both teleoperated demonstrations and the extensive and cheap in-the-wild demonstrations in policy learning, resulting in a more general and robust policy compared to training with even more teleoperated demonstrations.

*equal contribution.
†corresponding author.

NeurIPS 2023 Workshop on Robot Learning: Pretraining, Fine-Tuning, and Generalization with Large Scale Models, New Orleans, United States of America

and explore the ability of robots to effectively execute such tasks. To distinguish from the classical manipulation involving end-effectors, we refer to these actions as **whole-arm manipulation**. Since most whole-arm manipulation tasks require the coordinated collaboration of both limbs, we formalize them into the framework of the bimanual manipulation problem.

Whole-arm manipulation presents challenges for robots, including the risk of collisions due to extensive contact with the environment and the need for precise whole-arm movement. To address these challenges, we employ joint-level control through imitation learning during robot demonstrations. Recently, Zhao *et al.* [46] introduced an open-source low-cost ALOHA system which exhibits the capability to perform joint-level imitation learning through real-world teleoperated data. ALOHA system leverages two small, simple and modular bimanual robots ViperX [37] and WidowX [40] that are almost identical to each other, to establish a leader-follower framework for teleoperation. Due to the limited payload of the robots, they focus more on fine-grained manipulation. Besides, their hardwares cannot be seamlessly adapted to other robots commonly employed for laboratory research or industrial purposes. Similarly, while several literatures [8, 15, 17, 19, 45] also designed special exoskeletons for certain humanoid robots or robot arms, the cross-robot transferability of their exoskeletons remain a challenge.

To address the above issues, we develop *AirExo*, an *open-source*, *low-cost*, *robust* and *portable* dual-arm exoskeleton system that can be quickly modified for different robots. All structural components of *AirExo* are *universal* across robots and can be fabricated entirely through 3D printing, enabling easy assembly even for non-experts. After calibration with a dual-arm robot, *AirExo* can achieve precise joint-level teleoperations of the robot. Contributed to its portable property, *AirExo* enables *in-the-wild data collection for dexterous manipulation without needing a robot*. Humans can wear the dual-arm exoskeleton system, conduct manipulation in the wild, and collect demonstrations at scale. This breakthrough capability not only simplifies data collection but also extends the reach of whole-arm manipulation into unstructured environments, where robots can learn and adapt from human interactions. The one-to-one mapping of joint configurations also reduces the barriers of transferring policies trained on human-collected data to robots. Experiments show that with our in-the-wild learning framework, the policy can become more sample efficient for the expensive teleoperated demonstrations, and can acquire more high-level knowledge for task execution, resulting in a more general and robust strategy.

2 AirExo: An Open-Source, Portable, Adaptable, Inexpensive and Robust Exoskeleton

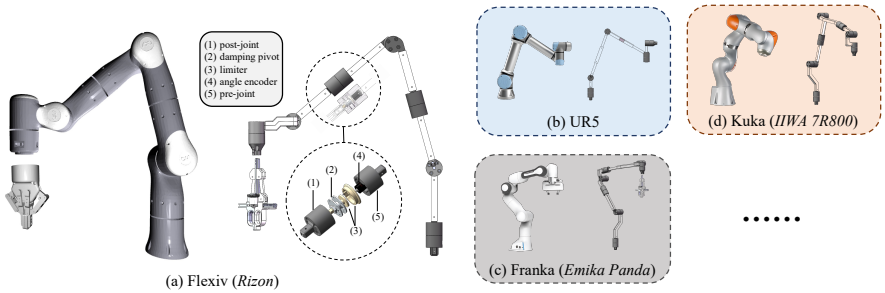


Figure 2: *AirExo* models for different types of robots. Notice that the internal structure of the joints is standardized, only the linkages are altered to accommodate different robotic arm configurations.

Exoskeleton We employ two Flexiv Rizon arms [12] for experiments in this paper. As a result, the structural design of *AirExo* is predominantly tailored to their specifications. Meanwhile, to ensure its universality, it can be easily modified for use with other robotic arms like UR5 [36], Franka [13] and Kuka [20], as depicted in Fig. 2. Based on the morphology of our robot system, *AirExo* is composed of two symmetrical arms, wherein the initial 7 degree-of-freedoms (DoFs) of each arm correspond to the DoFs of the robotic arm, and the last DoF corresponds to the end-effector of the robotic arm. Here, we design a two-finger gripper with 1 DoF as an optional end-effector for each arm. Overall, *AirExo* is capable of simulating the kinematics of the robot across its entire workspace, as well as emulating the opening and closing actions of the end-effectors.

To improve the wearable experience for operators and concurrently enhance task execution efficiency, we dimension *AirExo* to be 80% of the robot's size, based on the length of the human arm. *AirExo* utilizes a dual-layer joint structure with *pre-joints* and *post-joints* connected by a metal *damping pivot*, as illustrated in Fig. 2(a). *Angle encoders* with high precision (0.08 degrees) are mounted on the *pre-joints* to achieve accurate motion capture. A *limiter*, consisting of a dual-layer disc and several steel balls, sets angle limits for each joint, ensuring encoders remain unaffected by bending moments and reducing failures. In the end-effector of the exoskeleton, we design a handle and a scissor-like opening-closing mechanism to simulate the function of a two-fingered gripper, while also facilitating gripping actions by the operator. The two arms of the exoskeleton are affixed to a base, which is mounted on a vest. This allows the operator to wear it stably, and evenly distributing the weight of the exoskeleton across the back of the operator to reduce the load on the arms, thereby enabling more flexible arm motions. Additionally, an adjustable camera mount can be installed on the base for image data collection during operations.

Except the fasteners, damping pivots, and electronic components, all other components of *AirExo* are fabricated using PLA plastic through 3D printing. The prevalence of 3D-printed components allows the exoskeleton to be easily adapted to different robots. This adaptation entails adjusting the dimensions of certain components based on the target robot's specifications and subsequently reprinting and installing them, without modifying the internal structure. *AirExo* costs approximately $600 in total. Please refer to Appendix B for details about *AirExo*.

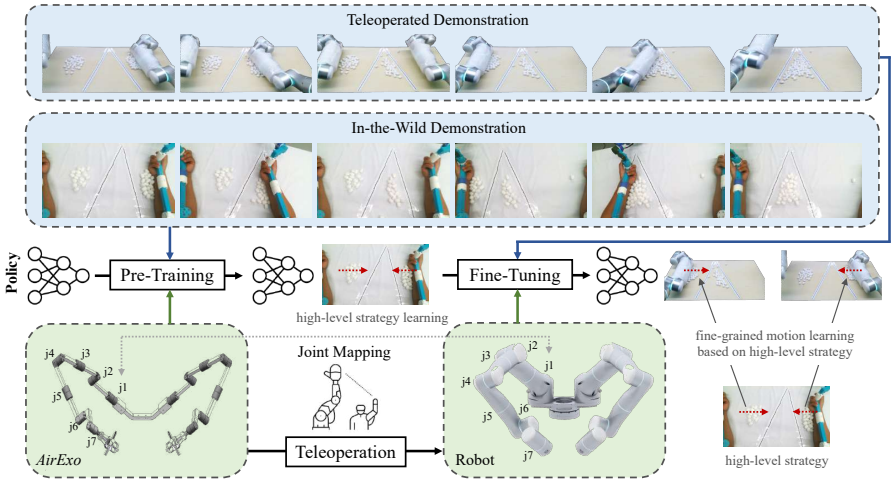


Figure 3: Overview of learning whole-arm manipulations in the wild with *AirExo*.

Learning in the Wild Our approach to learn whole-arm manipulation in the wild with *AirExo* is illustrated in Fig. 3. *AirExo* serves as a natural bridge for the kinematic gap between humans and robots. To address the domain gap between images, our approach involves a two-stage training process. In the first stage, we pre-train the policy using in-the-wild human demonstrations and actions recorded by the exoskeleton encoders. During this phase, the policy primarily learns the high-level task execution strategy from the large-scale and diverse in-the-wild human demonstrations. Then, in the second stage, the policy undergoes fine-tuning using teleoperated demonstrations with robot actions to refine the motions based on the previously acquired high-level task execution strategy.

We use the state-of-the-art bimanual imitation learning method ACT [46] for policy learning. Our experiments demonstrate that it can indeed learn the high-level strategy through the pre-training process and significantly enhance the evaluation performance of the robot and the sample efficiency of the expensive teleoperated demonstrations.

3 Experiments

In this section, we conduct experiments on 2 whole-arm tasks (***Gather Balls*** and ***Grasp from the Curtained Shelf***) to evaluate the performance of the proposed learning method. All demonstration data are collected by *AirExo*. Several baselines, including VINN [26], ConvMLP [44], BeT [31] and

ACT [46] are evaluated in the experiments. We also apply our proposed learning approach to ACT for learning from in-the-wild demonstrations. Please refer to the Appendix C for more details.

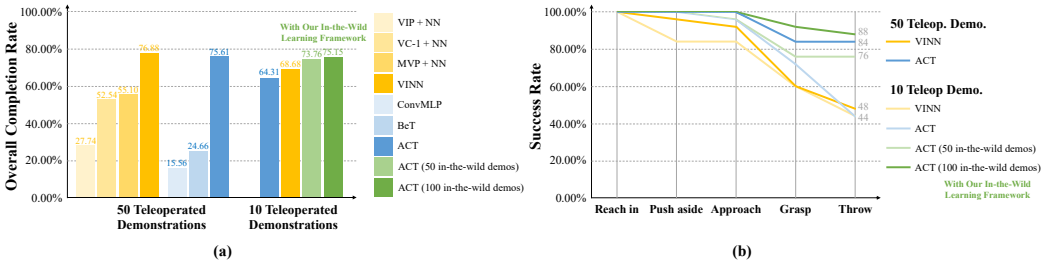


Figure 4: Experimental Results on the (a) ***Gather Balls*** task and (b) ***Grasp from the Curtained Shelf*** task.

Gather Balls In Fig. 4(a), experimental results reveal that VINN and ACT excel among non-parametric and parametric methods respectively, when trained with 50 teleoperated demonstrations. However, using only 10 teleoperated demonstrations leads to performance degradation for both methods. Nonetheless, employing our in-the-wild learning framework, assisted by in-the-wild demonstrations, enables ACT to match the performance of 50 teleoperated demonstrations with just 10, showcasing the enhanced sample efficiency of our framework for teleoperated demonstrations. We then delve into the experimental results to provide more insights about why and how our learning framework works. Please refer to the Appendix C.1 for details.

Grasp from the Curtained Shelf Results are shown in Fig. 4(b). Similar to the ***Gather Balls*** task, as the number of teleoperated demonstrations decreases, VINN and ACT both show reduced success rates, particularly in the later "throw" phase. However, our in-the-wild learning framework substantially boosts the success rates of ACT in the "grasp" and "throw" phases. Even with just 10 demonstrations lasting approximately 3 minutes, ACT outperforms the original 50 demonstrations lasting over 20 minutes. This underscores how our in-the-wild framework enhances policy learning for better success rates in multi-stage tasks.

Robustness Analysis We design three kinds of disturbances in the ***Grasp from the Curtained Shelf*** task to explore whether in-the-wild learning improves the robustness of the policy. The results shown in Tab. 1 demonstrate that our in-the-wild learning framework can leverage diverse in-the-wild demonstrations to make the learned policy more robust and generalizable to various environmental disturbances.

Disturbances	w/wo i.t.w. learning	Success / All
Novel Object	✘	4 / 8
	✔	**7** / 8
Different Background	✘	2 / 8
	✔	**6** / 8
Visual Distractors	✘	4 / 8
	✔	**8** / 8

Table 1: Results of the robustness experiments on the ***Grasp from the Curtained Shelf*** task.

4 Conclusion

In this paper, we develop *AirExo*, an open-source, low-cost, universal, portable, and robust exoskeleton, for both joint-level teleoperation and learning whole-arm manipulations in the wild. Our proposed in-the-wild learning framework decreases the demand for the resource-intensive teleoperated demonstrations. Experimental results demonstrate the effectiveness and robustness of the policies learned through this approach. In the future, we are excited to see our *AirExo* collecting large-scale human demonstrations in unstructured environments and facilitating robot learning.

References

[1] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. "Human-to-Robot Imitation in the Wild". In: *Robotics: Science and Systems (RSS)*. 2022.

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

[3] Anthony Brohan et al. "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control". In: *arXiv preprint arXiv:2307.15818* (2023).

[4] Annie S. Chen, Suraj Nair, and Chelsea Finn. “Learning Generalizable Robotic Reward Functions from "In-The-Wild" Human Videos”. In: *Robotics: Science and Systems (RSS)*. 2021.

[5] Cheng Chi et al. “Diffusion Policy: Visuomotor Policy Learning via Action Diffusion”. In: *Robotics: Science and Systems (RSS)*. 2023.

[6] Danny Driess et al. “PaLM-E: An Embodied Multimodal Language Model”. In: *International Conference on Machine Learning (ICML)*. Vol. 202. PMLR, 2023, pp. 8469–8488.

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

[8] Fabian Falck, Kawin Larppichet, and Petar Kormushev. “DE VITO: A Dual-Arm, High Degree-of-Freedom, Lightweight, Inexpensive, Passive Upper-Limb Exoskeleton for Robot Teleoperation”. In: *Towards Autonomous Robotic Systems: 20th Annual Conference, TAROS 2019, London, UK, July 3–5, 2019, Proceedings, Part I 20*. Springer. 2019, pp. 78–89.

[9] Hao-Shu Fang et al. “AnyGrasp: Robust and Efficient Grasp Perception in Spatial and Temporal Domains”. In: *IEEE Transactions on Robotics (TRO)* (2023).

[10] Hao-Shu Fang et al. “RH20T: A Robotic Dataset for Learning Diverse Skills in One-Shot”. In: *RSS 2023 Workshop on Learning for Task and Motion Planning*. 2023.

[11] Hao-Shu Fang et al. “Robust Grasping across Diverse Sensor Qualities: The GraspNet-1Billion Dataset”. In: *The International Journal of Robotics Research* (2023), p. 02783649231193710.

[12] *Flexiv Rizon Robot*. URL: https://www.flexiv.com/en/technology/robot.

[13] *Franka Emika Panda*. URL: https://www.franka.de/research.

[14] Jean-Bastien Grill et al. “Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 21271–21284.

[15] Yasuhiro Ishiguro et al. “Bilateral Humanoid Teleoperation System Using Whole-Body Exoskeleton Cockpit TABLIS”. In: *IEEE Robotics and Automation Letters* 5.4 (2020), pp. 6419–6426.

[16] 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.

[17] Heecheol Kim, Yoshiyuki Ohmura, and Yasuo Kuniyoshi. “Robot Peels Banana with Goal-Conditioned Dual-Action Deep Imitation Learning”. In: *arXiv preprint arXiv:2203.09749* (2022).

[18] Heecheol Kim, Yoshiyuki Ohmura, and Yasuo Kuniyoshi. “Transformer-Based Deep Imitation Learning for Dual-Arm Robot Manipulation”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2021, pp. 8965–8972.

[19] Heecheol Kim et al. “Training Robots Without Robots: Deep Imitation Learning for Master-to-Robot Policy Transfer”. In: *IEEE Robotics and Automation Letters* 8.5 (2023), pp. 2906–2913.

[20] *Kuka IIWA 7R800*. URL: https://www.kuka.com/en-de/products/robot-systems/industrial-robots/lbr-iiwa.

[21] Junjia Liu et al. “Robot Cooking with Stir-Fry: Bimanual Non-Prehensile Manipulation of Semi-Fluid Objects”. In: *IEEE Robotics and Automation Letters* 7.2 (2022), pp. 5159–5166.

[22] Yecheng Jason Ma et al. “VIP: Towards Universal Visual Reward and Representation via Value-Implicit Pre-Training”. In: *International Conference on Learning Representations (ICLR)*. 2023.

[23] Arjun Majumdar et al. “Where are We in the Search for an Artificial Visual Cortex for Embodied Intelligence?” In: *ICRA 2023 Workshop on Pretraining for Robotics*. 2023.

[24] 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.

[25] Suraj Nair et al. “R3M: A Universal Visual Representation for Robot Manipulation”. In: *Conference on Robot Learning (CoRL)*. PMLR. 2022, pp. 892–909.

[26] Jyothish Pari et al. “The Surprising Effectiveness of Representation Learning for Visual Imitation”. In: *Robotics: Science and Systems (RSS)*. 2022.

[27] Dean A Pomerleau. “ALVINN: An Autonomous Land Vehicle in a Neural Network”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 1. 1988.

[28] Pragathi Praveena et al. “Characterizing Input Methods for Human-to-Robot Demonstrations”. In: *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 344–353.

[29] Ilija Radosavovic et al. “Real-World Robot Learning with Masked Visual Pre-Training”. In: *Conference on Robot Learning (CoRL)*. PMLR. 2022, pp. 416–426.

[30] Erick Rosete-Beas et al. “Latent Plans for Task Agnostic Offline Reinforcement Learning”. In: *Conference on Robot Learning (CoRL)*. PMLR. 2022, pp. 1838–1849.

[31] Nur Muhammad Shafiullah et al. “Behavior Transformers: Cloning k Modes with One Stone”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 35. 2022, pp. 22955–22968.

[32] Anthony Simeonov et al. “Neural Descriptor Fields: SE(3)-Equivariant Object Representations for Manipulation”. In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2022, pp. 6394–6400.

[33] Shuran Song et al. “Grasping in the Wild: Learning 6DoF Closed-Loop Grasping from Low-Cost Demonstrations”. In: *IEEE Robotics and Automation Letters* 5.3 (2020), pp. 4978–4985.

[34] Kanata Suzuki et al. “In-Air Knotting of Rope Using Dual-Arm Robot Based on Deep Learning”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2021, pp. 6724–6731.

[35] Alexander Toedtheide et al. “A Force-Sensitive Exoskeleton for Teleoperation: An Application in Elderly Care Robotics”. In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 12624–12630.

[36] *UR5 Robot*. URL: https://www.universal-robots.com/products/ur5-robot.

[37] *ViperX 300 Robot Arm 6DoF*. URL: https://www.trossenrobotics.com/viperx-300-robot-arm-6dof.aspx.

[38] Chen Wang et al. “MimicPlay: Long-Horizon Imitation Learning by Watching Human Play”. In: *arXiv preprint arXiv:2302.12422* (2023).

[39] Thomas Weng et al. “FabricFlowNet: Bimanual Cloth Manipulation with a Flow-based Policy”. In: *Conference on Robot Learning (CoRL)*. PMLR. 2021, pp. 192–202.

[40] *WidowX 250 Robot Arm 6DoF*. URL: https://www.trossenrobotics.com/widowx-250-robot-arm-6dof.aspx.

[41] Fan Xie et al. “Deep Imitation Learning for Bimanual Robotic Manipulation”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 2327–2337.

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

[43] Andy Zeng et al. “Transporter Networks: Rearranging the Visual World for Robotic Manipulation”. In: *Conference on Robot Learning (CoRL)*. PMLR. 2020, pp. 726–747.

[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] Liang Zhao et al. “A Wearable Upper Limb Exoskeleton for Intuitive Teleoperation of Anthropomorphic Manipulators”. In: *Machines* 11.4 (2023), p. 441.

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

Appendices

A Related Works

Imitation Learning Imitation learning has been widely applied in robot learning to teach robots how to perform various tasks by observing and imitating demonstrations from human experts. One of the simplest methods in imitation learning is behavioral cloning [27], which learns the policy directly in a supervised manner without considering intentions and outcomes. Most approaches parameterize the policy using neural networks [2, 5, 31, 44, 46], while non-parametric VINN [26] leverages the weighted k-nearest-neighbors algorithm based on the visual representations extracted by BYOL [14] to generate the action from the demonstration database. This simple but effective method can also be extended to other visual representations [22, 23, 25, 29] for robot learning. In the context of imitation learning for bimanual manipulation, Xie *et al.* [41] introduced a paradigm to decouple the high-level planning model into the elemental movement primitives. Several literature have focused on designing special frameworks to solve specific tasks, such as knot tying [18, 34], banana peeling [17], culinary activities [21], and fabric folding [39]. Addressing the challenge of non-Markovian behavior observed in demonstrations, Zhao *et al.* [46] utilized the notion of action chunking as a strategy to enhance overall performance.

Teleoperation Demonstration data play a significant role in robotic manipulation, particularly in the methods based on imitation learning. For the convenience of subsequent robot learning, these demonstration data are typically collected within the robot domain. A natural approach to gather such demonstrations is human teleoperation [24], where a human operator remotely controls the robot to execute various tasks. Teleoperation methods can be broadly categorized into two classes based on their control objectives: one aimed at manipulating the end-effectors of the robots [2, 7, 10, 16, 30, 44] and one focused on regulating the complete poses of the entire robots, such as exoskeletons [8, 15, 17, 35, 45] and a pair of leader-follower robots [46]. For whole-arm manipulation tasks, we need to control the full pose of the robots, which makes exoskeletons a relatively favorable option under this circumstance.

Learning Manipulation in the Wild Despite the aforementioned teleoperation methods allow us to collect robotic manipulation data, the robot system is usually expensive and not portable, posing challenges to collect demonstration data at scale. To address this issue, previous research has explored the feasibility of learning from interactive human demonstrations, *i.e.* in-the-wild learning for robotic manipulation [1, 4, 19, 28, 33, 42]. In contrast to the costly robot demonstrations, in-the-wild demonstrations are typically cheap and easy to obtain, allowing us to collect a large volume of such demonstrations conveniently.

Typically, there are two primary domain gaps for learning manipulation in the wild: (1) the gap between human-operated images and robot-operated images, and (2) the gap between human kinematics and robot kinematics. The former gap can be solved through several approaches: by utilizing specialized end-effectors that match the end-effectors of the robots [19, 42]; by initially pre-training with in-the-wild data and subsequently fine-tuning with robot data [33]; or by applying special image processing technique to generate agent-agnostic images [1]. The latter gap is currently addressed by applying structure from motion algorithms [33, 42], adopting a motion tracking system [28], or training a pose detector [1, 38] to extract the desired poses. However, these methods are not suitable for whole-arm dexterous manipulation, since motion tracking usually focuses on the end-effector, and pose detector is vulnerable to visual occlusions and does not map to the robot kinematics.

Therefore, in this paper, we develop a low-cost and portable exoskeleton to serve as a bridge between human motion and robot motion. It can be applied not only to the teleoperation of robots but also as a powerful tool for learning manipulation in the wild.

B *AirExo*

Design Objectives From the preceding discussions, we summarize the following 5 key design objectives of an exoskeleton: (1) affordability; (2) adaptability; (3) portability; (4) robustness and (5) maintenance simplicity. *AirExo* is designed based on these objectives.

(1) **Affordability**. The exoskeleton system should be priced at a low level that ensures affordability for a broad spectrum of laboratories and even individual enthusiasts.

(2) **Adaptability**. The exoskeleton system should be readily adjustable to accommodate various robots without necessitating any modifications the internal joint structure.

(3) **Portability**. The exoskeleton system should exhibit a lightweight and ergonomic construction, facilitating maneuverability and an extensive array of motions.

(4) **Robustness**. The exoskeleton system should possess robust durability, enabling it to endure extended operational periods dedicated to demonstration data collection.

(5) **Maintenance Simplicity**. The components comprising the exoskeleton system should be engineered with an emphasis on simplicity. Assembly ought to be achievable without the requirement of specialized tools, and during maintenance, only a minimal number of components need to be disassembled.

Calibration and Teleoperation Since *AirExo* shares the same morphology with the dual-arm robot except for the scale, the calibration process can be performed in a quite straightforward manner. After positioning the robot arms at a specific location like a fully extended position, and aligning the exoskeleton to match the robot posture, we can record the joint positions $\{q_i^{(c)}\}_{i=1}^d$ and the encoder readings $\{p_i^{(c)}\}_{i=1}^d$ of *AirExo*, where d denotes the DoFs. Consequently, during teleoperation, we only need to fetch the encoder readings $\{p_i\}_{i=1}^d$ and transform them into the corresponding joint positions $\{q_i\}_{i=1}^d$ using Eqn. (1), and let the robot moves to the desired joint positions:

$$q_i = \min\left(\max\left(q_i^{(c)} + k_i(p_i - p_i^{(c)}), q_i^{\min}\right), q_i^{\max}\right), \tag{1}$$

where $k_i \in \mathbb{R}$ is the coefficient controlling direction and scale, and $q_i^{\min}, q_i^{\max}$ denote the joint angle limits of the robotic arms. Typically, we set $k = \pm 1$, representing the consistency between the encoder direction of the exoskeleton and the joint direction of the robot. For grippers, we can directly map the angle range of the encoders to the opening and closing range of the grippers for teleoperation.

After calibration, the majority of angles within the valid range of the robot arms can be covered by the exoskeleton. Given that the workspaces of most tasks fall within this coverage range, we can teleoperate the robot using the exoskeleton conveniently and intuitively. If a special task t needs a wider operation range, we can simply scale the exoskeleton range using coefficients k_i, and apply task-specific joint constraint $[q_i^{t,\min}, q_i^{t,\max}]$ instead of original kinematic constraint in Eqn. (1) for better teleoperation performance.

Learning in the Wild Here, we add more details of our in-the-wild learning framework with *AirExo*. For in-the-wild whole-arm manipulation learning, we install a camera (or cameras under multi-camera settings) on the camera mount of *AirExo* in roughly the same position(s) as the camera(s) on the robot. Using this configuration, images from both teleoperated demonstrations and in-the-wild demonstrations exhibit a relatively similar structure, which is advantageous for policy learning.

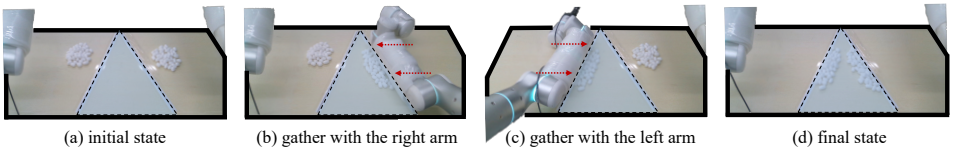
As previously discussed in Sec. 2, we resize the exoskeleton to ensure its wearability. Some concerns may arise regarding whether this scaling adjustment could impact the policy learning process. Here, we argue that it has a minimal effect on our learning procedure. Firstly, the core kinematic structure, essential for our learning framework, remain unaffected by the resizing. Thus human demonstrations preserve the fundamental dynamics of the system. Secondly, our approach does not impose strict alignment requirements between human demonstration images and robot images. We find that similar visual-action pairs collected by our exoskeleton effectively support the pretraining stage, without demanding precise visual matching between human and robot demonstrations.

C Experiments

In this section, we furnish additional details pertaining to the experiments as well as present more experimental results.

C.1 *Gather Balls*

Task Two clusters of cotton balls are randomly placed on both sides of the tabletop (40 balls per cluster). The goal is to gather these balls into the designated central triangular area using both arms. The process of this contact-rich task is illustrated in Fig. 5.



(a) initial state (b) gather with the right arm (c) gather with the left arm (d) final state

Figure 5: Definition of ***Gather Balls*** task. The goal is to gather the balls into the central triangular area, which is highlighted in light blue. The red dashed arrows denote the motions of the robot arms. Sponge paddings are used to envelop the external surface of the robot arms to diminish the mechanical failures arising from contacts.

Metrics We consider the percentage of balls being allocated within the central triangular area as the task completion rate c (if a ball is precisely on the line, it is considered a half), including both the completion rates of the left arm and the right arm. Simultaneously, task success is defined as the task completion rate exceeding a certain threshold δ. In this experiment, we set $\delta = 40\%, 60\%, 80\%$. We also record the collision rate to gauge the precision of the operations.

Methods We employ VINN [26] and its variants that alter the visual representations [22, 23, 29] as non-parametric methods. Other methods include ConvMLP [44], BeT [31] and ACT [46]. All of them are designed for joint-space control or can be easily adapted for joint-space control. We apply our proposed learning approach to ACT for learning from in-the-wild demonstrations. For all methods, we carefully select the hyper-parameters to ensure better performance.

Protocols The evaluation is conducted on a workstation equipped with an Intel Core i9-10980XE CPU. The time limit is set as 60 seconds per trial. Given that all methods can operate at approximately 5Hz, resulting in a total of 300 steps for the evaluation, the time constraint proves sufficient for the task. We conduct 50 consecutive trials to ensure stable and accurate results, calculating the aforementioned metrics.

# Demos		Method	Completion Rate c (%) ↑			Success Rate (%) ↑		
Teleoperated	In-the-Wild		Overall	Left	Right	$c \geq 80$	$c \geq 60$	$c \geq 40$
50	-	VIP [22] + NN	27.74	0.02	55.45	0	0	36
50	-	VC-1 [23] + NN	52.54	32.53	72.55	4	42	74
50	-	MVP [29] + NN	55.10	58.55	62.00	12	62	76
50	-	VINN [26]	**76.88**	75.73	78.03	**58**	**84**	94
50	-	ConvMLP [44]	15.56	2.35	28.78	0	0	2
50	-	BeT [31]	24.66	7.38	41.95	0	2	32
50	-	ACT [46]	75.61	94.63	56.60	54	70	**100**
10	-	VINN [26]	68.68	60.28	77.08	36	76	88
10	-	ACT [46]	64.31	91.95	36.68	24	60	**96**
10	50	ACT [46]	73.76	88.83	58.70	**62**	72	88
10	100	ACT [46]	**75.15**	75.63	74.68	56	**80**	88

Table 2: Experimental results on the ***Gather Balls*** task.

Results and Analyses The detailed experimental results on the ***Gather Balls*** task are shown in Tab. 2. We then delve into the experimental results to provide more insights about why and how our learning framework works. When analyzing the failure cases of different methods in the experiments in Fig. 6(a), we find that the ACT policy trained solely on teleoperated demonstrations exhibits an issue of imbalance between accuracies of two arms, with better learning outcomes for the left arm. This imbalance becomes more pronounced as the number of teleoperated demonstrations decreases

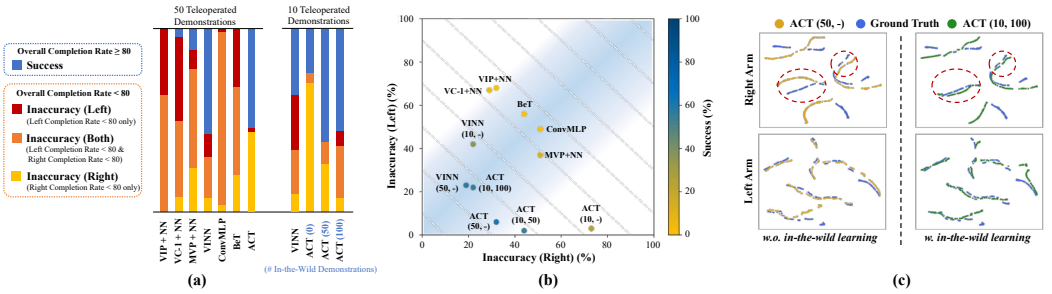


Figure 6: Analyses of methods on the ***Gather Balls*** task. Here we define the overall completion rate over 80% as success. **(a)** We analyze the failure causes of each method in every trial. **(b)** We amortize the inaccuracy (both) rate evenly into the inaccuracy (left) and inaccuracy (right) rates, and draw a comparison plot of failure modes for different methods. (x, y) means the policy is trained with y in-the-wild demonstrations then x teleoperated demonstrations. The dashed lines represent contour lines with the same success rate, and the regions with light blue background imply a more balanced policy between left and right arms. **(c)** t-SNE visualizations of the ground-truth actions and the policy actions w/wo in-the-wild learning on the validation set.

to 10. With the help of the in-the-wild learning stage, the policy becomes more balanced between two arms even with fewer teleoperated demonstrations, as shown in Fig. 6(b). From Fig. 6(c), we also observe that the policy focuses more on learning the motions of the right arm when cooperated with in-the-wild learning, as highlighted in red dashed circles, while keeping the accurate action predictions on the left arm. We believe that this is attributed to the extensive, diverse, and accurate in-the-wild demonstrations provided by *AirExo*, enabling the policy to acquire high-level strategy knowledge during the pre-training stage. Consequently, in the following fine-tuning stage, it can refine its actions based on the strategy, thus avoiding learning actions blindly from scratch.

C.2 *Grasp from the Curtained Shelf*

Task A cotton toy is randomly placed in the center of a shelf with curtains. The goal is to grasp the toy and throw it into a bin. To achieve it, the robot needs to use its right arm to push aside the transparent curtain first, and maintain this pose during the following operations. The process of this multi-stage task is illustrated in Fig. 7.

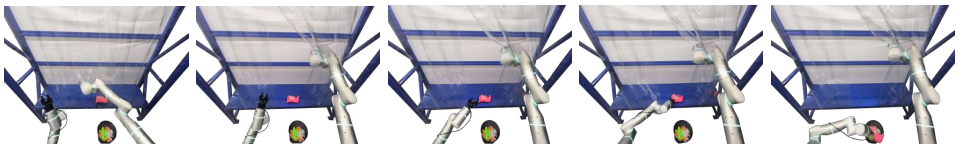


Figure 7: Definition of the ***Grasp from the Curtained Shelf*** task. The robot needs to (a) reach in its right arm to the transparent curtain and (b) push aside the curtain, then (c) approach the object with its left arm, (d) grasp the object and finally (e) throw the object.

Metrics, Methods, and Protocols We calculate the average success rate at the end of each stage as metrics. Based on the experimental results on the ***Gather Balls*** task, we select VINN [26] and ACT [46] as methods in experiments, as well as ACT equipped with our in-the-wild learning framework. The evaluation protocols are the same as the ***Gather Balls*** task, except that the time limit is 120 seconds (about 400 steps) and the number of trials is 25.

Results The detailed experimental results on the ***Grasp from the Curtained Shelf*** task are shown in Tab. 3.

D Future Works

In the future, we will investigate how to better address the image gap between in-the-wild data in the human domain and teleoperated data in the robot domain, enabling robots to learn solely through in-the-wild demonstrations with *AirExo*, thus further reducing the learning cost.

# Demos		Method	Success Rate (%) ↑				
Teleoperated	**In-the-Wild**		Reach in	Push aside	Approach	Grasp	Throw
50	-	VINN [26]	**100**	96	92	60	48
50	-	ACT [46]	**100**	**100**	**100**	**84**	**84**
10	-	VINN [26]	**100**	84	84	60	44
10	-	ACT [46]	**100**	**100**	96	72	44
10	50	ACT [46]	**100**	**100**	96	76	76
10	100	ACT [46]	**100**	**100**	**100**	**92**	**88**

Table 3: Experimental results on the ***Grasp from the Curtained Shelf*** task.