HoMMI: Learning Whole-Body Mobile Manipulation from Human Demonstrations

Xiaomeng Xu[1,2] Jisang Park[1] Han Zhang[1] Eric Cousineau[2] Aditya Bhat[2] Jose Barreiros[2]
Dian Wang[1] Jeannette Bohg[1] Shuran Song[1]

[1]Stanford University [2]Toyota Research Institute
https://hommi-robot.github.io

Figure 1. **Whole-Body Mobile Manipulation Interface (HoMMI).** (a) We extend UMI with egocentric sensing to enable scalable ***mobile*** manipulation with ***active perception***. (b) However, the new egocentric view creates a substantial embodiment gap in both observation and action space, making policy transfer difficult. (c) We bridge this embodiment gap by carefully redesigning the visual and action representations and integrating them with a constraint-aware whole-body controller. Together, HoMMI is able to learn diverse mobile manipulation skills directly from human demonstrations, without ***any*** robot teleoperation data.

Abstract

We present Whole-Body Mobile Manipulation Interface (HoMMI), a data collection and policy learning framework that learns whole-body mobile manipulation directly from robot-free human demonstrations. We augment UMI interfaces with egocentric sensing to capture the global context required for mobile manipulation, enabling portable, robot-free, and scalable data collection. However, naively incorporating egocentric sensing introduces a larger human-to-robot embodiment gap in both observation and action spaces, making policy transfer difficult. We explicitly bridge this gap with a cross-embodiment hand-eye policy design, including an embodiment agnostic visual representation; a relaxed head action representation; and a whole-body controller that realizes hand-eye trajectories through coordinated whole-body motion under robot-specific physical constraints. Together, these enable long-horizon mobile manipulation tasks requiring bimanual and whole-body coordination, navigation, and active perception.

1. Introduction

Achieving generalizable and effective mobile manipulation requires seamless **whole-body coordination**, which consists of coordinating diverse *sensory* inputs (e.g., egocen-

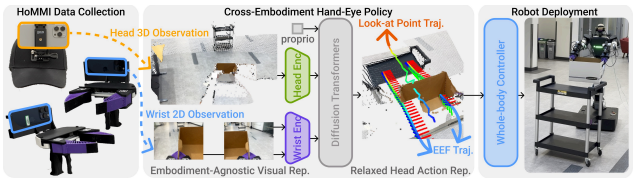

Figure 2. **HoMMI System Overview.**

tric head-mounted cameras to eye-in-hand wrist cameras) and complex *action* spaces (e.g., between the arms, torso, head, and base movements). Manually programming such intricate coordination for the vast variety of real-world tasks is prohibitively difficult, making learning from human a promising alternative.

However, existing human demonstration paradigms mostly rely on robot teleoperation, which is expensive, slow, and unintuitive to deploy for mobile manipulators across diverse real-world settings. Handheld data collection devices such as UMI [1] offer a more scalable solution. They essentially learn end-effector motions through handheld grippers with wrist-mounted camera observations, allowing portable and robot-free demonstration collection. However, wrist-centric sensing provides only local views around the end-effectors and often under-observes the global context needed for navigation, bimanual coordination, and task progress tracking.

Adding an egocentric view (i.e., head-mounted camera) is a natural solution to fill this gap. By capturing the broader workspace, the spatial relationship between hands, as well as humans' active perception behaviors, egocentric views provide critical information that wrist cameras lack. However, ***naively incorporating egocentric sensing into UMI framework introduces a larger human-to-robot embodiment gap***, including:

- *Visual gap:* Human and robot arms differ in appearance, and egocentric viewpoints vary due to height discrepancies between human and robot embodiments.
- *Kinematic gap:* Humans and robots differ in body morphology and neck degrees of freedom. Directly regressing and tracking both hands and head 6-DoF trajectories often yield infeasible robot motions.

As a result, prior egocentric systems either rely on additional teleoperation data for action grounding [3, 6], or restrict the application domain to fixed-base bimanual manipulation without whole-body coordination [4, 5]. This paper aims to ***scale mobile manipulation learning by augmenting the UMI framework with egocentric observation, while explicitly bridging the embodiment gap.*** Our system highlights the following key technical contributions:

- **HoMMI Data Collection System:** We extend the bimanual UMI framework with a head-mounted camera. Using the iPhone ARKit, the system enables synchronous capture of multi-view video and 6-DoF poses within a unified global coordinate frame.

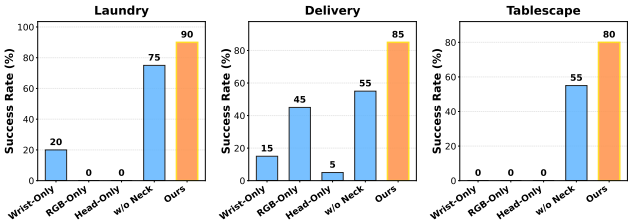

Figure 3. **Quantitative Results.** `Ours` consistently outperforms baselines across all three long-horizon mobile manipulation tasks.

- **Embodiment-Agnostic Vision Representations:** To bridge the observation gap, we use a 3D visual representation for egocentric observations. This allows us to use embodiment-agnostic coordinate frames (i.e., end-effector frame), and remove embodiment-specific observations (e.g., demonstrator's arms and body), mitigating appearance and viewpoint mismatches.
- **Relaxed Head Action Representation:** Since our egocentric representation is view-agnostic, we represent the robot gaze as a "3D look-at point" to bridge the kinematic gap. Compared with directly copying the 6-DoF human head poses, which is often kinematically incompatible with robot, this relaxed action representation enables *effective* transfer of active perception strategies to robots with disparate heights and joint constraints, without sacrificing the end-effector tracking accuracy.
- **Constraint-Aware Whole-Body Control:** We design a whole-body controller that coordinates whole-body motions to *precisely* track end-effector trajectories for accurate manipulation, while respecting the constraints in a bimanual mobile robot for stable and safe motions.

Together, these ideas enable a scalable, in-the-wild human demonstration collection that is directly transferable to real robots. We demonstrate that our system achieves precise, long-horizon, and spatially complex whole-body mobile manipulation tasks, including active search, manipulation, and navigation across large workspaces.

2. Evaluation

We evaluate whether long-horizon mobile manipulation can be learned *directly* from human demonstrations and transferred to a real mobile manipulator. We compare HoMMI to these baselines and ablations, with results in Fig. 3:

- `Wrist-Only(UMI)`: the original UMI [1, 2] setup, using wrist RGBs as input and gripper poses as output.
- `RGB-Only(UMI+Ego)`: naively adding head RGB to the UMI design and predicting gripper and 6-DoF head actions directly.
- `Head-Only`: removing wrist RGBs from `Ours` policy observation and only using the 3D head observation.
- `w/o Active Neck`: running `Ours` policy but disabling head motion control.

References

[1] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. 2

[2] Huy Ha, Yihuai Gao, Zipeng Fu, Jie Tan, and Shuran Song. Umi-on-legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. In *Conference on Robot Learning*, pages 5254–5270. PMLR, 2025. 2

[3] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13226–13233. IEEE, 2025. 2

[4] Justin Yu, Yide Shentu, Di Wu, Pieter Abbeel, Ken Goldberg, and Philipp Wu. Egomi: Learning active vision and whole-body manipulation from egocentric human demonstrations. *arXiv preprint arXiv:2511.00153*, 2025. 2

[5] Qiyuan Zeng, Chengmeng Li, Jude St John, Zhongyi Zhou, Junjie Wen, Guorui Feng, Yichen Zhu, and Yi Xu. Activeumi: Robotic manipulation with active perception from robot-free human demonstrations. *arXiv preprint arXiv:2510.01607*, 2025. 2

[6] Lawrence Y Zhu, Pranav Kuppili, Ryan Punamiya, Patcharapong Aphiwetsa, Dhruv Patel, Simar Kareer, Sehoon Ha, and Danfei Xu. Emma: Scaling mobile manipulation via egocentric human data. *IEEE Robotics and Automation Letters*, 2026. 2