HoMMI: Learning Whole-Body Mobile Manipulation from Human Demonstrations

Anonymous CVPR submission

Paper ID *****

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., egocentric 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

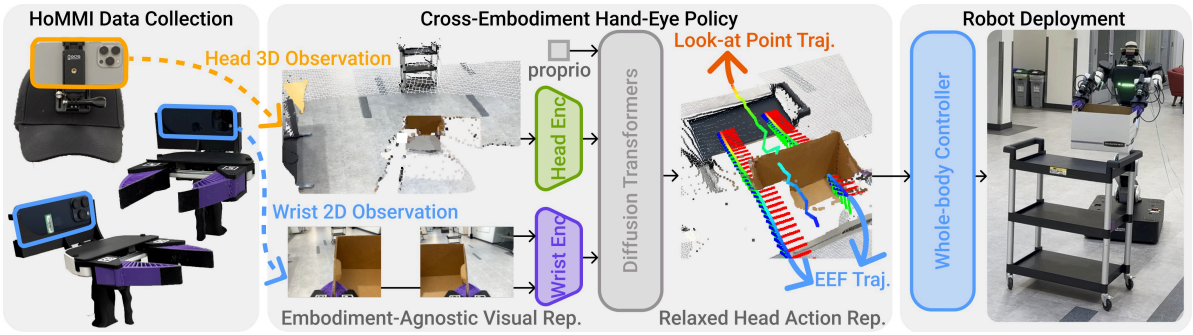

Figure 2. **System Overview.** We learn whole-body mobile manipulation from human demonstrations with an intuitive data collection interface (§ 3), a cross-embodiment policy design with an embodiment-agnostic visual representation and a relaxed head action representation (§ 4), and a whole-body controller that achieves hand-eye tracking through whole-body motions respecting physical constraints (§ 5).

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 [2] 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 [5, 12], or restrict the application domain to fixed-base bimanual manipulation without whole-body coordination [9, 11]. 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.
- **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. Design Objectives

The goal of this paper is to design a general learning from demonstration framework for whole-body mobile manipulation for diverse manipulation tasks. To meet this requirement, we target the following system capabilities:

- *Scalability*: fast, intuitive, and portable demonstration interface for data collection in diverse environments.
- *Transferability*: overcoming both visual and kinematic embodiment gaps from human demonstrators to robots.
- *Whole-body coordination*: efficiently coordinating whole-body action to realize both *precise* end-effector

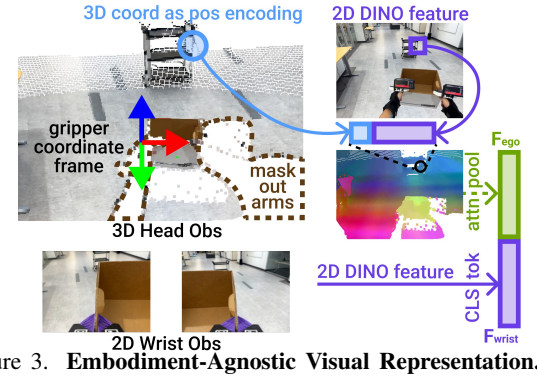

Figure 3. **Embodiment-Agnostic Visual Representation.** We use a 3D representation for egocentric observations that allows using an embodiment-agnostic gripper coordinate frame, and masking out embodiment-specific arms and body observations.

tracking for accurate manipulation and *effective* active perception to gather task-relevant information.

As shown in Fig. 2, we achieve scalability through an intuitive data collection interface (§ 3), transferability through a cross-embodiment hand-eye policy (§ 4), and whole-body motion through a whole-body controller (§ 5) executing policy outputs under physical constraints.

3. HoMMI Data Collection Interface

To enable scalable, robot-free demonstration data collection for bimanual mobile manipulation, we adapt the UMI gripper design while extending it with an egocentric view and head motion capture. Concretely, the data collection system uses three iPhones: two mounted on the grippers and one mounted on a cap (Fig. 2 left). We leverage Apple's ARKit multi-device collaboration to establish a shared coordinate frame across phones. During each demonstration, we record RGB video, depth maps, 6-DoF poses, and gripper widths at 60 Hz on all three iPhones, producing synchronized multimodal trajectories that are directly consumable by our downstream policy learning pipeline (§ 4).

4. Cross-embodiment Hand-Eye Policy

Leveraging the collected data, we train an end-to-end visuomotor policy based on Diffusion Policy [1, 3]. At each time step t, the policy conditions on a short observation window $O_t = o_{t-T_o+1}, \ldots, o_t$ and predicts a horizon of actions $A_t = a_{t+1}, \ldots, a_{t+T_p}$. However, naively adding head RGB and directly predicting head pose substantially enlarges the embodiment gap, often leading to deployment failures. We therefore introduce three key designs: (1) a 3D visual representation, (2) a 3D look-at point action representation, and (3) a gripper-centric observation-action frame. The center of Fig. 2 shows an overview of our policy.

4.1. 3D Visual Repr. to Mitigate the Visual Gap

Head-mounted RGB cameras often exhibit larger viewpoint and appearance differences between the human and robot compared to wrist-mounted cameras. Consequently, instead of directly feeding head RGB to the policy, we lift the egocentric observations into 3D and encode them with geometry-aware tokens, inspired by Adapt3R [7]. As shown in Fig. 3, for each head camera frame, we first obtain a pointmap, then patchify and downsample it via nearest neighbor interpolation s.t. each 16×16 patch corresponds to one 3D point. We then process the RGB frame by extracting a DINO-v3 ViT patch feature [6, 8] for each patch. These features are lifted to 3D by concatenating them with a sinusoidal encoding of the corresponding 3D point, tying appearance to geometry and making the feature robust to head pose and height changes. To further reduce the appearance mismatch, we mask out arm points by transforming the pointmaps into left/right gripper frames and discarding points with $z < 0$, since arms originate behind the grippers. Finally, we use an attention pooling layer to process all tokens and obtain a head observation embedding.

4.2. 3D Look-at Point Action Repr. to Mitigate the Kinematic Gap

Mobile robots have different kinematics than human demonstrators (e.g., shorter torso and fewer degrees of freedom in the neck). As a result, directly mimicking 6-DoF head poses from human data can easily produce infeasible motions. We instead control head motion via a *3D look-at point* $\ell_t \in \mathbb{R}^3$ (Fig. 4). This relaxed representation preserves active perception intent while respecting kinematic constraints (Fig. 5a).

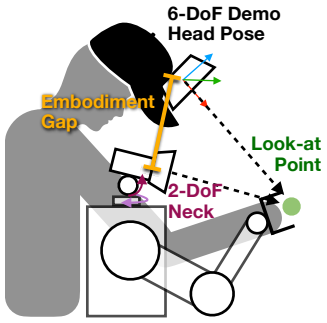

Figure 4. **Look-at Point Action Representation.** To bridge the kinematic gap (e.g., height and neck DoF), we relax the head action constraint by representing the robot gaze as a "3D look-at point". This representation allows effective active perception for gathering task-relevant information without over-constraining the robot to mimic human head motions exactly.

During training, the look-at point is computed as the intersection of the center camera ray with the scene pointmap. At inference, the head controller converts ℓ_t to a feasible head orientation by constructing a rotation whose forward axis points toward ℓ_t. Let $c_t \in \mathbb{R}^3$ be the current head position and let $R_t^{\text{cur}} = \begin{bmatrix} x_t & y_t & z_t \end{bmatrix} \in \mathbb{R}^{3\times3}$ be the current head orientation, where x_t denotes the current head x-axis. We define the desired viewing direction as a unit vector pointing from the current position to the look-at point, $\hat{d}_t = \frac{\ell_t - c_t}{\|\ell_t - c_t\|}$. We then project the current x-axis onto the plane orthogonal to $\hat{d}_t$, $x'_t = x_t - (x_t^\top \hat{d}_t)\hat{d}_t, \hat{x}_t = \frac{x'_t}{\|x'_t\|}$, and construct the remaining axis $\hat{y}_t = \hat{d}_t \times \hat{x}_t$. The target head rotation is then $R_t = \begin{bmatrix} \hat{x}_t & \hat{y}_t & \hat{d}_t \end{bmatrix}$. If $\|x'_t\|$ is near zero, we replace x_t with a fixed world-up vector before projection. This yields a feasible head command without constraining the policy to robot-specific pose limits.

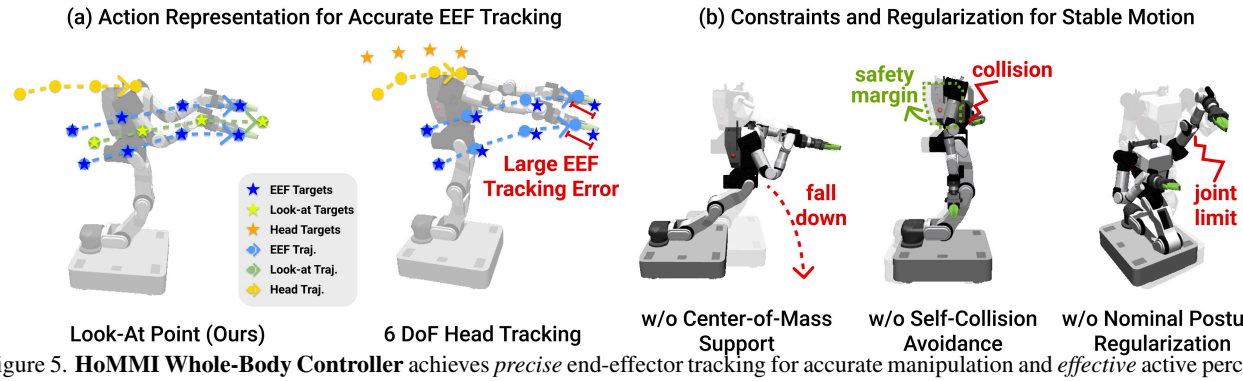

Figure 5. **HoMMI Whole-Body Controller** achieves *precise* end-effector tracking for accurate manipulation and *effective* active perception for information gathering. To do so, it uses (a) a relaxed head look-at point action representation that allows accurate bimanual end-effectors SE(3) tracking, circumventing the infeasibility and increased error associated with simultaneous 6-DoF head-hand tracking. Additionally, we apply (b) constraints and regularization to ensure stability and prevent disastrous behaviors that would otherwise occur.

4.3. Gripper-Centric Frame for Spatial Awareness

Hand-eye coordination requires a reference frame that keeps observations and actions in-distribution. Egocentric frames shift with head motion and embodiment differences (height, neck DoF, camera placement), hurting transfer from human demonstration to robot. We express observations and actions in a gripper-centric frame by transforming gripper poses, head pointmaps, and look-at points to the left-gripper frame, so the policy reasons in a consistent spatial frame. This anchors observation and action to the manipulators, improving spatial awareness and reducing cross-embodiment mismatch over an egocentric frame that drifts with out-of-distribution (OOD) head motion.

5. Constraint-Aware Whole-body Controller

Our policy outputs end-effector poses and look-at points; a whole-body controller solves joint actions and base motions for end-effector tracking. The controller must achieve: accuracy (low tracking error), smoothness (non-jerky motion), stability (no falls or self-collisions), and human-likeness (similar range of motion as the demonstrator).

To satisfy these requirements, we implement a differential whole-body IK solver using `Mink` [10] with (i) high-weight bimanual SE(3) tracking terms to prioritize accuracy, (ii) temporal command interpolation combined with posture and velocity regularization to encourage smooth motions, (iii) explicit constraints and tasks such as torso upright orientation, center-of-mass (CoM) support, and self-collision avoidance, to ensure stability; and (iv) regularization toward a nominal "human" posture and a balanced allocation between arm motion and base motion to produce human-like behavior (Fig. 5b).

Concretely, let $\Delta q \in \mathbb{R}^{n_v}$ be the velocity DoFs, define the objective function $f(\Delta q) = C_{\text{ee}}(\Delta q) + C_{\text{nominal}}(\Delta q) + C_{\text{current}}(\Delta q) + C_{\text{com}}(\Delta q)$. The costs include (1) C_{ee} end-effector pose tracking (primary task); (2) C_{nominal} a nominal posture task to bias toward a preset human-like configuration; (3) C_{current} a current posture task to discourage sudden posture changes; and (4) C_{com} a CoM-over-base task to keep the body mass supported by the base. At each timestep, we solve for Δq using a constrained quadratic program,

$$\begin{aligned} \min_{\Delta q \in \mathbb{R}^{n_v}} \quad & f(\Delta q) + \lambda \left\| \Delta q \right\|_2^2 \\ \text{s.t.} \quad & G_{\text{cfg}} \Delta q \le h_{\text{cfg}}, \quad G_{\text{joint-vel}} \Delta q \le h_{\text{joint-vel}} \\ & G_{\text{base-vel}} \Delta q \le h_{\text{base-vel}}, \quad G_{\text{coll}} \Delta q \le h_{\text{coll}} \\ & A_{\text{upright}} \Delta q = 0 \end{aligned}$$

where λ is the damping coefficient. The inequality constraints $G_j \Delta q \le h_j$ encode configuration bounds G_{cfg}, joint velocity bounds $G_{\text{joint-vel}}$, base velocity bounds $G_{\text{base-vel}}$, and collision avoidance limits G_{coll}. Finally, the equality constraint $A_{\text{upright}} \Delta q = 0$ enforces a zero-sum constraint on the three torso joints for an upright posture.

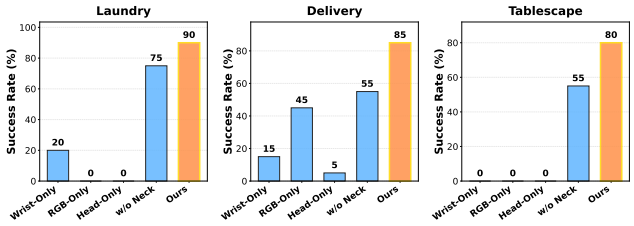

Figure 6. **Quantitative Results.** `Ours` consistently outperforms baselines across all three long-horizon mobile manipulation tasks.

6. 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. 6:

- `Wrist-Only (UMI)`: the original UMI [2, 4] 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] Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, et al. A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*, 2025. 3

[2] 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, 4

[3] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44 (10-11):1684–1704, 2025. 3

[4] 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. 4

[5] 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

[6] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 3

[7] Albert Wilcox, Mohamed Ghanem, Masoud Moghani, Pierre Barroso, Benjamin Joffe, and Animesh Garg. Adapt3r: Adaptive 3d scene representation for domain transfer in imitation learning. In *9th Annual Conference on Robot Learning*, 2025. 3

[8] Xiaomeng Xu, Yanchao Yang, Kaichun Mo, Boxiao Pan, Li Yi, and Leonidas Guibas. Jacobinerf: Nerf shaping with mutual information gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16498–16507, 2023. 3

[9] 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

[10] Kevin Zakka. Mink: Python inverse kinematics based on MuJoCo, 2025. 4

[11] 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

[12] 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