

LEGATO: Cross-Embodiment Imitation Using a Grasping Tool

Mingyo Seo^{1,2}, H. Andy Park², Shenli Yuan², Yuke Zhu^{1†}, and Luis Sentis^{1,2†}

Abstract—Cross-embodiment imitation learning enables policies trained on specific embodiments to transfer across different robots, unlocking the potential for large-scale imitation learning that is both cost-effective and highly reusable. This paper presents LEGATO, a cross-embodiment imitation learning framework for visuomotor skill transfer across varied kinematic morphologies. We introduce a handheld gripper that unifies action and observation spaces, allowing tasks to be defined consistently across robots. We train visuomotor policies on task demonstrations using this gripper through imitation learning, applying transformation to a motion-invariant space for computing the training loss. Gripper motions generated by the policies are retargeted into high-degree-of-freedom whole-body motions using inverse kinematics for deployment across diverse embodiments. Our evaluations in simulation and real-robot experiments highlight the framework’s effectiveness in learning and transferring visuomotor skills across various robots. More information can be found on the project page: <https://ut-hcrl.github.io/LEGATO>.

Index Terms—Imitation Learning, Transfer Learning, Whole-Body Motion Planning and Control

I. INTRODUCTION

RECENT advancements in robot hardware—from wheeled manipulators to humanoid robots [1–5]—have greatly increased access to diverse robotic platforms. To fully leverage these advancements in supporting human activities, robots must autonomously perform a wide range of complex tasks. Deep imitation learning has shown promise in training autonomous policies for sensorimotor skills, reducing the need for extensive human programming compared to traditional rule-based approaches. It has yielded impressive results in complex robotic systems [6, 7] and across diverse dexterous manipulation tasks [8–10]. However, such an approach typically requires demonstration data from a specific target robot, which limits scalability due to high hardware costs and the intensive workload involved in operating the robot during demonstrations. Additionally, individualized training for each robot restricts cross-embodiment applications, as data cannot be transferred to different robot systems, even for similar tasks.

To enable scalable demonstration collection, pioneering works have introduced data collection tools that allow humans to directly manipulate during demonstrations [11–15]. These approaches enable training visuomotor policies that are deployable to specified target robots, reducing human workload and avoiding the costs and risks associated with using real

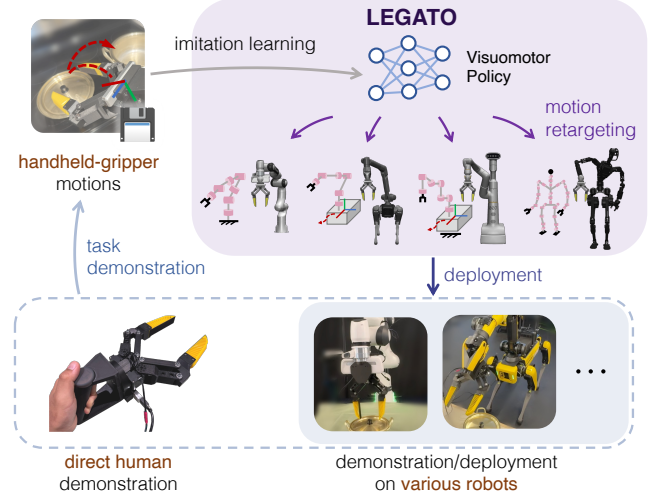


Fig. 1: Overview of LEGATO. LEGATO addresses the challenge of transferring visuomotor skills across diverse robot embodiments. We present a cross-embodiment imitation learning framework using a versatile handheld grasping tool that ensures consistent physical interactions across different embodiments. Visuomotor policies trained on demonstrations by humans or teleoperated robots using the tool can be deployed across various robots equipped with the same gripper. Motion retargeting enables the execution of trajectories on different robots without requiring robot-specific training data.

robots for data collection. However, these methods require either designing specialized data collection tools for specific robot gripper mechanisms or replacing robots’ original grippers with customized hardware. This limitation restricts the applicability of these tools for robots with diverse gripper mechanisms. For example, Universal Manipulation Interface [15] is designed specifically for the Schunk WSG-50 gripper with its parallel-jaw mechanism but is incompatible with grippers employing other mechanisms, such as hinge types. Additionally, variations in control latency and trajectory-tracking errors across robots, absent in human demonstrations, complicate policy transfer between embodiments. Chi et al. [15] addressed this by compensating for control and observation latencies, while Song et al. [11] used fine-tuning through trial and error on the target robot system. However, these strategies are difficult to generalize across diverse robots, as control latency varies between platforms due to differences in hardware and controllers.

Our key idea to address hardware differences across robots is to integrate a handheld grasping tool that can be shared across various robot embodiments for performing the same tasks (see Figure 1). We name our method LEGATO (Learning with a Handheld Grasping Tool). In our method, the LEGATO Gripper, a custom-designed handheld gripper, acts as a versatile and adaptive tool, representing tasks through its trajectories and grasping actions to ensure consistency across embodiments during both demonstration and deploy-

Manuscript received: August 22, 2024; Revised: November 21, 2024; Accepted: January 7, 2025. This paper was recommended for publication by Editor Jens Kober upon evaluation of the Associate Editor and Reviewers’ comments. This work was partially supported by the AI Institute and the Office of Naval Research (N00014-22-1-2204).

¹Mingyo Seo, Yuke Zhu, and Luis Sentis are with the University of Texas at Austin. ²Mingyo Seo, H. Andy Park, Shenli Yuan, and Luis Sentis are with the AI Institute. [†]Yuke Zhu and Luis Sentis contributed equally as advisers.

Correspondence: mingyo@utexas.edu

Digital Object Identifier (DOI): see top of this page.

ment. This handheld gripper—compatible with various robot grippers—enables a single visuomotor policy to be applied across diverse robot systems without requiring modifications to the original systems. Sharing the gripper across robots and directly manipulating human demonstrators ensures consistent, actively actuated grasping actions. In contrast, other data collection methods often create discrepancies in grasping actions because demonstrations rely on passive actuation by demonstrators, while robots require active actuation during deployment.

Our framework incorporates low-level motion retargeting through inverse kinematic (IK) optimization, tailored to each robot, along with a high-level transferable visuomotor policy (see Figure 2). Commands from the visuomotor policy for the handheld gripper are converted into whole-body motions using IK, enabling adaptation across robotic systems with only kinematic information, thereby avoiding extensive robot-specific training. While the IK optimization adapts gripper motions to each robot, variations in hardware and kinematics introduce differences in control latency and errors. To address this, we incorporate regularization on the gripper’s trajectories in a motion-invariant space during training, preventing bias toward the demonstration embodiment and effectively learning motor skills. This approach ensures that gripper trajectories from the visuomotor policy can be consistently translated into whole-body robot motions, regardless of control latency and IK response differences. Together, these components enable the learned visuomotor policy in LEGATO to be effectively transferred across diverse robotic systems.

We validate our approach through simulation and real-robot experiments, demonstrating its cross-embodiment transferability. In simulation, visuomotor policies trained on human demonstrations are successfully deployed across various embodiments, including a tabletop manipulator, a wheeled robot, a quadruped, and a humanoid. We further demonstrate the reusability of demonstration data by transferring a policy trained on one robot to another. In real-world tabletop manipulator setups, our method achieves a 72% success rate in complex manipulation tasks through policy transfer from direct human demonstrations.

II. RELATED WORK

A. Skill Learning Across Embodiments

Cross-embodiment policy transfer enables collecting demonstration data from easier or less costly embodiments and reusing it across different robot embodiments. One approach involves learning from videos of direct human demonstrations, which has been extensively explored in earlier works [16–21]. However, robot deployment is limited by its reliance on third-person visual observations and the absence of real-world physical interaction data.

Another approach integrates tools specifically designed for demonstration collection. One method involves leader-follower systems, which have been shown to be successful [22, 23]. However, this method requires designing leader hardware with a kinematic structure identical to a target robot, enabling its joint states to be directly mapped onto the robot. Alternatively, recent works use handheld tools to record on-hand

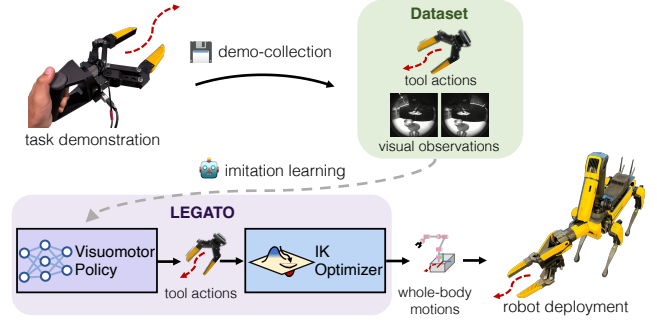


Fig. 2: LEGATO’s cross-embodiment learning pipeline. During data collection, the LEGATO Gripper records its trajectories, grasping actions, and visual observations captured by its egocentric stereo camera. A visuomotor policy is then trained on these demonstrations through imitation learning. During deployment, the visuomotor policy’s outputs are retargeted to the robots’ whole-body motions through IK optimization.

visual observations and corresponding motions from demonstrations [11–13, 15, 24]. While these approaches simplify tool design, they still require a system-specific tool or modifications to a robot’s original hardware to align the end-effector’s physical interactions and the recorded visual observations.

Unlike prior methods, we aim to generalize cross-embodiment learning by incorporating an adaptable handheld gripper and flexible kinematics-based motion retargeting. As a result, our learned policies are deployable across various types of robots without requiring robot-specific hardware for demonstrations.

B. Whole-body Motion Retargeting

Motion retargeting enables the generation of practical motions across different embodiments. A substantial body of research exists on motion retargeting between similar morphologies, such as from humans to humanoid robots, employing either model-based methods [25–28] or data-driven approaches [29–31]. In contrast, translating human movements into equivalent motions for varying target morphologies presents a challenge due to the inherently ambiguous nature of the task. Recent studies have investigated motion retargeting across embodiments with diverse morphologies [32–34]. While these studies have successfully demonstrated motion retargeting from humanoid to non-humanoid morphologies, they often rely on either embodiment-specific or task-specific models.

Our method shares similarities with previous works [27, 28] in utilizing kinematic optimization for motion retargeting. Unlike these works, however, we use a lower-dimensional action space based on the motions of the handheld gripper. This approach enables flexible motion retargeting across diverse robot embodiments with varying morphologies.

C. Trajectory Representations

To effectively transfer motions across diverse embodiments, motion primitives offer a practical solution. These primitives facilitate the assembly of elemental motions, as extensively explored in the literature [35–37]. However, their adaptability is inherently limited by design, making them less effective for novel tasks or environments. Recent advancements, such

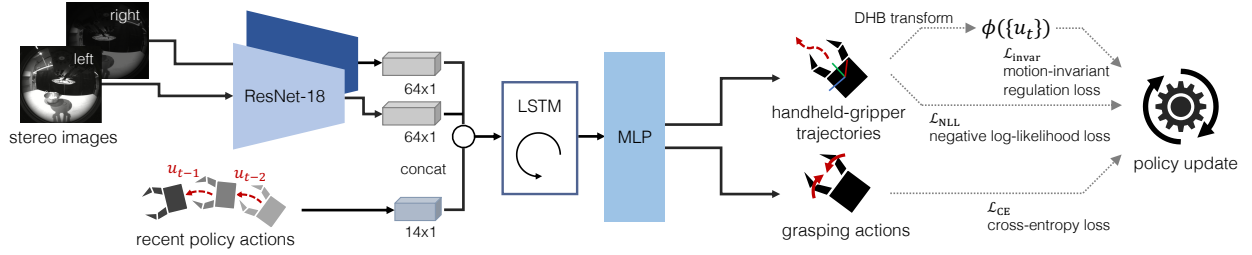


Fig. 3: High-level visuomotor policy architecture. The trained policies generate desired handheld-gripper trajectories and grasping actions u_t at 10 Hz from ego-centric stereo camera observations and previous policy actions. These action and observation spaces, defined in the handheld-gripper frame, remain consistent across various robot platforms. To learn actions on handheld-gripper trajectories, we apply two action losses: the negative log-likelihood loss $\mathcal{L}_{NLL}$ for the distribution in $SE(3)$ and the L2 loss $\mathcal{L}_{invar}$ in the DHB motion-invariant space. The grasping actions are trained using the cross-entropy loss $\mathcal{L}_{CE}$.

as encoder-decoder frameworks that project trajectories into a latent space, have significantly improved the learning of motion primitives [38–40]. Encoding motions into learned latent spaces facilitates the transfer of human motions to simulated humanoid models [41] and adapts them to robot embodiments with varying kinematic structures [34]. Nevertheless, these approaches often encounter challenges in achieving broad generalization across different scenarios.

To address these challenges, our method employs motion representations as training regularization elements rather than directly using them for motion generation. Specifically, we utilize the Denavit-Hartenberg Bidirectional (DHB) invariant representation [42], which offers invariance to rotational-translational shifts and scaling. The regularization in this motion-invariant space ensures robust alignment with demonstrated trajectories, enhancing policy generalizability across various embodiments.

III. METHOD

Here, we introduce LEGATO, a cross-embodiment imitation learning framework comprising a visuomotor policy at the high level and motion retargeting at the low level, as illustrated in Figure 2. The visuomotor policy is trained through imitation learning on task demonstrations, collected either from humans directly using the tool or from teleoperated robots holding the tool. The consistency of the action space of handheld gripper motions and ego-centric visual observations across different robots enables deployment to various embodiments. The low-level motion retargeting realizes these gripper trajectories as whole-body motions across different robot platforms.

A. Problem Formulation

We model the problem of cross-embodiment manipulation as a discrete-time Markov Decision Process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma, \rho_0)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}(\cdot|s, a)$ is the transition probability, $R(s, a, s')$ is the reward function, $\gamma \in [0, 1]$ is the discount factor, and $\rho_0(\cdot)$ is the initial state distribution. Our objective is to learn a closed-loop visuomotor policy $\pi(a_t|s_t)$ that maximizes the expected return $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})]$ across different robot systems. In our problem, $\mathcal{S}$ is the space of visual observations captured by the handheld gripper’s egocentric cameras, the robot’s proprioceptive feedback, and previous actions, while $\mathcal{A}$ is the space of the robot’s joint-space commands. These two spaces vary across different robot systems, reflecting the

diversity in sensory capabilities and actuation mechanisms inherent to each robot platform. $R(s, a, s')$ is the reward function designed for the manipulation task, and π is a closed-loop policy deployed on the robot.

To handle the complexity of visuomotor skills and ensure that the policy π is deployable across various robot systems, we decompose the policy into a two-level hierarchy. At the high level is a cross-embodiment visuomotor policy π_H , implemented as neural networks that compute target motion trajectories of the handheld gripper u . We train π_H through imitation learning with demonstrations collected by manipulating the handheld gripper. At the low level, we use a whole-body motion optimizer π_L , which determines target joint configurations to follow the trajectories u established by π_H through IK. This eliminates the need for additional training on target robot systems. With this hierarchical structure, the entire policy can be represented as $\pi(a_t|s_t) = \pi_L(a_t|s_t, u_t)\pi_H(u_t|s_t)$.

B. Actions Based on the Handheld Gripper

We incorporate the LEGATO Gripper, which can be shared across different robot systems. Inspired by Noguchi et al. [43], each robot uses its own gripper to hold the handheld gripper, integrating it as part of the embodiment. The handheld gripper maintains a consistent shape and viewpoint, unifying visual observation and action spaces across embodiments. This reduces the complexity of the cross-embodiment problem, streamlining the mapping of handheld-gripper motions and coordinating whole-body motions to execute them.

We define the action space as the differential pose in $SE(3)$ between consecutive time steps. This action space is suitable for generalizable whole-body manipulation on floating-base robots, as it eliminates reliance on a fixed reference frame. Actions are thus represented in the current handheld-gripper frame. Differential pose actions, sampled from a Gaussian Mixture Model [44], capture the multimodal nature of human demonstrations and essential motion information within the motion-invariant space, as described in Section III-D.

C. Whole-body Motion Retargeting

In this section, we detail our approach for mapping the trajectory actions of the handheld gripper u_t into whole-body robot motions a_t . We employ an optimization-based IK method to handle the constrained IK problem. This enables our motion optimizer π_L to map commands effectively to robot motions by leveraging degree-of-freedom (DOF) redundancy

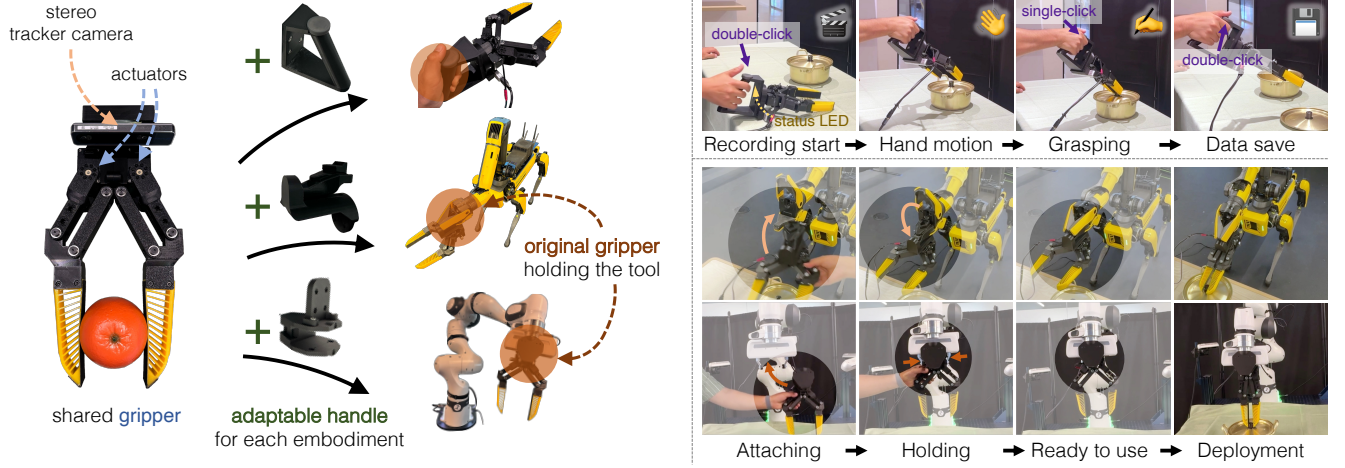


Fig. 4: LEGATO Gripper design. The LEGATO Gripper is designed for both human demonstration collection and robot deployment. **(Left)** It features a shared actuated gripper with adaptable handles, ensuring reliable human handling and consistent grasping across robots while minimizing components. **(Right top)** A human demonstrator can directly perform tasks by carrying the LEGATO Gripper in hand. The design includes a simple yet intuitive button interface with a status LED, allowing data recording to start and end with a double-click and grasping actions to trigger with a single click. **(Right bottom)** The LEGATO Gripper is easily installed on various robots, securely held by their original grippers, and is ready for immediate use.

while respecting actuation bounds and other constraints. Our approach addresses kinematic differences, constraints, and the diversity in DOFs across robot embodiments by using kinematic redundancy to satisfy joint and other constraints during deployment, without requiring additional robot-specific demonstrations.

The motion retargeting formulation is expressed as the following quadratic programming (QP) problem:

$$\begin{aligned} \min_{a_t} \quad & \sum \frac{1}{2} a_t^\top H a_t \\ \text{subject to} \quad & \\ & J_i(q_t) a_t = \dot{x}_i^{\text{des}}, \\ & \underline{L}_i \leq C_i a_t \leq \bar{L}_i, \end{aligned} \quad (1)$$

where $\underline{L}_i$, $\bar{L}_i$ and C_i define the velocity-level constraints for joint positions q , velocities $\dot{q}$, and accelerations $\ddot{q}$, along with other Cartesian limits (e.g., virtual walls and collision distance bounds). H is a positive-definite weighting matrix. $\dot{x}_i^{\text{des}}$ represents the desired velocities at the prioritized tasks, formulated in the following order:

$$\begin{aligned} J_1 &= J_{\text{grip}}, \quad \dot{x}_1^{\text{des}} = K_{\text{grip}}(x_{\text{des}} - x(q)), \\ J_2 &= I_{n_j}, \quad \dot{x}_2^{\text{des}} = K_{\text{bias}}(q_{\text{bias}} - q), \end{aligned} \quad (2)$$

where J_{grip} and I_{n_j} represent the Jacobians for gripper pose control and maintaining the configuration bias q_{bias} of n_j DOFs, respectively, with the control gains K_{grip} and K_{bias} . x_{des} represents the target pose of the handheld gripper, as determined by the visuomotor policy outputs u_t .

To solve the QP problem with the hierarchical priorities and the inequality constraints, we utilize the extended Saturation in the Null Space (eSNS) algorithm [45]. The constrained optimization problem of Equation 1 is formulated as follows:

$$\begin{aligned} \text{maximize} \quad & \sum_i c_i \\ \text{subject to} \quad & \\ & J_i(q_t) a_t = c_i \dot{x}_i^{\text{des}}, \\ & \underline{L}_i \leq C_i a_t \leq \bar{L}_i. \end{aligned} \quad (3)$$

Here, c_i represents scaling factors within the range $[0,1]$ that allow for scaling speed of the prioritized tasks defined in Equation 2, while accurately tracking their trajectories. This optimization problem, maximizing the sum of scaling factors, yields the closest mapping motions while meeting all constraints. Our framework ensures that robot movements not only follow the prescribed hierarchy but also adhere to the constraints, guaranteeing flexible task execution and robust motion generalization.

D. Training of Visuomotor Policies

Task demonstrations can be performed on any embodiment capable of executing the tasks within its joint space. The collected demonstration dataset $\mathcal{D}$ consists of state-action pairs $\mathcal{D} = \{(s_i, u_i)\}_{i=1}^N$. Here, N represents the total number of data points. The observations s_i comprise stereo images from the handheld gripper's onboard camera and a history buffer of previous actions $u_{k:k=t-T:t-1}$. The visual observations are provided as stereo grayscale images with a resolution of 128×128 pixels. The demonstration commands u_i include the subsequent setpoints for the handheld gripper (6D) and the grasping actions (1D).

We train our policy π_H using a deep imitation learning algorithm, specifically using a behavioral cloning policy with LSTM networks [46, 47], as shown in Figure 3. The visuomotor policy employs two separate ResNet18-based image encoders [48], trained end-to-end. The encoded features are flattened and processed by two-layer LSTM networks, each with 400 hidden units. The policy outputs are generated by a three-layer Multi-Layer Perceptron (MLP), with each layer containing 2048 hidden units. For hand trajectories, the policy outputs Gaussian Mixture Model (GMM) parameters to determine the next target pose based on spatial and rotational differences from the previous frame, using a 5-mode GMM. Grasping actions are provided as binary classifications for opening and closing the gripper.

For imitation learning, we employ behavioral cloning with the following training loss:

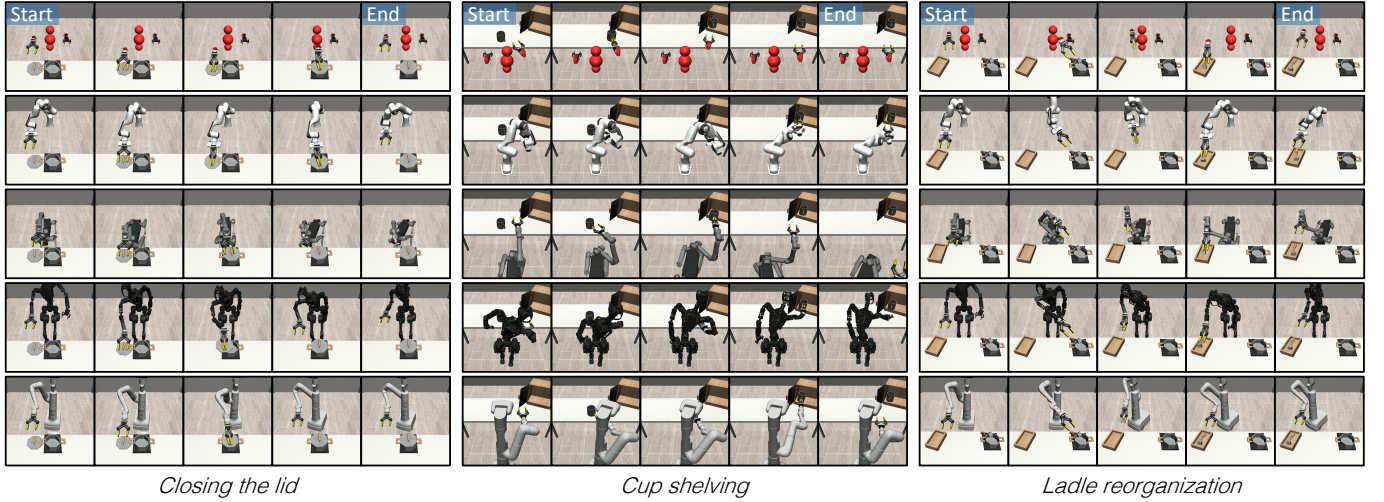


Fig. 5: Timelapse of deploying LEGATO in simulation. We trained visuomotor policies using demonstrations from the *Abstract* embodiment and deployed them on robots with diverse morphologies, from the top: *Abstract*, *Panda*, *Spot*, *GR-1*, and *Google Robot*. The timelapse of deploying these policies reveals consistent time steps. The tracking performance of the IK motion retargeting varies with morphology, leading to domain gaps across embodiments. Despite these challenges, LEGATO achieves successful deployment on various robots.

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \mathcal{L}_{\text{invar}} + \mathcal{L}_{\text{CE}}. \quad (4)$$

Here, $\mathcal{L}_{\text{NLL}}$ and $\mathcal{L}_{\text{invar}}$, associated with the trajectories of the handheld gripper, represent the negative log-likelihood loss and the regularization term in the motion-invariant space, respectively. Additionally, $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss for binary grasping actions. The term $\mathcal{L}_{\text{invar}}$ is calculated as the L2 loss in the motion-invariant space, specifically using the DHB invariant transformation [42], $\phi(\cdot)$ (see Appendix for details). This transformation is applied to the differential poses of the gripper in SE(3) from the policy outputs u_t and the demonstrations $\hat{u}_t$:

$$\mathcal{L}_{\text{invar}} = \sum \|\phi(u_t, \{\hat{u}_k\}_{k=t-T}^{t-1}) - \phi(\{\hat{u}_k\}_{k=t-T}^t)\|^2. \quad (5)$$

Incorporating $\mathcal{L}_{\text{invar}}$ into the training loss influences the distribution of handheld-gripper motions, preventing the policy from being biased by embodiment-specific properties like tracking errors and control latency. Unlike SE(3), these invariants represent motion by breaking it down into magnitude and directional changes, unaffected by embodiment-specific factors such as viewpoints, reference frames, pose offsets, or scales. Leveraging motion invariance ensures the policy captures essential motion information from demonstrations without bias toward specific embodiments. This is critical for cross-embodiment learning, as it makes the policy robust to domain mismatches across different robot embodiments.

IV. EXPERIMENTS

In this section, we demonstrate the feasibility and effectiveness of LEGATO for cross-embodiment transfer of visuomotor policies, both in MuJoCo simulation [49] and real-world settings. Leveraging the scalability and ease of simulation environments, we employ them to investigate the following research questions: 1) How does the regularization in the motion-invariant space impact the training of cross-embodiment policies? 2) How do differences in morphology

and controllability affect the task capacity of robot embodiments?

A. LEGATO Gripper

The LEGATO Gripper design described in Section III-B is implemented on real hardware. It is designed for intuitive use in both direct human demonstrations and robot operations, as illustrated in Figure 4. During human demonstrations, a user carries the LEGATO Gripper and controls the grasping actions. Robots, on the other hand, can hold the tool with their original grippers without needing any hardware modifications.

To enable easy attachment and replacement, the design is modularized, particularly the handle parts that robots grip, as presented in Figure 4. Thus, only the handle parts need replacement for different robots, allowing all other core components to be shared across robots. These handle parts can be easily designed using CAD models of target robots provided by manufacturers, ensuring adaptability to diverse robot systems. The LEGATO Gripper features two pairs of parallel four-bar linkage mechanisms, each actuated by one DOF, enabling a wider and more flexible range of opening distances. The tool employs fingertips of compliant fin ray mechanisms for high compliance and adaptability during contact. Made from 3D-printed parts, with TPU 95A for the fingertips and PLA for other components, the LEGATO Gripper is significantly lighter than typical commercial grippers. These key features facilitate the concept of integrating the tool as part of the embodiment, even under robots' limited payload capacities, while supporting generalizable grasping actions.

The LEGATO Gripper is equipped with a Realsense T265 [50] camera (or an alternative stereo tracking camera such as the SeerSense XR50 [51]), with fisheye stereo cameras and an IMU for streaming visual observations and estimating the handheld gripper's motions via visual odometry. During demonstrations, both stereo images and visual odometry data are recorded to form observation-action pairs. In contrast, during robot deployment, only stereo images are streamed.

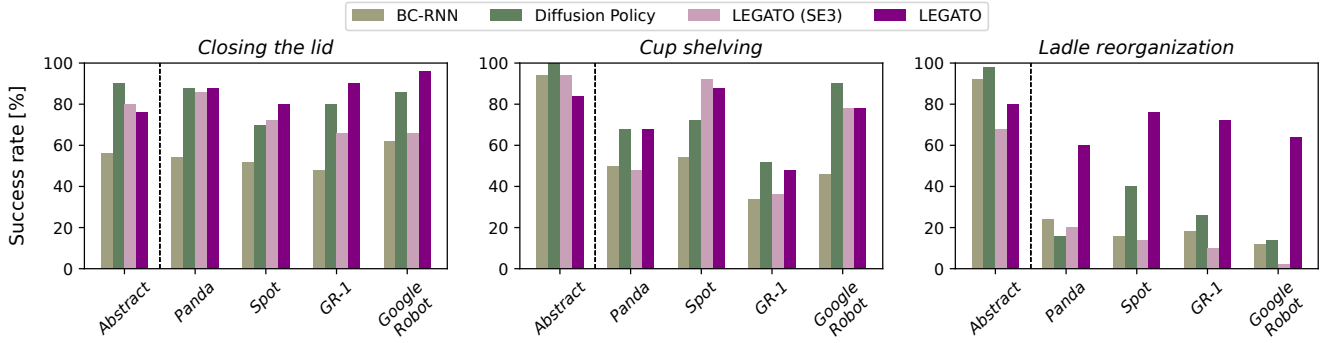


Fig. 6: Quantitative results in simulation. We report success rates on 50 trials of our LEGATO policies compared to baselines. On average, LEGATO outperforms all other methods in cross-embodiment deployment by 28.9%, 10.5%, and 21.1%, compared to BC-RNN, Diffusion Policy, and LEGATO (SE3), respectively. Notably, unlike the baselines that achieve high success rates only on specific robot embodiments, typically the *Abstract* embodiment used for training, LEGATO demonstrates consistent success rates across different embodiments.

In simulation, the mechanism and geometry of the real hardware are modeled. Visual observations are emulated by adapting the properties of the tracker camera used in the hardware design. The LEGATO Gripper is attached to each robot’s original gripper with an offset.

B. Experimental Setup

We designed the following three realistic manipulation tasks to study the cross-embodiment deployment of our method for both simulation and real robot systems.

- **Closing the lid:** A robot grasps the lid and places it on the pot within reach. This task requires precise manipulation to accurately close the lid.
- **Cup shelving:** A robot places the cup into the shelf. In addition to requiring a large workspace, this task involves complex collision-free motion to position the cup against the shelf’s non-convex shapes.
- **Ladle reorganization:** A robot picks up the ladle from the pot and places it into the utensil organizer. This task involves complex manipulation and handling objects potentially out of view due to limited visibility and occlusion.

A task is considered successful if a robot accomplishes the designated goals within a specific time limit. Across all baselines and tasks, the initial states of robots are consistent, and the initial poses of objects, other than the table, are uniformly randomized.

C. Quantitative Evaluation in Simulation

In these experiments, we demonstrate the effectiveness of our domain-transfer method across different embodiments. To systematically evaluate our method, we use embodiments representing various kinematic morphologies, as shown in Figure 1: Franka Emika *Panda* [52], Boston Dynamics *Spot* [2], Fourier *GR-1* [5], and *Google Robot* [53].

- **Abstract:** an idealized embodiment designed for simulation that can manipulate hands along continuous trajectories without speed or workspace limitations. Human motion commands are directly mapped to the simulation, replicating direct human demonstration in real-world settings.
- **Panda:** a 7-DOF tabletop manipulator, used to demonstrate the impact of redundant DOFs on motion retargeting compared to the robots listed below.

- **Spot:** a quadrupedal robot with 6 DOFs in its arm and 6 DOFs in its body pose. To demonstrate achieving an extensive workspace through whole-body motion alone, the robot’s locomotion is not considered; instead, the leg joints track the robot’s body within a limited SE(3) range.
- **GR-1:** a humanoid robot with 7 DOFs per arm and 3 DOFs each for the head and torso. Similar to *Spot*, the robot’s locomotion is not considered, and the leg joints are fixed. This demonstrates the application of our method to highly redundant DOF systems.
- **Google Robot:** a wheeled mobile robot with 7 DOFs in its arm and 3 DOFs in planar base motion, used to show how handheld-gripper trajectories map to mobile manipulation.

We focus on demonstrations from the *Abstract* embodiment, where commands from human demonstrators are directly mapped, enabling task performance without being affected by tracking errors that occur in robots. The trained policies are then transferred directly to the robots without any adaptation, using the pre-defined IK optimizer, as shown in Figure 5.

Figure 6 reports quantitative evaluations of our simulated tasks, comparing our model with the following baselines.

- **BC-RNN:** a baseline that uses recurrent neural networks to learn manipulation skills from teleoperated demonstrations, as introduced by Mandlekar et al. [47]. It employs flat GMM outputs and is trained with the negative log-likelihood loss.
- **Diffusion Policy:** a baseline that employs a receding-horizon visual-motor policy based on a diffusion model, as introduced by Chi et al. [54]. In particular, we adapt the Velocity Diffusion Policy. Due to the requirement for precise tracking and state estimation of the end-effector pose with respect to the global frame, the Position Diffusion Policy is not applicable in our settings.
- **LEGATO (SE3):** a variant of our final model that excludes the motion-invariant regularization loss $\mathcal{L}_{\text{invar}}$ from the training loss, with all other components unchanged. This variant is designed to evaluate the impact of the regularization in the motion-invariant space on cross-embodiment deployment.

All baselines are trained on the same dataset of 150 task demonstrations using the *Abstract* embodiment with identical policy inputs, as shown in Figure 3. They output handheld gripper trajectories realized by the same IK optimizer, using consistent parameters within a single robot.

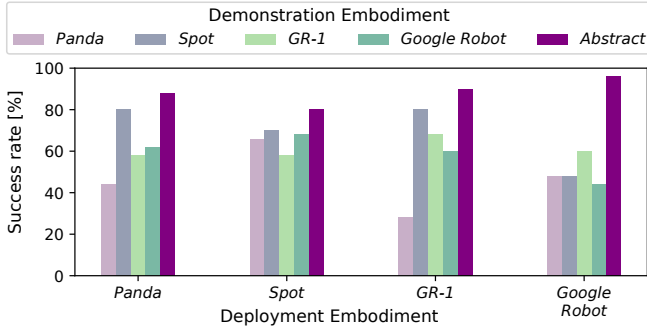


Fig. 7: **Simulation evaluation across different demonstration embodiments.** We show success rate changes on 50 trials from policies trained on demonstrations of different embodiments.

As shown in Figure 6, LEGATO outperforms the three baseline methods across various robot embodiments. The success rates for the *Abstract* embodiment indicate that all baselines are well-trained within that domain. However, the higher success rates of LEGATO (SE3) and Diffusion Policy, which uses the state-of-the-art diffusion model and performs best in the *Abstract* domain, are followed by a notable decline when applied to other robots. This suggests that exclusive training in SE(3) may introduce domain bias toward the training embodiment. In contrast, the performance improvement of LEGATO across diverse embodiments suggests that its regularization in the motion-invariant space enhances policy robustness across different robot domains, even with a simple model architecture. In addition, LEGATO and the baselines show reduced performance overall in the *Cup shelving* task with the *Panda* and *GR-1* robots, as well as in the *Ladle reorganization* task with the *Panda* robot. This underscores the limitations of fixed-base robots for tasks requiring extensive motion, highlighting the need for whole-body manipulation. Supplementary evaluations for further discussion are provided in Appendix.

D. Varying Demonstration Embodiments

We investigate how different embodiments used in demonstrations affect policy performance in simulation. We collected 150 trajectories by teleoperating each robot in Section IV-C for the *Closing the lid* task and trained policies on these demonstrations. Figure 7 shows the success rates for deploying these policies across the robots. The policies trained on the *Abstract* embodiment performed best, indicating that joint controller latency and the IK motion retargeting affect demonstration quality. Among the policies trained on robot teleoperation, demonstrations on the *Spot* robot yield the highest success rates, whereas those on the *Panda* robot yield the lowest. This suggests that whole-body motion capability and redundancy influence task demonstration quality, as the *Spot* robot has full 6 DOFs in its body, while the *Panda* robot has less redundancy. Notably, policies trained on the *Panda* and *Google Robot* exhibit lower success rates when deployed on the robot used for demonstrations compared to other robots. This highlights that the trajectory tracking capabilities constrained by the deployment embodiments affect task complexity.

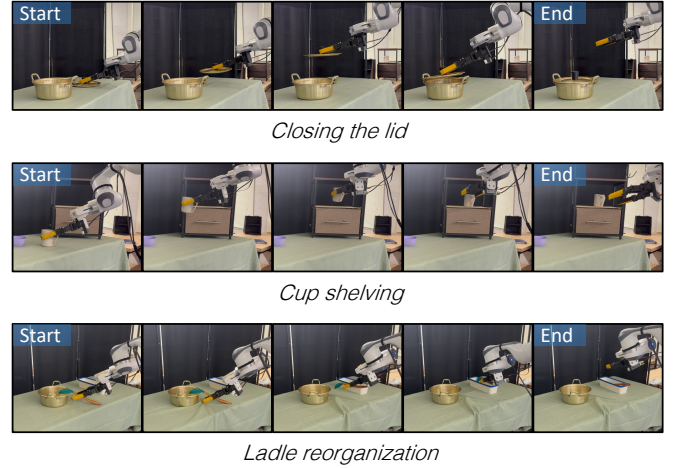


Fig. 8: **Timelapse of deploying LEGATO in the real-robot system.** We trained visuomotor policies on direct human demonstrations and successfully deployed them on the *Panda* robot system. The timelapse of deploying these policies demonstrates consistent time intervals.

E. Real-Robot Experiments

We aim to validate the robustness of our method against variations caused by controller limitations and sensor inaccuracies in real-world settings. We collected direct human demonstrations through the LEGATO Gripper to train policies and deployed them, particularly on the *Panda* robot, the most challenging due to its limited workspace at the IK motion retargeting, as shown in Section IV-C. During the evaluation of each task, the same visuomotor policy was attempted for 20 trials on the *Panda* robot. Our method succeeded in 16 trials of the *Closing the lid* task, 13 trials of the *Cup shelving* task, and 14 trials of the *Ladle reorganization* task, respectively (see Figure 8). More videos are provided on our project website.

V. CONCLUSION

We present LEGATO, a cross-embodiment learning framework for transferring visuomotor skills across diverse robot morphologies. By using a handheld gripper for consistent observations and actions, our framework enables visuomotor policies to transfer across embodiments without hardware modifications. Handheld-gripper trajectories from the visuomotor policies are mapped to whole-body robot motions through IK optimization. Although trained on specific embodiments, regularization in a motion-invariant space allows these policies to adapt easily to different robots, managing variations in control latency and tracking errors. Our current focus is on whole-body manipulation, employing redundant DOFs for flexibility and an extended workspace through coordinated body movements, though our method is limited to non-walking scenarios. Future work will incorporate loco-manipulation, integrating manipulation with walking to enable legged robots to perform diverse tasks with larger workspaces. Additionally, while our method currently relies on designing handles specific to target robots, this can be addressed by identifying grasping location to enable the use of universal handles across robots. We also believe our approach is adaptable to a wide range of tools and applications beyond the LEGATO Gripper design.

Acknowledgements This work was conducted during Mingyo Seo’s internship at the AI Institute. We thank Rutav Shah and Minkyung Kim for providing feedback on this manuscript. We thank Osman Dogan Yirmibesoglu for designing the fin ray style compliant fingers and helping with hardware prototyping. We thank Mitchell Pryor and Fabian Parra for their support with the real Spot demonstration.

REFERENCES

- [1] C. C. Kemp, A. Edsinger, H. M. Clever, and B. Matulevich, “The design of stretch: A compact, lightweight mobile manipulator for indoor human environments,” in *IEEE International Conference on Robotics and Automation*, 2022.
- [2] *Boston Dynamics Spot*, <https://bostondynamics.com/products/spot>.
- [3] *Unitree B1*, <https://shop.unitree.com/products/unitree-b1>.
- [4] *Agility Robotics Digit*, <https://agilityrobotics.com/robots>.
- [5] *GRI - Fourier Intelligence*, <https://fourierintelligence.com/gr1>.
- [6] F. Xie, A. Chowdhury, M. De Paolis Kaluza, L. Zhao, L. Wong, and R. Yu, “Deep imitation learning for bimanual robotic manipulation,” in *Advances in Neural Information Processing Systems*, 2020.
- [7] M. Seo *et al.*, “Deep imitation learning for humanoid locomotion through human teleoperation,” in *IEEE-RAS International Conference on Humanoid Robots*, 2023.
- [8] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, “Learning fine-grained bimanual manipulation with low-cost hardware,” in *Robotics: Science and Systems*, 2023.
- [9] A. Mandlekar *et al.*, “Mimicgen: A data generation system for scalable robot learning using human demonstrations,” in *Conference on Robot Learning*, 2023.
- [10] K. Sridhar, S. Dutta, D. Jayaraman, J. Weimer, and I. Lee, “Memory-consistent neural networks for imitation learning,” in *International Conference on Learning Representations*, 2024.
- [11] 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*, 2020.
- [12] S. Young, D. Gandhi, S. Tulsiani, A. Gupta, P. Abbeel, and L. Pinto, “Visual imitation made easy,” in *Conference on Robot Learning*, 2021.
- [13] J. Pari, N. M. Shafullah, S. P. Arunachalam, and L. Pinto, “The surprising effectiveness of representation learning for visual imitation,” in *Robotics: Science and Systems*, 2022.
- [14] M. Seo, R. Gupta, Y. Zhu, A. Skoutnev, L. Sentis, and Y. Zhu, “Learning to walk by steering: Perceptive quadrupedal locomotion in dynamic environments,” in *IEEE International Conference on Robotics and Automation*, 2023.
- [15] C. Chi *et al.*, “Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots,” in *Robotics: Science and Systems*, 2024.
- [16] L. Smith, N. Dhawan, M. Zhang, P. Abbeel, and S. Levine, “Avid: Learning multi-stage tasks via pixel-level translation of human videos,” in *Robotics: Science and Systems*, 2020.
- [17] K. Zakka, A. Zeng, P. Florence, J. Tompson, J. Bohg, and D. Dwibedi, “Xirl: Cross-embodiment inverse reinforcement learning,” in *Conference on Robot Learning*, 2022.
- [18] K. Shaw, S. Bahl, and D. Pathak, “Videodex: Learning dexterity from internet videos,” in *Conference on Robot Learning*, 2023.
- [19] C. Wang *et al.*, “Mimicplay: Long-horizon imitation learning by watching human play,” in *Conference on Robot Learning*, 2023.
- [20] M. Xu, Z. Xu, C. Chi, M. Veloso, and S. Song, “Xskill: Cross embodiment skill discovery,” in *Conference on Robot Learning*, 2023.
- [21] J. Li *et al.*, “Okami: Teaching humanoid robots manipulation skills through single video imitation,” in *Conference on Robot Learning*, 2024.
- [22] H. Fang *et al.*, “Low-cost exoskeletons for learning whole-arm manipulation in the wild,” in *IEEE International Conference on Robotics and Automation*, 2024.
- [23] H. Kim, Y. Ohmura, A. Nagakubo, and Y. Kuniyoshi, “Training robots without robots: Deep imitation learning for master-to-robot policy transfer,” *IEEE Robotics and Automation Letters*, 2023.
- [24] 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 *Conference on Robot Learning*, 2024.
- [25] S. Tachi *et al.*, “Telexistence cockpit for humanoid robot control,” *Advanced Robotics*, 2003.
- [26] K. Ayusawa and E. Yoshida, “Motion retargeting for humanoid robots based on simultaneous morphing parameter identification and motion optimization,” *IEEE Transactions on Robotics*, 2017.
- [27] L. Penco *et al.*, “Robust real-time whole-body motion retargeting from human to humanoid,” in *IEEE-RAS International Conference on Humanoid Robots*, 2018.
- [28] K. Darvish *et al.*, “Whole-body geometric retargeting for humanoid robots,” in *IEEE-RAS International Conference on Humanoid Robots*, 2019.
- [29] X. B. Peng, E. Coumans, T. Zhang, T.-W. E. Lee, J. Tan, and S. Levine, “Learning agile robotic locomotion skills by imitating animals,” in *Robotics: Science and Systems*, 2020.
- [30] T. He *et al.*, “Learning human-to-humanoid real-time whole-body teleoperation,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024.
- [31] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, “Humanplus: Humanoid shadowing and imitation from humans,” in *Conference on Robot Learning*, 2024.
- [32] S. Kim, M. Sorokin, J. Lee, and S. Ha, “Human motion control of quadrupedal robots using deep reinforcement learning,” in *Robotics: Science and Systems*, 2022.
- [33] T. Li, J. Won, A. Clegg, J. Kim, A. Rai, and S. Ha, “Ace: Adversarial correspondence embedding for cross morphology motion retargeting from human to nonhuman characters,” in *SIGGRAPH Asia*, 2023.
- [34] Y. Yan, E. V. Mascaro, and D. Lee, “ImitationNet: Unsupervised human-to-robot motion retargeting via shared latent space,” in *IEEE-RAS International Conference on Humanoid Robots*, 2023.
- [35] M. Saveriano, F. J. Abu-Dakka, A. Kramberger, and L. Petermel, “Dynamic movement primitives in robotics: A tutorial survey,” *The International Journal of Robotics Research*, 2023.
- [36] A. Paraschos, C. Daniel, J. R. Peters, and G. Neumann, “Probabilistic movement primitives,” in *Advances in Neural Information Processing Systems*, 2013.
- [37] Y. Zhou, J. Gao, and T. Asfour, “Learning via-point movement primitives with inter-and extrapolation capabilities,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019.
- [38] M. Noseworthy, R. Paul, S. Roy, D. Park, and N. Roy, “Task-conditioned variational autoencoders for learning movement primitives,” in *Conference on robot learning*, 2020.
- [39] J. Uraim, M. Ginesi, D. Tateo, and J. Peters, “Imitationflow: Learning deep stable stochastic dynamic systems by normalizing flows,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020.
- [40] B. Lee, Y. Lee, S. Kim, M. Son, and F. C. Park, “Equivariant motion manifold primitives,” in *Conference on Robot Learning*, 2023.
- [41] Z. Luo *et al.*, “Universal humanoid motion representations for physics-based control,” in *International Conference on Learning Representations*, 2024.
- [42] D. Lee, R. Soloperto, and M. Saveriano, “Bidirectional invariant representation of rigid body motions and its application to gesture recognition and reproduction,” *Autonomous Robots*, 2018.
- [43] Y. Noguchi, T. Matsushima, Y. Matsuo, and S. S. Gu, “Tool as embodiment for recursive manipulation,” *arXiv preprint arXiv:2112.00359*, 2021.
- [44] Z. Wang *et al.*, “Critic regularized regression,” in *Advances in Neural Information Processing Systems*, 2020.
- [45] M. D. Fiore, G. Meli, A. Ziese, B. Siciliano, and C. Natale, “A general framework for hierarchical redundancy resolution under arbitrary constraints,” *IEEE Transactions on Robotics*, 2023.
- [46] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, 1997.
- [47] A. Mandlekar *et al.*, “What matters in learning from offline human demonstrations for robot manipulation,” in *Conference on Robot Learning*, 2022.
- [48] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [49] E. Todorov, T. Erez, and Y. Tassa, “Mujoco: A physics engine for model-based control,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.
- [50] *Intel RealSense SDK*, <https://github.com/IntelRealSense/librealsense>.
- [51] *SeerSense™ XR50 Module*, <https://www.xvisiotech.com/product/seersense-xr50/>.
- [52] *Franka Robotics*, <https://franka.de/>.
- [53] A. Herzog *et al.*, “Deep rl at scale: Sorting waste in office buildings with a fleet of mobile manipulators,” in *Robotics: Science and Systems*, 2023.
- [54] C. Chi *et al.*, “Diffusion policy: Visuomotor policy learning via action diffusion,” *The International Journal of Robotics Research*, 2024.
- [55] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu, “Viola: Imitation learning for vision-based manipulation with object proposal priors,” in *Conference on Robot Learning*, 2022.

APPENDIX

A. Motion-Invariant Regulation Loss

The motion-invariant transform $\phi(\cdot)$, used to compute $\mathcal{L}_{\text{invar}}$ in Equation 5, follows the DHB motion-invariant framework [42]. Given trajectories $\{u_k\}_{k=t-T}^t$ with $T \geq 2$, we compute the relative positions p_k and orientations r_k of the gripper with respect to the initial frame at $t - T$, where p_{t-T} and r_{t-T} are at the origin.

The differences $\Delta \mathbf{p}_k = \mathbf{p}_{k+1} - \mathbf{p}_k$ and $\Delta \mathbf{r}_k = \mathbf{r}_{k+1} - \mathbf{r}_k$ represent the linear and angular trajectory changes between $k + 1$ and k . The initial linear frames are defined as:

$$\begin{aligned}\hat{\mathbf{x}}_{p,k} &= \frac{\Delta \mathbf{p}_k}{\|\Delta \mathbf{p}_k\|}, \\ \hat{\mathbf{y}}_{p,k} &= \frac{\hat{\mathbf{x}}_{p,k} \times \hat{\mathbf{x}}_{p,k+1}}{\|\hat{\mathbf{x}}_{p,k} \times \hat{\mathbf{x}}_{p,k+1}\|}, \\ \hat{\mathbf{z}}_{p,k} &= \hat{\mathbf{x}}_{p,k} \times \hat{\mathbf{y}}_{p,k}.\end{aligned}$$

Similarly, the initial angular frames are:

$$\begin{aligned}\hat{\mathbf{x}}_{r,k} &= \frac{\Delta \mathbf{r}_k}{\|\Delta \mathbf{r}_k\|}, \\ \hat{\mathbf{y}}_{r,k} &= \frac{\hat{\mathbf{x}}_{r,k} \times \hat{\mathbf{x}}_{r,k+1}}{\|\hat{\mathbf{x}}_{r,k} \times \hat{\mathbf{x}}_{r,k+1}\|}, \\ \hat{\mathbf{z}}_{r,k} &= \hat{\mathbf{x}}_{r,k} \times \hat{\mathbf{y}}_{r,k}.\end{aligned}$$

The directions of the axes in both frames are chosen to prevent discontinuities across time steps.

In the DHB transformation, the motion of a rigid body is separated into position and orientation components using two frames. Two invariants are the norms of the relative positions and orientations between these frames:

$$\begin{aligned}m_{p,k} &= \|\Delta \mathbf{p}_k\|, \\ m_{r,k} &= \|\Delta \mathbf{r}_k\|.\end{aligned}$$

These invariants, m_p and m_r , describe the translation of the linear and angular frames. Four additional values describe their rotation:

$$\begin{aligned}\theta_{p,k}^1 &= \arctan \left(\frac{\hat{\mathbf{x}}_{p,k} \times \hat{\mathbf{x}}_{p,k+1}}{\hat{\mathbf{x}}_{p,k} \cdot \hat{\mathbf{x}}_{p,k+1}} \cdot \hat{\mathbf{y}}_{p,k} \right), \\ \theta_{p,k}^2 &= \arctan \left(\frac{\hat{\mathbf{y}}_{p,k} \times \hat{\mathbf{y}}_{p,k+1}}{\hat{\mathbf{y}}_{p,k} \cdot \hat{\mathbf{y}}_{p,k+1}} \cdot \hat{\mathbf{x}}_{p,k+1} \right), \\ \theta_{r,k}^1 &= \arctan \left(\frac{\hat{\mathbf{x}}_{r,k} \times \hat{\mathbf{x}}_{r,k+1}}{\hat{\mathbf{x}}_{r,k} \cdot \hat{\mathbf{x}}_{r,k+1}} \cdot \hat{\mathbf{y}}_{r,k} \right), \\ \theta_{r,k}^2 &= \arctan \left(\frac{\hat{\mathbf{y}}_{r,k} \times \hat{\mathbf{y}}_{r,k+1}}{\hat{\mathbf{y}}_{r,k} \cdot \hat{\mathbf{y}}_{r,k+1}} \cdot \hat{\mathbf{x}}_{r,k+1} \right).\end{aligned}$$

This process produces the linear and angular invariant values ($m_{p,k}, \theta_{p,k}^1, \theta_{p,k}^2$) and ($m_{r,k}, \theta_{r,k}^1, \theta_{r,k}^2$), as established in the original work.

To ensure continuity, the computed frame rotations are transformed using $\sin(\cdot)$ and $\sin(2\cdot)$. The final transformation applied in our regularization loss thus yields 10 variables of length $T - 1$:

$$\phi(\{u_k\}_{k=t-T}^t) = \left\{ \begin{bmatrix} m_{p,k} \\ \sin(\theta_{p,k}^1) \\ \sin(2\theta_{p,k}^1) \\ \sin(\theta_{p,k}^2) \\ \sin(2\theta_{p,k}^2) \\ m_{r,k} \\ \sin(\theta_{r,k}^1) \\ \sin(2\theta_{r,k}^1) \\ \sin(\theta_{r,k}^2) \\ \sin(2\theta_{r,k}^2) \end{bmatrix} \right\}_{k=t-T}^{t-2}.$$

When computing $\mathcal{L}_{\text{invar}}$, we transform two types of trajectories: 1) $\phi(\{\hat{u}_k\}_{k=t-T}^t)$, the transformed values from the demonstration trajectories, and 2) $\phi(u_t, \{\hat{u}_k\}_{k=t-T}^{t-1})$, the transformed values from the given previous trajectories $\{\hat{u}_k\}_{k=t-T}^{t-1}$ and the predicted target u_t at time t . By calculating the L2 loss between these two transformed values and using it as a training loss, the predicted trajectories u_t are aligned with the demonstration trajectories in the motion-invariant space, given $\{\hat{u}_k\}_{k=t-T}^{t-1}$.

B. Supplementary Evaluation in Simulation

We provide additional quantitative evaluations in simulation to further discuss cross-embodiment visuomotor policies. Specifically, we aim to address: 1) a comparison with existing cross-embodiment learning frameworks that utilize different action spaces, and 2) the applicability of the motion-invariant regularization to varied neural network architectures and its impact on their performance.

a) *Comparison with the Diffusion Policy Using Relative-Trajectory Actions:* To compare our method with existing works aimed at cross-embodiment learning of ego-centric visuomotor policies, we adapted the Diffusion Policy with the action space of relative end-effector trajectories, as used in Universal Manipulation Interface [15], to our setting, where the visuomotor policy outputs gripper trajectories based on previous actions and visual observations. We refer to this baseline as Diffusion Policy (Relative Trajectory). This baseline serves as a reference for Universal Manipulation Interface [15], but without the latency compensation process. As described in Section I, generalizing latency compensation across various robot embodiments is challenging because it requires fine-tuning for each target robot system. Therefore, we exclude the latency compensation process in our evaluation.

We used the same setup as the quantitative evaluation in Section IV-C, utilizing the same dataset of 150 task demonstrations with the *Abstract* embodiment. In Figure 9, we report the success rates for deploying Diffusion Policy (Relative Trajectory) with LEGATO and the Diffusion Policy baseline used in Section IV-C. For clarification, the Diffusion Policy baseline from Section IV-C is referred to as Diffusion Policy (Velocity). In our work, we considered only the action space of the handheld gripper's differential poses to minimize the impact of visual odometry errors during demonstration collection and to eliminate reliance on specified frames other than the handheld gripper's pose. This ensures suitability across

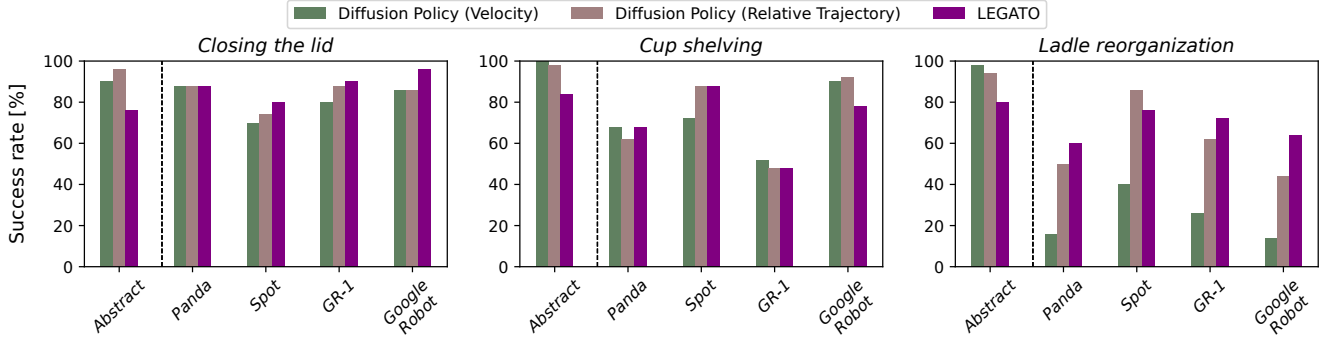


Fig. 9: Comparison with different types of Diffusion Policy baselines We report success rates from 50 trials of our LEGATO policies compared to Diffusion Policy baselines, using different action spaces: end-effector velocities [54] and relative end-effector trajectories [15]. Although Diffusion Policy (Relative Trajectory) outperformed Diffusion Policy (Velocity) in cross-embodiment settings, LEGATO policies achieved the highest success rates across most cross-embodiment settings, except for deployment on the *Google Robot* for the *Cup shelving* task and the *Spot* robot for the *Ladle reorganization* task.

various robot platforms, including floating-base robot systems. Unlike our setting, training with the action space of relative trajectories requires additional geometric information, such as the initial gripper pose for a sequence of receding-horizon actions, to generate the training dataset. This information is not incorporated into other baselines or the LEGATO framework. Nevertheless, LEGATO achieved higher performance than Diffusion Policy (Relative Trajectory) in cross-embodiment settings for the *Closing the lid* and *Ladle reorganization* tasks, though not in deployment within the same domain. As noted by Chi et al. [15], Diffusion Policy (Relative Trajectory) outperformed Diffusion Policy (Velocity) in their evaluation, and this was also observed in our setting.

b) Impact of Training with the Motion-Invariant Regularization: We provide an additional ablation study on the complementary use of the motion-invariant regularization loss to enhance cross-embodiment transferability in policies, as shown in Figure 10. Specifically, we applied the motion-invariant regularization loss $\mathcal{L}_{\text{invar}}$, as described in Section III-D and Appendix A, to BC-RNN [47], Diffusion Policy (Velocity) adapted from the Velocity Diffusion Policy introduced by Chi et al. [54], and Diffusion Policy (Relative Trajectory) adapted from the Diffusion Policy using relative end-effector trajectories as the action space in Universal Manipulation Interface [15]. Similar to the quantitative evaluation in Section IV-C, all baselines are trained on the same dataset of 150 task demonstrations using the *Abstract* embodiment. Additionally, the same IK optimizer with consistent parameters is used within a single robot.

As outlined in Equation 4 of Section III-D, the motion-invariant regularization loss are added to the original loss functions for each baseline. In BC-RNN, the motion-invariant regularization loss is added to the negative log-likelihood between the predicted and demonstration actions. For the Diffusion Policy baselines, the motion-invariant regularization loss is integrated with the L2-based DDPM loss during the denoising process. The motion-invariant regularization loss is adapted for the sequence of receding-horizon actions used in the Diffusion Policy baselines and is defined as:

$$\mathcal{L}_{\text{invar}} = \sum \|\phi(\{u_k\}_{k=t}^{t+P}, \{\hat{u}_k\}_{k=t-T}^{t-1}) - \phi(\{\hat{u}_k\}_{k=t-T}^{t+P})\|^2,$$

where P is the prediction horizon, and the sequence of actions

of length $P+1$ is transformed into the motion-invariant space.

Our findings indicate that incorporating the motion-invariant regularization during training generally reduces success rates when deploying policies on the *Abstract* embodiment but enhances performance in cross-embodiment settings with different robot embodiments, regardless of the neural network architectures. This highlights the applicability of leveraging motion invariance across various neural network architectures and its effectiveness for cross-embodiment learning.

C. Implementation Details

The visuomotor policy π_H predicts target poses for the handheld gripper at 10 Hz. The IK optimizer π_L realizes these target poses by retargeting them into whole-body motions, updating target joint positions and body orientation at 100 Hz. In simulation, we applied low-level PD control for each joint and body at 500 Hz. For the *Spot* robot, we additionally computed joint positions for the legs by solving IK analytically based on the target body pose. For the *Google Robot*, body motion was controlled similarly to other arm joints with PD control, though using high gains. In real robot setups, we controlled the robots through APIs provided by the manufacturers. For quantitative evaluation on the *Panda* robot, we used JOINT_IMPEDANCE mode via Deoxys [55] for joint position control. In the demonstration on the *Spot* robot, we directly streamed one-point trajectories for arm joint positions and body poses through Boston Dynamics’ Spot SDK.

D. Demonstrations in Simulation

Task demonstrations in simulation use the same tracking camera setup as in real-robot evaluations—a Realsense T265 [50] camera. To replicate real-world human demonstration behaviors, visual odometry data from the tracking camera is mapped to simulated handheld gripper motions in the *Abstract* embodiment or to IK commands for teleoperated simulation robots. The button interface for triggering grasp actions and recording data is kept consistent with the real-world setup. However, unlike real-world demonstrations, simulation does not require physical interaction with the handheld gripper. Therefore, shared gripper components were removed, and a simplified handle was used to reduce the workload on the human demonstrators.

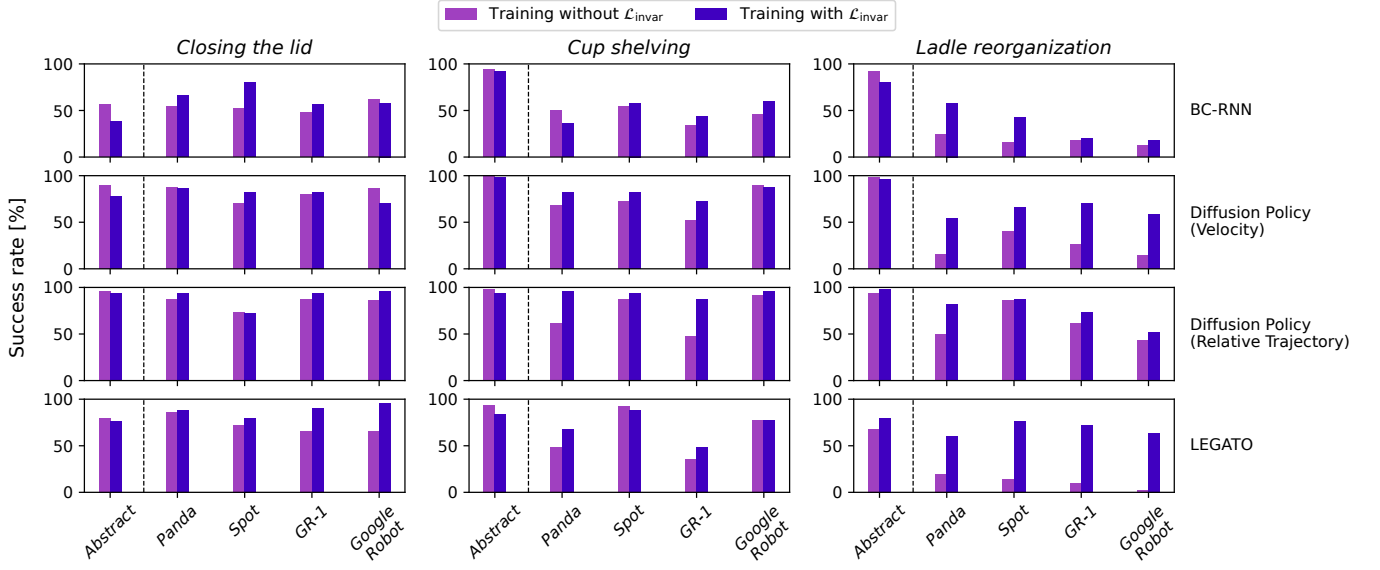


Fig. 10: Impact of training with the motion-invariant regularization We evaluate the success rates of policies trained with and without the motion-invariant regularization, $\mathcal{L}_{\text{invar}}$, over 50 trials across varied architectures. Each policy is trained on demonstrations from the *Abstract* embodiment. While the success rates on the *Abstract* embodiment (used for demonstrations) decrease with the motion-invariant regularization, the success rates improve in cross-embodiment settings, where policies are deployed to other robots.