

HoMMI: Learning Whole-Body Mobile Manipulation from Human Demonstrations

Xiaomeng Xu^{1,2} Jisang Park¹ Han Zhang¹ Eric Cousineau² Aditya Bhat² Jose Barreiros²
Dian Wang¹ Jeannette Bohg¹ Shuran Song¹

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



Fig. 1: **Whole-Body Mobile Manipulation Interface (HoMMI)**. (a) We extend UMI with egocentric sensing to enable scalable *mobile* manipulation with *active perception* – capabilities that cannot be achieved with the original UMI. (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.

I. 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 [16, 50], or restrict the application domain to fixed-base bimanual manipulation without whole-body coordination [46, 44]. 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. By integrating the iPhone ARKit, the system enables synchronous capture of multi-view video and 6-DoF poses within a unified and globally consistent 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.

II. RELATED WORK

A. Data Collection Interfaces for Robot Learning

Robot learning from demonstrations traditionally relies on teleoperation [35, 39, 41, 37, 4], 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, 48] 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 [27] or VR headsets [46, 44], 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.

B. Robot Learning from Egocentric Demonstrations

Egocentric human demonstrations offer a scalable data source for learning bimanual manipulation. Prior works leverage large-scale human videos [18, 43, 3] or utilize wearable devices for scalable data collection [16, 24, 50, 14, 26, 25]. 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 [37, 46, 44, 8]. However, these approaches assume a robot with a customized 6-DoF neck to directly mimic 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.

C. Learning Mobile Manipulation From Demonstrations

Mobile manipulation couples long-range navigation with precise manipulation, making it challenging to learn from human demonstrations. While learning decoupled navigation-manipulation strategies [36, 32, 42] 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 [12, 31]. While effective, these pipelines 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 [10, 9, 15] attempt to ease the collection of such coordinated bimanual demonstrations, their dependence on robot teleoperation creates a bottleneck for

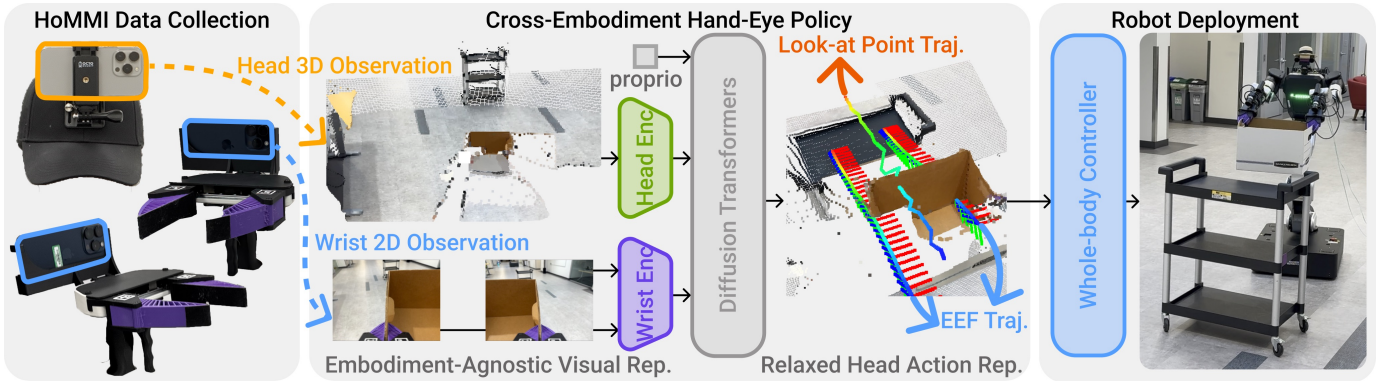


Fig. 2: **HoMMI System Overview.** We learn whole-body mobile manipulation from human demonstrations with an intuitive data collection interface (§ IV), a cross-embodiment policy design with an embodiment-agnostic visual representation and a relaxed head action representation (§ V), and a whole-body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

data scalability. Alternative approaches explore in-the-wild data collection with wearable devices [50], learning from human videos [1], or automated data generation through simulation [19], yet these methods still require robot teleoperation data for fine-tuning. In contrast, HoMMI allows mobile manipulation directly from robot-free human demonstrations.

III. 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*: able to efficiently coordinate whole-body action to realize both *precise* end-effector tracking for accurate manipulation and *effective* active perception to intentionally gather task-relevant information.

Fig. 2 shows an overview of our system. We achieve scalability through an intuitive data collection interface (§ IV), transferability through a cross-embodiment hand-eye policy trained on the collected demonstrations (§ V), and whole-body motion through a whole-body controller (§ VI-B) that executes policy outputs under the robot’s physical constraints.

IV. 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 60Hz on all three iPhones, producing synchronized multimodal trajectories that are directly consumable by our downstream policy learning pipeline (§ V). The interface is designed to be intuitive and lightweight, providing direct visual and haptic feedback to

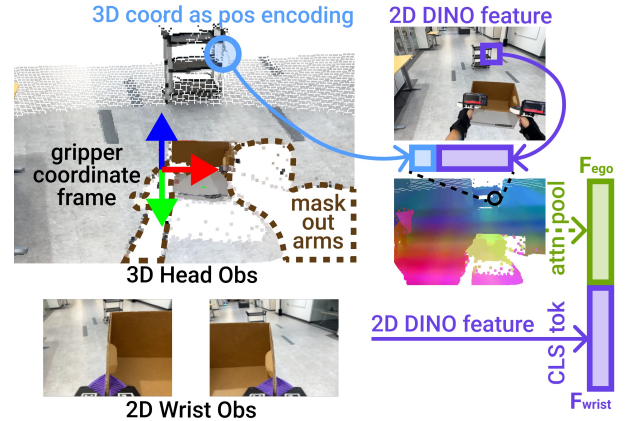


Fig. 3: **Embodiment-Agnostic Visual Representation.** We use a 3D representation for egocentric observations that allows using an embodiment-agnostic gripper coordinate frame, and masking out embodiment-specific arms and body observations.

the operator and avoiding the motion-sickness often associated with VR-based data collection [44, 46, 37].

V. 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 $O_t = o_{t-T_o+1}, \dots, o_t$ and predicts a horizon of actions $A_t = a_{t+1}, \dots, a_{t+T_p}$. However, naively adding the head RGB image to the observation and directly predicting the head pose as part of the action substantially enlarges the human-robot embodiment gap, often leading to failures in deployment. We therefore introduce three key algorithmic designs that overcome these transferability challenges, including (1) a 3D visual representation, (2) a 3D look-at point action representation, and (3) a gripper-centric frame shared by observations and actions. The center of Fig. 2 shows an overview of our policy.

A. 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 [34]. Specifically, for each head camera frame, we first obtain a pointmap (from iPhone depth or stereo depth estimation [33] 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 [28, 38] for each patch. These patch features are further lifted to 3D by concatenating them with a sinusoidal encoding of the corresponding 3D point in the downsampled pointmap, tying appearance feature to 3D 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. In the end, 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 entire observation embedding includes the 3D representation mentioned 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} .

B. 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).

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 $R_t^{cur} = [x_t \ y_t \ z_t] \in \mathbb{R}^{3 \times 3}$ 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

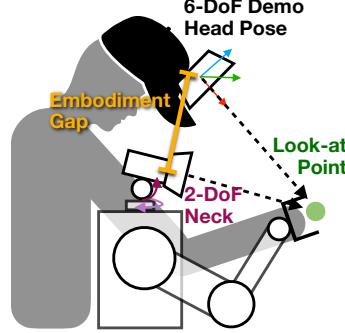


Fig. 4: **Look-at Point Action Representation.** To bridge the kinematic gap (e.g., height and neck DoF), we relax the head action constraint by representing the robot gaze as a “3D look-at point”. This representation allows effective active perception for gathering task-relevant information without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

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 $R_t = [\hat{x}_t \ \hat{y}_t \ \hat{d}_t]$. 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.

C. Gripper-Centric Frame for Spatial Awareness

In our system, 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), which hurts transfer from human demonstration to robot. We therefore express all observations and actions in a gripper-centric frame by transforming gripper poses (both proprioception and action), head pointmaps, and look-at points to the left-gripper frame, so the policy always reasons in a consistent spatial frame centered at the manipulator. This anchors observation and action to the manipulators that execute the task, improving spatial awareness for 3D representations and reducing cross-embodiment mismatch compared to an egocentric frame that drifts with out-of-distribution (OOD) head motion.

VI. ROBOT SYSTEM

A. Bimanual Mobile Manipulator Hardware Setup

We build a mobile bi-manipulation platform targeting generalizability, observability, and transferability of the learned 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 industrial-grade wide-angle cameras (FLIR BFS-PGE-23S3C-CS) to capture egocentric context. To align 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.

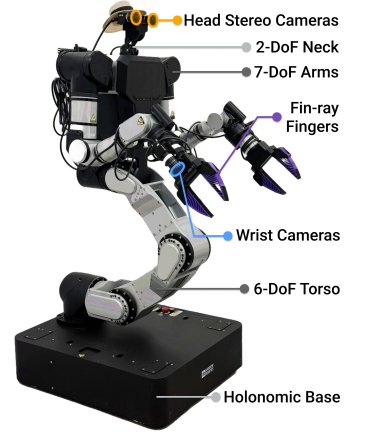


Fig. 6: **HoMMI Robot Hardware** features a high DoF bimanual mobile manipulator with customized cameras and fingers that match the HoMMI data collection hardware.

B. Constraint-Aware Whole-body Controller

Since our policy outputs end-effector poses and head look-at points, we need a whole-body controller to solve whole-body joint actions and base motions to achieve the cartesian space end-effector trajectory commands. Specifically, the whole-body controller needs to meet these requirements: accuracy

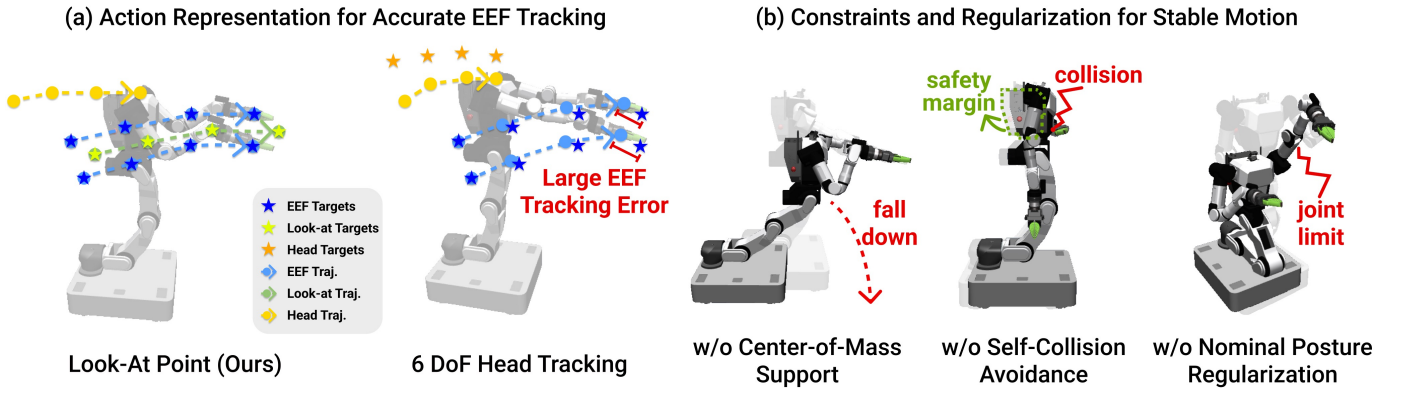


Fig. 5: **HoMMI Whole-Body Controller** is designed to achieve *precise* end-effector tracking for accurate manipulation and *effective* active perception for information gathering. To do so, it uses (a) a relaxed head look-at point action representation that allows accurate bimanual end-effectors SE(3)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

(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 [45] 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 $f(\Delta q) = C_{ee}(\Delta q) + C_{nominal}(\Delta q) + C_{current}(\Delta q) + C_{com}(\Delta q)$. 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,

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

where λ is the damping coefficient. The inequality constraints $G_j \Delta q \leq h_j$ encode configuration bounds G_{cfg} , joint velocity bounds $G_{joint-vel}$, base velocity bounds $G_{base-vel}$, and collision avoidance limits G_{coll} . Finally, the equality constraint $A_{upright} \Delta q = 0$ enforces a zero-sum constraint on the three torso joints for an upright posture. Together, these tasks and constraints balance accuracy, smoothness, stability, and human-likeness in a single optimization. Importantly, the look-at point is controlled separately from the whole-body IK and

is excluded from the IK objective to avoid interfering with bimanual EEF tracking. As described in Sec. V-B, the neck orientation is computed to point toward the policy-predicted look-at point from the robot’s current head position.

We run this IK solver asynchronously at 100 Hz to bridge the 10 Hz policy loop and the 500 Hz robot control loop. The policy produces a stream of target end-effector poses with specified command durations (0.1 s). At each IK tick, we compute an interpolated target by linearly blending the previous and current targets based on the elapsed fraction of the command duration. This reduces discontinuities at policy update boundaries and improves tracking smoothness.

C. Asynchronous Policy Inference

Mobile manipulation cannot be paused for inference without introducing base jerks and tracking errors, so we decouple perception, policy inference, and whole-body control. We run a *detached policy server* that receives timestamped observations, performs inference, and returns a timestamped action chunk; and a *real-time execution bridge* that aligns observations across sensors to prepare timestamped observations for the policy, receives actions from the policy, filters stale actions, and streams time-aligned targets to the whole-body controller.

At each inference cycle, the bridge collects a history of camera frames and proprioception, corrects each camera stream by a measured latency, and then uses the *latest* camera timestamp as the anchor. Proprioception is interpolated to these anchor timestamps. This yields a *synchronized observation window*, similar to the latency matching in UMI [5]. The policy server takes in these observations, runs policy inference, and outputs a horizon of actions whose timestamps are anchored to the last observation time. The bridge then discards any actions whose timestamps fall *before* the earliest feasible execution time, given inference time and execution latency, and updates the *scheduled action buffer* with the remaining actions. These buffered actions are streamed at 10 Hz to provide latency-matched targets to the whole-body controller.

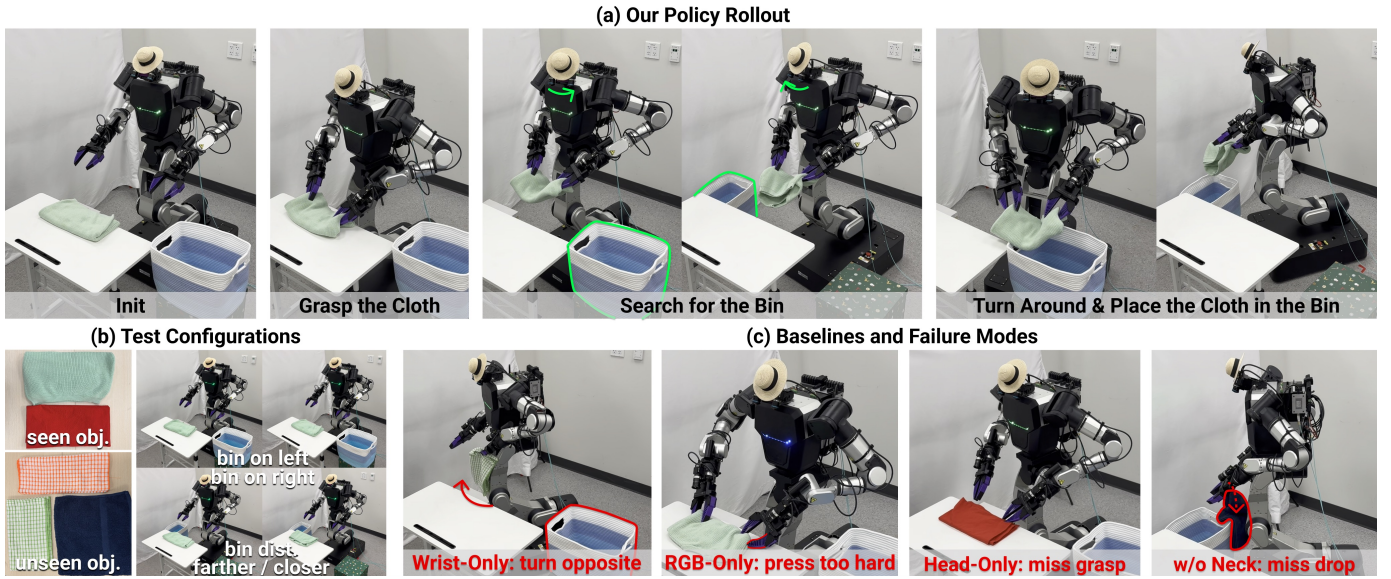


Fig. 7: **Laundry Task.** (a) Our cross-embodiment hand-eye policy rollout, highlighting our system’s capability of whole-body coordination and active perception. (b) Different test scenarios with different objects and bin locations. (c) Typical failure cases of the baselines.

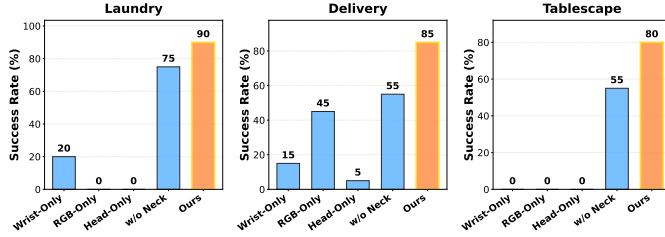


Fig. 8: **Quantitative Results.** Ours consistently outperforms baselines across all three long-horizon mobile manipulation tasks.

VII. EVALUATION

We evaluate whether long-horizon bimanual mobile manipulation can be learned *directly* from robot-free human demonstrations 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. Required for all tasks.
- **Bimanual / 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 the following baselines and ablations:

- **Wrist-Only (UMI):** the original UMI [5, 12] setup, using only wrist RGBs as input and gripper trajectories 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 [37], 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.

A. 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 task 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 until it locates it.

Data Collection: We collected 200 demonstrations with randomized bin locations, initial configurations, and objects.

Test Scenarios: As shown in Fig. 7b, we ran evaluation across 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 not grasping firmly enough and inaccurately placing the cloth in the bin. We hypothesize that these issues are due to the lack of global context and spatial

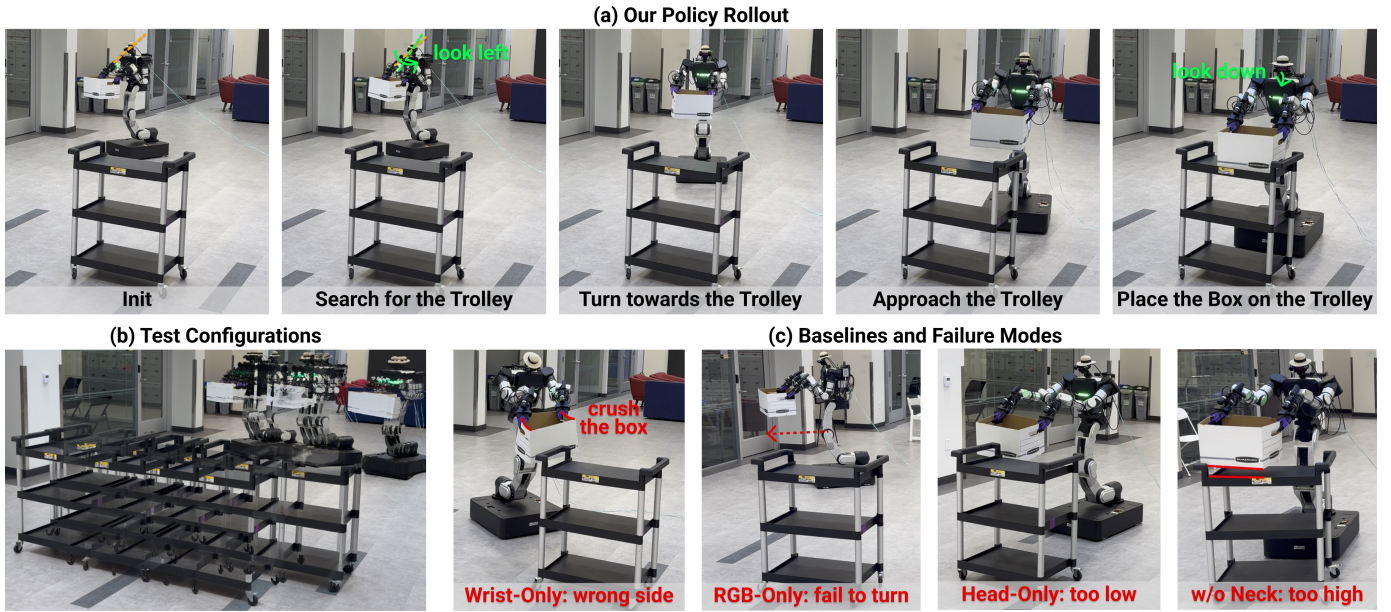


Fig. 9: **Delivery Task.** (a) Our policy rollout, demonstrating long-horizon navigation over a large workspace and active perception. (b) Different test scenarios with different trolley locations and initial base positions and orientations. (c) Typical failure cases of the baselines.

information from wrist views. (2) RGB-Only consistently fails to 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 only grasping one edge which leads to slip off later. 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.

B. 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 need to maintain a stable distance to avoid crushing or tearing the box. They also need to coordinate heights to lift the box up and then lower it for accurate placement. Long-horizon navigation: The robot needs to navigate a large workspace (6×6m) and accurately approach the trolley in randomized locations. Active perception: The trolley may initially be out of view when the robot is rotated to face the other way, requiring the robot to search for the trolley, rotate, and then 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 in total, consisting of 5 trolley locations and 4 different 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 correct direction towards the trolley. It also reactively adjusts the robot’s approaching direction midway if the initial alignment is inaccurate. The remaining failures are due to slight misalignment at the end after long navigation.

Typical baseline failure modes are shown in Fig. 9c. (1) Wrist-Only achieves 15%, the policy frequently approaches from an incorrect side or misaligns during placement, demonstrating that navigation and approach require 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 active head motion commands are unachievable by the whole-body IK due to kinematic infeasibility. (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.

C. Tablescape Task

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

Capability: Bimanual coordination: two hands need to precisely coordinate rotation and height to grasp the edges of the mat, and consistently maintain a stable distance and height to unfold it. Whole-body coordination: The robot needs to coordinate the base, torso, and arm motions to navigate across the workspace and significantly adjust the height of the grippers throughout the task.

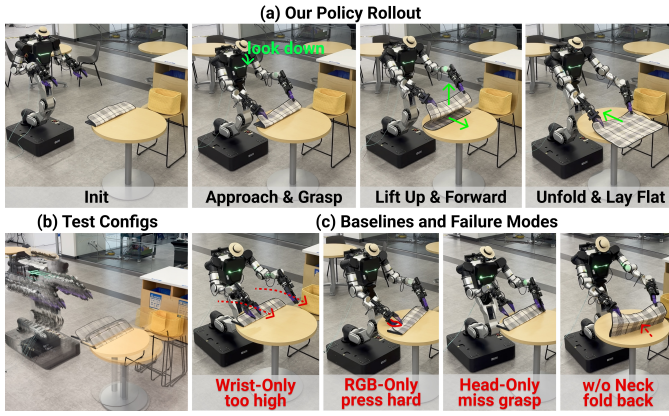


Fig. 10: **Tablescape Task.** (a) Our policy rollout, demonstrating precise bimanual and whole-body coordination. (b) Different test scenarios with different initial base positions and mat placement. (c) Typical failure cases of the baselines.

Data Collection: We collected 115 demonstrations with varying initial standing locations and mat placements on the table. **Test Scenarios:** We ran 20 rollouts in total, including 5 initial base initializations and 2 mat configurations, and tried twice for each configuration (Fig. 10b).

Performance: Quantitative results are shown in Fig. 8 (right). Ours achieves an 80% success rate, demonstrating robust recovery behaviors. When the mat is not perfectly aligned with the table or folds back on the first attempt, the robot lifts the mat again and retries until it successfully unfolds. Occasional failure cases 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 resulting from missing the grasp and failing to recover when the mat folds back, potentially due to the inability to actively adjust the viewpoint for better alignment observation.

D. Findings Summary

F1: Wrist-only sensing under-observes global task context and bimanual coordination. Wrist-Only policy exhibits poor performance on all tasks that require search, navigation, and alignment, which depend on the wider scene. It is not capable of actively searching for task-relevant context in the scene due to its limited field of view. It is brittle in long-horizon navigation, drifting easily and unable to recover from failures due to the lack of global task progress. It also lacks spatial awareness of the other hand, which causes failures in coordinating both hands for precise bimanual manipulation. On the contrary, HoMMI augments UMI with *egocentric sensing*, providing global context and active perception behaviors crucial for mobile manipulation.

F2: Head-mounted camera alone is insufficient. While being the most common camera configuration for humanoid

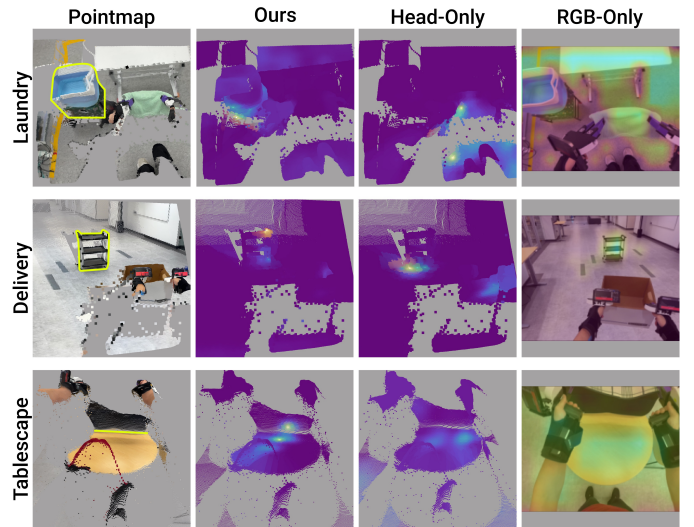


Fig. 11: **Egocentric Attention Comparison.** We visualize attention maps for egocentric observations with yellow representing higher attention values. Ours exhibits clean attention highlighted around task-relevant objects, while baselines' attentions are less informative.

design [4, 23, 11], the Head-Only baseline that relies solely on a head-mounted camera fails in grasping and alignment. HoMMI combines head camera views with wrist camera views, which provide essential local contact cues for fine-grained manipulation. We also find that having wrist visual observations as policy input and jointly finetuning the vision encoder on both wrist and egocentric images helps the policy learn cleaner egocentric attention that is more focused on task-relevant objects (Fig. 11 Ours v.s. Head-Only).

F3: Naively adding egocentric RGB can degrade performance under embodiment mismatch. Directly feeding the head RGB to the policy and regressing the head motion leads to brittle grasping and unstable motions, yielding a 0% success rate on two tasks, indicating a significant OOD shift due to viewpoint/appearance mismatch. Tracking 6-DoF head and hand trajectories together often leads to large tracking errors and violates the robot's kinematic constraints (Fig. 5). HoMMI bridges the visual gap by leveraging an embodiment-agnostic 3D egocentric visual representation, and bridges the kinematic gap through a relaxed 3D look-at point head action representation that allows the whole-body controller to achieve precise end-effector tracking and effective active perception.

F4: Active head control effectively gathers task-relevant information and maintains policy observability. Disabling head motion reduces success, particularly when it is required to actively search for the object and precisely place or align objects. This supports that our look-at point based active head control *effectively* imitates human's active perception behavior to gather task-relevant information and aligns the egocentric view more closely with the training distribution.

F5: Our cross-embodiment hand-eye policy learns task-relevant attention. As shown in Fig. 11, Ours yields egocentric attention maps highlighted on task-relevant objects and contacts, demonstrating the effectiveness of our observation representation and gripper coordinate frame design. This also

potentially helps to mitigate the visual embodiment gap, as the policy attends more to task-relevant regions than OOD observation points.

F6: Our system allows transfer from human demonstrators with different body proportions. Demonstrations were collected from two operators with different heights (167 cm and 182 cm) and both transferred successfully. This supports that our 3D look-at point action representation mitigates the height difference among operators.

VIII. 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 policy design with an embodiment-agnostic visual representation and a relaxed look-at point head action representation. A whole-body controller then achieves precise end-effector tracking and effective active perception by coordinating the robot’s whole-body motions while respecting physical constraints. Extensive real world experiments demonstrate that HoMMI allows versatile and challenging mobile tasks. We hope the proposed data collection hardware, learning framework, and robot system setup will encourage and facilitate future research on democratizing and scaling mobile manipulation.

IX. LIMITATIONS AND FUTURE WORK

Our policy uses a short observation history, which can limit recovery in long-horizon tasks and under partial observability; incorporating longer-term memory [30, 13] is a natural next step. We rely on vision-only sensing; adding force/tactile sensing and explicit compliance control could improve contact-rich manipulation and safety [7, 20, 51, 21]. Finally, while we align camera placement and gripper geometry between data collection and deployment, there remains a hardware embodiment gap; better co-design, including generative hardware design [40, 17], may further improve policy transferability.

X. ACKNOWLEDGMENTS

The authors would like to thank Calder Phillips-Grafflin, Aimee Goncalves, and Andrew Beaulieu from TRI for their help with the RB-Y1 hardware setup. Austin Patel for his help with the iPhone data collection app, Maximilian Du for his help with the fisheye camera calibration, and his feedback on the manuscript. Kosei Tanada, Vitor Guizilini, Paarth Shah, Eric Dusel, Sam Creasey, Hillel Hochshtein, Benjamin Burchil, Mengchao Zhang, Mark Zolotas, and Naveen Kuppaswamy from TRI for their helpful discussions. Phoebe Horgan, Allison Henry, Richard Denitto, Maya Angeles, Owen Pfannenstiehl, Mariah Smith-Jones, and Gordon Richardson from TRI for their help with data collection. Yifan Hou, Huy Ha, Hojung Choi, all REALab members, and Jinghan Sun for their helpful discussions and feedback on the manuscript. This work was supported in part by the NSF Award #2143601, #2037101, and #2132519, and Toyota Research Institute. The views

and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

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.
- [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] 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.
- [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.
- [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.
- [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.
- [7] Hojung Choi, Yifan Hou, Chuer Pan, Seongheon Hong, Austin Patel, Xiaomeng Xu, Mark R Cutkosky, and Shuran Song. In-the-wild compliant manipulation with umi-ft. *arXiv preprint arXiv:2601.09988*, 2026.
- [8] 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.
- [9] 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*.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streaming2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2568–2577, 2025.
- [14] Ryan Hoque, Peide Huang, David J Yoon, Mouli Siva-
purapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.
- [15] 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.
- [16] 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.
- [17] Byungchul Kim, Tsun-Hsuan Wang, and Daniela Rus. Generative-ai-driven jumping robot design using diffusion models. In *2025 International Conference on Robotics and Automation (ICRA)*, 2025.
- [18] 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.
- [19] 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*.
- [20] Fangchen Liu, Chuanyu Li, Yihua Qin, Jing Xu, Pieter Abbeel, and Rui Chen. Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface. *arXiv preprint arXiv:2504.06156*, 2025.
- [21] Yun Liu, Xiaomeng Xu, Weihang Chen, Haocheng Yuan, He Wang, Jing Xu, Rui Chen, and Li Yi. Enhancing generalizable 6d pose tracking of an in-hand object with tactile sensing. *IEEE Robotics and Automation Letters*, 9(2):1106–1113, 2023.
- [22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [23] 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.
- [24] Ryan Punamiya, Dhruv Patel, Patcharapong Aphiwetsa, Pranav Kuppli, 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.
- [25] Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y Zhu, et al. Egoverse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [30] Ajay Sridhar, Jennifer Pan, Satvik Sharma, and Chelsea Finn. Memer: Scaling up memory for robot control via experience retrieval. *arXiv preprint arXiv:2510.20328*, 2025.
- [31] Priya Sundareshan, 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.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] Xiaomeng Xu, Dominik Bauer, and Shuran Song. RoboPanoptes: The All-Seeing Robot with Whole-body Dexterity. In *Proceedings of Robotics: Science and Systems*, 2025.
- [40] Xiaomeng Xu, Huy Ha, and Shuran Song. Dynamics-guided diffusion model for sensor-less robot manipulator design. In *Conference on Robot Learning*, pages 4446–4462. PMLR, 2025.
- [41] 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.
- [42] 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.
- [43] Ruihan Yang, Qinxu 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.
- [44] 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.
- [45] Kevin Zakka. Mink: Python inverse kinematics based on MuJoCo, December 2025. URL <https://github.com/kevinzakka/mink>.
- [46] 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.
- [47] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.
- [48] Zhaxizhuom Zhaxizhuoma, Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Peng 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.
- [49] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019.
- [50] 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.
- [51] Xinyue Zhu, Binghao Huang, and Yunzhu Li. Touch in the wild: Learning fine-grained manipulation with a portable visuo-tactile gripper. *arXiv preprint arXiv:2507.15062*, 2025.

I. POLICY TRAINING DETAILS

Observations and actions. We use a short observation history of $T_o=2$ steps and predict an action horizon of $T_p=32$ steps at 20 Hz (downsampled from 60 Hz demonstrations). Observations include left/right wrist RGB (224×224), head RGB and pointmap (512×512), and proprioception (end-effector poses and gripper widths). Actions are 23-dimensional, including 9-dimensional left/right gripper poses, a 3-dimensional look-at point, and gripper widths ($23 = 2 \times 9 + 3 + 2$). Each gripper pose is represented by a 3-dimensional vector for position and a 6-dimensional vector for orientation, using the first two columns of the rotation matrix [49].

Model. We use Diffusion Policy [6] with a Diffusion Transformers backbone (DiT) [2]. The diffusion model is conditioned on global observation embeddings and predicts noise with a DDIM scheduler [29]. We use 100 training timesteps, 16 inference steps, input perturbation 0.1, and train with 8 diffusion noise samples per observation. The DiT uses embedding dimension 768, depth 10, 12 heads, MLP ratio 4, and RMS norm [47]. The observation encoder finetunes dinov3-vitb16 backbones [28] for wrist and head RGBs with shared weights.

Optimization. We use AdamW [22] with a cosine learning rate schedule starting at a learning rate of 7.5×10^{-5} for the diffusion model and 7.5×10^{-6} for finetuning the vision backbone, weight decay 1×10^{-6} , and betas (0.95, 0.999). We train our policy and all baselines for 500 epochs.

II. WHOLE-BODY CONTROLLER DETAILS

A. Temporal Command Interpolation

To ensure smooth trajectory tracking despite the lower policy frequency, the controller performs temporal interpolation of the target commands at each IK iteration. Given a new 6-DoF pose command received from the policy loop at time t_0 with a duration T , we compute a progress fraction $\alpha(t) = \min(1, \frac{t-t_0}{T})$. We derive the intermediate pose by interpolating its components: the position $\mathbf{p}$ uses linear interpolation,

$$\mathbf{p}_{\text{interp}}(t) = (1 - \alpha)\mathbf{p}_{\text{prev}} + \alpha\mathbf{p}_{\text{cmd}}, \quad (1)$$

whereas the orientation is interpolated using spherical linear interpolation between the previous and target rotations. This interpolation approach ensures that the high-frequency whole-body controller receives continuous targets, effectively eliminating jitter in the resulting motion.

B. Whole-body IK

Our differential whole-body IK solver maps the interpolated Cartesian targets to generalized joint velocity Δq by solving a differential IK problem. We use the `daqp` solver to handle the optimization, with the damping coefficient λ set to 10^{-6} for numerical stability. The objective function $f(\Delta q)$ consists of the following individual costs:

- **Bimanual SE(3) Tracking** (C_{ee}): We penalize the Cartesian tracking error of both end-effectors in differential form:

$$C_{\text{ee}} = \sum_{i \in \{\text{L}, \text{R}\}} \|\mathbf{J}_i \Delta q - \mathbf{v}_i\|_{\mathbf{W}_{\text{ee}}}^2,$$

where $\mathbf{J}_i$ is the Jacobian and $\mathbf{v}_i$ is the Cartesian command, with $\mathbf{W}_{\text{ee}}$ being a diagonal matrix of positional weights w_p and rotational weights w_o .

- **Nominal Posture Regularization** (C_{nominal}): We regularize the solution toward a nominal posture q_{nom} to encourage human-like, consistent body configurations. This nominal posture is defined by the robot's initial pose, which is set to match the human posture observed during data collection:

$$C_{\text{nominal}} = \|(q + \Delta q) - q_{\text{nom}}\|_{\mathbf{W}_{\text{nom}}}^2,$$

where $\mathbf{W}_{\text{nom}}$ is a diagonal matrix whose entries are set to $w_{\text{nom}, \text{torso}}$ for torso joints and $w_{\text{nom}, \text{arm}}$ for arm joints.

- **Current Posture Regularization** (C_{current}): To ensure motion smoothing in velocity space, we penalize any deviation from the current joint configuration and mobile base pose:

$$C_{\text{current}} = \|\Delta q\|_{\mathbf{W}_{\text{curr}}}^2,$$

where $\mathbf{W}_{\text{curr}}$ is a diagonal matrix whose entries are set to the weight w_{curr} for all joints, and $w_{\text{base}, \text{pos}}$ and $w_{\text{base}, \text{ori}}$ for the base translational and rotational components, respectively.

- **CoM-over-base Support** (C_{com}): We maintain the upper-body mass centered over the mobile base through the following objective and relative XY displacement constraints:

$$C_{\text{com}} = \|p_{\text{torso}}^{\text{xy}}(q + \Delta q) - p_{\text{base}}^{\text{xy}}(q + \Delta q) - r_{\star}^{\text{xy}}\|_{\mathbf{W}_{\text{com}}}^2$$

s.t. $|p_{\text{torso}}^x - p_{\text{base}}^x - r_{\star}^x| \leq b_x, \quad |p_{\text{torso}}^y - p_{\text{base}}^y - r_{\star}^y| \leq b_y,$

where $\mathbf{W}_{\text{com}}$ is a diagonal matrix with weights w_{com} applied to the x and y dimensions, and b_x, b_y are the permitted displacement bounds.

In addition to these cost terms, the solver must satisfy several physical and safety constraints, expressed as $G_j \Delta q \leq h_j$ and $A \Delta q = b$, to maintain the robot's operational integrity:

- **Configuration Bounds** (G_{cfg}): These constraints ensure the next state $q + \Delta q$ remains within the physical joint position limits specified in the robot model.
- **Joint Velocity Limits** ($G_{\text{joint-vel}}$): Joint velocities follow the limits specified in the model, scaled by a safety factor of 0.9. Specifically, `torso_0`, `torso_4`, and `torso_5` are fixed to zero velocity to maintain a consistent torso orientation.
- **Base Velocity Limits** ($G_{\text{base-vel}}$): The mobile base velocity is also subject to the 0.9 safety scale, with maximum limits set to 1.0 m/s for translation and 1.0 rad/s for rotation.
- **Collision Avoidance** (G_{coll}): Collision avoidance is enforced over selected geometry groups, including (*base+torso+head*, *arms*) and inter-arm pairs. The solver maintains a minimum separation $d_{\text{safe}} = 0.01$ m, activated within an influence distance $d_{\text{inf}} = 0.02$ m.
- **Upright Posture** (A_{upright}): An equality constraint $A_{\text{upright}} \Delta q = 0$ ensures that the sum of velocities for `torso_1`, `torso_2`, and `torso_3` remains zero for

TABLE I: Whole-body IK parameters for the three evaluation tasks

Symbol	Laundry	Delivery	Tablespace
w_p	10000	10000	10000
w_o	10000	10000	10000
$w_{\text{nom,torso}}$	50	1000	200
$w_{\text{nom,arm}}$	50	1000	10
w_{curr}	50	1000	10
$w_{\text{base,pos}}$	50	50	5000
$w_{\text{base,ori}}$	50	50	5000
w_{com}	100000	100000	100000
b_x, b_y	0.08 m	0.08 m	0.08 m

a vertical torso orientation throughout the motion. For the *delivery* task, the velocities of these three joints are individually fixed to zero to maximize payload stability.

The specific weight parameters used during evaluation are summarized in Table I. In our optimization hierarchy, maintaining the center-of-mass (C_{com}) over the base is given the highest priority to ensure global stability, followed by accurate bimanual end-effector tracking (C_{ee}). The relative priorities among the mobile base, torso, and arm configurations are tuned according to the specific requirements of each task to balance mobility, stability, and reachability. During our experiments, these parameter settings required only 1–2 rollout checks before being finalized.

III. ROBOT HARDWARE SETUP

Our platform is built on a commercially available Rainbow Robotics RB-Y1 bimanual mobile manipulator, which provides a holonomic base with mecanum wheels, a 6-DoF torso, two 7-DoF arms, and a 2-DoF neck. The stock parallel grippers are replaced with fin-ray fingers identical to the fingers of the UMI grippers. On the 2-DoF neck we mount a stereo pair of industrial wide-angle RGB cameras (FLIR BFS-PGE-23S3C-CS) using a custom neck bracket and pitch-joint stop that constrains the neck range for safe cabling. On each wrist, we install a custom clamp and rigidly attach a FLIR BFS-PGE-50S5C-C camera at a fixed offset relative to the gripper.

As shown in Fig. 12, our sensing and computing stack is organized around two fiber-connected subnetworks. All cameras connect via GigE Ethernet cables to a PoE (Power over Ethernet) GigE switch mounted in a backpack structure on the robot torso. This PoE switch uplinks to the external workstation through a 10 Gbps multimode fiber link, providing sufficient bandwidth for simultaneous multi-camera streaming. In parallel, a Ubiquiti router connects the robot’s internal control computers – including a user PC (U-PC) and a robot PC (R-PC) – to the workstation over a second fiber link, carrying all low-level robot state and command traffic. The external workstation runs our control stack, GigE camera drivers, and policy inference, and interfaces with the robot via these

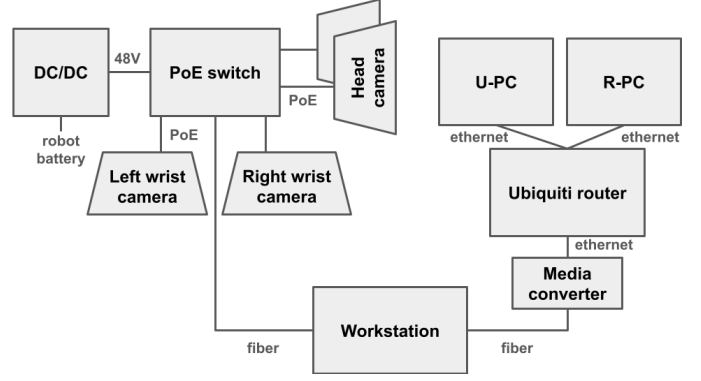


Fig. 12: Hardware Schematic.

two fiber uplinks. This architecture supports high-bandwidth multi-camera streaming and high-frequency robot closed-loop control.

IV. ADDITIONAL EXPERIMENTS

A. Depth quality sensitivity

Fig. 13 shows that the robot-view stereo pointcloud is clean and the policy attends to task-relevant regions (bin and cloth) during inference.

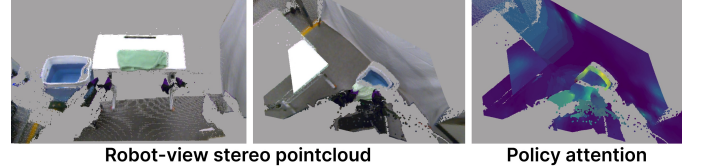


Fig. 13: High quality stereo pointcloud and informative attention.

We inject additional Gaussian noise into depth and evaluate the Laundry task under 3 noise levels (10 rollouts each). As shown in Tab. II, performance remains unchanged up to 1 cm standard deviation noise and degrades only under substantially larger perturbations.

noise std (mm)	0	2	10	20
success (%)	90	90	90	50

TABLE II: Depth quality sensitivity.

B. Robustness to unseen lightings and objects.

To further assess robustness, we additionally evaluated the Laundry task under 3 unseen lighting conditions and on 3 unseen objects (4 rollouts per case). Object variation yields $90.63\% \pm 12.10\%$ success, and lighting variation yields $93.75\% \pm 10.83\%$, indicating robustness of our policy.

seen and unseen object	s1	s2	u1	u2	u3	u4	u5	u6
success (%)	100	75	100	75	100	100	100	75

TABLE III: Generalization to unseen objects.

light (lux)	4300	3370	1840	810
success (%)	100	100	100	75

TABLE IV: Robustness to lighting variations.

C. iPhone-ARKit tracking error.

We measure the average tracking error against ground truth obtained from Motion Capture. The error is within 5.0 mm position / 0.8° rotation.