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., 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 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 [5] 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 [14, 38], or restrict the application domain to fixed-base bimanual manipulation without whole-body coordination [34, 36]. 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 head poses from humans, which is often kinematically incompatible with robot hardware, this relaxed action representation enables *effective* transfer of active perception strategies to robots with disparate heights and joint constraints, without sacrificing the tracking accuracy of end-effectors.
- **Constraint-Aware Whole-Body Control:** We design a whole-body controller that can coordinate whole-body motions to *precisely* track end-effector trajectories for accurate manipulation, while respecting the unique constraints in a bimanual mobile robot system 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. Related Work

2.1. Data Collection Interfaces for Robot Learning

Robot learning from demonstrations traditionally relies on teleoperation [4, 26, 28, 30, 31], which yields robot-native data with minimal embodiment gap but is slow, costly, and difficult to deploy for mobile manipulators in diverse environments. UMI [5, 37] addresses scalability by enabling in-the-wild data collection with a portable handheld system. While UMI minimizes the embodiment gap by using wrist-mounted cameras and relative end-effector control, its reliance on wrist-centric sensing fundamentally limits the observability of the global task context. Recent UMI extensions incorporate an external camera [20] or VR headsets [34, 36], but their stationary setups or motion sickness limit their application to fixed-base tasks. In contrast, HoMMI integrates a non-intrusive head-mounted camera into the UMI framework, enabling seamless and scalable deployment in dynamic mobile environments.

2.2. Learning from Egocentric Demonstrations

Egocentric human demonstrations offer a scalable data source for learning bimanual manipulation. Prior works leverage large-scale human videos [3, 15, 33] or utilize wearable devices for scalable data collection [12, 14, 18, 19, 38]. However, they still require co-training or fine-tuning with robot teleoperation data due to the large human-to-robot embodiment gap. In addition to learning bimanual manipulation, recent works further leverage egocentric demonstrations to learn active perception behaviors [7, 28, 34, 36]. However, these approaches assume a robot with a customized 6-DoF neck to directly mimic

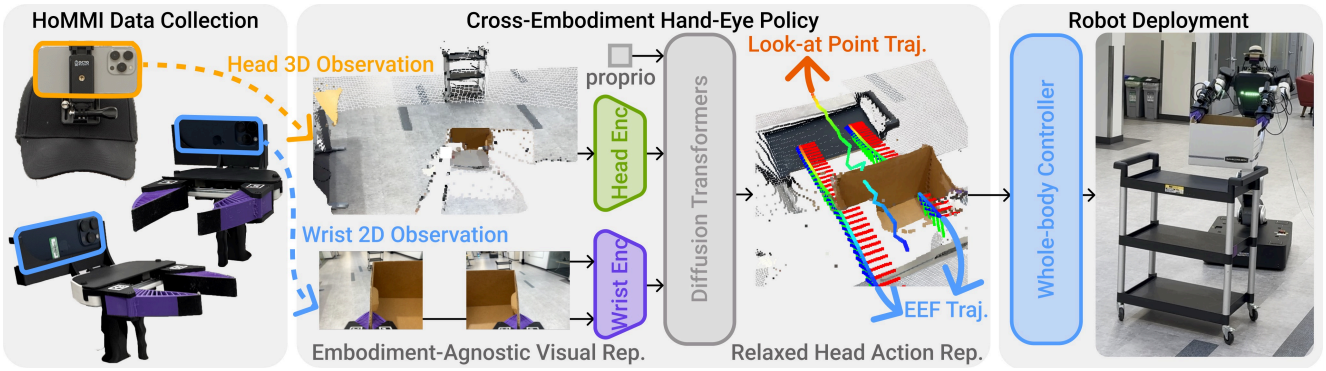

Figure 2. **System Overview.** We learn whole-body mobile manipulation from human demonstrations with an intuitive data collection interface (§ 4), a cross-embodiment policy design with an embodiment-agnostic visual representation and a relaxed head action representation (§ 5), and a whole-body controller that achieves hand-eye tracking through whole-body motions respecting physical constraints (§ 6.2).

human head motions, bypassing the kinematic and action-space gaps between human and robot heads. On the contrary, we leverage a 3D visual representation and a look-at point action abstraction to transfer active perception behaviors from human demonstrations to a standard bimanual mobile manipulator with only a 2-DoF neck.

2.3. Learning-based Mobile Manipulation

Mobile manipulation couples long-range navigation with precise manipulation, making it challenging to learn from human demonstrations. While learning decoupled navigation-manipulation strategies [23, 27, 32] simplifies the problem, these methods limit the ability to imitate end-to-end behaviors directly from human demonstrations. Recent works learn policies that predict end-effector commands, employing a whole-body controller to realize them through coordinated motion [11, 22]. While effective, they have primarily been demonstrated on *single-arm* platforms.

Scaling to the *bimanual* setting introduces distinct challenges, where two-arm coordination, base positioning, and active perception must be synchronized. Although low-cost whole-body interfaces [8, 9, 13] attempt to ease the collection of such coordinated bimanual demonstrations, their dependence on robot teleoperation creates a bottleneck for data scalability. Alternative approaches explore in-the-wild data collection with wearable devices [38], learning from human videos [1], or data generation through simulation [16], yet these methods still require robot teleoperation data for fine-tuning. In contrast, HoMMI allows mobile manipulation directly from robot-free human demonstrations.

3. 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 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 (§ 4), transferability through a cross-embodiment hand-eye policy (§ 5), and whole-body motion through a whole-body controller (§ 6.2) executing policy outputs under physical constraints.

4. 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 (§ 5). The interface is designed to be intuitive and lightweight, providing direct visual and haptic feedback to the operator and avoiding the motion-sickness often associated with VR-based data collection [28, 34, 36].

5. Cross-embodiment Hand-Eye Policy

Leveraging the collected data, we train an end-to-end visuomotor policy based on Diffusion Policy [2, 6]. 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

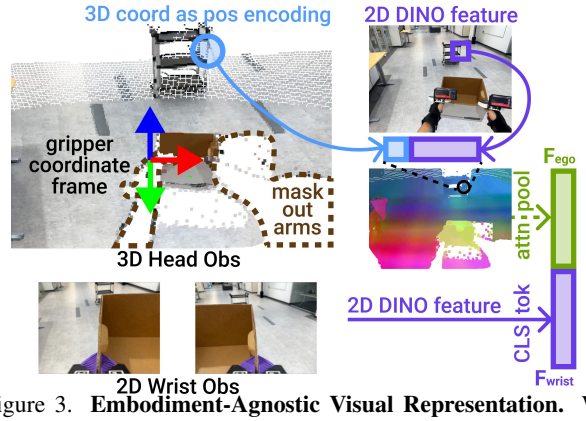

Figure 3. **Embodiment-Agnostic Visual Representation.** We use a 3D representation for egocentric observations that allows using an embodiment-agnostic gripper coordinate frame, and masking out embodiment-specific arms and body observations.

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.

5.1. 3D Visual Representation 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 [25]. Specifically, for each head camera frame, we first obtain a pointmap (from iPhone depth or stereo depth estimation [24] on the robot), and 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 [21, 29] 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.

Fig. 3 illustrates the visual representation of our policy. The observation embedding includes the 3D representation above, a 2D representation for wrist images, and proprioception. Concretely, we finetune a shared `dinov3-vitb16` encoder for wrist and head images. Wrist images are resized to 224×224 and represented by the CLS token features F_{wrist}. The egocentric image is resized to 512×512, split into $32 \times 32 = 1024$ image patches, augmented with 3D positional encoding, and downsampled to 512 tokens; attention pooling (with the arm attention mask) yields F_{ego}.

5.2. 3D Look-at Point Action Representation 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).

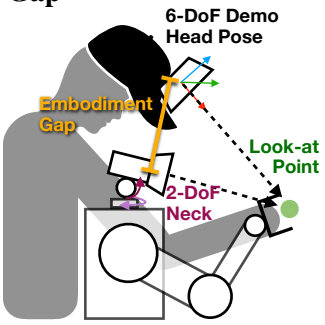

Figure 4. **Look-at Point Action Representation.** To bridge the kinematic gap (e.g., height and neck DoF), we relax the head action constraint by representing the robot gaze as a "3D look-at point". This representation allows effective active perception for gathering task-relevant information without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

5.3. 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 reduc-

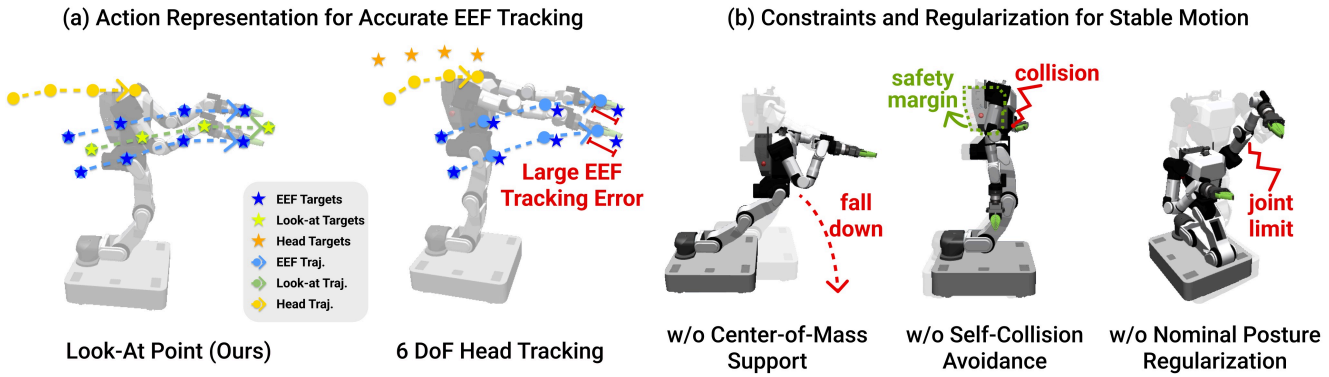

Figure 5. **HoMMI Whole-Body Controller** achieves *precise* end-effector tracking for accurate manipulation and *effective* active perception for information gathering. To do so, it uses (a) a relaxed head look-at point action representation that allows accurate bimanual end-effectors SE(3) tracking, 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.

ing cross-embodiment mismatch over an egocentric frame that drifts with out-of-distribution (OOD) head motion.

6. Robot System

6.1. Mobile Manipulator Hardware Setup

We build a mobile manipulation platform targeting generalizability, observability, and transferability of the policy (Fig. 6). We employ the `Rainbow Robotics RB-Y1` as a core platform, equipped with two 7-DoF arms and a 6-DoF torso on a holonomic base to support diverse mobile manipulation tasks. It also supports active perception via a 2-DoF neck, on which we install a stereo pair of wide-angle cameras (`FLIR BFS-PGE-23S3C-CS`) to capture egocentric context. To align the training and deployment setup, we mount fin-ray fingers identical to the UMI grippers on the end-effectors and mount wrist-mounted cameras (`FLIR BFS-PGE-50S5C-C`) at similar locations.

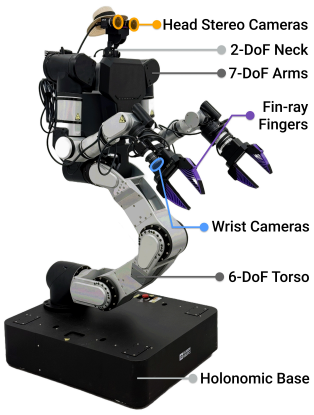

Figure 6. **HoMMI Robot Hardware** features a high DoF bimanual mobile manipulator with customized cameras and fingers that match the HoMMI data collection hardware.

6.2. 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` [35] 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,

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

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. The IK solver runs at 100 Hz to bridge the 10 Hz policy loop and the 500 Hz robot loop.

7. Evaluation

We evaluate whether long-horizon mobile manipulation can be learned *directly* from robot-free human demonstrations

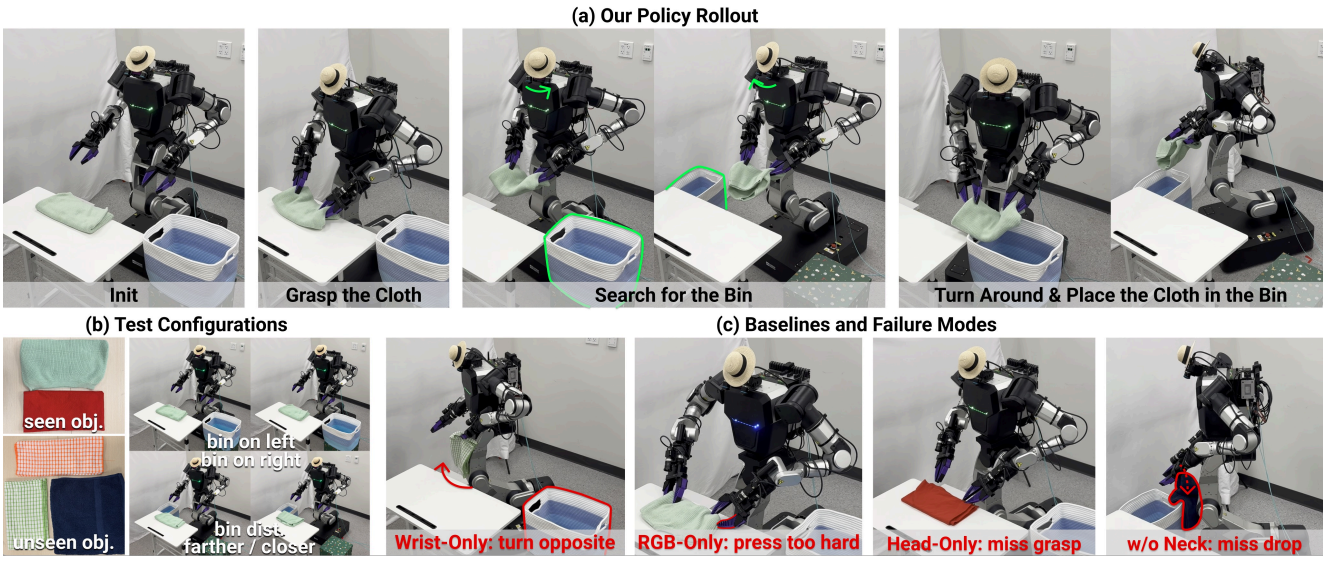

Figure 7. **Laundry Task.** (a) Our cross-embodiment hand-eye policy rollout, highlighting our system's capability of whole-body coordination and active perception. (b) Different test scenarios with different objects and bin locations. (c) Typical failure cases of the baselines.

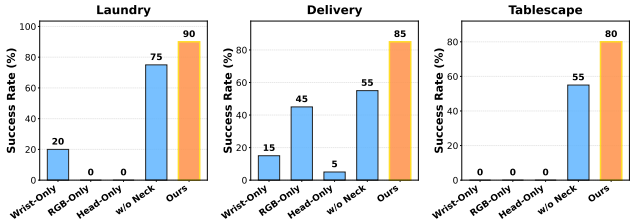

Figure 8. **Quantitative Results.** `Ours` consistently outperforms baselines across all three long-horizon mobile manipulation tasks.

and transferred to a real mobile manipulator. Specifically, our evaluation probes four core capabilities:

- **Cross-embodiment transfer**: deploying policies learned from robot-free human demonstrations on a robot with a different appearance and kinematics.
- **Whole-body coordination**: coordinating two arms, mobile base, torso, and head for mobile manipulation.
- **Long-horizon navigation**: moving through a large workspace and approaching targets whose locations vary.
- **Active perception**: intentionally controlling head motion to acquire task-relevant information that may initially be outside the field of view.

We compare HoMMI to these baselines and ablations:

- `Wrist-Only(UMI)`: the original UMI [5, 11] 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. This setup is similar to ViA [28], however, we use a wearable UMI device for data collection instead of teleoperating the same robot embodiment, which provides better scalability but also introduces additional challenges in cross-embodiment policy learning.
- `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.

7.1. Laundry Task

Task: As shown in Fig. 7a, the robot approaches a table, grasps a cloth with both hands, searches for a bin, navigates, and places the cloth in the bin. The success rate is measured by whether the cloth is placed in the bin in the end.

Capability: Bimanual coordination: The robot must grasp the cloth firmly with both hands. Whole-body coordination: The bin is placed to the side and lower than the table, thus requiring the robot to flexibly coordinate whole-body motion to navigate, rotate, and bend down to approach the bin and place the cloth into it. Active perception: The bin may be outside the camera view after grasping, thus requiring the robot to actively search for it by looking sideways.

Data Collection: We collected 200 demonstrations with randomized bin locations, robot and object configurations.

Test Scenarios: As shown in Fig. 7b, we ran a total of 20 rollouts, involving 5 objects (2 seen and 3 unseen) and 4 bin configurations (2 on the left and 2 on the right).

Performance: Quantitative results are shown in Fig. 8 (left). `Ours` achieves a 90% success rate. It flexibly coordinates whole-body motion to navigate the workspace and place the cloth in the bin, robustly searches the environment to find the bin, and always turns correctly. Occasional failures result from not grasping the cloth firmly enough, causing it to slip halfway. Fig. 7c shows the baselines' typical failure modes. (1) `Wrist-Only` policy's dominant failure is consistently turning to one side, regardless of the bin location, due to the bin not being visible from the wrist camera view. Other failure cases include inaccurately placing the cloth in the bin. We hypothesize that these issues are due to the lack of global context and spatial information from wrist views. (2) `RGB-Only` consistently fails to

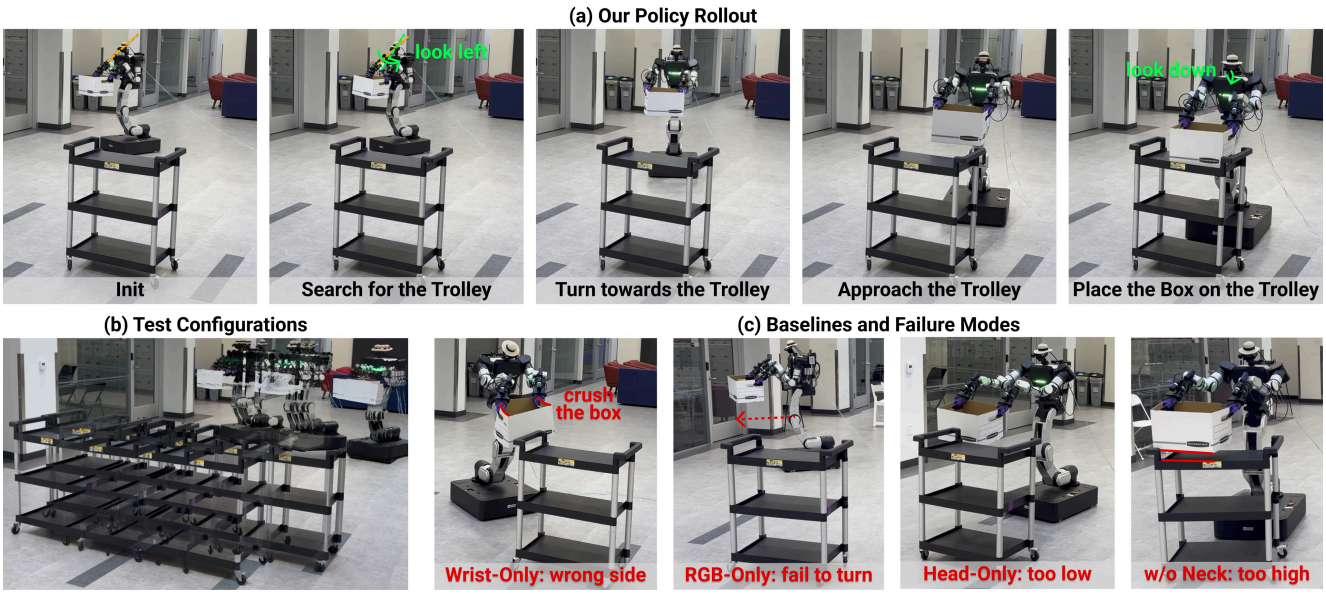

Figure 9. **Delivery Task.** (a) Our policy rollout, demonstrating long-horizon navigation over a large workspace and active perception. (b) Different test scenarios with different trolley locations and initial base positions and orientations. (c) Typical failure cases of the baselines.

grasp and presses the table too hard, triggering the robot's wrench safety guard, which we hypothesize is due to egocentric RGB having appearance and viewpoint mismatches in human and robot observations, causing the policy to go OOD. (3) `Head-Only`'s success rate is also 0%, failing due to missing the cloth when attempting to grasp it, and unstable grasps. Compared with `Ours`, this demonstrates that wrist cameras help provide local contact information that can improve grasping accuracy. (4) `w/o Active Neck` achieves a 75% success rate, mostly failing to accurately place the cloth into the bin. We hypothesize that the lack of active perception causes the view to be more OOD and the bin to be not fully in view.

7.2. Delivery Task

Task: As shown in Fig. 9a, the robot carries a box, searches for a trolley, navigates over a large workspace, and places the box onto the trolley. The task success rate is measured by whether the box is eventually placed onto the trolley.

Capability: Bimanual coordination: Two hands maintain a stable distance to avoid crushing the box and coordinate heights for accurate placement. Long-horizon navigation: The robot navigates a 6×6m workspace and accurately approaches the trolley in randomized locations. Active perception: The trolley may initially be out of view, requiring the robot to search, rotate, and navigate over.

Data Collection: We collected 166 demonstrations with varying trolley locations and initial standing locations.

Test Scenarios: As shown in Fig. 9b, we conducted 20 rollouts with 5 trolley locations and 4 initial robot base initializations (position + yaw).

Performance: Quantitative results are shown in Fig. 8 (middle). `Ours` achieves 85% success. The policy robustly performs visual servoing and long-horizon navigation, always approaching the trolley from the correct direction. It also reactively adjusts midway if initially misaligned. Failures are due to slight misalignment after long navigation. Typical baseline failure modes are shown in Fig. 9c. (1) `Wrist-Only` achieves 15%, frequently approaching from the wrong side or misaligning during placement, showing that navigation requires global context beyond wrist views. (2) `RGB-Only` achieves 45%. The policy consistently fails to turn towards the trolley when it is initially out of view because 6-DoF head motion commands are kinematically infeasible for the whole-body IK. (3) `Head-Only` achieves 5%, often colliding the box with the trolley because the gripper heights are too low. This highlights that egocentric context alone is insufficient for manipulation precision. (4) `w/o Neck` achieves 55%, often lifting the box too high during final placement due to the lack of a look down head motion.

7.3. Tablescape Task

Task: As shown in Fig. 10a, the robot approaches a table and grasps the two edges of a mat, lifts it up and moves forward to unfold it, and finally lays the mat flat on the table and retracts its hands.

Capability: Bimanual coordination: two hands coordinate rotation and height to grasp the mat edges and maintain stable distance to unfold it. Whole-body coordination: the robot coordinates base, torso, and arm motions to navigate and adjust gripper height throughout the task.

Data Collection: We collected 115 demonstrations with varying initial standing locations and mat placements.

Test Scenarios: We ran 20 rollouts in total, including 5 initial base initializations and 2 mat configurations, and tried

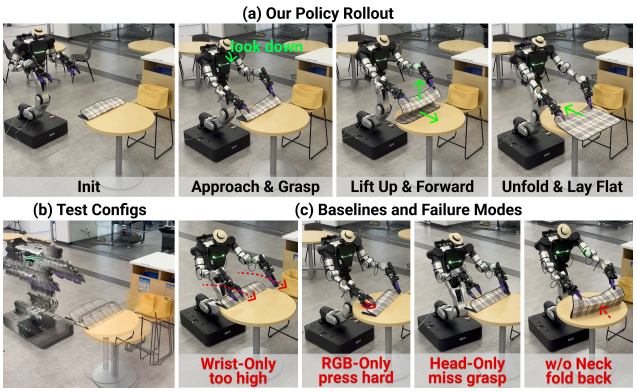

Figure 10. **Tablescape Task.** (a) Our policy rollout, demonstrating precise bimanual and whole-body coordination. (b) Different test scenarios with different initial base positions and mat placement. (c) Typical failure cases of the baselines.

twice for each configuration (Fig. 10b).

Performance: Quantitative results are shown in Fig. 8 (right). `Ours` achieves 80% success, demonstrating robust recovery. When the mat is misaligned or folds back, the robot retries until it unfolds. Occasional failures arise from slightly missing the grasp. Fig. 10c shows the failure modes of the baselines. (1) `Wrist-Only` grippers rotate too late and go well above the mat, which we hypothesize is due to the lack of global spatial context. (2) `RGB-Only` presses the grippers too hard against the table, potentially due to OOD head observations. (3) `Head-Only` misses contact with the mat, demonstrating the need for wrist cameras to provide local contact information. (4) `w/o Neck` achieves 55% success, with failures from missing the grasp and failing to recover when the mat folds back, due to the inability to actively adjust the viewpoint.

7.4. Findings Summary

F1: Wrist-only sensing under-observes global task context and bimanual coordination. `Wrist-Only` achieves poor performance across all tasks: it cannot actively search for task-relevant context due to its limited field of view, drifts in long-horizon navigation without global task progress, and lacks spatial awareness of the other hand, causing bimanual coordination failures. HoMMI augments UMI with *egocentric sensing*, providing global context and active perception crucial for mobile manipulation.

F2: Head-mounted camera alone is insufficient. While being the most common humanoid camera configuration [4, 10, 17], the `Head-Only` baseline fails in grasping and alignment. HoMMI combines head and wrist cameras, which provide essential local contact cues for fine-grained manipulation. Joint finetuning of the vision encoder on wrist and egocentric images also yields cleaner task-relevant egocentric attention (Fig. 11).

F3: Naively adding egocentric RGB can degrade performance under embodiment mismatch. Directly feeding head RGB and regressing 6-DoF head motion leads to 0% success on two tasks from significant OOD shift. HoMMI bridges the visual gap with an embodiment-agnostic 3D egocentric representation, and the kinematic gap through a relaxed look-at point representation, enabling precise end-effector tracking and effective active perception.

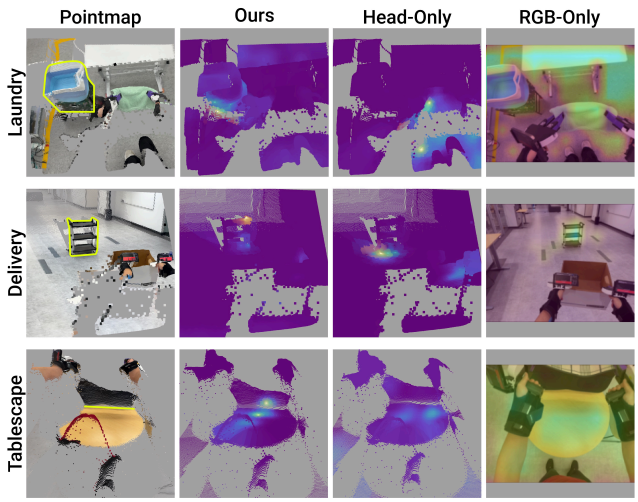

Figure 11. **Egocentric Attention Comparison.** `Ours` exhibits clean attention on task-relevant objects (yellow = higher attention), while baselines' attentions are less informative.

F4: Active head control effectively gathers task-relevant information and maintains policy observability. Disabling head motion reduces success, particularly when actively searching for the object and precisely placing or aligning objects. This supports that look-at point based head control imitates active perception to gather task-relevant information and keep the egocentric view in-distribution.

F5: Our cross-embodiment hand-eye policy learns task-relevant attention. As shown in Fig. 11, `Ours` yields egocentric attention highlighted on task-relevant objects, demonstrating the effectiveness of our observation representation and gripper-centric frame. This helps mitigate the visual embodiment gap, as the policy attends to task-relevant regions over OOD observation points.

8. Conclusion

We present HoMMI, a system that enables learning long-horizon whole-body mobile manipulation skills directly from robot-free human demonstrations. We employ a scalable data collection interface that augments bimanual UMI with egocentric sensing. To bridge the human-to-robot embodiment gap induced by egocentric sensing, we propose a cross-embodiment hand-eye policy design with an embodiment-agnostic visual representation and a relaxed look-at point head action representation. A whole-body controller then achieves precise bimanual end-effector tracking and effective active perception by coordinating the robot's whole-body motions while respecting its physical constraints. Extensive real world experiments demonstrate that HoMMI allows for versatile and challenging mobile manipulation tasks.

References

[1] Arpit Bahety, Arnav Balaji, Ben Abbatematteo, and Roberto Martín-Martín. Safemimic: Towards safe and autonomous human-to-robot imitation for mobile manipulation. In *RSS 2025 Workshop: Mobile Manipulation: Emerging Opportunities & Contemporary Challenges*, 2025. 3

[2] 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

[3] Xiongyi Cai, Ri-Zhao Qiu, Geng Chen, Lai Wei, Isabella Liu, Tianshu Huang, Xuxin Cheng, and Xiaolong Wang. In-n-on: Scaling egocentric manipulation with in-the-wild and on-task data. *arXiv preprint arXiv:2511.15704*, 2025. 2

[4] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. In *Conference on Robot Learning*, pages 2729–2749. PMLR, 2025. 2, 8

[5] 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, 6

[6] 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

[7] Ian Chuang, Jinyu Zou, Andrew Lee, Dechen Gao, and Iman Soltani. Look, focus, act: Efficient and robust robot learning via human gaze and foveated vision transformers. *arXiv preprint arXiv:2507.15833*, 2025. 2

[8] Shivin Dass, Wensi Ai, Yuqian Jiang, Samik Singh, Jiaheng Hu, Ruohan Zhang, Peter Stone, Ben Abbatematteo, and Roberto Martín-Martín. Telemoma: A modular and versatile teleoperation system for mobile manipulation. In *RSS 2024 Workshop: Data Generation for Robotics*. 3

[9] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation using low-cost whole-body teleoperation. In *8th Annual Conference on Robot Learning*, 2024. 3

[10] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. In *Conference on Robot Learning*, pages 2828–2844. PMLR, 2025. 8

[11] 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. 3, 6

[12] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. 2

[13] Yunfan Jiang, Ruohan Zhang, Josiah Wong, Chen Wang, Yanjie Ze, Hang Yin, Cem Gokmen, Shuran Song, Jiajun Wu, and Li Fei-Fei. BEHAVIOR robot suite: Streamlining real-world whole-body manipulation for everyday household activities. In *9th Annual Conference on Robot Learning*, 2025. 3

[14] 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

[15] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from in-the-wild human videos using data-editing. In *Human to Robot: Workshop on Sensorizing, Modeling, and Learning from Humans*, 2025. 2

[16] Chengshu Li, Mengdi Xu, Arpit Bahety, Hang Yin, Yunfan Jiang, Huang Huang, Josiah Wong, Sujay Garlanka, Cem Gokmen, Ruohan Zhang, et al. Momagen: Generating demonstrations under soft and hard constraints for multi-step bimanual mobile manipulation. In *RSS 2025 Workshop: Mobile Manipulation: Emerging Opportunities and Contemporary Challenges*. 3

[17] Boxiao Pan, Adam W Harley, Francis Engelmann, C Karen Liu, and Leonidas J Guibas. Lookout: Real-world humanoid egocentric navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24977–24988, 2025. 8

[18] Ryan Punamiya, Dhruv Patel, Patcharapong Aphiwetsa, Pranav Kuppili, Lawrence Y Zhu, Simar Kareer, Judy Hoffman, and Danfei Xu. Egobridge: Domain adaptation for generalizable imitation from egocentric human data. In *Human to Robot: Workshop on Sensorizing, Modeling, and Learning from Humans*, 2025. 2

[19] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J. Yoon, Ryan Hoque, Lars Paulsen, Ge Yang, Jian Zhang, Sha Yi, Guanya Shi, and Xiaolong Wang. Humanoid policy $\sim$ human policy. In *9th Annual Conference on Robot Learning*, 2025. 2

[20] Omar Rayyan, John Abanes, Mahmoud Hafez, Anthony Tzes, and Fares Abu-Dakka. Mv-umi: A scalable multi-view interface for cross-embodiment learning. *arXiv preprint arXiv:2509.18757*, 2025. 2

[21] 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. 4

[22] Priya Sundaresan, Rhea Malhotra, Phillip Miao, Jingyun Yang, Jimmy Wu, Hengyuan Hu, Rika Antonova, Francis Engelmann, Dorsa Sadigh, and Jeannette Bohg. Homer: Learning in-the-wild mobile manipulation via hybrid imitation and whole-body control. *arXiv preprint arXiv:2506.01185*, 2025. 3

[23] Shagun Uppal, Ananye Agarwal, Haoyu Xiong, Kenneth Shaw, and Deepak Pathak. Spin: Simultaneous perception interaction and navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18133–18142, 2024. 3

[24] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-

shot stereo matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5249–5260, 2025. 4

[25] 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. 4

[26] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163. IEEE, 2024. 2

[27] Haoyu Xiong, Russell Mendonca, Kenneth Shaw, and Deepak Pathak. Adaptive mobile manipulation for articulated objects in the open world. *arXiv preprint arXiv:2401.14403*, 2024. 3

[28] Haoyu Xiong, Xiaomeng Xu, Jimmy Wu, Yifan Hou, Jeannette Bohg, and Shuran Song. Vision in action: Learning active perception from human demonstrations. In *9th Annual Conference on Robot Learning*, 2025. 2, 3, 6

[29] 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. 4

[30] Xiaomeng Xu, Dominik Bauer, and Shuran Song. RoboPanoptes: The All-Seeing Robot with Whole-body Dexterity. In *Proceedings of Robotics: Science and Systems*, 2025. 2

[31] Xiaomeng Xu, Yifan Hou, Zeyi Liu, and Shuran Song. Compliant residual DAgger: Improving real-world contact-rich manipulation with human corrections. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. 2

[32] Jingyun Yang, Isabella Huang, Brandon Vu, Max Bajracharya, Rika Antonova, and Jeannette Bohg. Mobi-π: Mobilizing your robot learning policy. In *9th Annual Conference on Robot Learning*, 2025. 3

[33] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025. 2

[34] 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, 3

[35] Kevin Zakka. Mink: Python inverse kinematics based on MuJoCo, 2025. 5

[36] 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, 3

[37] Zhaxizhuom Zhaxizhuoma, Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan CHEN, Pingrui Zhang, et al. Fastumi: A scalable and hardware-independent universal manipulation interface with dataset. In *Conference on Robot Learning*, pages 3069–3093. PMLR, 2025. 2

[38] 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, 3