HDMI: Learning Interactive Humanoid Whole-Body Control from Human Videos

Haoyang Weng, Yitang Li, Nikhil Sobanbabu, Zihan Wang,
Zhengyi Luo, Tairan He, Deva Ramanan, and Guanya Shi

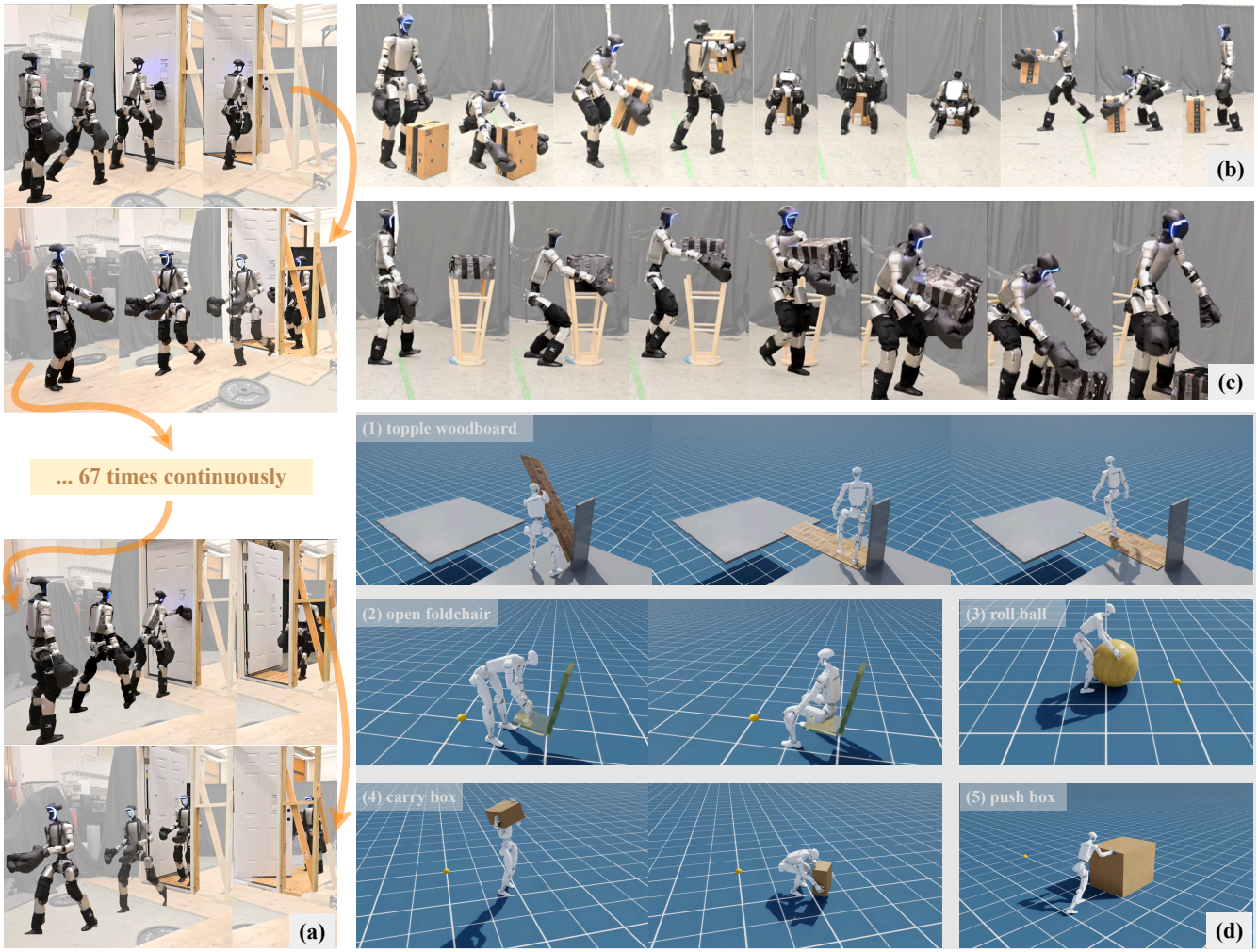


Fig. 1: HDMI enables humanoid robots to acquire diverse whole-body interaction skills directly from human videos. **(a)** Traversing doors: the robot successfully passes through a door for *67 consecutive trials* ($\sim$ 34 mins), and remains robust under terrain changes. **(b)** Moving a cardboard box: the robot kneels to grasp and relocate the box, demonstrating coordinated whole-body motion. **(c)** Carrying and dropping objects: the robot walks forward to pick up and drop a pile of foam mats. **(d)** A wide range of interaction tasks in simulation, including toppling a wood board, opening a foldable chair, rolling a ball, carrying a box, and pushing a box. Website: https://hdmi-humanoid.github.io/

***Abstract*— Enabling robust whole-body humanoid–object interaction (HOI) remains challenging due to motion data scarcity and the contact-rich nature. We present HDMI (HumanoiD iMitation for Interaction), a simple and general framework that learns whole-body humanoid-object interaction skills directly from monocular RGB videos. Our pipeline (i) extracts and retargets human and object trajectories from unconstrained videos to build structured motion datasets, (ii) trains a reinforcement learning (RL) policy to co-track robot and object states with three key designs: a unified object representation, a residual action space, and a general interaction reward, and (iii) zero-shot deploys the RL policies on real humanoid robots. Extensive sim-to-real experiments on a Unitree G1 humanoid demonstrate the robustness and generality of our approach: HDMI achieves 67 consecutive door traversals and successfully performs 6 distinct loco-manipulation tasks in the real world and 14 tasks in simulation. Our results establish HDMI as a simple and general framework for acquiring interactive humanoid skills from human videos.**

The authors are with the Robotics Institute, Carnegie Mellon University, USA. {hweng2, yitangl, nsobanba, zihanwa3, zhengyiluo5, tairanhe, deva, guanyas}@andrew.cmu.edu

I. Introduction

Humanoid robots hold immense potential for assisting humans in diverse environments due to their human-like morphology and versatility. To fully unleash their capabilities, enabling humanoids to robustly interact with objects and their environments is critical.

Learning from human motions has been a dominant pipeline in humanoid robotics, which leverages abundant human data to achieve agile locomotion [1–3] and dexterous manipulation [4–6]. However, there are significant limitations when extending these successes to whole-body, contact-rich humanoid-object interaction (HOI) tasks. We identify two key challenges: (1) Compared to free-space locomotion, human-object interaction data with 3D human and object motions is scarce. (2) Learning whole-body interaction tasks poses new challenges for RL training, such as guiding desired contact behavior under imperfect motion references and learning to balance with objects in challenging poses.

Prior humanoid-object interaction works have either relied on task-specific motion reference generation pipelines or manual reward engineering [7, 8] that limit generality or require high-level VLM or model-based planners [9, 10]. To overcome these limitations, we propose HDMI (**HumanoiD iMitation for Interaction**), a general framework for acquiring interactive skills directly from RGB videos. Our framework supports interaction with different body parts (hand, foot), and with different types of objects (articulated or rigid body, fixed-based, floating-based).

Our key insight is to bypass task-specific reward engineering by jointly tracking robot and object motions from video with an end-to-end RL control policy. Our pipeline has three stages: (i) extract and retarget human and object trajectories from RGB videos using pose estimation [11, 12] to build structured reference datasets; (ii) train a control policy with RL to co-track robot and object states, using proprioception, a phase variable, and object state in the robot's local frame; and (iii) deploy the learned policy directly on humanoid robot to execute interaction skills.

Our contribution is three-fold:

- We propose HDMI, a simple and general framework that learns interactive skills for humanoids directly from RGB videos. As far as we know, this is the first general framework that enables learning autonomous whole-body humanoid-object interaction skills directly from human videos.
- To facilitate stable and efficient training for complex humanoid-object interactions, we design *three simple and unified components*: a *unified object representation* for diverse objects, a *residual action space* for stable exploration of challenging poses, and a *unified interaction reward* that promotes robust and precise contact even with imperfect reference motions.
- We demonstrate the robustness and generality of our framework through extensive sim-to-real experiments on Unitree G1. As shown in Figure 1, the learned policy achieves *67 consecutive bi-directional door openings and traversals*, and we successfully train and deploy other *6 distinct loco-manipulation tasks* on real hardware and *14 tasks* in simulation.

II. Related Works

A. Humanoid Learning

RL has made remarkable progress for agile humanoid locomotion skills. Through large-scale training, humanoids have achieved robust walking and running [13, 14] , dynamic behaviors such as jumping [15], parkour [16], and expressive whole-body motions [3, 17]. In parallel, imitation learning and large-scale human demonstrations have produced dexterous manipulation policies [4–6], enabled by advances in data collection [18–20]. Despite this progress in locomotion and manipulation individually, relatively few works have addressed humanoid loco-manipulation, where the robot must simultaneously move and interact with objects in contact-rich settings [21–23]. Our work focuses on this challenging frontier, developing policies that couple locomotion with object interaction to enable robust, and expressive humanoid skills in the real world.

B. Humanoid Loco-Manipulation

Learning-based methods have recently extended from locomotion and manipulation in isolation to full humanoid loco-manipulation, yet existing approaches remain limited in generality and robustness. Some rely on reward engineering or trajectory optimization, achieving task-specific successes such as box transport [7, 8]. Others introduce specialized architectures, including skill blending [24] or decouple low-level tracking with high level task policies [3, 22, 23]. More recent efforts address robustness through dual-agent RL frameworks [25, 26], and adaptive policies for human–humanoid collaboration [27]. While these advances expand humanoid capabilities, their applicability is relatively narrow. Our framework instead introduces a general, dense, demonstration-driven objective that couples locomotion with object interaction, enabling adaptive and contact-stable behaviours that scale across diverse tasks.

C. Robot Learning from Human Videos

Recent advances in humanoid robot learning have increasingly turned to human videos as a scalable source of demonstration data. Despite the success of methods learning locomotion skills from video demonstrations [1, 28–30], these approaches fundamentally lack object interaction capabilities as they do not explicitly model object dynamics during training. Another line of work focuses on learning manipulation skills from video demonstrations [31–33]. However, these methods are typically constrained to only upper-body interactions, lacking potential to utilize the large workspace that can be achieved with whole body coordination. To address this limitation, HDMI learns whole-body interactions from monocular RGB videos, modeling human–object dynamics and co-tracking trajectories to unify locomotion and manipulation.

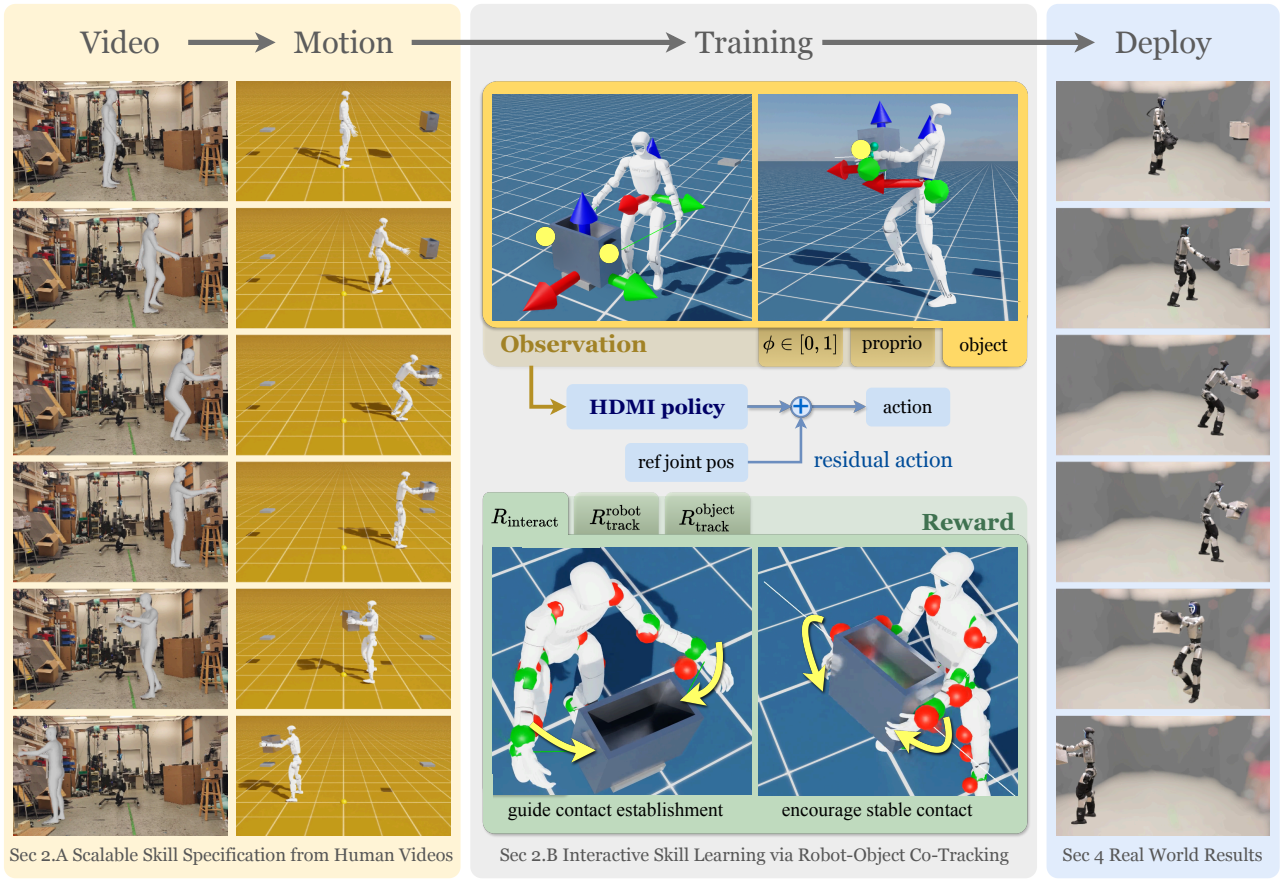


Fig. 2: HDMI is a general framework for interactive skill learning. Monocular RGB videos are processed into a structured dataset as reference trajectories (section III-A), which are used to train an interaction centric policy via robot-object co-tracking (section III-B). The trained policies are succesfully deployed to real world humanoids (section IV).

III. Method

HDMI is a general framework for interactive skill acquisition. We first retarget and process monocular RGB videos of human interacting with objects into a structured dataset of reference trajectories (section III-A). Then these reference trajectories are used to train a interaction centric policy via robot-object co-tracking (section III-B).

A. Scalable Skill Specification from Human Videos

A primary challenge in teaching robots complex interaction skills is the difficulty of specifying the desired behavior. To address this, we leverage monocular RGB videos of humans interacting with objects as an abundant and scalable data source. We use GVHMR [11] and LocoMujoco [12] for smpl-pose estimation and retargeting. Next, we post-process and annotate the object trajectory and contact signals to produce a structured dataset of reference motions, in which each frame provides a reference state $\{\mathbf{s}_t^{\text{ref}}\}$ together with desired contact points defined in the object's local frame $\{\mathbf{p}_t^{\text{contact}}\}$. The reference state at frame t is $\mathbf{s}_t^{\text{ref}} = (\mathbf{s}_t^{\text{robot}}, \mathbf{s}_t^{\text{obj}}, \mathbf{c}_t)$ where $\mathbf{s}^{\text{robot}}$ consists of $\mathbf{p}^{\text{robot}}$, $\mathbf{q}^{\text{robot}}$ and θ_t^{ref}, representing the robot's reference root position, root orientation and reference joint positions. $\mathbf{s}^{\text{obj}}$ consists of $\mathbf{p}^{\text{obj}}$ and $\mathbf{q}^{\text{obj}}$, representing the object's position and orientation. For articulated objects (e.g., doors or folding chairs), $\mathbf{s}_t^{\text{obj}}$ additionally includes their joint state θ^{obj}. $\mathbf{c}_t \in \{0,1\}$ is a binary signal indicating whether contact is intended at frame t.

B. Interactive Skill Learning via Robot-Object Co-Tracking

We formulate interactive skill learning as a robot–object co-tracking problem. Specifically, given a processed reference motion $\{\mathbf{s}_t^{\text{ref}}, \mathbf{p}_t^{\text{contact}}\}$, our goal is to train a whole-body control policy that simultaneously follows the reference trajectory of both the robot and the object at each timestep using reinforcement learning.

Following established tracking works [1, 2, 34], we use a DeepMimic-style [34] training: (1) *Reference state initialization*: during each episode, both the robot and object are initialized from a random frame s_t^{ref} in the reference motion, with additional small random perturbations to enhance robustness. (2) *Phase variable observation*: policy receives a phase variable $\phi \in [0,1]$, where $\phi = 0$ represents the start of a motion and $\phi = 1$ represents the end. This time phase variable alone is proven to be sufficient for single-motion tracking [1, 34]. (3) *Tracking-error-based termination*: we terminate the episode when robot or object states deviate too much from the reference motion. We train the policy with tracking and regularization rewards and use PPO to optimize

it. A detailed list of rewards and terminations can be found in table I and table II.

However, learning robust whole-body object interaction propose new challenges. We address the key challenges by introducing three targeted solutions:

*1) **Unified Object Representation**:* To allow our framework to generalize across diverse objects with different geometries and types, we design a unified representation for object observations. At each timestep, the policy receives the object's pose expressed relative to the robot's local root frame. This spatially invariant formulation facilitates generalization and can be naturally distilled to on-board sensory inputs such as RGB or depth images.

To further guide the interaction, we also provide the policy with reference contact points $\mathbf{p}^{\text{contact}}$, which specify desired robot–object contact locations as shown in fig. 3. These points are also transformed into the robot's root frame. Together with the robot's proprioceptive state $\mathbf{s}_t^{\text{proprio}}$ and the phase variable $\phi \in [0, 1]$, this unified observation (fig. 2 top) allows our framework to be applied to different object types without architectural changes.

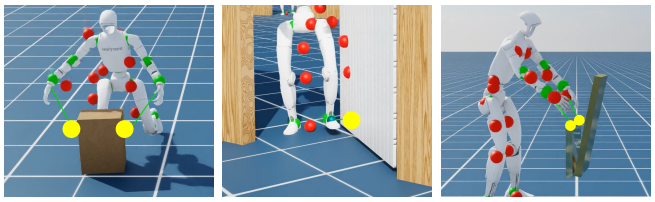


Fig. 3: Reference contact position (yellow dot) in three different tasks. Policy observes the positions of these contact points in root frame, both during training and deployment.

*2) **Residual Action Space**:* Learning complex motions often involves challenging poses, such as kneeling, that are far from the robot's default standing pose. At the beginning of training, the exploration of a standard policy is centered around the default pose. When an episode is initialized in a kneeling reference pose, the policy's first action often causes the robot to abruptly pop up toward a standing pose, leading to immediate instability and uninformative training samples.

To overcome this, we employ a residual action space. Instead of learning the absolute joint target θ_t^{target} directly, the policy learns to output a corrective offset $\mathbf{a}_t$, which is added to the joint positions from the reference kinematic trajectory θ_t^{ref} (fig. 2 middle) forming: $\theta_t^{\text{target}} = \theta_t^{\text{ref}} + \mathbf{a}_t$. This grounds the initial exploration to be centered around the current reference pose. For example, if the reference motion kneels down on the ground, due to exploration in the residual action space, the policy learns how to balance itself, while in the direct action space, the policy needs to learn to output large offsets to the zero pose. This targeted exploration leads to improved sample efficiency and significantly faster convergence to the desired behavior, especially for complex motions that deviate significantly from a standard standing pose.

*3) **Unified Interaction Reward**:* Reference trajectories obtained via video retargeting are purely kinematic and often lack precise contacts or contain penetration artifacts. Relying solely on motion-tracking rewards is therefore insufficient. To address this gap, we introduce a unified contact-promoting reward $R_{\text{interaction}}$, which encourages the policy to establish and maintain stable contact when the reference indicates the intended interaction ($\mathbf{c}_t = 1$).

For each active end-effector i, the reward combines (i) a position term that encourages alignment with the target contact point, and (ii) a force term that promotes stable yet bounded contact forces:

$$R_{\text{contact},i} = \exp\left(-\frac{\|\mathbf{p}_{\text{eef},i} - \mathbf{p}_{\text{target},i}\|_2}{\sigma_{\text{pos}}}\right) \cdot \max\left(\exp\left(\frac{\|\mathbf{F}_{\text{contact},i}\|_2 - F_{\text{thres}}}{\sigma_{\text{frc}}}\right), 1\right) \tag{1}$$

where $\mathbf{p}_{\text{eef},i}$ and $\mathbf{p}_{\text{target},i}$ are the positions of the end-effector and its target contact point on the object, respectively. $\mathbf{F}_{\text{contact},i}$ is the contact force between the eef and the object. The position term rewards proximity between end-effector and object contact points. The force term rewards sufficient but not excessive contact force, capped by a threshold F_{thres} to ensure safety at deployment. The overall interaction reward is averaged across all active end-effectors, gated by the contact signal.

$$R_{\text{interaction}} = \left(\frac{1}{N_c}\sum_{i=1}^{N_c} R_{\text{contact},i} \cdot \mathbf{c}_{t,i}\right) \tag{2}$$

where N_c represents the number of end-effectors desired to have contact.

Domain Randomization: To improve the robustness of the trained policy, we randomize the robot and object's inertial and friction properties during training.

TABLE I: Reward Functions.

Reward	Weight	Reward	Weight
Robot Tracking		**Regularization Penalties**	
Body Local Pose	2.0	Action Rate L2	0.1
Root Global Pose	1.0	Joint Position Limits L1	10.0
Body Global Velocity	1.0	Joint Velocity Penalty	5.0e-4
Joint Tracking	1.0	Joint Torque Limits L1	0.01
Object Tracking & Interaction		Feet Impact Force L2	1.0
Object Pose	2.0	Feet Slip	0.5
Contact Reward	5.0	Feet Air Time	5.0

TABLE II: Termination Conditions.

Termination Condition	Threshold	Min Steps
Robot Tracking Error		
Root Global Pose Error	0.5m, 1.2rad	25
Body Local Pose Error	0.5m, 1.2rad	25
Object Tracking Error		
Object Pose Error	0.5m, 1.2rad	25
Contact Loss		
Lost Contact	Pos 0.2m & Force 1.0N	25

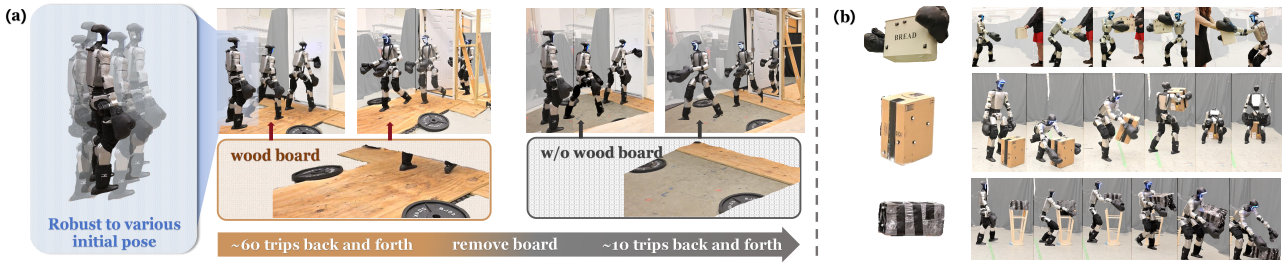


Fig. 4: Demonstrations on challenging real-world tasks. **(a)** Door opening and traversal: the robot adapts its footsteps to different initial poses and terrain variations (with/without wooden board), successfully completing 67 consecutive trips. **(b)** Box loco-manipulation: the policy enables versatile whole-body coordination for grasping, lifting, and transporting objects of varied shapes, sizes, and weights.

IV. Real-World Experiments

We evaluate HDMI on five real-world interaction tasks. These tasks collectively test the robot's ability to handle contact-rich interactions, combine locomotion and manipulation, and perform long-horizon whole-body coordination.

Experiment Setup. We train all policies in IsaacSim with same set of hyperparameters and directly deploy them on a Unitree G1 humanoid. To obtain inputs for the policy, we attach mocap markers on the robot's pelvis and on the manipulated objects, to acquire the global pose of each root link. Object poses are then transformed into the robot's root frame to serve as policy observations.

All reference motions are derived from RGB videos, except for suitcase manipulation, which is retargeted from the Omomo dataset [35]. While Omomo offers high-quality human–object interaction data, many motions involve prehensile manipulation (e.g., moving a monitor or lamp) that requires dexterous hands, which is outside the scope of our study.

A. Case I: Door Open and Traversal

This task tests the policy's ability to handle different contact interactions in sequence. The robot must push a door open with its hand, walk through, turn around, kick the door open with its foot, pass through again, and return to start.

The policy completed 67 continuous door runs (34 minutes) before failure. It remained robust under terrain changes, successfully performing $\sim$ 7 runs after the wooden floorboard was removed. As shown in fig. 4(a), the robot started each round with a random positional offset of 10–30 cm. Despite this variation, it adapted its footsteps and precisely raised its arm to make contact at the correct location.

These results highlight the policy's robustness to environmental variation and adaptability to positional changes.

B. Case II: Box Loco-Manipulation

These tasks evaluate whole-body coordination to manipulate box-like objects of varying heights and weights (fig. 4 b). The robot must integrate whole-body motions, such as kneeling, grasping, carrying, turning, and placing objects.

1) Suitcase manipulation: The robot executed 7 consecutive successful runs with smooth transitions between kneeling, lifting, and walking with load.
2) Bread box carrying: The policy completed 2 full trials. The main challenge was the rapid 180° turn, which occasionally caused leg collisions.
3) Foam mats relocation: The robot walks forward, grasps the mats, sidesteps, and places them down successfully.

These results shows HDMI enables seamless whole-body coordination between grasping, locomotion, and manipulation across various object properties and heights.

C. Case III: Truman's Bow

This complex, multi-stage task requires executing a long sequence of behaviors: climbing a staircase, performing a Truman-style bow, sitting on the stair, waving, jumping off, and walking back.

Due to the risks of stair climbing, human operators occasionally applied light support to the robot's back. The learned policy successfully executed the entire sequence 3 times continuously. This demonstrates the framework's robust capability for versatile and complex motions, encompassing dynamic and contact rich object-scene interactions (e.g., climbing and sitting on stairs) and precise full-body pose control for expressive gestures (e.g. bowing, waving hands).

This highlights HDMI's ability to produce highly contact-rich, long-horizon whole-body behaviors in the real world.

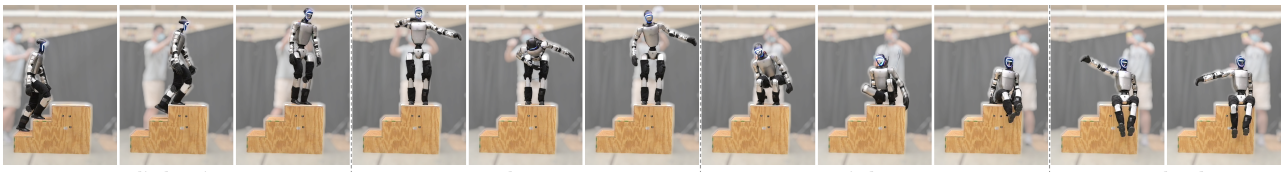


Fig. 5: **Truman's Bow.** This demonstration highlights long and continuous sequence of diverse, contact-rich behaviors.

V. Simulation Ablations

We conduct a set of experiments to analyze the contribution of key components and design choices within our framework. All simulation experiments are conducted in IsaacSim. Each evaluation is initialized at the start of the reference and is considered successful if policy finishes the task without triggering any termination. Mean and standard deviation of metrics are computed over 4096 parallel simulation environments.

A. Interaction Reward

We study the effectiveness of two design choices that guide learning towards desired contact behavior: **interaction reward** and **contact based termination** (table II, *Lost Contact*). We compare three training variants:

- **w/ interaction rew, w/ contact term**: Incorporates the interaction reward and contact-based termination.
- **w/o interaction rew, w/ contact term**: Only removes the interaction reward.
- **w/o interaction rew, w/o contact term**: Removes both.

Unless otherwise specified, contact-based terminations are removed for evaluations in this experiment.

For the majority of tasks, removing these two components does not significantly impact the final performance. However, we identified two specific types of tasks where these components are crucial, and provide a detailed analysis below.

Interaction reward handles imperfect references: Reference motions sometimes fails to precisely establish a grasp due to imperfect retargeting. Interaction reward enables the policy to deviate from an imperfect reference to achieve a proper grasp. Conversely, without it, the agent rigidly follows the flawed reference and fails to interact (fig. 7b top).

To further validate that the reward's importance lies in handling imperfection, we use a successful policy to collect a perfect reference motion to train. As shown in fig. 7a and fig. 7b, training succeeds even without interaction reward, confirming that the interaction reward specifically addresses challenges posed by imperfect reference motions.

Interaction reward guides precise contact locations: Pushing box requires the policy to accurately position its L-shaped end effector on the box's edge. Without interaction reward (fig. 7c right), the policy fails to achieve such precision. It frequently places end effectors on the vertical surface of the box, which results in highly unstable contact.

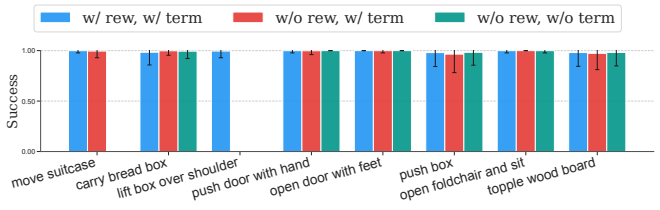
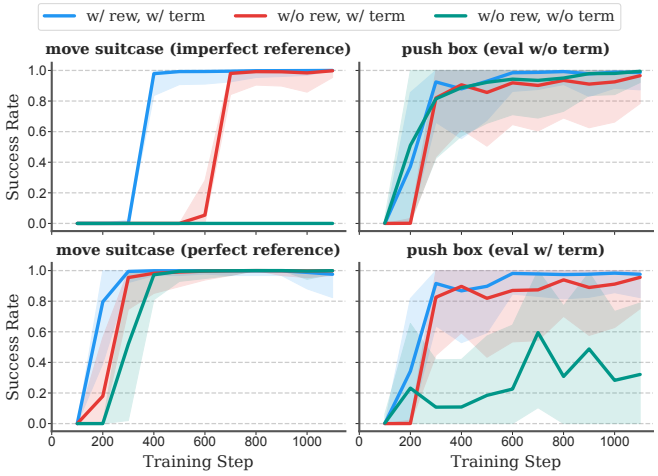
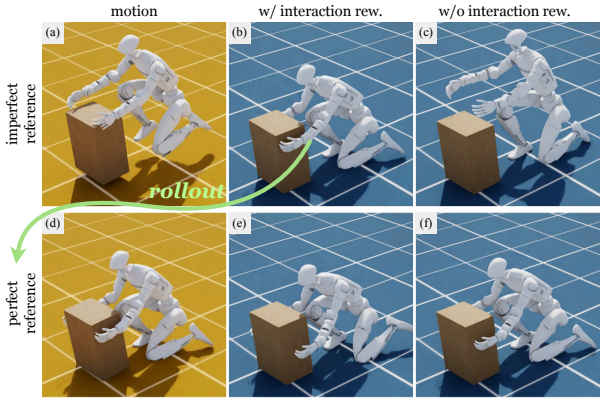


Fig. 6: **Final success rate across 8 tasks.** For most tasks removing contact reward and contact based termination actually does not affect final success rate.

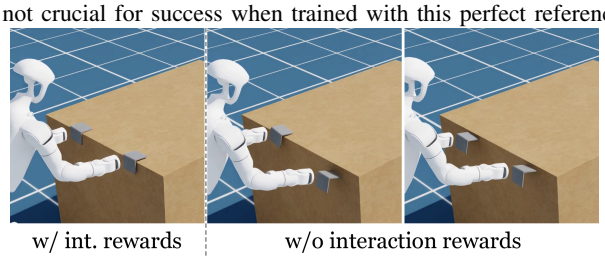
This demonstrates interaction reward is essential for learning policies that can maintain stable and precise contact.



(a) (left) For move suitcase, interaction reward is crucial for task success when reference is imperfect. (right) For push box, while all policies show comparable success rates without contact-based termination (eval w/o term), enabling it causes a significant drop for the policy trained without our interaction reward. This reveals that our reward is essential for learning to maintain a stable contact.



(b) **Interaction reward guides contact establishment with imperfect reference.** With imperfect motion (a), interaction reward drives the policy to deviate from the flawed reference to establish a grasp (b). Without it, the policy rigidly mimics the reference and fails (c). Rollouting (b) we collect a perfect reference (d), interaction reward is not crucial for success when trained with this perfect reference.



(c) **Interaction reward guides precise contact locations.** (left) Interaction reward guides the policy to accurately place its L-shaped end effector on the box's edge. (right) Without it, the policy fail to achieve precise contact, frequently placing the end effectors on the vertical surface of the box, leading to an unstable interaction.

Fig. 7: Ablation study on interaction reward and contact-based termination.

B. *Residual Action Space*

We study the impact of two design choices on exploration and convergence: **residual action space** and **body tracking error based termination** (table II, *Body Local Pose Error*). We compare three variants for training:

- **w/ residual action, w/ track term**: Full method that incorporates both residual action space and body tracking error termination.
- **w/o residual action, w/ track term**: Removes residual action space; policy outputs are defined relative to the default joint position instead of reference joint position.
- **w/o residual action, w/o track term**: Further disables body tracking error termination. Root error termination is retained in training to avoid ineffective samples of the robot lying on the ground.

Body tracking error based termination is removed for evaluations in this experiment.

As shown in fig. 8, the full method (**w/ residual action, w/ track term**) consistently achieves the lowest joint and body tracking errors across 8 tasks. Residual action space allows faster convergence and higher performance: without it (**w/o residual action, w/ track term** and **w/o residual action, w/o track term**), training converges much slower and fails to reach the same level of performance (fig. 9a).

For move suitcase task, without both components (**w/o residual action, w/o track term**) the policy cannot learn the intended kneeling motion fig. 9b. Instead, it converges to a suboptimal strategy of bending at the waist while keeping both feet flat.

This gap arises because residual action space grounds exploration around the reference motion. Without it, actions are defined relative to the default standing pose, thus initial exploration is centered around the standing pose; when an episode is initialized from kneeling, this leads to abrupt "pop up" behaviors (fig. 9c, bottom row), producing low-quality samples. In contrast, residual policies explore locally around the reference pose (top row), enabling stable training, faster convergence, and successful acquisition of challenging skills.

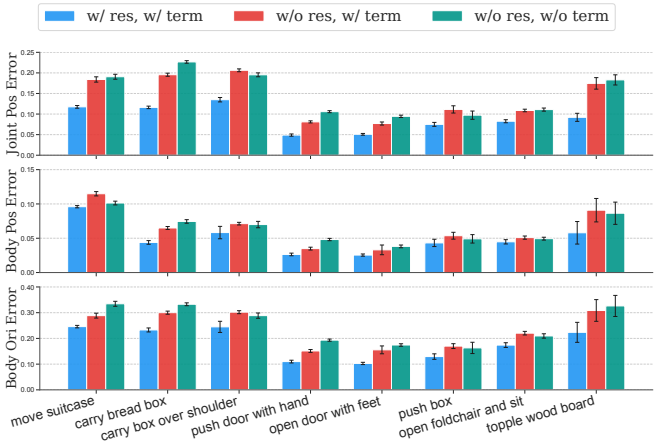
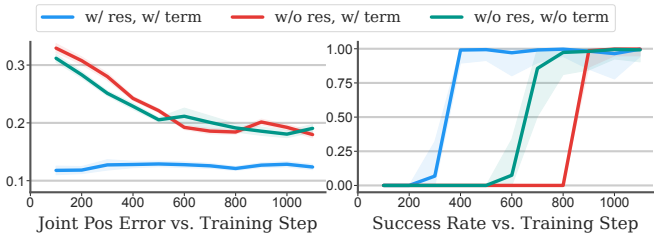
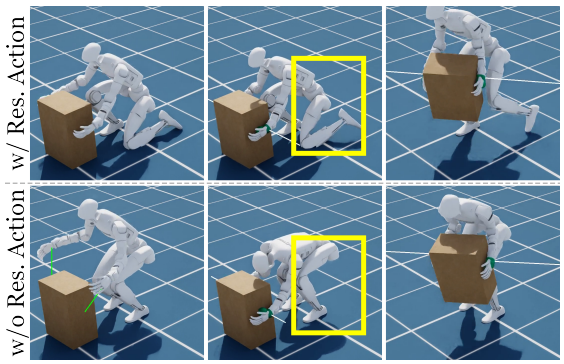


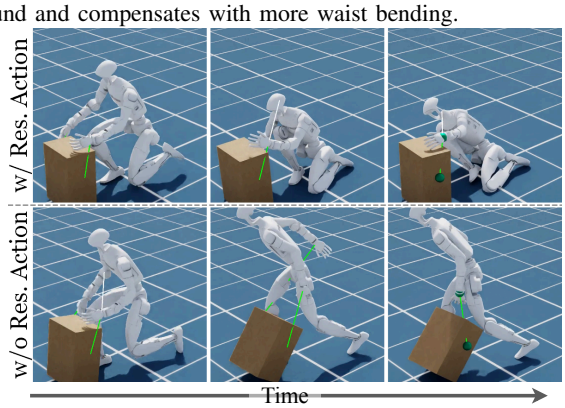
Fig. 8: **Joint and body tracking errors across 8 tasks.** Incorporating a residual action space (**w/ residual action, w/ track term**) consistently achieves the lowest tracking errors. Removing it (**w/o residual action, w/ track term** and **w/o residual action, w/o track term**) overall increases the tracking errors. Early termination based on body tracking error (**w/o residual action, w/ track term**) marginally decreases tracking error.



(a) **Residual actions enable stable and efficient training.** With residual action space, the policy achieves low and stable tracking errors from the beginning, and quickly converges to a high success rate. Policies without it start with higher errors and converge much slower.



(b) **Residual actions enable the learning of challenging poses.** With residual actions (top row), the policy successfully executes a kneeling motion to grasp the suitcase. Without them (bottom row), it converges to a suboptimal strategy: keeping both feet flat on the ground and compensates with more waist bending.



(c) **Residual actions anchor exploration to the reference motion.** When initialized in a kneeling pose, the residual policy (top row) explores around the reference pose. In contrast, the standard policy's exploration is centered on the default pose, causing the robot to abruptly "pop up" and immediately lose balance (bottom row), which generates uninformative training data.

Fig. 9: Ablation study on residual action space and early termination for body tracking errors.

VI. Limitations and Future Directions

Our framework, HDMI, enables humanoids to acquire diverse object interaction skills from human videos. While we have demonstrated its effectiveness across 14 simulated tasks, two key limitations remain:

Dependence on Mocap. The current system relies on ground-truth motion capture data (e.g., object poses). A critical next step is to develop policies that operate directly from on-board sensing modalities, such as cameras, to enable deployment in uninstrumented environments.

One Policy per Skill. At present, a separate specialist policy is trained for each task. An important future direction is to leverage the data from multiple skills to train a unified generalist model, capable of performing a wide range of interactions.

VII. Acknowledgement

We would like to thank Guanqi He for constructing the door setup. We are also grateful to Haotian Lin, Chaoyi Pan, Yuanhang Zhang, and Wenli Xiao for their valuable discussions. Guanya Shi holds concurrent appointments as an Assistant Professor at Carnegie Mellon University and as an Amazon Scholar. This paper describes work performed at Carnegie Mellon University and is not associated with Amazon.

References

[1] T. He et al., *Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills*, 2025. arXiv: 2502.01143 [cs.RO].

[2] Q. Liao, T. E. Truong, X. Huang, G. Tevet, K. Sreenath, and C. K. Liu, *Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion*, 2025. arXiv: 2508.08241 [cs.RO].

[3] M. Ji et al., "Exbody2: Advanced expressive humanoid whole-body control," *arXiv preprint arXiv:2412.13196*, 2024.

[4] T. Z. Zhao et al., "Aloha unleashed: A simple recipe for robot dexterity," *arXiv preprint arXiv:2410.13126*, 2024.

[5] Octo Model Team et al., "Octo: An open-source generalist robot policy," in *Proceedings of Robotics: Science and Systems*, 2024.

[6] L. Wang, X. Chen, J. Zhao, and K. He, "Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers," *arXiv preprint arXiv:2409.20537*, 2024.

[7] F. Liu et al., *Opt2skill: Imitating dynamically-feasible whole-body trajectories for versatile humanoid loco-manipulation*, 2025. arXiv: 2409.20514 [cs.RO].

[8] J. Dao, H. Duan, and A. Fern, *Sim-to-real learning for humanoid box loco-manipulation*, 2023. arXiv: 2310.03191 [cs.RO].

[9] R.-Z. Qiu et al., "Wildlma: Long horizon loco-manipulation in the wild," *arXiv preprint arXiv:2411.15131*, 2024.

[10] Z. Su et al., *Hitter: A humanoid table tennis robot via hierarchical planning and learning*, 2025. arXiv: 2508.21043 [cs.RO].

[11] Z. Shen et al., "World-grounded human motion recovery via gravity-view coordinates," in *SIGGRAPH Asia 2024 Conference Papers*, 2024.

[12] F. Al-Hafez, G. Zhao, J. Peters, and D. Tateo, "Loco-mujoco: A comprehensive imitation learning benchmark for locomotion," in *6th Robot Learning Workshop, NeurIPS*, 2023.

[13] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control," *The International Journal of Robotics Research*, p. 02 783 649 241 285 161, 2024.

[14] X. Gu et al., "Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning," *arXiv preprint arXiv:2408.14472*, 2024.

[15] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Robust and versatile bipedal jumping control through reinforcement learning," *arXiv preprint arXiv:2302.09450*, 2023.

[16] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," *arXiv preprint arXiv:2406.10759*, 2024.

[17] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," *arXiv preprint arXiv:2402.16796*, 2024.

[18] S. P. Arunachalam, S. Silwal, B. Evans, and L. Pinto, "Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation," in *2023 ieee international conference on robotics and automation (icra)*, IEEE, 2023, pp. 5954–5961.

[19] C. Chi et al., "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," *arXiv preprint arXiv:2402.10329*, 2024.

[20] R.-Z. Qiu et al., *Humanoid policy human policy*, 2025. arXiv: 2503.13441 [cs.RO].

[21] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang, *Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit*, 2025. arXiv: 2502.13013 [cs.RO].

[22] T. He et al., "Learning human-to-humanoid real-time whole-body teleoperation," *arXiv preprint arXiv:2403.04436*, 2024.

[23] T. He et al., "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," *arXiv preprint arXiv:2406.08858*, 2024.

[24] Y. Kuang et al., *Skillblender: Towards versatile humanoid whole-body loco-manipulation via skill blending*, 2025. arXiv: 2506.09366 [cs.RO].

[25] Y. Zhang et al., *Falcon: Learning force-adaptive humanoid loco-manipulation*, 2025. arXiv: 2505.06776 [cs.RO].

[26] Z. Ding et al., *Jaeger: Dual-level humanoid whole-body controller*, 2025. arXiv: 2505.06584 [cs.RO].

[27] G. C. R. Bethala et al., *H2-compact: Human-humanoid co-manipulation via adaptive contact trajectory policies*, 2025. arXiv: 2505.17627 [cs.RO].

[28] J. Mao et al., *Learning from massive human videos for universal humanoid pose control*, 2024. arXiv: 2412.14172 [cs.RO].

[29] J. Wang et al., *Hil: Hybrid imitation learning of diverse parkour skills from videos*, 2025. arXiv: 2505.12619 [cs.GR].

[30] A. Allshire et al., "Visual imitation enables contextual humanoid control," *arXiv preprint arXiv:2505.03729*, 2025.

[31] J. Li et al., "Okami: Teaching humanoid robots manipulation skills through single video imitation," in *8th Annual Conference on Robot Learning (CoRL)*, 2024.

[32] Y. Zhu, A. Lim, P. Stone, and Y. Zhu, "Vision-based manipulation from single human video with open-world object graphs," *arXiv preprint arXiv:2405.20321*, 2024.

[33] H. Zhou, R. Wang, Y. Tai, Y. Deng, G. Liu, and K. Jia, *You only teach once: Learn one-shot bimanual robotic manipulation from video demonstrations*, 2025. arXiv: 2501.14208 [cs.RO].

[34] X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Transactions On Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.

[35] J. Li, J. Wu, and C. K. Liu, "Object motion guided human motion synthesis," *ACM Trans. Graph.*, vol. 42, no. 6, 2023.