Pixel-UMI: Learning Environment-Aware Whole-Body Manipulation from Sensorized Human Demonstrations

Jason Jingzhou Liu*, Tony Tao*, Deepak Pathak†, Ruslan Salakhutdinov†

Carnegie Mellon University

*Equal contribution †Equal advising

https://pixel-robot.com

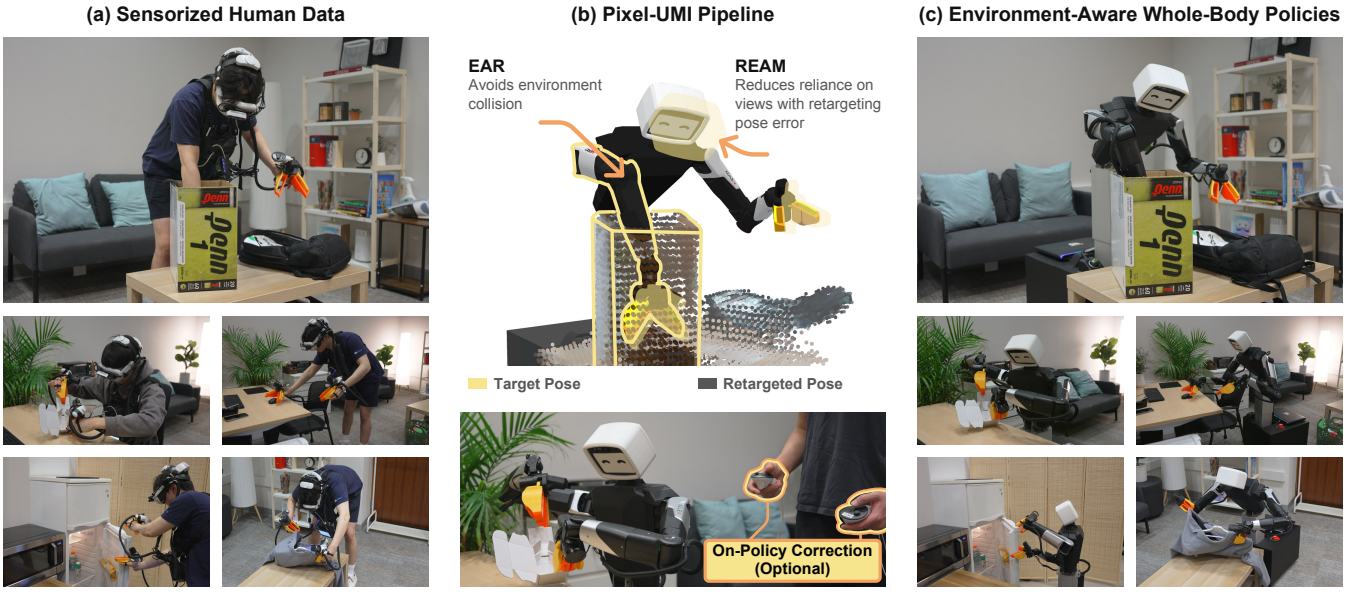


Fig. 1. **Pixel-UMI overview.** (a) Humans collect whole-body mobile manipulation demonstrations with Pixel-UMI. (b) Environment-Aware Retargeting (EAR) uses offline scene reconstruction to convert human demonstrations into collision-aware whole-body trajectories for policy training. Retargeting-Error-Aware Masking (REAM) uses residual pose errors to reduce reliance on camera views the robot cannot reproduce accurately. Optional on-policy corrections further improve performance on contact-rich tasks. (c) Pixel-UMI enables long-horizon, contact-rich whole-body manipulation.

***Abstract*— Sensorized human demonstrations offer scalable data for mobile manipulation, but embodiment differences complicate policy learning. We present Pixel-UMI, a pipeline for learning environment-aware whole-body policies from human data. We first introduce Environment-Aware Retargeting (EAR), which combines reconstructed scene geometry with demonstrated head and gripper trajectories to determine robot posture and generate collision-aware joint and base targets. Delay-calibrated action assignment converts these trajectories into actions. However, kinematic differences can leave residual pose errors, creating viewpoint mismatches between human demonstrations and robot observations. We therefore introduce Retargeting-Error-Aware Masking (REAM), which uses these errors during training to reduce reliance on mismatched views at deployment. Finally, because estimated actions cannot fully capture contact-rich interactions, we collect on-policy human corrections through DAgger to improve performance. We evaluate Pixel-UMI on five challenging tasks spanning articulated-object manipulation, deformable-object handling, constrained-space reaching, whole-body collision avoidance, and contact-rich bimanual coordination, achieving 87.0% average task progress. Video results are available at https://pixel-robot.com/pixel-umi**

I. Introduction

Mobile manipulation requires coordinated motion of the arms, torso, head, and base. Imitation learning provides a way to acquire these behaviors from demonstrations. Teleoperation is a widely used approach to collecting high-quality demonstrations [1], [2], but is challenging for mobile manipulation: operators must coordinate many degrees of freedom, and collection across environments requires deploying a robot at each site.

Sensorized human data collection offers a more direct way to gather demonstrations by giving people a manipulation interface matched to the robot's [3], [4]. The Universal Manipulation Interface (UMI), for example, records images and motion through handheld grippers that the demonstrator uses to interact with objects. This allows people to coordinate their own bodies and receive physical feedback from contact, making demonstration more intuitive than teleoperation while allowing collection across environments without transporting a robot. Extensions to mobile manipulation complement the grippers with a head-mounted camera, providing the broader visual context needed for navigation and whole-body coordination [5], [6].

Transferring these demonstrations to a robot requires aligning the demonstrated gripper and camera frames with their robot counterparts so that the manipulation and visual observations remain consistent across embodiments [3], [5].

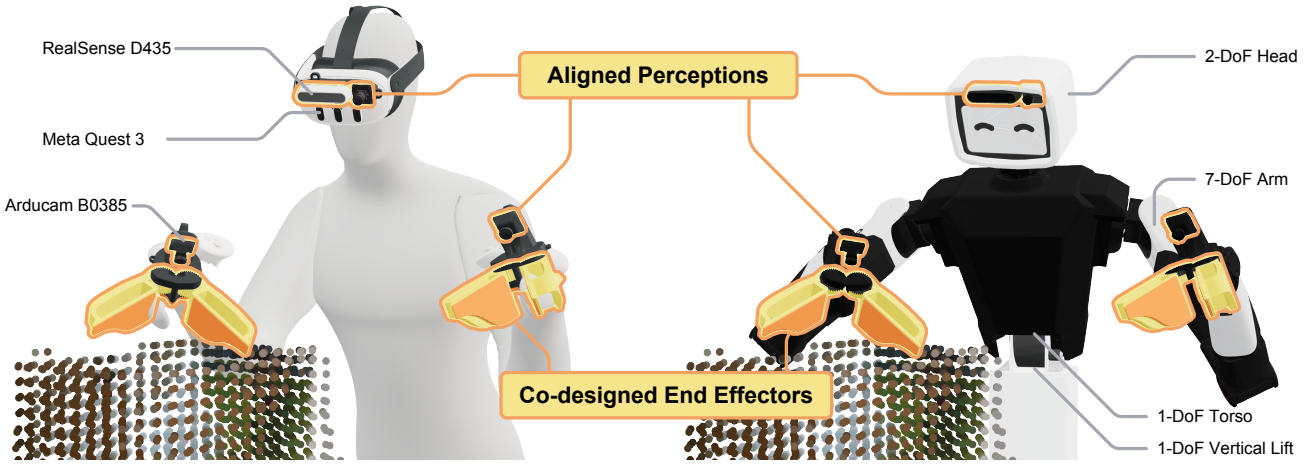


Fig. 2. **Co-designed hardware for Pixel-UMI.** The sensorized human interface (left) and Pixel robot (right) use matched grippers and corresponding head and wrist camera layouts to align manipulation and visual observations.

Although task-frame matching can suffice for tabletop manipulation with limited redundancy, it leaves the posture of a whole-body mobile manipulator underdetermined: many arm, torso, and base configurations can realize similar gripper and head targets. These choices affect whether the robot can continue following the demonstrated motion within the scene. When retrieving a tennis-ball canister from a deep box, for example, the elbow must lift above the rim to allow the gripper to descend. Without an appropriate elbow posture, contact with the rim can obstruct the arm and prevent continued gripper tracking.

To support continued task-frame tracking, the robot must choose a body configuration that accommodates the surrounding geometry. To this end, we introduce Environment-Aware Retargeting (EAR), which uses scene geometry to resolve the robot's posture while matching the demonstrated gripper and head motion. EAR reconstructs static scene geometry from the episode's depth sequence, using temporal evidence to separate dynamic objects without requiring a separate scene scan. We then use environment-aware geometric fabrics [7] to retarget the demonstrated motion against this geometry offline, producing body joint trajectories and base commands. The policy learns these posture choices directly, avoiding the need to combine scene estimation with collision-aware whole-body control during policy rollout.

However, fundamental differences in body shape and kinematics between humans and robots can still prevent exact head and gripper tracking, creating viewpoint mismatches between human training images and robot observations [5]. Retargeting alone cannot fully eliminate this mismatch, but we can mitigate its impact during policy training. We therefore introduce Retargeting-Error-Aware Masking (REAM). Because retargeting occurs before policy training, its pose errors reveal when a demonstrated viewpoint is difficult for the robot to reproduce. REAM uses these errors as geometric proxies for visual mismatch in percentile-based camera masking: views with higher-ranked errors are more likely to have their tokens excluded during training. This encourages the policy to draw on other available observations when a view is poorly aligned, while retaining access to all cameras at deployment.

Finally, sensorized human data collection without force or tactile sensors records motion states but not the interaction forces underlying human actions [3], [8], [9]. Motion states alone do not determine the robot commands needed to reproduce these interactions, so policies trained on this supervision can struggle with contact-rich manipulation. Nevertheless, learning from these demonstrations provides a useful policy prior for task motion and whole-body coordination, making on-policy correction practical. We therefore collect on-policy correction data through DAgger [10] to improve performance on contact-rich whole-body tasks.

We call our combined pipeline Pixel-UMI and evaluate it on the co-designed Pixel platform [11]. We summarize our key contributions as follows:

- EAR converts head and gripper demonstrations into environment-aware whole-body joint and base supervision. Across three tasks requiring whole-body coordination, it increases average task progress by 64.41 pp.
- REAM uses percentile-ranked retargeting errors to guide camera masking and reduce dependence on mismatched views. On the same three tasks, adding REAM improves average task progress by 14.05 pp.
- We evaluate on-policy correction through DAgger to address limitations of state-based UMI supervision. On two challenging, long-horizon, contact-rich bimanual whole-body tasks, two DAgger rounds improve average task progress by 72.95 pp.

II. Hardware

Pixel-UMI uses the co-designed robot and sensorized human interface of Pixel (Figure 2). The human-worn grippers match the robot grippers in manipulation geometry, visual appearance, and wrist-camera placement. Together with aligned head RGB cameras, this co-design reduces the physical and visual gaps between embodiments to support the conversion of sensorized human demonstrations into high-quality robot training data. Each human-worn gripper is



Fig. 3. **Static scene reconstructions.** Episode-level static geometry reconstructed from sensorized human demonstrations across five tasks, shown with the retargeted robot. These point clouds provide the environment geometry for downstream collision-aware retargeting.

equipped with an encoder to record gripper width. Meta Quest controllers track the gripper poses relative to the head, while a Meta Quest 3 headset worn by the demonstrator tracks the head pose in the world frame. We refer readers to the Pixel system paper [11] for further hardware details.

We additionally equip the human headset with a depth camera to record metric depth throughout each demonstration. We use these observations for offline metric 3D reconstruction to support collision-aware retargeting from human demonstrations to robot demonstrations. We note that during robot policy rollout, visual observations come from the head and wrist RGB cameras; the robot-side depth camera is not used.

III. PIXEL-UMI

We present Pixel-UMI, a pipeline for learning whole-body robot policies from sensorized human demonstrations. Our goal is to address the fundamental human-to-robot gaps involved in converting these demonstrations into high-quality robot training data.

Firstly, the recorded head and gripper poses leave the robot's whole-body posture underdetermined due to kinematic redundancy. To address this, we introduce Environment-Aware Retargeting (EAR). We reconstruct scene geometry offline and use collision avoidance to guide the robot's whole-body posture while tracking the demonstrated motion. However, the resulting trajectories still only describe robot states, not robot actions. We therefore assign pseudo-actions using calibrated joint tracking delays, so that the training targets better account for the robot's response.

Secondly, differences in human and robot kinematics and body geometry can still prevent exact tracking. As a result, the head and wrist camera views during robot policy rollout can differ from those recorded in the human demonstrations. To alleviate this visual gap, we introduce Retargeting-Error-Aware Masking (REAM), which uses retargeting errors during training to encourage the policy to rely less on poorly matched camera views.

Finally, remaining human-to-robot gaps can still limit policy performance on long-horizon, contact-rich manipulation tasks. We therefore collect a small amount of on-policy correction data through DAgger to improve policy performance on these tasks.

A. Preliminaries

For a synchronized human demonstration, let I_t^c denote the RGB image from camera $c \in \{H, L, R\}$, where H, L, and R denote the head, left gripper, and right gripper, respectively, and let D_t denote the synchronized head-camera depth image. We use the same indices for task frames, with $T_t^j \in \mathrm{SE}(3)$ denoting the pose of frame $j \in \{H, L, R\}$ in a common world frame, obtained from the Meta Quest 3 headset and controller tracking. The left and right gripper states g_t^L, g_t^R are derived from their encoder readings. An episode contains T such samples and a task instruction ℓ.

We represent the robot configuration as $x_t = (b_t, q_t)$, where $b_t = (x_t^b, y_t^b, \psi_t)$ is the planar base pose and $q_t \in \mathbb{R}^{20}$ contains lift, torso, arm, head, and gripper coordinates. The learned action is

$$a_t = [v_{x,t}^B, v_{y,t}^B, \omega_t, q_t^{\mathrm{cmd}}] \in \mathbb{R}^{23}, \tag{1}$$

with planar velocities expressed in the robot base frame and absolute position targets for the body and grippers. Thus, the policy directly specifies joint-space targets, eliminating the need for an online whole-body controller (WBC) to choose motion in the task-space null space.

B. Environment-Aware Retargeting (EAR)

*1) **Metric reconstruction from the demonstration**:* Our goal is to reconstruct a metric 3D scene for each collected episode, providing the environment geometry for obstacle avoidance during retargeting. The Meta Quest 3 tracking system provides head and gripper trajectories in a common world frame, with synchronized depth images from the head-mounted camera at each timestep. However, potential pose drift or tracking discontinuities can cause depth point clouds to misalign when combined using the recorded poses alone. To address this, we perform offline point-cloud registration and pose refinement, using geometric alignment across observations to refine the head trajectory before fusing the corrected point clouds into a scene. The reconstruction uses observations from the full recording, including a brief floor-calibration observation when needed; a separate scan of the workspace is not required.

We convert each depth image into a metric point cloud in the world frame using the calibrated camera intrinsics, head-to-camera transform, and recorded head pose. Before registration, we filter points near the cameras and tracked grippers to reduce the influence of the demonstrator and nearby manipulated objects. We then identify candidate static surfaces through temporal persistence: after voxelizing the clouds in the recorded world frame, we retain points whose voxels are observed in multiple frames within a local temporal window. This favors surfaces that remain

at approximately the same world location across nearby observations. We downsample these candidates and select keyframes to balance temporal coverage, camera motion, and geometric overlap.

We align overlapping keyframes using point-to-plane iterative closest point (ICP) registration [12], [13]. To limit unreliable registration constraints, we discard alignments with insufficient overlap or too few matched points, large distances between matched points, or excessive pose corrections relative to the VR initialization. We then refine the camera trajectory through pose-graph optimization [14], combining the accepted ICP constraints with relative-motion priors from the recorded VR poses of consecutive keyframes. We identify possible tracking discontinuities as abrupt changes in vertical or total camera translation that exceed displacement and velocity thresholds and are outliers relative to the episode's motion, measured using a median-based outlier score. For these intervals, we reduce the weights on the corresponding VR relative-motion residuals in the pose-graph objective, allowing geometric alignment to have greater influence on the refined trajectory. We interpolate the resulting pose corrections to all frames.

Finally, we establish a ground-referenced world frame for retargeting. EAR solves for absolute whole-body joint configurations with the mobile base supported by the floor, so the head and gripper targets must be expressed at the correct metric heights above that floor. An arbitrary vertical offset in the default VR fixed frame would otherwise shift these targets relative to the robot's base and change the resulting body posture. We estimate the ground plane from observed floor points and vertically shift the head and gripper trajectories and reconstructed scene together so that ground level corresponds to $z = 0$ in the fixed retargeting frame. For tasks whose demonstrations may not reveal the floor, such as seated box folding at a table, we require the demonstrator to look down and record a brief floor observation before beginning the task.

*2) **Separating static structure from dynamic geometry**:* Given the reconstructed scene, we remove the objects being manipulated so that collision avoidance does not prevent the contact required by the task. We make the simplifying assumption that these objects move or deform during the demonstration, and use this motion to separate them from static environmental structure.

We compare aligned depth observations across the episode using occupancy and free-space evidence [15]. Each depth measurement indicates a surface at its endpoint and empty space along the viewing ray before it. If another frame sees through a previously occupied location to a farther surface, this provides evidence that the earlier surface moved [16]. We retain candidate voxels as static when they have repeated surface observations across time and few conflicting free-space observations, excluding transient geometry and regions with insufficient evidence. Objects whose motion is not observed may remain in the static map. Figure 3 shows the resulting static geometry used for collision-aware retargeting.

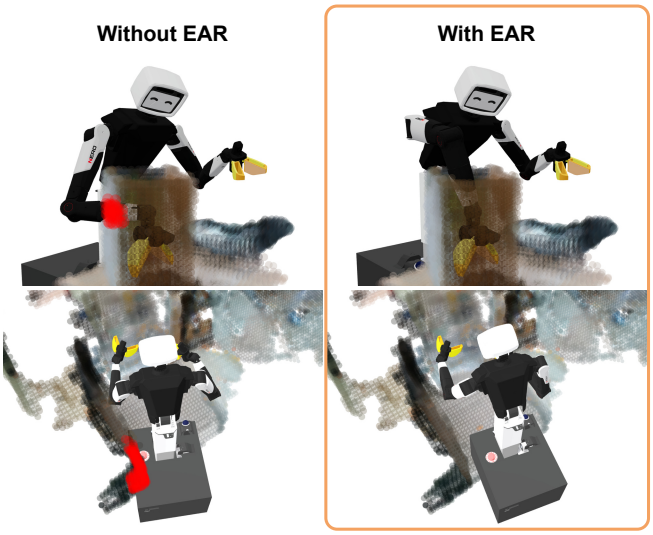


Fig. 4. **Effect of collision-aware retargeting.** Without EAR (left), retargeted robot postures intersect the surrounding scene, with collisions highlighted in red. EAR (right) uses reconstructed geometry to select whole-body postures that avoid these obstacles and support continued tracking of the demonstrated head and gripper motion.

*3) **Collision-aware whole-body retargeting**:* Given the static scene represented as a voxelized point cloud and the corrected head and gripper trajectories, our goal is to solve for robot joint and base configurations that track the demonstrated poses while avoiding the surrounding geometry. We use geometric fabrics [7] as an offline retargeting solver. This framework composes motion objectives defined in different task spaces through metric-weighted acceleration updates, allowing tracking, collision avoidance, and joint-limit avoidance to jointly shape whole-body motion. Our formulation combines head and gripper pose attraction, environment and self-collision repulsion, joint-limit repulsion, nominal-posture regularization, and mobile-base and joint-priority terms.

We represent the robot's collision geometry with spheres attached to its links. For a sphere with center $c_s(x)$ and radius r_s, its clearance from the static scene $\mathscr{S}$ is

$$d_s(x, \mathscr{S}) = \min_{p \in \mathscr{S}} \|c_s(x) - p\|_2 - r_s. \tag{2}$$

As this clearance decreases, repulsive fabric terms increasingly influence the robot's posture, including links whose motion is not specified by the head and gripper targets. We solve sequentially along the demonstrated trajectory, using the preceding robot configuration to initialize each step. Figure 4 illustrates how incorporating scene geometry changes the whole-body posture to support continued task-frame tracking.

In our formulation, collision avoidance acts as a soft constraint, with repulsion balanced against pose tracking through the fabric metrics. This differs from QP-based differential inverse kinematics formulations such as Mink [17] and Pink [18], whose collision-avoidance limits and barriers impose hard inequalities in the QP. We choose this soft formulation to accommodate reconstruction noise, such as isolated

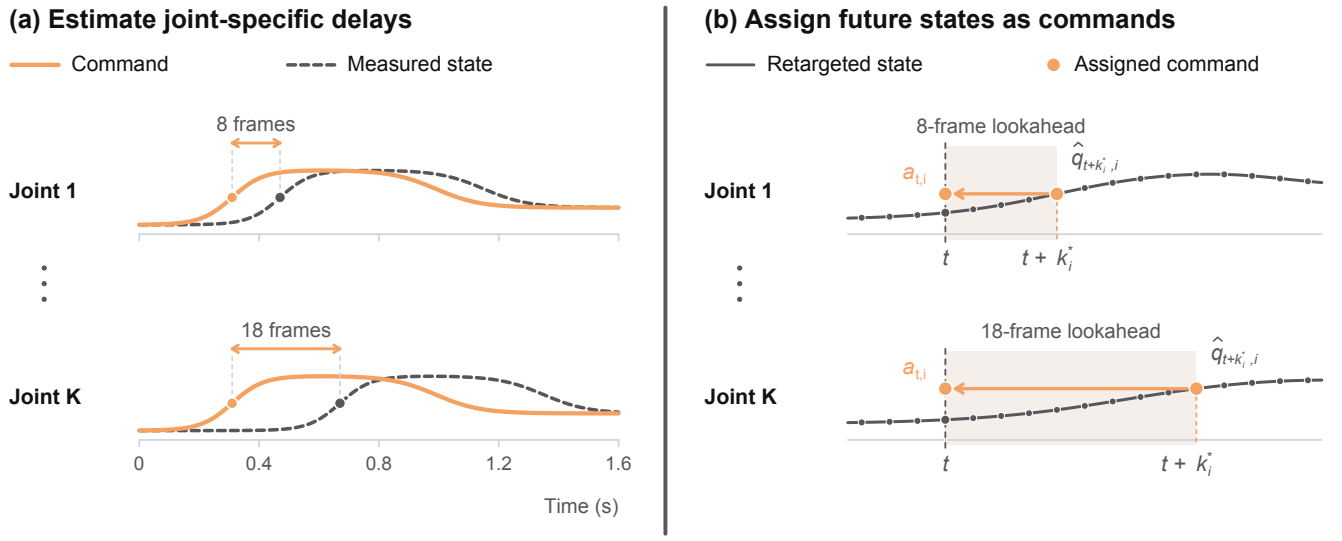


Fig. 5. **Dynamics-aware action assignment.** (a) Joint-specific delays are estimated by aligning robot commands and measured states. (b) These delays determine how far ahead to sample the retargeted state when constructing each joint's command label.

floating points: these points contribute repulsion without becoming inviolable geometric boundaries. The solver can therefore trade off tracking and clearance locally when the reconstructed geometry contains spurious obstacles.

*4) **Dynamics-aware action assignment**:* Retargeting produces whole-body state trajectories, but behavior cloning requires robot actions. We first time-scale the trajectories to reduce motion speed and resample their associated images. Sunday likewise reports using slower human demonstrations and policy rollouts [19]. We then derive base velocities by differentiating the base poses and expressing them in the robot frame. For body joints, next-state action labels ignore differences in how quickly each joint tracks its target. We therefore calibrate a tracking delay for each joint and use it to construct pseudo-action labels, illustrated in Figure 5.

We estimate these delays using actual robot trajectories, such as teleoperation data, with recorded joint positions and corresponding position targets. For joint i, let $a^n_{t,i}$ denote the target issued at time t in trajectory n, and $q^n_{t,i}$ the measured position. We select the offset that best aligns each target with the state reached later:

$$k_i^* = \underset{k\in\{0,\dots,K\}}{\operatorname{argmin}} \sum_n \sum_t \left(q^n_{t+k,i} - a^n_{t,i}\right)^2, \tag{3}$$

where we sum over trajectories and timestamps with valid future states for all candidate offsets. Calibration trajectories need not be paired with human demonstrations.

For the retargeted joint trajectory $\hat{q}_t$, we assign

$$a_{t,i} = \hat{q}_{t+k_i^*,i}, \tag{4}$$

advancing each joint's target by its calibrated delay. UMI also compensates for deployment latency [3]; here, we construct per-joint action labels for whole-body trajectories.

We highlight that this approximation accounts for tracking lag in free-space motion, but contact can prevent a joint from reaching its target. A target beyond a contact surface may generate force without further motion, so the state trajectory alone cannot determine the appropriate action. We address this gap through on-policy correction (Section III-D).

The resulting joint targets and base velocities form the action representation in Equation 1. The policy learns scene-dependent whole-body posture directly and sends these actions to the robot's low-level controllers during rollout, without requiring online scene reconstruction or task-space retargeting.

C. Retargeting-Error-Aware Masking (REAM)

Differences in human and robot kinematics and body geometry can prevent exact tracking of the demonstrated head and gripper poses. During drink retrieval, for example, the human can lean and reorient the head in ways that the robot cannot reproduce, illustrated in Figure 6. Consequently, the human images used to train the policy can differ from the views available during robot rollout. A policy that relies on these mismatched views may therefore transfer poorly.

We introduce REAM to alleviate this visual gap during policy training. Because offline retargeting provides both the demonstrated and retargeted poses throughout each episode, we can identify when and where tracking errors occur. We use these errors to increase the probability of masking visual tokens from poorly aligned cameras, encouraging the policy to rely less on those views when predicting actions.

For a given camera and timestep, we omit the camera and time indices for clarity. Let (p,R) and $(\hat{p},\hat{R})$ denote the demonstrated and retargeted positions and orientations of the associated head or gripper frame. We measure their discrepancy as

$$e_p = \|p-\hat{p}\|_2, \qquad e_R = \left\|\operatorname{Log}(R\hat{R}^\top)^\vee\right\|_2, \tag{5}$$

where e_p is the position error and e_R is the rotation error in radians. Here, Log denotes the matrix logarithm on $\mathrm{SO}(3)$, and $(\cdot)^\vee$ maps a skew-symmetric matrix to its corresponding

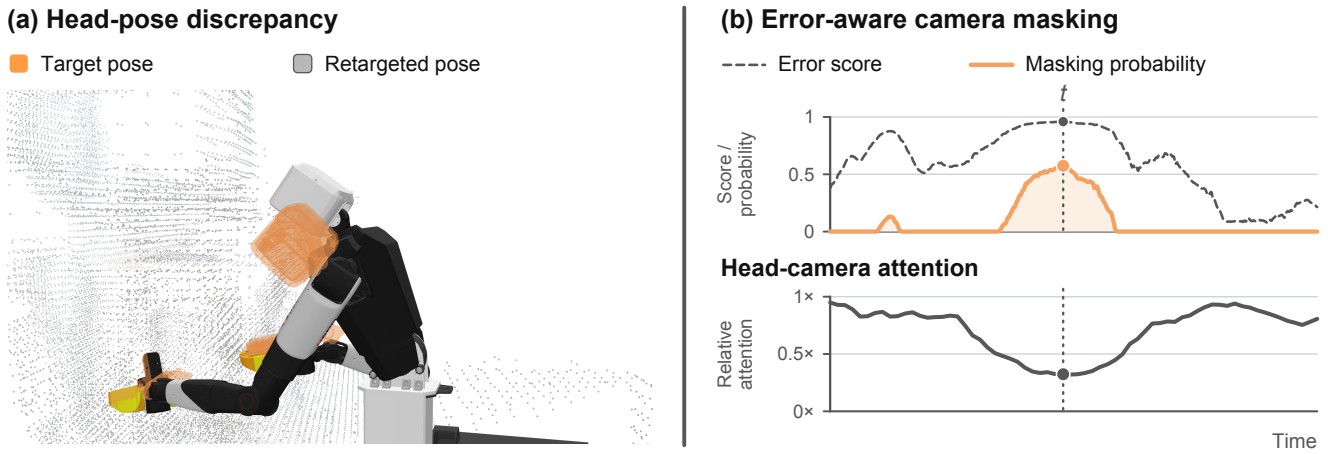


Fig. 6. **Retargeting-Error-Aware Masking (REAM).** (a) Head-pose mismatch between the demonstrated target (orange) and retargeted pose (gray). (b) Retargeting error and masking probability (top), with measured relative head-camera attention (bottom), for a transformed Retrieve Drink demonstration used for training. Attention decreases during the high-error interval.

rotation vector. These errors serve as geometric proxies for camera viewpoint mismatch.

We compare each error with the errors observed in the training set. We pool position errors across cameras and timesteps into a sorted list, and construct a separate list for orientation errors. Let $F_p(e)$ and $F_R(e)$ return the percentile rank of error e in these lists, normalized to $[0,1]$. Both mappings are computed from the training data and shared across cameras.

We average the two ranks to obtain a combined error score $r \in [0,1]$:

$$r = \frac{F_p(e_p) + F_R(e_R)}{2}. \tag{6}$$

We map this score to a masking probability:

$$\pi = \begin{cases} 0, & r \le \rho, \\ p_{\max}\left(\dfrac{r-\rho}{1-\rho}\right)^{\gamma}, & r > \rho. \end{cases} \tag{7}$$

The threshold $\rho \in [0,1)$ determines when masking begins. Above this threshold, the probability increases with the error score, reaching $p_{\max}$ at $r = 1$; $\gamma > 0$ controls the shape of this increase.

We apply this probability when sampling camera masks for the flow-matching policy described in the Pixel system paper [11]. For each training sample, let $\pi_{t,c}$ denote the masking probability for camera c at observation time t. We draw a Bernoulli mask for each camera using this probability. A masked camera has all its visual tokens across the observation window excluded as attention keys and values. For cross-attention, we implement this with an additive mask:

$$\mathrm{Attn}(Q,K,V;M_t) = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d}} + M_t\right)V, \tag{8}$$

where Q contains action-token queries, K and V contain context keys and values, and d is the key dimension. We set columns of M_t associated with a masked camera to $-\infty$, making their attention weights zero; all other entries are zero. The action targets used for supervision remain unchanged.

Figure 6 shows the retargeting error, masking probability, and measured relative head-camera attention for a transformed Retrieve Drink demonstration from the training set. Higher masking probabilities coincide with lower relative attention to the head camera. This masking encourages the policy to predict actions with less reliance on views that may become out of distribution during rollout due to camera-pose mismatch. During rollout, we retain all views without using human reference trajectories or online retargeting errors.

D. On-Policy Correction via DAgger

A policy trained on EAR-transformed demonstrations with REAM can already achieve strong performance on various tasks, but may still struggle with challenging long-horizon, contact-rich manipulation such as box folding and jacket hanging, illustrated in Figure 7. EAR and REAM reduce the human-to-robot gap, but our sensorized human interface records states rather than actions. These demonstrations therefore cannot fully specify the robot actions needed for complex contact interactions.

Nevertheless, these demonstrations provide useful motion priors. The policy already produces approximately the desired motion, making it easier to provide targeted action corrections than to teleoperate the entire task without policy assistance. We use this policy to collect on-policy corrections via DAgger [10], [20], [21].

Our policy predicts joint-space actions, whereas the VR controllers provide relative end-effector corrections. We apply these corrections to the robot's current end-effector poses and pass the resulting targets to the same WBC used for teleoperation in the Pixel system paper [11]. The WBC converts these targets into joint-space actions while using the policy's joint space predictions for posture regularization.

To support a smooth return to policy rollout after an intervention, we use real-time chunking (RTC) [22]. We retain a short history of the corrected actions and supply its final steps as a fixed prefix for the next policy prediction, with the prefix length determined by the estimated inference

delay. The flow-matching sampler holds this prefix fixed while generating the remainder of the action chunk. We resume rollout from the generated suffix, conditioning the continuation on the recent corrected motion.

We record complete rollout trajectories, including policy-generated actions and human corrections. After each round, we aggregate these robot observation–action pairs with the processed human demonstrations and robot data from earlier rounds, and train a new policy from scratch.

IV. Experiments

We evaluate Pixel-UMI on five mobile manipulation tasks to address four questions:

1) Does EAR improve task performance compared with retargeting that does not account for the environment?
2) Which action representation is most effective, and does delay-calibrated action assignment improve policy performance?
3) Which observation space is most effective, and can REAM improve transfer in the presence of retargeting errors?
4) How can we effectively use on-policy correction data collected through DAgger?

We study the first three questions on three tasks using policies trained only on human demonstrations. For the fourth, we use two additional contact-rich tasks to evaluate learning from on-policy correction data.

A. Tasks and Evaluation Metric

Figure 7 illustrates the human demonstrations and policy rollouts for all the tasks. The demonstration counts listed below refer to human demonstrations.

1) **Arrange Mouse (100 demonstrations).** The robot must retrieve a computer mouse and place it on a mouse pad on a desk while avoiding a chair positioned in front of the desk. This task evaluates whole-body coordination and collision avoidance while moving through cluttered environments, requiring the robot to dynamically adjust its posture during retrieval and placement.
2) **Pack Tennis Canister (178 demonstrations).** The robot must reach into a deep cardboard box to retrieve a tennis-ball canister, open a bag, and place the canister inside. This task evaluates constrained-space reaching and bimanual manipulation, requiring the robot to maneuver its elbow around the box rim while reaching inside.
3) **Retrieve Drink (219 demonstrations).** The robot must open a refrigerator, reach inside to retrieve a drink can, and close the door while avoiding nearby shelving. This task evaluates articulated-object manipulation and robustness through mismatches in head viewpoint.
4) **Fold Box (393 demonstrations).** The robot must fold a cardboard blank into a box, insert the side tabs, and crimp the edges to secure it. This task evaluates precise, contact-rich bimanual manipulation through a sequence of folding, insertion, and crimping operations.
5) **Hang Jacket (261 demonstrations).** The robot must pick up a hanger, insert it into a jacket, wrap the garment around the hanger, and hang it on a rack. This task evaluates deformable-object manipulation, bimanual coordination, and grasp transitions across multiple stages of manipulation.

For each experiment, we conduct 20 evaluation trials across four randomized initial configurations. We report *task progress* on a scale from 0% to 100%, with discrete task-completion stages weighted according to their difficulty.

B. Policy Details

We use single-task policies for all ablation experiments. For Arrange Mouse, Pack Tennis Canister, and Retrieve Drink, we train policies using only human demonstrations. For Fold Box and Hang Jacket, we first train on human demonstrations, then add 60 successful on-policy robot correction demonstrations per task in each of two rounds.

The policy uses pretrained DINOv3 [23] ViT-B/16 visual backbones and a flow-matching [24] action decoder, following the Pixel system paper [11]. Unless otherwise stated, the policy receives synchronized RGB images from the head camera and both wrist cameras, along with robot joint states. Head images are processed at 384×384 and wrist images at 224×224.

We initialize the action decoder from scratch and jointly fine-tune the pretrained visual backbones for 150,000 optimizer updates.

The policy predicts a chunk of 50 actions spanning one second at a control rate of 50 Hz. Each 23-dimensional action contains 20 absolute position targets for the lift, torso, arms, head, and grippers, together with the planar base velocities (v_x^B, v_y^B, ω) expressed in the robot base frame. These commands are sent directly to the low-level controllers during rollout, without an online task-space whole-body controller.

We use real-time chunking (RTC) [22] as in the system paper. This compensates for inference latency and smooths transitions between consecutive action chunks.

C. Environment Awareness

EAR optimizes robot motion offline to avoid collisions with the environment. We evaluate whether this supervision leads to collision avoidance during policy rollout and improves task performance. We choose three tasks that impose different geometric constraints. Pack Tennis Canister requires an arm configuration that keeps the elbow clear of the box rim while the gripper reaches inside. Arrange Mouse requires whole-body coordination during motion, as the robot must dynamically adjust its posture to avoid the chair while retrieving and placing the mouse. Retrieve Drink requires the arms to reach inside the refrigerator without collision while the base remains clear of the adjacent cabinet. We evaluate the following two experimental conditions:

EAR only. We use trajectories retargeted with environment collision avoidance enabled.

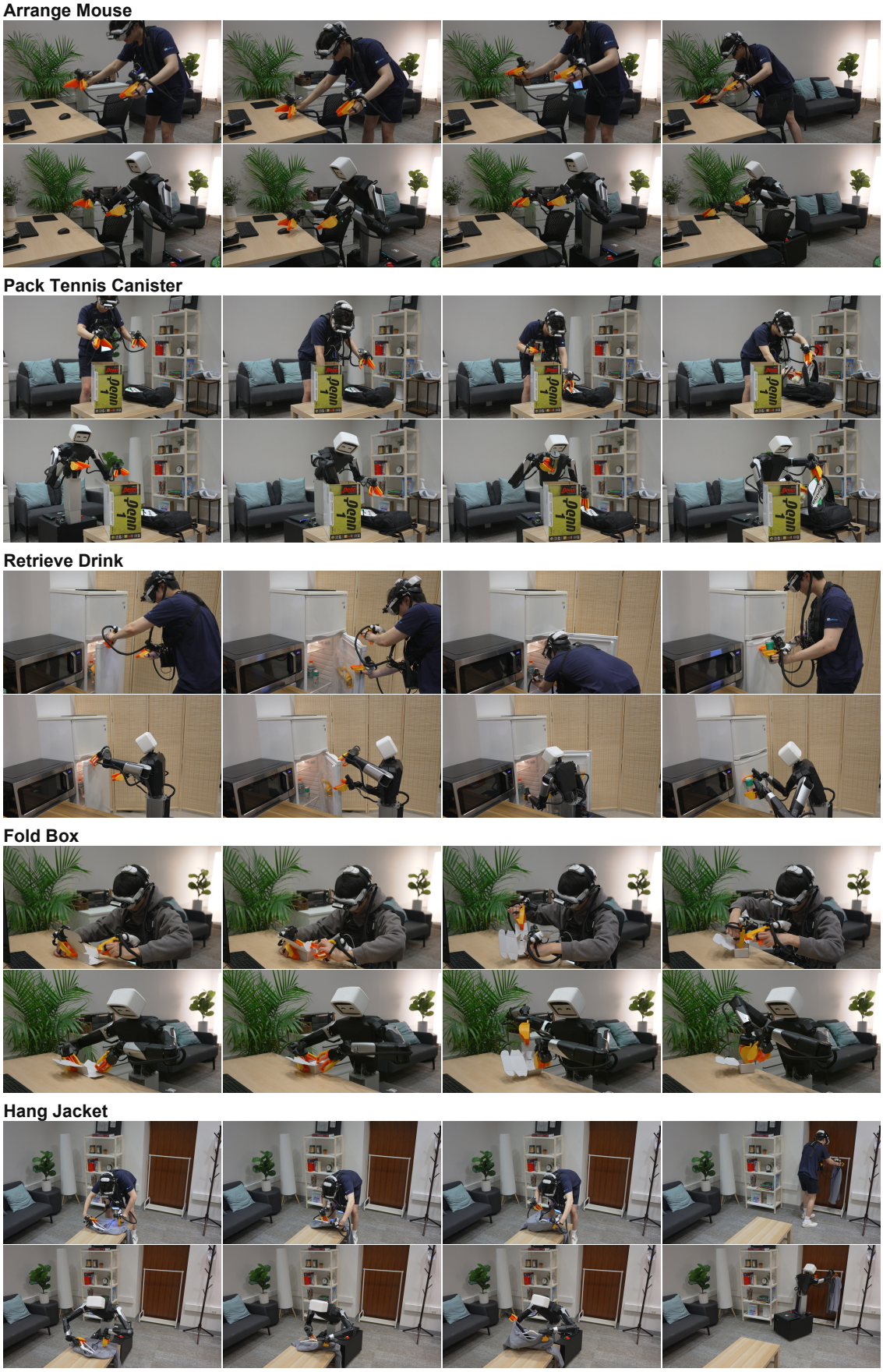


Fig. 7. We evaluate EAR, REAM, and Action Assignment across three challenging mobile manipulation tasks and evaluate on-policy correction on a further two tasks. Video results are available at https://pixel-robot.com/pixel-umi

No environment awareness. We use trajectories retargeted without environment collision avoidance.

Both conditions retain self-collision avoidance and use absolute joint-space control with the full observation space described above. REAM is disabled in both conditions.

D. Action Representation and Assignment

We first compare three action representations on Arrange Mouse, Pack Tennis Canister, and Retrieve Drink to assess how the action space affects transfer of EAR's whole-body motion.

Joint Abs. The policy predicts absolute joint-position targets together with planar base velocities, directly specifying the robot's whole-body posture.

Joint Rel. The policy expresses joint targets as relative offsets while retaining the same base-velocity commands.

EEF-Rel Base. Following the relative-base representation in the Pixel system paper [11], the policy predicts head and bimanual end-effector target poses in a shared, ground-referenced pseudo-base frame for each action chunk. We retain joint angles for the gripper. An online whole-body controller converts these targets into robot motion. This tests whether task-space targets suffice without directly learning EAR's posture choices.

Based on this comparison, we select *Joint Abs* and test whether accounting for joint response delays improves performance. We hold the retargeted state trajectories fixed and compare two action-labeling schemes:

Calibrated labels. We assign joint targets using per-joint lookahead settings calibrated from previously collected robot teleoperation data, following Section III-B.4.

Next-state labels. We use the next retargeted joint state as the action, providing a reference without per-joint delay calibration.

All action-representation and assignment conditions use EAR and the full observation space, with REAM disabled.

E. Observation Space and Retargeting Error Aware Masking

Retargeting cannot always reproduce the demonstrated head and wrist poses exactly, creating viewpoint mismatches between human training images and robot observations at deployment. We investigate how camera observations and head-tracking objectives affect policy performance on Arrange Mouse, Pack Tennis Canister, and Retrieve Drink. We compare four conditions:

All cameras (EAR only). We use standard head and wrist retargeting and train with all three camera views, with REAM disabled.

Wrist only. We retarget only the wrist poses and use only the two wrist cameras during training and deployment, with REAM disabled. This tests whether removing the head view and its tracking objective improves transfer despite the loss of visual context.

Relaxed head tracking. We reduce the head-pose tracking weight during retargeting while retaining all three cameras, with REAM disabled. This tests whether favoring wrist tracking at the expense of head alignment improves transfer.

REAM. We retain standard retargeting and all three cameras, but mask camera observations during training according to their retargeting pose errors. This tests whether error-guided masking can mitigate viewpoint mismatch while preserving visual context at deployment.

All conditions use EAR, *Joint Abs* actions, and calibrated action assignment. For REAM, we use an 80th-percentile cutoff ($\rho = 0.8$) and a maximum masking probability of 90% ($p_{\max} = 0.9$) in Equation 7. The masking probability is zero at or below the cutoff and increases quadratically for higher error-rank scores. Masking is applied only during training; all three cameras remain available at deployment.

F. On-Policy Correction

We evaluate whether DAgger [10] improves policies trained on human demonstrations and which parts of the collected robot data are most useful. We study Fold Box and Hang Jacket, which require long-horizon, contact-rich bimanual manipulation.

Starting from a human-trained policy, an operator corrects rollout errors using the procedure in Section III-D. We collect 60 successful corrected robot rollouts per task in each of two rounds, for a total of 120 per task. We compare three training conditions:

Full rollouts. We combine human demonstrations with complete corrected robot rollouts, including both policy-executed and operator-controlled segments.

Interventions only. We combine human demonstrations with only the operator-controlled segments of the same rollouts. Comparison with *Full rollouts* tests the value of retaining policy-executed segments alongside corrections.

Robot-only demos. We train on all complete robot rollouts from both rounds without human demonstrations, testing whether human data remain useful once robot correction data are available.

After each round, we retrain *Full rollouts* and *Interventions only* from scratch on the original human dataset plus their respective robot data accumulated across all completed rounds. The policy trained after round one is used to collect round two. All conditions use uniform sampling.

V. Results

Videos of results are available at https://pixel-robot.com/pixel-umi

A. EAR Improves Whole-Body Manipulation

As shown in Figure 8, EAR improves task progress across all three tasks, increasing average progress by 64.41 percentage points over retargeting without environment awareness. The largest gains occur in Arrange Mouse and Pack Tennis Canister, where nearby obstacles constrain the body configurations needed to complete the task.

In Pack Tennis Canister, the policy trained without environment awareness hits the box rim with its elbow as the gripper descends. This collision prevents the robot from reaching the canister, resulting in zero task progress. EAR instead supplies training trajectories that keep the elbow

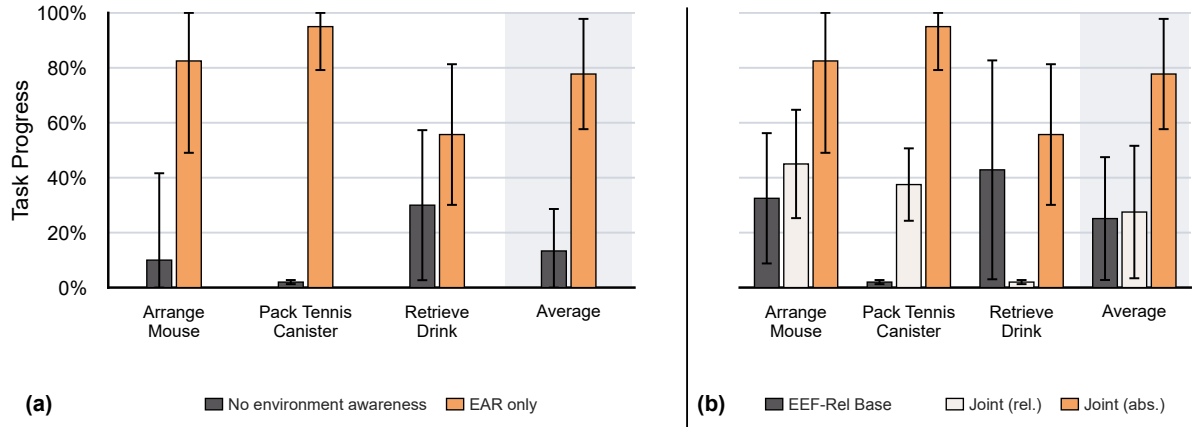


Fig. 8. **EAR and action-representation ablations.** (a) EAR only versus retargeting without environment awareness. (b) EEF-Rel Base versus Joint Rel and Joint Abs. Both panels omit REAM.

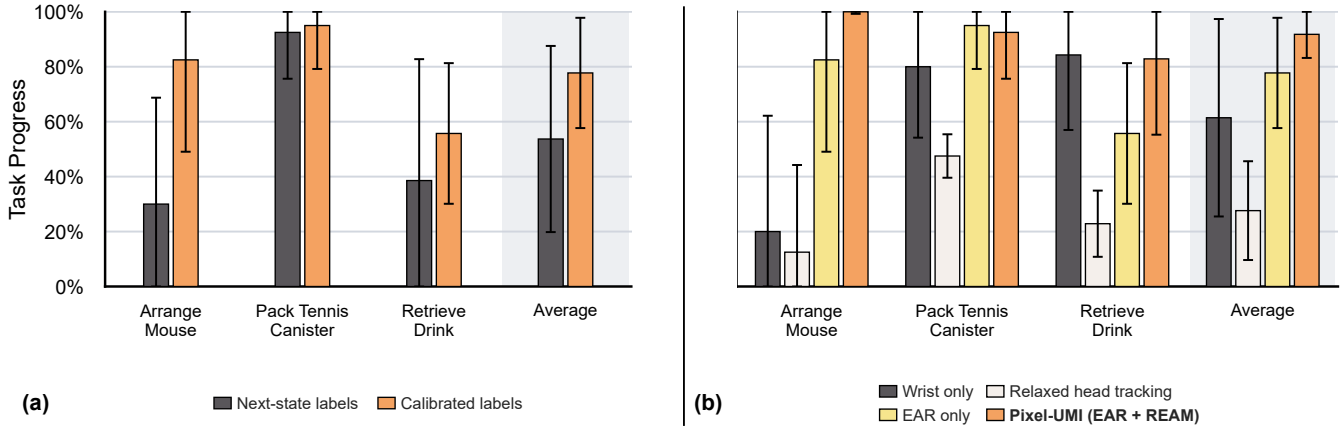


Fig. 9. **Action-assignment and REAM ablations.** (a) Robot-calibrated command labels versus next-state labels, with REAM omitted. (b) Pixel-UMI (EAR + REAM) versus EAR only, wrist-only observations, and relaxed head tracking.

clear of the rim while allowing the gripper to reach into the box. In Arrange Mouse, the policy trained without environment awareness often collides with the chair, either while approaching the mouse or after grasping it and moving toward the mouse pad. Avoiding these collisions requires coordinating the robot's body throughout both retrieval and placement. EAR also improves Retrieve Drink, where the robot must reach into the refrigerator while remaining clear of nearby shelving and cabinets.

The qualitative comparison in Figure 4 illustrates how incorporating scene geometry during retargeting changes the robot's whole-body posture. These scene-aware configurations provide the policies with a strong prior for moving around obstacles while continuing the demonstrated manipulation. The results support using environment geometry during offline retargeting to teach policies the coordinated body motion needed for manipulation in constrained spaces.

B. Absolute Joint Targets and Delay Calibration Improve Policy Performance

With EAR enabled, Joint Abs outperforms EEF-Rel Base and Joint Rel across all three tasks in Figure 8, achieving 77.74% average task progress compared with 25.12% and 27.50%, respectively. EEF-Rel Base specifies head and end-effector poses but leaves the remaining body configuration to an online whole-body controller. These targets underspecify the posture needed to avoid nearby obstacles, and the controller often produces motions that collide with the environment. In Pack Tennis Canister, this representation fails to preserve the elbow configuration needed to reach inside the box, resulting in zero task progress.

Joint Rel retains joint-level supervision from EAR's environment-aware trajectories. However, errors accumulate across relative commands, causing the robot to drift away from the demonstrated trajectories and quickly encounter states outside the training distribution. This is particularly problematic for tasks requiring precise positioning. In Retrieve Drink, the gripper must enter a small notch in the refrigerator door to open it. With Joint Rel, accumulated errors often cause the gripper to miss this notch, preventing further progress. Joint Abs directly specifies the target joint configuration at each step, helping preserve both the demonstrated posture and the precision needed for these interactions.

Holding Joint Abs fixed, we next compare calibrated action assignment with next-state labels. Figure 9 shows

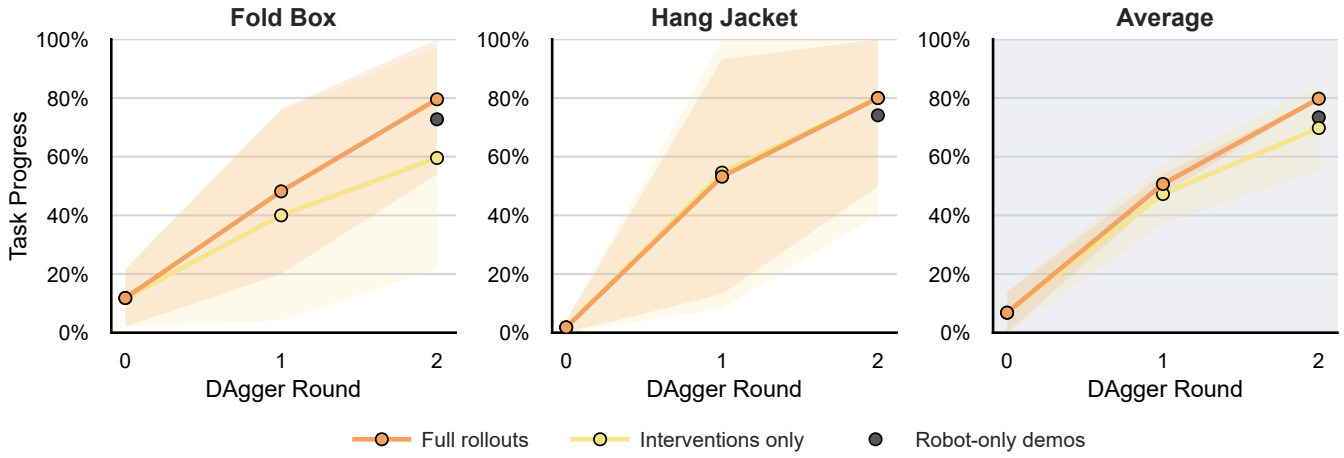


Fig. 10. **Policy improvement through DAgger.** We compare training on complete corrected rollouts with intervention segments only. Gray markers show policies trained on robot demonstrations alone.

that calibration increases average task progress by 24.05 percentage points, with the clearest gains on Arrange Mouse and Retrieve Drink. Pack Tennis Canister benefits relatively little. We hypothesize that faster motions amplify tracking errors caused by joint response delays. Arrange Mouse requires large, coordinated motions while avoiding the chair, making accurate timing across joints particularly important. For slower, posture-dominated motions, such as reaching into the box in Pack Tennis Canister, these dynamic differences may have less impact on policy performance.

C. REAM Improves Performance with All Camera Views

The observation ablations in Figure 9 highlight the importance of head-camera context for whole-body manipulation. Although the wrist-only condition performs well on Pack Tennis Canister and Retrieve Drink, it achieves only 20% task progress on Arrange Mouse, compared with 82.5% for EAR only using all cameras. During parts of Arrange Mouse, the arms are positioned in front of the chair, leaving this obstacle outside the wrist-camera views. The head camera provides the broader scene context needed to coordinate motion around it. Consequently, relying only on wrist observations produces uneven performance across tasks with different visual-context requirements.

Retaining all cameras without REAM performs well on Arrange Mouse and Pack Tennis Canister, but underperforms the wrist-only condition on Retrieve Drink. The robot's kinematic constraints prevent accurate reproduction of the demonstrated head poses, creating a mismatch between human training images and robot observations during rollout. In Retrieve Drink, EAR-only failures frequently occur at the task stage corresponding to the largest head-pose errors in the demonstrations. This correspondence is consistent with our hypothesis that viewpoint mismatch contributes to the performance degradation.

To examine this tradeoff, we relax the head-tracking objective to prioritize wrist tracking while retaining all camera observations. Performance degrades further across all three tasks relative to EAR only. These results suggest that prioritizing wrist tracking at the expense of head alignment does not resolve the observation mismatch when the policy continues to receive the head view.

Combining all cameras with REAM achieves the highest average task progress among the observation conditions, at 91.79%, an improvement of 14.05 percentage points over EAR only. REAM recovers most of the performance lost on Retrieve Drink while retaining the head-camera context needed for Arrange Mouse; performance on Pack Tennis Canister remains comparable. This provides strong performance across tasks without removing a camera from the policy.

Figure 6 provides a qualitative example from a Retrieve Drink training demonstration. Relative attention to the head camera decreases during the interval with the largest head-pose error. This aligns with the intended role of REAM: reducing reliance on poorly matched views at particular moments in a trajectory while preserving access to useful context from the available cameras.

D. DAgger Improves Performance Across Rounds

On Fold Box and Hang Jacket, each round of DAgger improves policy performance, as shown in Figure 10. With Full rollouts, average task progress across the two tasks increases from 6.82% for the initial human-trained policy to 50.68% after the first round and 79.77% after the second. Interventions only also improves substantially, reaching 47.27% and 69.77% after the corresponding rounds. Retaining complete corrected rollouts therefore yields higher average performance than training only on intervention segments.

We observe that each round enables the policy to complete additional task stages, allowing subsequent corrections to address failures later in the task. This progression helps explain the approximately linear increase in task progress over the two evaluated rounds: successive rounds extend how far the policy can execute the task successfully. We expect longer tasks with more subtask stages to require additional rounds of correction to achieve comparable progress.

Robot-only demos also performs strongly, achieving 73.41% average task progress using the same collected robot rollouts without human demonstrations in the final training dataset, compared with 79.77% when human demonstrations are retained. However, these rollouts were collected by correcting policies initially trained on human demonstrations. Collecting successful robot demonstrations for these tasks is often difficult without this initial policy, so human demonstrations remain valuable for making correction-data collection practical.

VI. Related Work

A. Sensorized Human Demonstrations and Embodiment Transfer

Sensorized human interfaces enable demonstration collection without requiring a robot at the collection site. Gripper-based interfaces reduce the embodiment gap by aligning the demonstration tool and camera placement with the robot's manipulation interface, with subsequent designs improving tracking, ergonomics, and collection scalability [25], [26], [27], [3], [28], [29]. Wearable interfaces offer alternative tradeoffs: exoskeletons align recorded motion with robot kinematics [30], [31], [32], while instrumented gloves capture natural hand motion for subsequent retargeting [33], [34]. Egocentric bare-hand capture further reduces instrumentation, using human–robot co-training to connect human observations and motion to robot behavior [35]. Related industrial efforts also emphasize scalable collection of sensorized-interaction data [4], [36].

Beyond hand-centric manipulation, interfaces that jointly record head and hand motion provide broader scene context and demonstrations of active perception [6], [37]. Mobile manipulation systems combine human motion with robot data or whole-body control to coordinate navigation and manipulation [38], [5]. Humanoid extensions explore both additional pelvis and foot tracking to specify body motion and simpler head-and-hand interfaces that delegate the remaining coordination to learned controllers [39], [40].

Transferring these demonstrations requires aligning both observations and actions across embodiments. Visual adaptation methods replace human appearance with robot imagery [31], [32], mask embodiment-specific features or use geometry-aware representations [35], [5], and reproject observations toward the robot viewpoint [40]. Motion adaptation addresses robot kinematic constraints through retargeting and constrained control [38], [5], while latency compensation and controller-aware trajectory adjustment account for differences between recorded motion and executed actions [3], [40]. Force-sensing interfaces additionally capture interaction wrenches for learning force-aware and compliant manipulation [8], [9].

Our work complements these approaches through scene-aware whole-body retargeting, retargeting-error-guided camera masking, and delay-calibrated action assignment, followed by on-policy correction.

B. Learning for Whole-Body Mobile Manipulation

Behavior cloning naturally extends to mobile manipulation, provided that suitable demonstration data can be collected. Whole-body teleoperation provides demonstrations of coordinated base and arm motion, enabling learned policies to perform diverse household tasks [1], [41], [2]. Action-chunking and diffusion-based policies provide effective architectures for learning visuomotor behavior from these demonstrations [42], [43]. Simulation can further expand source demonstrations across new task configurations while accounting for reachability and camera visibility [44].

Human demonstrations provide a complementary source of supervision without requiring mobile robot teleoperation. Several approaches learn task-space policies from human demonstrations and use inverse kinematics or whole-body controllers to execute their predictions [45], [46], [5]. Other methods combine human mobile-manipulation data with stationary robot demonstrations [38], or extend human demonstration learning to humanoids through hierarchical policies and learned motion controllers [39], [40]. Offline retargeting instead converts recorded human motion into robot joint trajectories before policy training [47].

Reinforcement learning provides another route to acquiring coordinated mobile-manipulation behavior. Simulation-trained policies learn base placement and navigation conditioned on manipulation goals [48], [49], coupled locomotion and arm control [50], and visual coordination of movement, manipulation, and perception [51], [52]. Hybrid approaches combine a demonstration-trained visual planner with an RL controller [53]. COWBOY combines environment-aware whole-body control for reaching with local RL experts for object interaction, then distills their behavior into a single visual whole-body policy [54]. Beyond simulation, autonomous real-world RL has also enabled navigation and grasping to improve through robot interaction [55].

Our work focuses on learning whole-body policies from sensorized human demonstrations, using environment-aware retargeting to construct robot motion supervision and on-policy correction to improve performance.

C. Continual Improvement of Robot Policies

Once an initial policy is available, further interaction can improve its reliability and task performance. Some reinforcement learning approaches learn residual corrections around a frozen behavior-cloning policy [56], [57], or combine learned action edits with value-based action selection [58], [59]. Other approaches directly fine-tune diffusion policies [60], [61] or train lightweight actor–critic heads using pretrained VLA representations [62]. RL can also generate data for subsequent supervised refinement, using residual specialists to collect recovery trajectories that are distilled back into a generalist policy [63].

Interactive imitation learning improves policies through expert feedback on their own execution. DAgger addresses distribution shift by collecting expert action labels at states visited by the learned policy and iteratively training on the aggregated dataset [10]. Subsequent methods make this

process more practical by selectively requesting supervision. Human-gated approaches let operators decide when to take control [20], while robot-gated approaches use novelty and estimated risk to request interventions within a supervision budget [64].

For manipulation, the correction interface determines what feedback humans can provide. Compliant bilateral teleoperation supports smooth handovers between autonomous execution and human control [65], while kinesthetic interfaces allow humans to provide residual motion and force corrections during ongoing policy execution [66]. Correction protocols also shape the behaviors learned: combining recovery to familiar states with corrective task completion teaches policies to retry and adapt during long-horizon manipulation [21].

Another question is how to learn from rollouts containing both autonomous rollouts and human corrections. Weighted imitation methods balance these data sources and use intervention signals to prioritize informative behavior [67], [68]. Preference-based approaches additionally use these signals to reinforce corrective actions and suppress failure-prone behavior [69]. Distributed systems extend this improvement loop across robot fleets, coupling experience collection with frequent policy updates through either interactive imitation learning or RL [70]. We study interactive improvement for whole-body policies initialized from sensorized human demonstrations, comparing aggregation of complete corrected rollouts with intervention segments alone.

VII. Conclusion and Limitations

We present Pixel-UMI, a pipeline for learning whole-body mobile manipulation from sensorized human demonstrations. Our results highlight the value of absolute joint-space actions for preserving body configurations that head and end-effector targets leave unspecified, and of environment geometry for determining those configurations. Together with calibrated action assignment and REAM, this approach supports challenging mobile and dexterous manipulation, achieving 87.0% average task progress across five tasks. For contact-rich tasks, on-policy corrections progressively extend the stages the policy can complete, increasing average task progress from 6.82% to 79.77% over two rounds. These findings demonstrate how human demonstrations can provide useful whole-body motion supervision and a starting point for further improvement through robot interaction.

Several limitations remain. Delay-calibrated action assignment provides approximate command labels but cannot recover the contact-dependent actions underlying demonstrated motion. In particular, motion alone does not reveal the forces applied when contact prevents further movement. Force/torque or tactile sensing, combined with robot dynamics and controller models, could improve action inference and reduce reliance on corrective robot data.

Retargeting quality also depends on the reconstructed environment. Our reconstruction uses a single head-mounted depth camera, leaving occluded or unobserved surfaces incomplete; additional viewpoints and sensors could improve coverage and accuracy. The static-scene formulation further limits how moving obstacles and intended contacts are represented. Finally, although retargeting generates whole-body trajectories, the demonstrated motion targets specify only the head and grippers. Intentional contacts involving other body parts, such as using an elbow to hold a door open or a hip to close a drawer, are therefore not explicitly captured. Supporting these behaviors would require richer body-motion or contact supervision and retargeting objectives that account for the intended interactions.

Acknowledgment

We would like to thank Philip Han and Ray Liu for their help with data collection. We would also like to thank Kenneth Shaw, Arthur Allshire, Ritvik Singh, Andrew Wang, Zheyuan Hu, Peiqi Liu, Sri Anumakonda, Steven Oh, Adam Kan, Haoran Huan, Jiahui Yang, Hengkai Pan, Lili Chen, Mihir Prabhudesai, Wenli Xiao, and Tairan He for their thoughtful discussions. This work was supported by AFOSR FA9550-23-1-0747, ONR MURI N00014-22-1-2773 and ONR MURI N00014-24-1-2748.

References

[1] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile ALOHA: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," in *Conference on Robot Learning*, 2024. [Online]. Available: https://mobile-aloha.github.io/

[2] Y. Jiang, R. Zhang, J. Wong, C. Wang, Y. Ze, H. Yin, C. Gokmen, S. Song, J. Wu, and L. Fei-Fei, "BEHAVIOR robot suite: Streamlining real-world whole-body manipulation for everyday household activities," in *Proceedings of The 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 305, 2025, pp. 1246–1281. [Online]. Available: https://proceedings.mlr.press/v305/jiang25b.html

[3] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *Robotics: Science and Systems*, 2024. [Online]. Available: https://umi-gripper.github.io/

[4] Sunday Team, "ACT-1: A robot foundation model trained on zero robot data," Sunday Robotics research blog, Nov. 2025. [Online]. Available: https://www.sunday.ai/blog/no-robot-data

[5] X. Xu, J. Park, H. Zhang, E. Cousineau, A. Bhat, J. Barreiros, D. Wang, J. Bohg, and S. Song, "HoMMI: Learning whole-body mobile manipulation from human demonstrations," *arXiv preprint arXiv:2603.03243*, 2026. [Online]. Available: https://arxiv.org/abs/2603.03243

[6] Q. Zeng, C. Li, J. St. John, Z. Zhou, J. Wen, G. Feng, Y. Zhu, and Y. Xu, "ActiveUMI: Robotic manipulation with active perception from robot-free human demonstrations," *arXiv preprint arXiv:2510.01607*, 2025. [Online]. Available: https://arxiv.org/abs/2510.01607

[7] M. Xie, K. Van Wyk, A. Li, M. A. Rana, Q. Wan, D. Fox, B. Boots, and N. Ratliff, "Geometric fabrics for the acceleration-based design of robotic motion," *arXiv preprint arXiv:2010.14750*, 2020. [Online]. Available: https://arxiv.org/abs/2010.14750

[8] W. Liu, J. Wang, Y. Wang, W. Wang, and C. Lu, "ForceMimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation," in *IEEE International Conference on Robotics and Automation*, 2025. [Online]. Available: https://arxiv.org/abs/2410.07554

[9] H. Choi, Y. Hou, C. Pan, S. Hong, A. Patel, X. Xu, M. R. Cutkosky, and S. Song, "In-the-wild compliant manipulation with UMI-FT," *arXiv preprint arXiv:2601.09988*, 2026. [Online]. Available: https://arxiv.org/abs/2601.09988

[10] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 15, 2011, pp. 627–635. [Online]. Available: https://proceedings.mlr.press/v15/ross11a.html

[11] T. Tao, J. J. Liu, R. Salakhutdinov, and D. Pathak, "Pixel: A co-designed full-stack robot system for learning general-purpose mobile manipulation," 2026, preprint. [Online]. Available: https://pixel-robot.com/pixel/

[12] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[13] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image and Vision Computing*, vol. 10, no. 3, pp. 145–155, 1992.

[14] S. Choi, Q.-Y. Zhou, and V. Koltun, "Robust reconstruction of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5556–5565. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2015/html/Choi_Robust_Reconstruction_of_2015_CVPR_paper.html

[15] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An efficient probabilistic 3D mapping framework based on octrees," *Autonomous Robots*, 2013. [Online]. Available: https://octomap.github.io/

[16] J. Schauer and A. Nüchter, "The peopleremover—removing dynamic objects from 3-D point cloud data by traversing a voxel occupancy grid," *IEEE Robotics and Automation Letters*, 2018. [Online]. Available: https://robotik.informatik.uni-wuerzburg.de/telematics/download/RA-L_2018/schauer-ral18-peopleremover.pdf

[17] K. Zakka, "Mink: Python inverse kinematics based on MuJoCo," Software, version 1.1.0, 2026. [Online]. Available: https://github.com/kevinzakka/mink

[18] S. Caron, Y. De Mont-Marin, R. Budhiraja, S. H. Bang, I. Domrachev, S. Nedelchev, P. Du, J. Vaillant, B. Wingo, S. Patapati, D. San José Pro, and N. G. Marticorena Vidal, "Pink: Python inverse kinematics based on Pinocchio," Software, version 4.4.0, 2026. [Online]. Available: https://github.com/pink-kinematics/pink

[19] Sunday Robotics, "Technology," Technical overview, accessed October 19, 2026. [Online]. Available: https://www.sunday.ai/technology

[20] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer, "HG-DAgger: Interactive imitation learning with human experts," *arXiv preprint arXiv:1810.02890*, 2019. [Online]. Available: https://arxiv.org/abs/1810.02890

[21] Z. Hu, R. Wu, N. Enock, J. Li, R. Kadakia, Z. Erickson, and A. Kumar, "RaC: Robot learning for long-horizon tasks by scaling recovery and correction," *arXiv preprint arXiv:2509.07953*, 2025. [Online]. Available: https://arxiv.org/abs/2509.07953

[22] K. Black, A. Z. Ren, M. Equi, and S. Levine, "Training-time action conditioning for efficient real-time chunking," *arXiv preprint arXiv:2512.05964*, 2025. [Online]. Available: https://arxiv.org/abs/2512.05964

[23] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, "DINOv3," *arXiv preprint arXiv:2508.10104*, 2025. [Online]. Available: https://arxiv.org/abs/2508.10104

[24] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2023. [Online]. Available: https://arxiv.org/abs/2210.02747

[25] S. Song, A. Zeng, J. Lee, and T. Funkhouser, "Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4978–4985, 2020. [Online]. Available: https://arxiv.org/abs/1912.04344

[26] S. Young, D. Gandhi, S. Tulsiani, A. Gupta, P. Abbeel, and L. Pinto, "Visual imitation made easy," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 155. PMLR, 2021, pp. 1992–2005. [Online]. Available: https://proceedings.mlr.press/v155/young21a.html

[27] N. M. M. Shafiullah, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto, "On bringing robots home," *arXiv preprint arXiv:2311.16098*, 2023. [Online]. Available: https://arxiv.org/abs/2311.16098

[28] Z. Zhaxizhuoma, K. Liu, C. Guan, Z. Jia, Z. Wu, X. Liu, T. Wang, S. Liang, P. Chen, P. Zhang, H. Song, D. Qu, D. Wang, Z. Wang, N. Cao, Y. Ding, B. Zhao, and X. Li, "FastUMI: A scalable and hardware-independent universal manipulation interface with dataset," in *Proceedings of the 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 305. PMLR, 2025, pp. 3069–3093. [Online]. Available: https://proceedings.mlr.press/v305/zhaxizhuoma25a.html

[29] T. Ohkawa, J. Arima, Y. Noguchi, M. Tateno, M. Sugiura, T. Okubo, K. Ikeuchi, Y. Shin, H. Nishizawa, N. Kanazawa, Y. Wakayama, D. Fukunaga, K. Makihara, T. Motoda, F. Erich, Y. Domae, T. Matsushima, Y. Okumatsu, and K. Ota, "YUBI: Yielding universal bidigital interface for bimanual dexterous manipulation at scale," 2026. [Online]. Available: https://arxiv.org/abs/2606.10244

[30] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu, "AirExo: Low-cost exoskeletons for learning whole-arm manipulation in the wild," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 15 031–15 038. [Online]. Available: https://doi.org/10.1109/ICRA57147.2024.10610799

[31] H. Fang, C. Wang, Y. Wang, J. Chen, S. Xia, J. Lv, Z. He, X. Yi, Y. Guo, X. Zhan, L. Yang, W. Wang, C. Lu, and H.-S. Fang, "AirExo-2: Scaling up generalizable robotic imitation learning with low-cost exoskeletons," *arXiv preprint arXiv:2503.03081*, 2025. [Online]. Available: https://arxiv.org/abs/2503.03081

[32] M. Xu, H. Zhang, Y. Hou, Z. Xu, L. Fan, M. Veloso, and S. Song, "DexUMI: Using human hand as the universal manipulation interface for dexterous manipulation," *arXiv preprint arXiv:2505.21864*, 2025. [Online]. Available: https://arxiv.org/abs/2505.21864

[33] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and K. Liu, "DexCap: Scalable and portable mocap data collection system for dexterous manipulation," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, Jul. 2024. [Online]. Available: https://www.roboticsproceedings.org/rss20/p043.html

[34] T. Tao, M. K. Srirama, J. J. Liu, K. Shaw, and D. Pathak, "DexWild: Dexterous human interactions for in-the-wild robot policies," in *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, Jun. 2025. [Online]. Available: https://www.roboticsproceedings.org/rss21/p075.html

[35] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "EgoMimic: Scaling imitation learning via egocentric video," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 13 226–13 233. [Online]. Available: https://arxiv.org/abs/2410.24221

[36] Generalist Team, "GEN-0: Embodied foundation models that scale with physical interaction," *Generalist AI Blog*, Nov. 2025, first-party technical blog; not peer reviewed. [Online]. Available: https://generalistai.com/blog/gen-0

[37] J. Yu, Y. Shentu, D. Wu, P. Abbeel, K. Goldberg, and P. Wu, "EgoMI: Learning active vision and whole-body manipulation from egocentric human demonstrations," 2025. [Online]. Available: https://arxiv.org/abs/2511.00153

[38] L. Y. Zhu, P. Kuppili, R. Punamiya, P. Aphiwetsa, D. Patel, S. Kareer, S. Ha, and D. Xu, "EMMA: Scaling mobile manipulation via egocentric human data," *IEEE Robotics and Automation Letters*, vol. 11, no. 3, pp. 3087–3094, 2026. [Online]. Available: https://doi.org/10.1109/LRA.2026.3653320

[39] R. Nai, B. Zheng, J. Zhao, H. Zhu, S. Dai, Z. Chen, Y. Hu, Y. Hu, T. Zhang, C. Wen, and Y. Gao, "Humanoid manipulation interface: Humanoid whole-body manipulation from robot-free demonstrations," *arXiv preprint arXiv:2602.06643*, 2026. [Online]. Available: https://arxiv.org/abs/2602.06643

[40] Z. Zhao, Y. Zhao, G. Zhang, C. Liu, M. Zheng, and W. Lian, "HALOMI: Learning humanoid loco-manipulation with active perception from human demonstrations," 2026. [Online]. Available: https://arxiv.org/abs/2606.18772

[41] J. Wu, W. Chong, R. Holmberg, A. Prasad, Y. Gao, O. Khatib, S. Song, S. Rusinkiewicz, and J. Bohg, "TidyBot++: An open-source holonomic mobile manipulator for robot learning," in *Proceedings of the 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270. PMLR, 2025, pp. 3729–3741. [Online]. Available: https://proceedings.mlr.press/v270/wu25a.html

[42] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proceedings of Robotics: Science and Systems*, 2023. [Online]. Available: https://roboticsproceedings.org/rss19/p016.html

[43] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. C. Burchfiel, and S. Song, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, Jul. 2023.

[44] C. Li, M. Xu, A. Bahety, H. Yin, Y. Jiang, H. Huang, J. Wong, S. Garlanka, C. Gokmen, R. Zhang, W. Liu, J. Wu, R. Martín-Martín, and L. Fei-Fei, "MoMaGen: Generating demonstrations under soft and hard constraints for multi-step bimanual mobile manipulation," in *International Conference on Learning Representations*, 2026. [Online]. Available: https://arxiv.org/abs/2510.18316

[45] H. Ha, Y. Gao, Z. Fu, J. Tan, and S. Song, "UMI-on-Legs: Making manipulation policies mobile with manipulation-centric whole-body controllers," in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270, 2025, pp. 5254–5270. [Online]. Available: https://proceedings.mlr.press/v270/ha25a.html

[46] M. Seo, H. A. Park, S. Yuan, Y. Zhu, and L. Sentis, "LEGATO: Cross-embodiment imitation using a grasping tool," *IEEE Robotics and Automation Letters*, vol. 10, no. 3, pp. 2854–2861, 2025. [Online]. Available: https://arxiv.org/abs/2411.03682

[47] Z. Chen, C. Kong, C. Zhang, Y. Yang, L. Y. Zhu, S. Kousik, and D. Xu, "WARP: Whole-body retargeting for learning from offline human demonstrations," *arXiv preprint arXiv:2606.29940*, 2026. [Online]. Available: https://arxiv.org/abs/2606.29940

[48] D. Honerkamp, T. Welschehold, and A. Valada, "N^2M^2: Learning navigation for arbitrary mobile manipulation motions in unseen and dynamic environments," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3601–3619, 2023. [Online]. Available: https://arxiv.org/abs/2206.08737

[49] S. Jauhri, J. Peters, and G. Chalvatzaki, "Robot learning of mobile manipulation with reachability behavior priors," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8399–8406, 2022. [Online]. Available: https://arxiv.org/abs/2203.04051

[50] Z. Fu, X. Cheng, and D. Pathak, "Deep whole-body control: Learning a unified policy for manipulation and locomotion," in *Proceedings of The 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 205, 2023, pp. 138–149. [Online]. Available: https://proceedings.mlr.press/v205/fu23a.html

[51] M. Liu, Z. Chen, X. Cheng, Y. Ji, R.-Z. Qiu, R. Yang, and X. Wang, "Visual whole-body control for legged loco-manipulation," in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270, 2025, pp. 234–257. [Online]. Available: https://proceedings.mlr.press/v270/liu25b.html

[52] S. Uppal, A. Agarwal, H. Xiong, K. Shaw, and D. Pathak, "SPIN: Simultaneous perception, interaction and navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 133–18 142. [Online]. Available: https://arxiv.org/abs/2405.07991

[53] Z. He, K. Lei, Y. Ze, K. Sreenath, Z. Li, and H. Xu, "Learning visual quadrupedal loco-manipulation from demonstrations," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 9102–9109. [Online]. Available: https://arxiv.org/abs/2403.20328

[54] J. Yang, J. J. Liu, R. M. Liu, A. E. Wang, K. Shaw, D. Pathak, and R. Salakhutdinov, "COWBOY: A scalable sim-to-real framework for learning contextual whole-body manipulation," in *Conference on Robot Learning*, 2026. [Online]. Available: https://arxiv.org/abs/2610.28948

[55] C. Sun, J. Orbik, C. M. Devin, B. H. Yang, A. Gupta, G. Berseth, and S. Levine, "Fully autonomous real-world reinforcement learning with applications to mobile manipulation," in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 164, 2022, pp. 308–319. [Online]. Available: https://proceedings.mlr.press/v164/sun22a.html

[56] L. Ankile, A. Simeonov, I. Shenfeld, M. Torne, and P. Agrawal, "From imitation to refinement—residual RL for precise assembly," *arXiv preprint arXiv:2407.16677*, 2024. [Online]. Available: https://arxiv.org/abs/2407.16677

[57] L. Ankile, Z. Jiang, R. Duan, G. Shi, P. Abbeel, and A. Nagabandi, "Residual off-policy RL for finetuning behavior cloning policies," *arXiv preprint arXiv:2509.19301*, 2025. [Online]. Available: https://arxiv.org/abs/2509.19301

[58] P. Dong, Q. Li, D. Sadigh, and C. Finn, "EXPO: Stable reinforcement learning with expressive policies," in *International Conference on Learning Representations*, 2026. [Online]. Available: https://arxiv.org/abs/2507.07986

[59] P. Dong, K.-H. Hung, T. Gao, D. Sadigh, and C. Finn, "EXPO-FT: Sample-efficient reinforcement learning finetuning for vision-language-action models," *arXiv preprint arXiv:2605.25477*, 2026. [Online]. Available: https://arxiv.org/abs/2605.25477

[60] A. Z. Ren, J. Lidard, L. L. Ankile, A. Simeonov, P. Agrawal, A. Majumdar, B. Burchfiel, H. Dai, and M. Simchowitz, "Diffusion policy policy optimization," in *International Conference on Learning Representations*, 2025. [Online]. Available: https://arxiv.org/abs/2409.00588

[61] K. Lei, H. Li, D. Yu, Z. Wei, L. Guo, Z. Jiang, Z. Wang, S. Liang, and H. Xu, "RL-100: Performant robotic manipulation with real-world reinforcement learning," *arXiv preprint arXiv:2510.14830*, 2025. [Online]. Available: https://arxiv.org/abs/2510.14830

[62] C. Xu, J. T. Springenberg, M. Equi, A. Amin, A. Esmail, S. Levine, and L. Ke, "RL Token: Bootstrapping online RL with vision-language-action models," *arXiv preprint arXiv:2604.23073*, 2026. [Online]. Available: https://arxiv.org/abs/2604.23073

[63] W. Xiao, H. Lin, A. Peng, H. Xue, T. He, Y. Xie, F. Hu, J. Wu, Z. Luo, L. J. Fan, G. Shi, and Y. Zhu, "Self-improving vision-language-action models with data generation via residual RL," in *International Conference on Learning Representations*, 2026. [Online]. Available: https://arxiv.org/abs/2511.00091

[64] R. Hoque, A. Balakrishna, E. Novoseller, A. Wilcox, D. S. Brown, and K. Goldberg, "ThriftyDAgger: Budget-aware novelty and risk gating for interactive imitation learning," in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 164, 2022, pp. 598–608. [Online]. Available: https://proceedings.mlr.press/v164/hoque22a.html

[65] P. Wu, Y. Shentu, Q. Liao, D. Jin, M. Guo, K. Sreenath, X. Lin, and P. Abbeel, "RoboCopilot: Human-in-the-loop interactive imitation learning for robot manipulation," *arXiv preprint arXiv:2503.07771*, 2025. [Online]. Available: https://arxiv.org/abs/2503.07771

[66] X. Xu, Y. Hou, Z. Liu, and S. Song, "Compliant residual DAgger: Improving real-world contact-rich manipulation with human corrections," in *Advances in Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=cjcm5LYVWm

[67] A. Mandlekar, D. Xu, R. Martín-Martín, Y. Zhu, L. Fei-Fei, and S. Savarese, "Human-in-the-loop imitation learning using remote teleoperation," *arXiv preprint arXiv:2012.06733*, 2020. [Online]. Available: https://arxiv.org/abs/2012.06733

[68] H. Liu, S. Nasiriany, L. Zhang, Z. Bao, and Y. Zhu, "Robot learning on the job: Human-in-the-loop autonomy and learning during deployment," in *Robotics: Science and Systems*, 2023. [Online]. Available: https://ut-austin-rpl.github.io/sirius/

[69] W. Xia, Y. Yang, H. Wu, X. Ma, T. Kong, and D. Hu, "Human-assisted robotic policy refinement via action preference optimization," in *Advances in Neural Information Processing Systems*, 2025. [Online]. Available: https://arxiv.org/abs/2506.07127

[70] M. Pan, S. Feng, Q. Zhang, X. Li, J. Song, C. Qu, Y. Wang, C. Li, Z. Xiong, Z. Chen, Y. Liu, and J. Luo, "SOP: A scalable online post-training system for vision-language-action models," *arXiv preprint arXiv:2601.03044*, 2026. [Online]. Available: https://arxiv.org/abs/2601.03044