

Humanoid Hanoi: Investigating Shared Whole-Body Control for Skill-Based Box Rearrangement

Minku Kim[†], Kuan-Chia Chen[†], Aayam Shrestha, Li Fuxin, Stefan Lee, and Alan Fern

Abstract—We investigate a skill-based framework for humanoid box rearrangement that enables long-horizon execution by sequencing reusable skills at the task level. In our architecture, all skills execute through a shared, task-agnostic whole-body controller (WBC), providing a consistent closed-loop interface for skill composition, in contrast to non-shared designs that use separate low-level controllers per skill. We find that naively reusing the same pretrained WBC can reduce robustness over long horizons, as new skills and their compositions induce shifted state and command distributions. We address this with a simple data aggregation procedure that augments shared-WBC training with rollouts from closed-loop skill execution under domain randomization. To evaluate the approach, we introduce *Humanoid Hanoi*, a long-horizon Tower-of-Hanoi box rearrangement benchmark, and report results in simulation and on the Digit V3 humanoid robot, demonstrating fully autonomous rearrangement over extended horizons and quantifying the benefits of the shared-WBC approach over non-shared baselines. Project page: https://osudrl.github.io/Humanoid_Hanoi/

I. INTRODUCTION

Long-horizon humanoid box rearrangement requires transforming an initial configuration of stacked boxes into a target configuration under placement and stacking constraints. This demands reliable composition of locomotion and manipulation skills over extended horizons. In practice, long-horizon execution compounds small errors and reveals failure modes that are rarely visible in isolated skill evaluations or short-horizon demonstrations. In this paper, we investigate a control architecture aimed at improving robustness and long-horizon task success for humanoid box rearrangement.

To solve box rearrangement, a humanoid must flexibly sequence locomotion and manipulation behaviors based on the current scene and intermediate outcomes, motivating skill- or stage-based long-horizon architectures [1, 10, 19]. Some learning-based systems train such skills end-to-end and deploy them as modular components [4]. However, naively composing independently trained skills is brittle. Composition can induce state and command distributions that differ from isolated training. Furthermore, switching between skill-specific low-level controllers or objectives can change closed-loop dynamics at skill boundaries, increasing the risk of transient instability. Finally, training skills in isolation often fails to exploit shared whole-body structure, leading to duplicated effort and reduced

[†] Contributed equally to this work.

* This work is supported by the NSF Award 2321851 and DARPA contract HR0011-24-9-0423.

The authors are with the Collaborative Robotics and Intelligent Systems Institute, Oregon State University, Corvallis, Oregon, 97331, USA {kimminku, chenku3, aayam.shrestha, Fuxin.Li, leestef, Alan.Fern}@oregonstate.edu

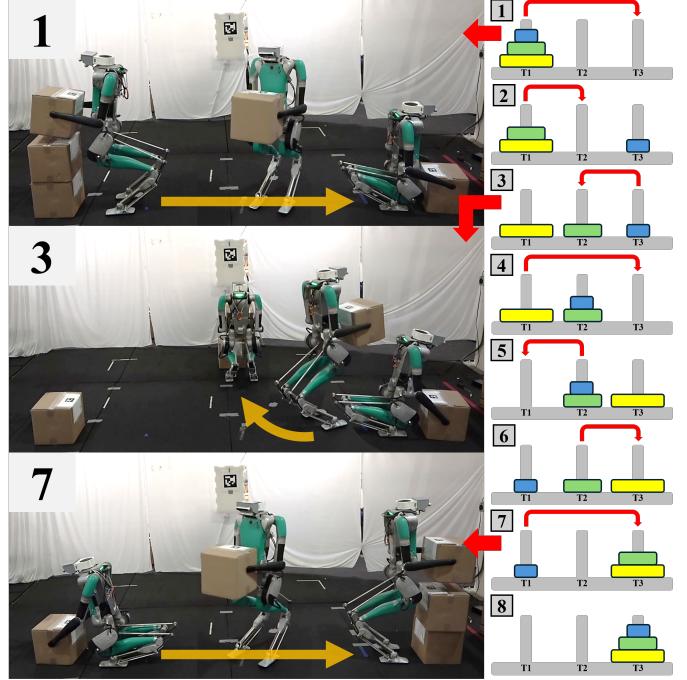


Fig. 1: *Humanoid Hanoi*, a problem instance from the Tower-of-Hanoi box rearrangement benchmark. The robot moves boxes between three towers (T1–T3) while respecting stacking constraints. The panels illustrate representative stages of a successful hardware execution (5+ min), with the corresponding symbolic state shown on the right. This benchmark stresses long-horizon autonomy by requiring repeated skill chaining with precise placement under constraints.

scalability as new skills are added.

We address this with a modular control framework in which independently trained skills are composed at runtime and executed through an always-on, shared, task-agnostic whole-body controller (WBC). By maintaining a unified low-level controller, skill composition does not alter the underlying closed-loop control structure and instead corresponds only to changes in the high-level command distribution. This enables skill reuse, simplifies composition, and avoids skill-dependent low-level control. While this architecture is natural, it has seen limited systematic exploration or demonstration for long-horizon humanoid tasks such as box rearrangement.

A key challenge in modular skill composition is maintaining robustness as the skill library grows and composed execution induces new state and command distributions. Long-horizon execution under disturbances, domain shift, and sim-to-real transfer can degrade closed-loop performance. Prior work often makes skill-specific modifications of the low-level

controller (e.g., residual policies or per-skill fine-tuning), but this introduces skill-dependent low-level control and complicates scalable composition. We instead treat robustness as a maintenance problem of the shared WBC. We refine the shared WBC via data aggregation, augmenting training with closed-loop composed rollouts under domain randomization and continuing optimization under the original WBC objective.

To evaluate long-horizon humanoid box rearrangement, we introduce *Humanoid Hanoi* (Fig. 3), a Tower-of-Hanoi-style benchmark that captures core challenges of obstacle-free box rearrangement while defining a broad distribution of instances. The benchmark stresses repeated skill reuse with precise placement and naturally exposes off-nominal robot and box configurations induced by imperfect locomotion and placement over many steps. We evaluate in simulation and on the Digit V3 humanoid robot and include a failure analysis that categorizes dominant error modes and suggests directions for improving long-horizon robustness.

In summary, the main contributions of this paper are:

- **Shared-WBC skill composition:** We investigate a modular skill-based architecture that sequences reusable skills at the task level while executing all skills through a single shared, task-agnostic WBC.
- **Shared-WBC coverage expansion:** We study rollout-based data aggregation to expand shared-WBC training coverage under skill-induced distribution shift, comparing against per-skill residual and task-objective fine-tuning baselines.
- **Humanoid Hanoi benchmark:** We introduce *Humanoid Hanoi*, a long-horizon Tower-of-Hanoi benchmark with task success and precision metrics that expose off-nominal states from accumulated execution error. We report simulation and hardware results with a failure-mode analysis, and release the benchmark publicly¹.

II. RELATED WORK

A. Learning-Based Humanoid Loco-Manipulation

Recent progress in learning-based humanoids has largely focused on learning task-specific policies or pipelines using teleoperation/mocap demonstrations and/or reinforcement learning (e.g., [7, 16, 11, 13, 20]). These systems often produce impressive behaviors such as picking up an object and walking, or executing a fixed manipulation routine, but the learned controller is typically tied to a particular task definition and horizon. Even when the behavior can be informally decomposed into subtasks (e.g., approach, grasp, transport, place), the controller is trained and executed as a single task policy and does not directly support arbitrary resequencing of its implicit “skills”. As a result, variations requiring different ordering or repetition generally call for additional training or redesign, rather than simply recombining reusable components.

In contrast, compositional architectures aim to learn reusable skills that can be flexibly sequenced to solve a range of long-horizon goals. Surprisingly, with few exceptions, learning-based humanoid work has not demonstrated

¹https://github.com/osudrl/Humanoid_Hanoi

robust, flexible sequencing of independently learned loco-manipulation skills over long horizons involving tens of skill invocations executed over minutes of operation. One partial exception is box loco-manipulation in [4], which learns and composes separate pickup, locomotion, and put-down behaviors (see below). Another example is ViRAL [9], which demonstrates extended operation by repeatedly executing a learned behavior for moving an object between tables. While robust over repetition, it does not demonstrate more arbitrary sequencing of distinct skills.

B. Humanoid Box Loco-Manipulation

Box pickup, transport, and stacking are canonical contact-rich loco-manipulation problems that have been studied extensively in humanoid robotics. Early systems relied on model-based planning and control, demonstrating lifting and transporting heavy objects via carefully planned whole-body motions [8] and multi-contact motion generation for dynamic lifting [3]. More recent optimization-based whole-body control methods address multi-contact dynamics and reactive coordination of locomotion and manipulation [1, 10]. These approaches offer strong performance in structured settings but are computationally expensive and sensitive to modeling error.

Box loco-manipulation behaviors have also served as representative demonstrations in learning-based humanoid systems (e.g., [16, 19, 11, 20]). However, these works typically do not focus on task-oriented rearrangement benchmarks with explicit long-horizon success and precision metrics for learned behaviors, nor do they emphasize reuse of skills across varied instance distributions. Closest to our setting is [4], which learns separate whole-body policies for pickup, locomotion, and put-down end-to-end. Because each skill is a distinct whole-body controller, stable composition can require additional transition machinery (e.g., specialized transition training and runtime interpolation). In contrast, we execute all skills through a single shared, task-agnostic WBC, so composition changes only the high-level directive stream and supports scalable long-horizon skill reuse.

III. SYSTEM OVERVIEW

We study long-horizon humanoid box rearrangement, where the environment contains multiple boxes of varying dimensions and masses, and the goal specifies a desired pose for each box. Each target pose may correspond to a placement on the floor or on another box, allowing goals that include stacking constraints and multi-step rearrangement. We assume the robot can perceive each box’s SE(3) pose and size, which are used to define task goals and provide inputs to the skills.

Fig. 2 shows our skill-based control architecture. At the high level, the system executes a small library of reusable loco-manipulation skills. We use a *GoTo* locomotion skill that drives the robot to a desired SE(2) base pose, with two variants: *GoTo* (unloaded) and *GoTo-with-box* (carrying a box). The manipulation library includes box *Pickup* and box *Place* (including stacking). Each skill is implemented as

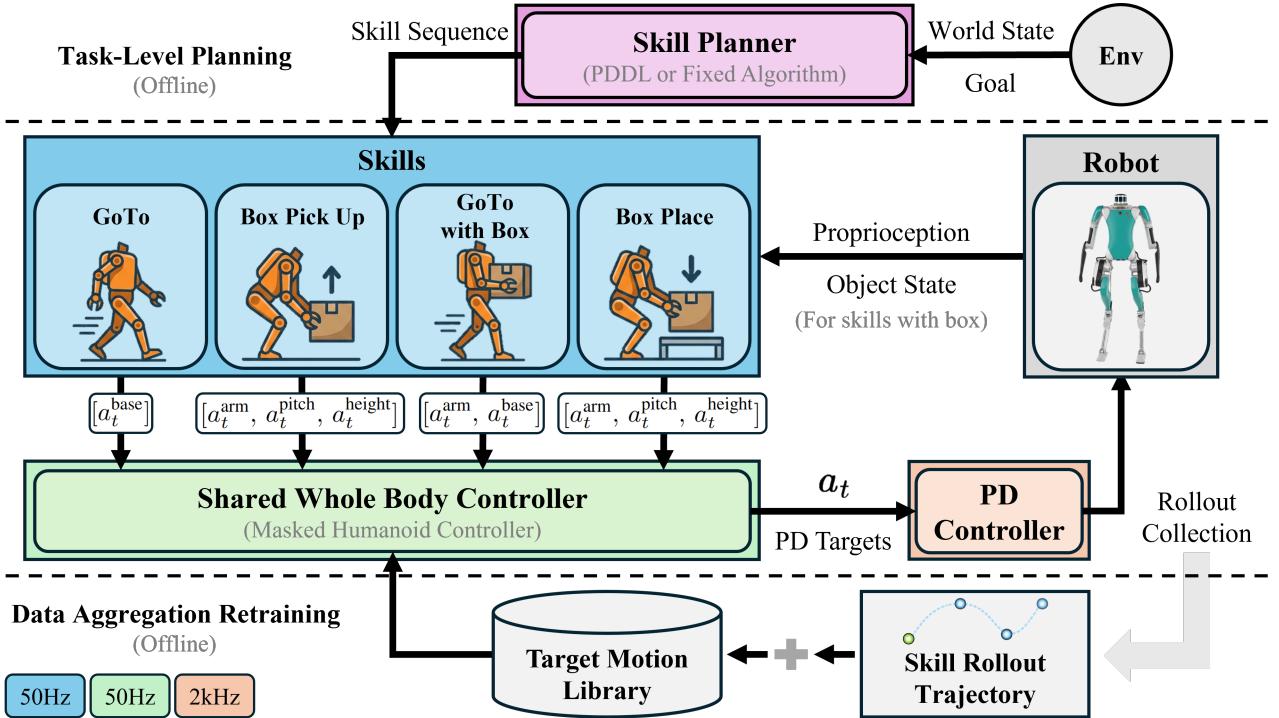


Fig. 2: Independently trained high-level skills generate task-level commands that are executed through a shared, task-agnostic whole-body controller (WBC). The WBC produces joint-level PD targets that are tracked by a low-level PD controller on the robot. Closed-loop rollouts from composed execution are aggregated to retrain the shared controller, improving robustness while preserving a unified control interface. The base action a_t^{base} specifies base locomotion commands, including a stand bit, planar velocity targets, and yaw rate.

an independent policy that maps observations to task-level commands rather than directly actuating the robot.

All skill commands are routed through an always-on, shared, task-agnostic whole-body controller (WBC) that provides a unified low-level control interface. Given the robot’s proprioceptive state and a skill’s command, the WBC outputs joint-level PD setpoints that are tracked by a PD controller. By keeping the WBC fixed across skills, composing skills do not switch low-level control laws, and skill composition corresponds primarily to changes in the high-level command.

Finally, our architecture is agnostic to how skills are selected and sequenced. In this paper, we focus on long-horizon execution and robustness and use simple task-level procedures to generate skill sequences for the Humanoid Hanoi benchmark (i.e., the Tower-of-Hanoi recursion driven by perceived box poses). We also implemented a symbolic planner interface (PDDL [2]) that can map discrete rearrangement goals to skill sequences. While this interface can solve Hanoi instances, we do not evaluate planning performance beyond this setting.

The following sections describe the skill interfaces, the shared WBC, and a rollout-based WBC extension procedure that maintains robustness as new skills are added.

IV. SHARED WHOLE-BODY CONTROLLER

For whole-body control, we use the pre-trained Masked Humanoid Controller (MHC) [6] as the shared WBC used by all skills. The MHC is a learned policy trained to execute *partially specified motion directives*, where a directive consists

of a target motion sequence paired with a binary mask indicating which pose components are active constraints. The pose representation includes the root state (position, orientation, linear and angular velocity) and joint positions, enabling a unified control interface at varying levels of specificity.

This masking mechanism allows different skills to constrain different subsets of the body while the WBC autonomously completes unspecified degrees of freedom in a dynamically consistent manner. For example, directives can specify only root linear and angular velocity for locomotion, additionally include torso pitch and height for stability, or further incorporate upper-body joint targets during manipulation. At runtime, the WBC takes the robot’s proprioceptive state and the masked motion directive produced by the active skill and outputs joint-level PD position targets for all actuated joints, which are tracked by a fixed-gain PD controller. Using a single shared WBC keeps the low-level closed-loop control structure fixed across skills, and skill composition changes only the high-level directive stream rather than switching low-level controllers.

V. SKILL LEARNING

This section describes the learning setup and training details for the learned *Pickup*, *Place*, and *GoTo/GoTo-with-box* skills.

Policy Architecture and Training. Each skill is a two-layer LSTM policy (64-D hidden state) running at 50 Hz that observes robot proprioception and skill-specific conditioning inputs (e.g., box pose/dimensions and target poses). The policy outputs a skill-dependent WBC motion directive, which the shared WBC converts (with proprioception) into PD position

	Parameters	Range
Dynamics Randomization	Body Mass	[0.75, 1.25] \times Default
	Joint Damping	[0.5, 3.5] \times Default
	Center of Mass Position	[0.95, 1.05] \times Default
	Friction Coefficient	[0.8, 1.2] \times Default
Box Randomization	Mass (kg)	[0.0, 3.5]
	Size XYZ (m)	[0.15, 0.5]
	X Displacement (m)	[0.3, 0.6]
	Y Displacement (m)	[-0.1, 0.1]
	Z Displacement (m)	[0.085, 0.95]
	Yaw Rotation ($^{\circ}$)	[-18, 18]
	Sliding Friction	[0.1, 1.0]
	Rolling Friction	[0.01, 0.1]
Communication	Delay	[2, 4]ms

TABLE I: Domain randomization parameters.

targets for the 20 actuated joints, tracked by a fixed-gain PD controller at 2kHz. Skills are trained with reinforcement learning (RL) in MuJoCo [18] on Digit V3 using PPO [15] with dynamics randomization [14] (Table I) for robustness. Unless otherwise specified, rewards are weighted sums of bounded exponential terms $w \exp(-\alpha c)$. We highlight key components below and defer full specifications to the supplementary material.

Pickup Skill. We train a *Pickup* policy to grasp boxes with varying pose (yaw), dimensions, and mass and return to stable standing while holding the box. The policy observes robot proprioception, the initial box pose in the robot root frame, and box dimensions, and outputs a masked WBC directive consisting of arm joint targets plus base pitch and height commands. Training uses a phased pickup curriculum (*approach*, *contact*, *lift*, *stand*) with a distance-based trajectory-tracking reward [4] on end-effector motion and base height. The *approach* duration is adapted to the box height,

$$t_{\text{approach}} = 100 \times (0.9 - \text{box height}) + 30,$$

while *contact* and *lift* run for 25 and 35 steps, and *stand* is set by distance to the target standing height (0.8 m). We apply domain randomization over box properties, dynamics, and communication delay (Table I). A support table is removed 45 steps after *contact* to force full lifting, and episodes end on timeout, box drop, or invalid hand contact. In addition to tracking, the reward includes grasp and stability terms (hand-face contact bonus, foot-force balance, base pitch/stance regularization) and hardware-oriented smoothness terms (action-rate penalty, height bounds, and an extended terminal stand of 120 steps).

Place Skill. We train a *Place* policy to set a held box at an upright, yaw-only target pose while maintaining whole-body stability. The policy observes robot proprioception, current and target box poses in the robot root frame, and box dimensions, and outputs the same directive space as *Pickup*. Since *Place* is harder to train, we use a privileged critic with additional inputs (hand and base poses, hand/table contact forces, and box mass).

To encourage consistent contacts and dynamically feasible placement, we generate reference motions by reversing successful *pickup* rollouts. We collect 60k successful *Pickup* trajectories (including 10k at challenging heights: box heights

[70, 90] cm, sizes [13.5, 30] cm) and reverse recorded end-effector, base-height, contact, and box/table state sequences to obtain *Place* references (mapping pickup *lift/stand* to place *approach/place*). Episodes start from standing with a box in hand and we randomize target distance ([40, 50] cm) and support height ([0, 75] cm) under the same domain randomization as *Pickup* (Table I), and inject box pitch disturbances ($[-30^{\circ}, 5^{\circ}]$) while keeping the target upright. In addition to reference tracking, explicit base-pitch constraints were critical, which penalized excessive forward leaning during lowering.

GoTo Skills. The *GoTo* locomotion skill drives the robot to SE(2) targets (x, y, yaw) using the constellation reward for target-oriented locomotion [5]. We train two variants: *GoTo* for unloaded walking and *GoTo-with-Box* for walking while carrying a box. Both are conditioned on a local-frame goal $(\Delta x, \Delta y, \Delta \text{yaw})$ that is periodically resampled during training. *GoTo-with-Box* additionally observes the box dimensions and box pose relative to the base. Policies output WBC directives consisting of planar base velocity and heading commands, and *GoTo-with-Box* includes upper-body targets to stabilize transport; base height and pitch are fixed (0.85 m, 0 rad).

Episodes initialize at $(0, 0)$ with randomized yaw and *GoTo-with-Box* samples feasible robot/box initial states from a collected dataset. Both variants terminate on falls/instability (excess roll/pitch, low base height), self-collision, or excessive speed, and *GoTo-with-Box* also terminates on loss of box contact. Training uses dynamics randomization and external perturbations (Table I). Beyond the constellation objective, regularizers encourage stable foot pose, smooth actions, and low effort, with additional box-specific terms for transport.

VI. SHARED WBC COVERAGE EXPANSION

A key practical challenge with a shared WBC is maintaining high reliability as new skills induce state and command distributions that differ from those seen during pretraining. In our system, initially learned skills achieve high reward in simulation when executed in isolation, yet we observe systematic failures in composed execution and under sim-to-real stressors. For example, low-height pick-ups and other extreme configurations can produce execution errors even when the skill policy outputs reasonable commands.

We attribute these failures to limited coverage in the pre-trained WBC. More broadly, even a very strong pretrained controller cannot cover all configurations, contact modes, and command patterns that arise as new skills and tasks are introduced. As the skill library grows, long-horizon composition inevitably exposes corner cases outside the original training distribution, especially under disturbances and sim-to-real shift. This motivates WBC coverage expansion, where we treat the shared WBC as a maintained component that is incrementally updated to expand coverage as skills are added.

In particular, rather than changing the architecture or introducing skill-dependent low-level control, we expand the WBC training distribution to include motion data induced by

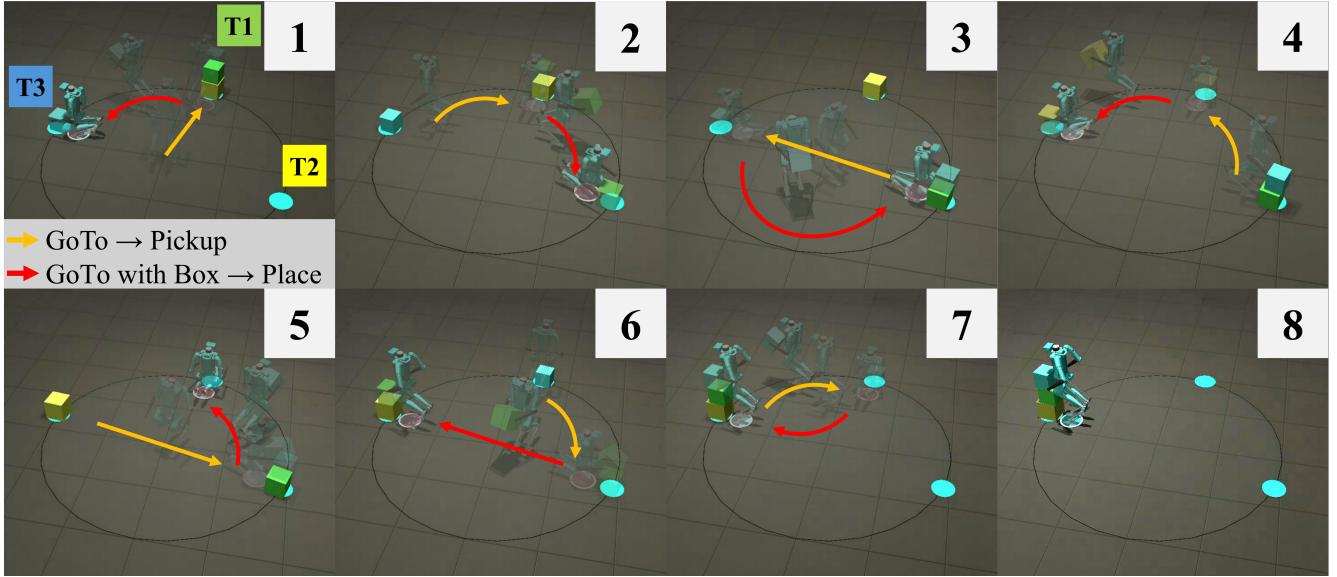


Fig. 3: Long-horizon *Humanoid Hanoi* execution. Sequential snapshots show a complete Tower-of-Hanoi-style box rearrangement episode. Transparent overlays visualize the executed robot trajectory over time, and **T1**, **T2**, and **T3** denote the target tower locations. The benchmark is divided into seven *moves*, each consisting of a sequence of *GoTo*, *Pickup*, *GoTo with Box*, and *Place* skills.

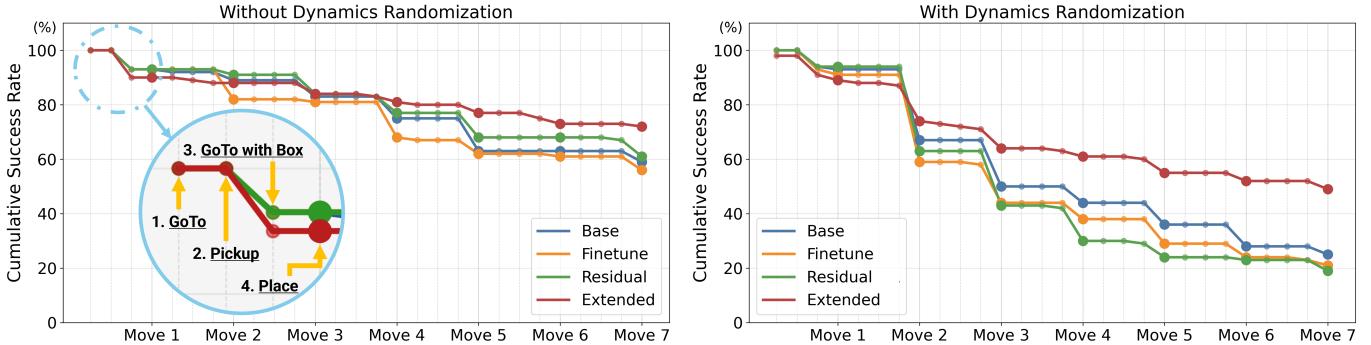


Fig. 4: Cumulative task success rates for the Humanoid Hanoi benchmark. Each *move* consists of four skills: *GoTo*, *Pickup*, *GoTo with Box*, and *Place*, shown as small markers, with large markers indicating move completion. Drops between consecutive points indicate failures in the subsequent skill (e.g., a drop from *Pickup* to *GoTo with Box* indicates failures during the *GoTo with Box* execution).

the learned skills, together with additional domain randomization and external disturbances. Intuitively, this robustifies the shared WBC in the parts of trajectory space that the new skills rely on, improving closed-loop reliability while preserving a unified control interface for skill composition.

Rollout-Based Data Aggregation. We implement coverage expansion via a simple data aggregation procedure that expands the reference motion set used to define the WBC training distribution. Let \mathcal{R} denote a set of reference trajectories. The MHC training process induces a distribution over masked motion directives, which we denote by $\mathcal{D}[\mathcal{R}]$. Concretely, sampling from $\mathcal{D}[\mathcal{R}]$ consists of (i) selecting a reference trajectory $r \in \mathcal{R}$ and a time index along r , (ii) applying a binary mask to specify which pose components are active constraints, and (iii) sampling additional randomized command targets (e.g., locomotion and upper-body targets) according to the original MHC command distribution [6]. This defines the distribution of directives used to train the shared WBC.

The pretrained MHC is trained under $\mathcal{D}[\mathcal{R}_0]$, where \mathcal{R}_0 contains reference trajectories from the AMASS dataset [12] and

trajectory-optimization-derived motions. When a new high-level skill π_i is trained, we execute the hierarchical system with the current shared WBC (without domain randomization) and record the resulting closed-loop *directive sequences* (targets and masks) issued to the WBC during successful rollouts. We treat these directive sequences as skill reference trajectories, forming a set \mathcal{S}_i , which we then expand incrementally:

$$\mathcal{R}_i = \mathcal{R}_{i-1} \cup \mathcal{S}_i, \quad (1)$$

so that the updated WBC training distribution becomes $\mathcal{D}[\mathcal{R}_i]$.

Finally, we continue training the shared WBC on $\mathcal{D}[\mathcal{R}_i]$ using the *same training objective* as in its original pretraining [6], while applying domain randomization and disturbance injection. This yields a single updated shared WBC that improves robustness on the skill-induced behaviors without introducing additional parameters, skill-specific losses, or architectural changes. In Sec. VIII, we quantify the benefits of this coverage expansion approach and compare against common alternatives that modify the low-level controller on a per-skill basis, such as residual policies [17] and skill-specific fine-tuning.

Metric	Skill	Base	Finetune	Residual	Extended
Full Height Distribution ($\mu \pm \sigma_{\bar{x}}$)					
Success Rate (%) \uparrow	Pickup	89.43 \pm 0.22	92.83 \pm 0.28	93.15 \pm 0.31	96.92 \pm 0.15
	Place	89.75 \pm 0.27	93.51 \pm 0.35	90.76 \pm 0.27	93.03 \pm 0.32
CoM Error (cm) \downarrow	Pickup	5.455 \pm 0.017	5.104 \pm 0.016	5.176 \pm 0.019	4.160 \pm 0.010
	Place	5.480 \pm 0.025	4.908 \pm 0.010	5.901 \pm 0.016	4.339 \pm 0.009
Force Distribution \downarrow	Pickup	0.4432 \pm 0.0008	0.4323 \pm 0.0008	0.4341 \pm 0.0009	0.3930 \pm 0.0006
	Place	0.4254 \pm 0.0013	0.4340 \pm 0.0005	0.4179 \pm 0.0010	0.3714 \pm 0.0007
Box Position Error (cm) \downarrow	Place	4.895 \pm 0.015	4.623 \pm 0.016	4.855 \pm 0.017	4.781 \pm 0.020
Box Yaw Error ($^{\circ}$) \downarrow	Place	2.64 \pm 0.35	3.71 \pm 0.37	2.34 \pm 0.26	2.48 \pm 0.37
Low Box Heights (< 0.2 m) (μ)					
Success Rate (%) \uparrow	Pickup	85.00	89.43	89.43	95.03
	Place	85.06	90.40	86.93	91.50
CoM Error (cm) \downarrow	Pickup	5.676	5.455	5.520	2.731
	Place	5.515	4.690	5.798	4.216
Force Distribution \downarrow	Pickup	0.4714	0.4705	0.4685	0.3854
	Place	0.4739	0.4556	0.4339	0.4065

TABLE II: Performance on *Pickup* and *Place* across box heights with dynamics randomization. **Top:** box and target heights sampled from full distribution (mean μ and standard error $\sigma_{\bar{x}}$ over 10 random seeds of 1000 episodes). **Bottom:** low box and target heights (< 0.2 m), reporting mean performance over 3000 episodes. **Gold** and **Silver** denote the best and second-best results, respectively.

VII. HUMANOID HANOI BENCHMARK

To systematically evaluate long-horizon box rearrangement, we introduce *Humanoid Hanoi*, a publicly-available benchmark inspired by the classical Tower of Hanoi puzzle. At a high level, *Humanoid Hanoi* is not intended as a narrow special case of box rearrangement. Rather, it is a compact benchmark that exercises many of the core elements of *obstacle-free* humanoid box rearrangement, where locomotion does not require obstacle avoidance and pickup/placement need only reason about the boxes being manipulated and their support surfaces. The key challenge is that long-horizon rearrangement performance is determined not only by single-skill competence, but also by robustness to the off-nominal states induced by prior actions. Small imperfections in earlier locomotion and placement naturally accumulate into a wide distribution of subsequent robot and box configurations, providing a realistic stress test for repeated skill reuse. We will release the benchmark environment in simulation so that any humanoid robot model can be tested on it.

The task requires the robot to move and stack three boxes under Tower-of-Hanoi-style rules: only one box may be moved at a time, and boxes have a fixed ordering (e.g., by size or index) such that higher-ordered boxes may not be placed on lower-ordered ones. Solving the task requires repeated sequencing and reuse of the same loco-manipulation skills over extended horizons, including (i) placement and stacking at different heights (including the floor), (ii) manipulation of boxes with different sizes and masses, (iii) pickup and placement from non-uniform robot and box poses induced by prior GoTo and placement errors, and (iv) varied locomotion paths, including rotation, both with and without a carried box.

Instance Distribution. Each episode is initialized by sampling a workspace radius uniformly from [1.5, 2.5] m. Three tower locations are placed on the circumference of the circle and oriented outward from the circle center, with a minimum pairwise separation of 0.9 m. Depending on the sampled radius, the tower locations may be nearly collinear or span

Upper Body Tracking	Base	Extended
Upper joint MAE (median [IQR]) \downarrow	0.1246 [0.1763]	0.0805 [0.1535]
Success Rate (%) \uparrow	95	100

TABLE III: Upper-body tracking performance comparing the *Base* WBC and the *Extended* WBC trained with data aggregation, on trajectories from the baseline training dataset.

a wide range of orientations, resulting in diverse geometric configurations and approach directions for locomotion.

Box properties are randomized at the start of each episode. Box sizes are sampled from three categories: *small* [0.26, 0.29] m, *medium* [0.29, 0.32] m, and *large* [0.32, 0.35] m. The sliding friction coefficient is uniformly sampled from [0.5, 0.7] to capture variability in real-world cardboard box interactions. The box mass is randomized within [0.5, 3.0] kg, including a 0.3 kg point mass attached at the center of the bottom to simulate boxes containing internal items. Standard dynamics randomization (DR) from Table I is applied to the robot model in the with-DR setting shown in Fig. 4, while the without-DR setting uses nominal robot parameters.

Evaluation. A trial is successful if all three boxes are stacked at the goal and satisfy the Hanoi stacking constraints. Each method is evaluated over 100 randomized episodes, and we report success rate and final box pose error metrics, which measure the box pose error relative to the ideal final tower. An example *Humanoid Hanoi* execution is shown in Fig. 3.

VIII. SIMULATION EXPERIMENTS

A. Approaches Compared

We compare approaches for *adapting the pretrained MHC* as new skills are added. Across methods, the learned high-level skill policies are fixed; only the whole-body controller (WBC) is updated (or not) to better cover the state/command distributions induced by the skills. The reason we chose this experiment is that, for humanoids, most of the balancing tasks fall within the realm of the WBC – the skill policies only output upper body arm commands, torso height, and pitch rotation, and they usually do a good job on those. However,

a WBC that is not trained with enough scenarios can easily lead the humanoid to lose balance, resulting in skill failure. Namely, we compare:

Base (Frozen MHC). *Base* uses the pretrained MHC as a frozen shared WBC with no skill-specific adaptation.

Finetune. *Finetune* adapts the WBC separately for each skill by fine-tuning from the pretrained MHC using the *skill's RL reward*, rather than the MHC tracking objective. This yields skill-specific WBCs that are switched with the active skill.

Residual. *Residual* freezes the pretrained MHC and learns a skill-specific residual policy optimized with the corresponding skill reward. This is similar to [20], except the target motions are generated by skill policies rather than mocap. These skill policies generate good trajectories on our tasks, especially without domain randomization. *Residual* uses an action scaling factor of 0.5, and reduced exploration noise for stability. One residual policy is learned for each skill.

Extended (Ours). *Extended* maintains a single shared WBC and expands its coverage via data augmentation (Section VI). We aggregate 9 reference trajectories each for the Pickup and Place skills via closed-loop skill executions spanning extremes of the box position distribution (lateral position and height). The GoTo skills were high-performing using the pre-trained MHC and therefore were not part of aggregation.

B. Individual Skill Evaluation

We evaluate the WBC baselines on the *Pickup* and *Place* skills across target heights, randomized box parameters, and under domain randomization (Table II). We omit *GoTo* comparisons since the base MHC already achieves strong locomotion performance and adaptation yields negligible differences. We report task success and stability metrics, including center-of-mass (CoM) displacement error and foot force distribution, and additionally report box position and yaw errors for *Place*. Force distribution is summarized by the mean and standard deviation of contact force magnitudes across both feet.

Table II shows that all adaptation methods outperform the frozen *Base* WBC. The proposed *Extended* WBC achieves the best or second-best performance across metrics and yields the largest stability gains. While *Finetune*, *Residual*, and *Extended* achieve similar success rates at medium and high target heights, *Extended* improves markedly for low heights (< 0.2 m), where balance margins are smallest (Table II).

These results suggest that *Extended* benefits primarily from training on closed-loop execution rollouts, which capture realistic tracking errors and reduce deployment distribution shift. By aggregating trajectories at extreme box positions and heights, *Extended* expands coverage in low-margin regimes that stress balance. In contrast, *Finetune* and *Residual* optimize task rewards, which can over-specialize for task completion without explicitly preserving the WBC stability objective. Notably, *Extended* achieves these gains while optimizing the task-agnostic MHC objective, eliminating the need to specify or tune task rewards for WBC maintenance. Meanwhile, Table III shows that *Extended* still preserves performance on upper body

Metric	Base	Finetune	Residual	Extended
No Dynamics Randomization ($\mu \pm \sigma$)				
x_{err} (cm) ↓	4.71 ± 3.46	5.79 ± 4.10	5.35 ± 3.73	5.47 ± 3.60
y_{err} (cm) ↓	5.26 ± 3.58	5.75 ± 4.27	5.24 ± 3.71	5.60 ± 3.69
θ_{err} (°) ↓	2.89 ± 2.74	2.65 ± 2.34	2.66 ± 2.57	3.59 ± 3.96
Dynamics Randomization ($\mu \pm \sigma$)				
x_{err} (cm) ↓	5.36 ± 4.12	5.72 ± 3.95	5.87 ± 5.28	5.42 ± 3.82
y_{err} (cm) ↓	5.23 ± 4.27	5.33 ± 3.57	5.87 ± 5.13	5.99 ± 3.93
θ_{err} (°) ↓	2.89 ± 3.08	2.59 ± 2.32	3.43 ± 6.18	3.30 ± 6.64

TABLE IV: Humanoid Hanoi placement accuracy for successful trials. Final box position and orientation errors for the *Place* skill, reported as mean μ and standard deviation σ .

tracking and even improves over *Base* on this task, despite the augmented data not being specifically collected for it.

C. Humanoid Hanoi Evaluation

We evaluate long-horizon skill composition on 100 random Humanoid Hanoi instances. Each instance is solved by executing the standard Tower-of-Hanoi recursion, which results in seven sequential *moves*, each consisting of four skills, shown in yellow and red arrows in Fig. 3.

Fig. 4 reports *task survival* as execution progresses: the y-axis shows the percentage of trials that have succeeded *up to* a given skill invocation (x-axis), so the final value of each curve corresponds to *complete-task* success. Without dynamics randomization (DR), *Extended* achieves approximately 70% complete-task success. Although its early-stage survival is slightly lower than *Residual* and *Base* through the third move, *Extended* degrades more slowly over the remainder of the horizon, yielding the highest overall completion rate. In contrast, while *Finetune* can improve performance at individual skill stages, its survival drops sharply under long-horizon composition and its complete-task success is worse than *Base*, suggesting limited generalization when skills are chained.

A similar trend holds under DR. While *Extended*'s complete-task success decreases to 49%, it maintains a substantial margin over the other methods throughout the horizon, indicating improved robustness to environmental variability and skill-induced distribution shift.

Table IV reports final box position and yaw errors for the *Place* skill, computed over successful Humanoid Hanoi trials. One can see that all approaches achieve similar errors that are not statistically significantly different from one another. *Extended* does not always achieve the lowest placement error, but note that *Extended* often succeeds in stabilizing and stacking even when the achieved pose deviates from the nominal target, whereas other methods fail in similar low-margin configurations and are therefore absent from the successful-trial statistics. Consequently, the reported *Extended* errors correspond to a broader and more challenging subset of successful executions.

D. Humanoid Hanoi Analysis of Failure Modes

We analyze failures on the Humanoid Hanoi benchmark to identify the dominant limitations of the current system and to motivate future improvements. Fig. 6 shows that most

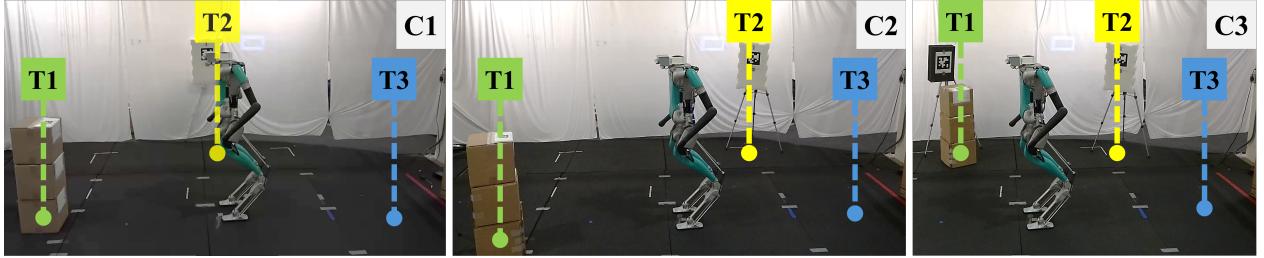


Fig. 5: Three Humanoid Hanoi configurations (C1–C3) where **T1**, **T2**, and **T3** denote the tower locations defining the stack positions.

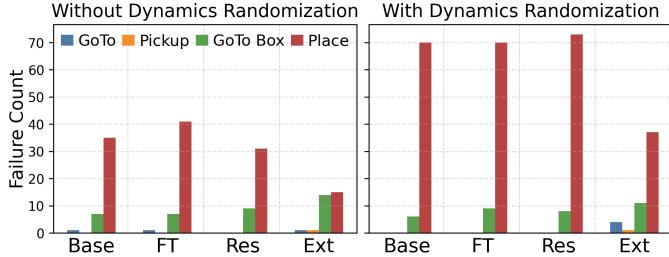


Fig. 6: Failure counts across skills in the Humanoid Hanoi benchmark for each baseline, shown with and without dynamics randomization.

failures occur during *Place*, which combines precise end-effector pose control with tight balance margins and stack stability constraints.

Placement-Induced Failures. A common failure pattern is unintended box pitch rotation during placement. Small control errors and contact transients can introduce orientation misalignment, leading to unstable support and eventual stack collapse. We also observe failures during post-placement stabilization: after releasing the box, the robot may take forward recovery steps that inadvertently contact and disturb the stack.

Upstream Errors Accumulate into Placement. Several failure modes originate before *Place* but become critical at placement time. First, off-centered grasps during *Pickup* introduce residual box pose errors in the gripper, which then propagate into placement errors and accumulate over multiple moves. Second, residual localization error after *GoTo-with-box* reduces placement margin. In our system, the robot transitions to *Place* once it is within 5 cm of the target SE(2) pose; remaining base pose errors at this point directly translate into placement offsets and can destabilize the stack in later stages.

Object-Unaware Locomotion. Finally, failures also occur during transitions from *Place* back to *GoTo*. Because *GoTo* is not object-aware, backward walking combined with in-place rotation can collide with the stack, displacing boxes and causing task failure.

IX. HARDWARE EXPERIMENTS

We evaluate sim-to-real transfer on the Digit V3 humanoid using AprilTags for box pose estimation and localization. We provide qualitative hardware videos for (i) *Pickup* and *Place* across variations in box size, mass, and target height, and (ii) stability comparisons between *Base* and *Extended*. See the supplementary material for setup details and videos.

Humanoid Hanoi. We evaluate long-horizon performance on three fixed Humanoid Hanoi workspace configurations (Fig. 5). For each configuration (C1)-(C3), we run 5 trials,

Configuration	Success Rate	Avg. Moves / Trial
C1	3/5	4.6
C2	2/5	3.8
C3	1/5	4.2
Overall	6/15 (40%)	4.2

TABLE V: Digit V3 real-robot Humanoid Hanoi results. Evaluated on three fixed configurations (C1–C3) with 5 trials each. We report the complete-task success rate and the average number of completed box moves per trial (including failed trials).

terminating on failure. Successful trials complete the full Hanoi sequence and take approximately 5 minutes. Table V shows the success rate and the average number of completed box moves for each configuration. The overall success rate is 40%, and the average number of completed moves is 4.2.

Failure Modes. Across the trials, failures fell into three categories: (i) perception/state estimation, (ii) navigation/alignment at skill handoffs, and (iii) physical interaction during manipulation/placement. The most common were AprilTag perception failures during *Pick* (2 trials), box flipping during *Place* from unintended contact (2), and yaw misalignment at handoff (2). Remaining failures included a large lateral localization error causing an unsuccessful grasp (1), an IMU fault that degraded walking and led to tag loss (1), and a box being kicked during placement/stand-up (1). Overall, these results indicate that hardware robustness is currently limited by tag robustness, base-pose accuracy entering *Place*, and clearance-aware body motion during placement and recovery.

X. CONCLUSION AND FUTURE WORK

We presented a skill-based humanoid box rearrangement framework that composes independently learned skills through a shared whole-body controller (WBC). We showed that refining the shared WBC via rollout-based data aggregation from closed-loop composed execution improves robustness under domain shift without introducing skill-specific low-level control. We evaluated the system on *Humanoid Hanoi* and demonstrated fully autonomous pickup, transport, and placement in simulation and on the Digit V3 humanoid.

While the shared-WBC architecture with rollout-based refinement is a practical approach to long-horizon execution, there is substantial headroom in full-task success and placement precision. Our failure analysis highlights directions for improvement, including object-aware locomotion and departures, stronger placement and stabilization behavior, and replacing external markers with onboard perception and state estimation (e.g., object detection and SLAM).

REFERENCES

- [1] Alphonsus Adu-Bredu, Grant Gibson, and Jessy Grizzle. Exploring kinodynamic fabrics for reactive whole-body control of underactuated humanoid robots. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10397–10404. IEEE, 2023.
- [2] Constructions Aeronautiques, Adele Howe, Craig Knoblock, ISI Drew McDermott, Ashwin Ram, Manuela Veloso, Daniel Weld, David Wilkins Sri, Anthony Barrett, Dave Christianson, et al. Pddl—the planning domain definition language. *Technical Report, Tech. Rep.*, 1998.
- [3] Hitoshi Arisumi, Jean-Rémy Chardonnet, Abderrahmane Kheddar, and Kazuhito Yokoi. Dynamic lifting motion of humanoid robots. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 2661–2667. IEEE, 2007.
- [4] Jeremy Dao, Helei Duan, and Alan Fern. Sim-to-real learning for humanoid box loco-manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16930–16936. IEEE, 2024.
- [5] Pranay Dugar, Mohitvishnu S Gadde, Jonah Siekmann, Yesh Godse, Aayam Shrestha, and Alan Fern. No more marching: Learning humanoid locomotion for short-range se (2) targets. In *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, pages 1–8. IEEE, 2025.
- [6] Pranay Dugar, Aayam Shrestha, Fangzhou Yu, Bart van Marum, and Alan Fern. Learning multi-modal whole-body control for real-world humanoid robots. In *Proceedings of the AAAI Symposium Series*, pages 650–657, 2025.
- [7] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*, 2024.
- [8] Kensuke Harada, Shuuji Kajita, Hajime Saito, Mitsuharu Morisawa, Fumio Kanehiro, Kiyoshi Fujiwara, Kenji Kaneko, and Hirohisa Hirukawa. A humanoid robot carrying a heavy object. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 1712–1717. IEEE, 2005.
- [9] Tairan He, Zi Wang, Haoru Xue, Qingwei Ben, Zhengyi Luo, Wenli Xiao, Ye Yuan, Xingye Da, Fernando Castañeda, Shankar Sastry, Changliu Liu, Guanya Shi, Linxi Fan, and Yuke Zhu. Viral: Visual sim-to-real at scale for humanoid loco-manipulation. *arXiv preprint arXiv:2511.15200*, 2025.
- [10] Junheng Li and Quan Nguyen. Multi-contact mpc for dynamic loco-manipulation on humanoid robots. *arXiv preprint arXiv:2209.08662*, 2022.
- [11] Fukang Liu, Zhaoyuan Gu, Yilin Cai, Ziyi Zhou, Hyunyoung Jung, Jaehwi Jang, Shijie Zhao, Sehoon Ha, Yue Chen, Danfei Xu, et al. Opt2skill: Imitating dynamically-feasible whole-body trajectories for versatile humanoid loco-manipulation. *IEEE Robotics and Automation Letters*, 2025.
- [12] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019.
- [13] Masaki Murooka, Takahiro Hoshi, Kensuke Fukumitsu, Shimpei Masuda, Marwan Hamze, Tomoya Sasaki, Mitsuhiro Morisawa, and Eiichi Yoshida. Tact: Humanoid whole-body contact manipulation through deep imitation learning with tactile modality. *IEEE Robotics and Automation Letters*, 2025.
- [14] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.
- [15] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [16] Mingyo Seo, Steve Han, Kyutae Sim, Seung Hyeon Bang, Carlos Gonzalez, Luis Sentis, and Yuke Zhu. Deep imitation learning for humanoid loco-manipulation through human teleoperation. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pages 1–8. IEEE, 2023.
- [17] Tom Silver, Kelsey Allen, Josh Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018.
- [18] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.
- [19] Chong Zhang, Wenli Xiao, Tairan He, and Guanya Shi. Wococo: Learning whole-body humanoid control with sequential contacts. *arXiv preprint arXiv:2406.06005*, 2024.
- [20] Siheng Zhao, Yanjie Ze, Yue Wang, C Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. Resmimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. *arXiv preprint arXiv:2510.05070*, 2025.

