Stanford Artificial Intelligence

January 2024

Mobile ALOHA:
Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation

Zipeng Fu[*1], **Tony Z. Zhao**[*1], **Chelsea Finn**[1]

[*]project co-leads, [1]Stanford University

https://mobile-aloha.github.io

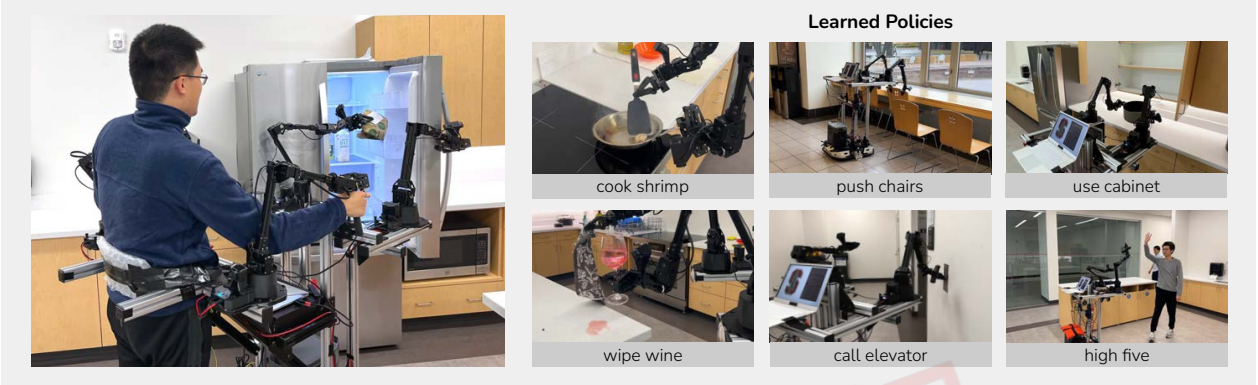

Figure 1: ***Mobile ALOHA*** . We introduce a low-cost mobile manipulation system that is bimanual and supports whole-body teleoperation. The system costs $32k including onboard power and compute. *Left:* A user teleoperates to obtain food from the fridge. *Right: Mobile ALOHA* can perform complex long-horizon tasks with imitation learning.

Abstract

Imitation learning from human demonstrations has shown impressive performance in robotics. However, most results focus on table-top manipulation, lacking the mobility and dexterity necessary for generally useful tasks. In this work, we develop a system for imitating mobile manipulation tasks that are bimanual and require whole-body control. We first present *Mobile ALOHA*, a low-cost and whole-body teleoperation system for data collection. It augments the *ALOHA* system [104] with a mobile base, and a whole-body teleoperation interface. Using data collected with *Mobile ALOHA*, we then perform supervised behavior cloning and find that co-training with existing *static ALOHA* datasets boosts performance on mobile manipulation tasks. With 50 demonstrations for each task, co-training can increase success rates by up to 90%, allowing *Mobile ALOHA* to autonomously complete complex mobile manipulation tasks such as sauteing and serving a piece of shrimp, opening a two-door wall cabinet to store heavy cooking pots, calling and entering an elevator, and lightly rinsing a used pan using a kitchen faucet.

1. Introduction

Imitation learning from human-provided demonstrations is a promising tool for developing generalist robots, as it allows people to teach arbitrary skills to robots. Indeed, direct behavior cloning can enable robots to learn a variety of primitive robot skills ranging from lane-following in mobile robots [67], to simple pick-and-place manipulation skills [12, 20] to more delicate manipulation skills like spreading pizza sauce or slotting in a battery [18, 104]. However, many tasks in realistic, everyday environments require whole-body coordination of both mobility and dexterous manipulation, rather than just individual mobility or manipulation behaviors. For example, consider the relatively basic task of putting away a heavy pot into a cabinet in Figure 1. The robot needs to first navigate to the cabinet, necessitating the mobility of the robot base. To open the cabinet, the robot needs to back up while simultaneously maintaining a firm grasp of the two door handles, motivating whole-body control. Subsequently, both arms need to grasp the pot handles and together move the pot into the cabinet, emphasizing the importance of bimanual coordination. Along a similar vein, cooking, cleaning, housekeeping, and even simply navigating an office using an elevator all require mobile manipulation and are often made easier with the added flexibility of two arms. In this paper, we study the feasibility of extending imitation learning to tasks that require whole-body control of bimanual mobile

robots.

Two main factors hinder the wide adoption of imitation learning for bimanual mobile manipulation. (1) We lack accessible, plug-and-play hardware for whole-body teleoperation. Bimanual mobile manipulators can be costly if purchased off-the-shelf. Robots like the PR2 and the TIAGo can cost more than $200k USD, making them unaffordable for typical research labs. Additional hardware and calibration are also necessary to enable teleoperation on these platforms. For example, the PR1 uses two haptic devices for bimanual teleoperation and foot pedals to control the base [93]. Prior work [5] uses a motion capture system to retarget human motion to a TIAGo robot, which only controls a single arm and needs careful calibration. Gaming controllers and keyboards are also used for teleoperating the Hello Robot Stretch [2] and the Fetch robot [1], but do not support bimanual or whole-body teleoperation. (2) Prior robot learning works have not demonstrated high-performance bimanual mobile manipulation for complex tasks. While many recent works demonstrate that highly expressive policy classes such as diffusion models and transformers can perform well on fine-grained, multi-modal manipulation tasks, it is largely unclear whether the same recipe will hold for mobile manipulation: with additional degrees of freedom added, the interaction between the arms and base actions can be complex, and a small deviation in base pose can lead to large drifts in the arm's end-effector pose. Overall, prior works have not delivered a practical and convincing solution for bimanual mobile manipulation, both from a hardware and a learning standpoint.

We seek to tackle the challenges of applying imitation learning to bimanual mobile manipulation in this paper. On the hardware front, we present *Mobile ALOHA*, a low-cost and whole-body teleoperation system for collecting bimanual mobile manipulation data. *Mobile ALOHA* extends the capabilities of the original *ALOHA* , the low-cost and dexterous bimanual puppeteering setup [104], by mounting it on a wheeled base. The user is then physically tethered to the system and backdrives the wheels to enable base movement. This allows for independent movement of the base while the user has both hands controlling *ALOHA* . We record the base velocity data and the arm puppeteering data at the same time, forming a whole-body teleoperation system.

On the imitation learning front, we observe that simply concatenating the base and arm actions then training via direct imitation learning can yield strong performance. Specifically, we concatenate the 14-DoF joint positions of *ALOHA* with the linear and angular velocity of the mobile base, forming a 16-dimensional action vector. This formulation allows *Mobile ALOHA* to benefit directly from previous deep imitation learning algorithms, requiring almost no change in implementation. To further improve the imitation learning performance, we are inspired by the recent success of pre-training and co-training on diverse robot datasets, while noticing that there are few to none accessible bimanual mobile manipulation datasets. We thus turn to leveraging data from *static* bimanual datasets, which are more abundant and easier to collect, specifically the *static ALOHA* datasets from [81, 104] through the RT-X release [20]. It contains 825 episodes with tasks disjoint from the *Mobile ALOHA* tasks, and has different mounting positions of the two arms. Despite the differences in tasks and morphology, we observe positive transfer in nearly all mobile manipulation tasks, attaining equivalent or better performance and data efficiency than policies trained using only *Mobile ALOHA* data. This observation is also consistent across different class of state-of-the-art imitation learning methods, including ACT [104] and Diffusion Policy [18].

The main contribution of this paper is a system for learning complex mobile bimanual manipulation tasks. Core to this system is both (1) *Mobile ALOHA*, a low-cost whole-body teleoperation system, and (2) the finding that a simple co-training recipe enables data-efficient learning of complex mobile manipulation tasks. Our teleoperation system is capable of multiple hours of consecutive usage, such as cooking a 3-course meal, cleaning a public bathroom, and doing laundry. Our imitation learning result also holds across a wide range of complex tasks such as opening a two-door wall cabinet to store heavy cooking pots, calling an elevator, pushing in chairs, and cleaning up spilled wine. With co-training, we are able to achieve over 80% success on these tasks with only 50 human demonstrations per task, with an average of 34% absolute improvement compared to no co-training.

2. Related Work

Mobile Manipulation. Many current mobile manipulation systems utilize model-based control, which involves integrating human expertise and insights into the system's design and architecture [9, 17, 33, 52, 93]. A notable example of model-based control in mobile manipulation is the DARPA Robotics Challenge [56]. Nonetheless, these systems can be challenging to develop and maintain, often