WEAVE: Learning Whole-Body Dexterous Loco-Manipulation from Human–Object Interactions

Liu Cao[1,†], Xingze Wu[3,†], Jingzhi Cui[1], Botian Xu[4], Mingzhi Pei[2], Ruoqu Chen[1,†], Mengdi Xu[1]
[1]Tsinghua IIIS [2]Tsinghua College AI
[3]Dalian University of Technology [4]The Chinese University of Hong Kong

***Abstract*—Learning humanoid–object interaction requires coordinating whole-body balance, locomotion, and dexterous hand contact to control both robot and object motion. Human demonstrations provide examples of coordinated interaction, but transferring these behaviors to humanoid robots requires learning how to establish and maintain effective contacts under different embodiments and dynamics. We present WEAVE, a unified framework for learning whole-body dexterous humanoid–object interaction from captured human demonstrations. WEAVE first converts captured human–object interactions into executable robot–object references through contact-aware retargeting and approach-motion completion. At its core is a contact- and geometry-aware policy that jointly commands 29 body joints and 12 actuated finger joints across multiple objects and interaction sequences. Evaluation across nine objects yields a 92.5% success rate on trained interactions and, without any additional training, 65.0% on sequences never seen during training. We additionally release $\sim 9{,}000$ physically executed rollouts spanning $\sim$23 hours, providing robot–object trajectories with contact annotations for downstream interaction-policy learning and physically consistent HOI motion generation. Our project website: https://xiaohu-art.github.io/Weave/.**

I. INTRODUCTION

Humanoid robots are expected to operate in environments built for humans, where many everyday tasks require the robot to manipulate objects—approaching, grasping, and transporting them while maintaining balance. Such contact-rich loco-manipulation is a core capability for general-purpose humanoids. Even a task as simple as moving a chair requires the robot to establish a stable multi-finger grasp, generate sufficient contact forces, and continuously adapt its whole-body posture to the resulting load. However, learning such behaviors with a unified policy remains challenging, as balance, locomotion, object motion, and dexterous grasping are tightly coupled: an inaccurate grasp alters the forces acting on the upper body, while poor whole-body coordination can destabilize both the robot and the object.

A primary challenge is **learning contact-rich robot-object interactions from human demonstrations**. Reference motion tracking [1]–[5] offers a route from demonstrations to physically feasible behavior, but supervises robot and object motion at the body and arm level and leaves finger-level hand–object contact unmodeled; dexterous interaction learning [6]–[11] does model such contact, yet targets simulated characters or fixed-base hands whose kinematics and actuation differ from a humanoid with underactuated hands. Human demonstrations describe coordinated body and object motion, however, differences in morphology and actuation change how

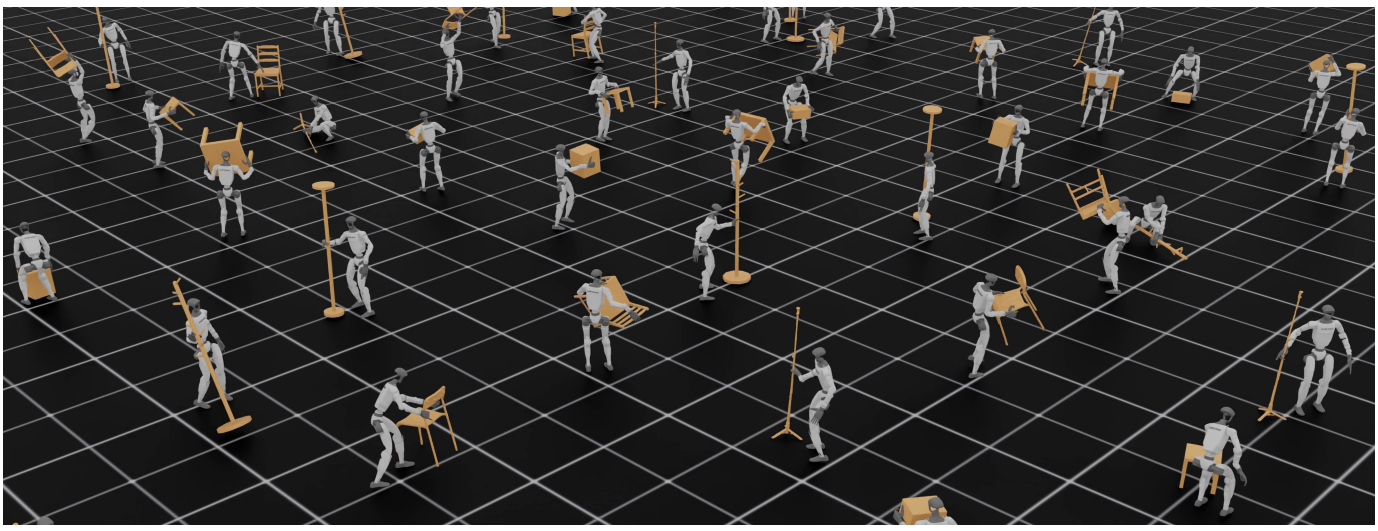

Figure 1. Policy rollout examples with **WEAVE.** We learn a unified policy for whole-body dexterous loco-manipulation from captured human–object interaction. Parallel simulation rollouts showcase the one policy coordinating locomotion and dexterous manipulation across diverse everyday objects.

†Work partially done at CocoMatrix.

the humanoid must establish hand–object contact, support the object, and maintain balance. The resulting control problem is also high-dimensional and heterogeneous: body joints govern balance and load bearing on timescales set by locomotion, while actuated finger joints govern contacts whose outcome turns on millimeter-scale placement, and both must be commanded by one policy. Even for a single interaction, matching demonstrated poses does not ensure that the robot can sustain the force and torque required to move the object. Learning therefore requires coupling robot–object motion tracking with the acquisition of contact-aware, closed-loop control under the robot's dynamics.

A secondary challenge is **learning diverse object interactions within a unified policy**. Across objects, differences in geometry affect the feasible contact locations and grasp configurations. Across interaction sequences, variations in approach direction, body posture, and object motion require different control behaviors, even for the same object. Prior dexterous interaction methods largely train one policy per object or per clip [9], which sidesteps this variation. The challenge is to learn a shared control strategy that captures common coordination patterns across interactions while adapting whole-body motion and finger-level contact to the specific object and reference sequence.

In this work, we investigate how to equip humanoids with whole-body dexterous interaction capability from captured human demonstrations. Our key idea is to make object geometry and hand–object contact explicit throughout the pipeline: references are constructed to preserve task-relevant contacts across the morphology gap, and a unified policy, conditioned on object geometry and trained jointly across objects and interaction clips, tracks robot and object motion while explicitly learning whole-body coordination and finger-level contact.

We present WEAVE, an end-to-end framework for learning whole-body dexterous loco-manipulation from human–object interaction. Given captured human–object interactions, which records full-body human motion while manipulating everyday objects such as chairs, boxes, and tables, WEAVE completes the missing approach and transition motions with whole-body motion generation [12], retargets the interaction to the robot with contact-aware retargeting, and trains the contact- and geometry-aware policy in simulation. As a result, the robot walks up to an object, grasps it with multi-fingered hands, lifts it, and carries it to a target configuration. We evaluate the learned policy and publicly release a curated dataset of simulation rollouts to support research on whole-body dexterous interaction and learning from robot demonstrations.

In summary, our contributions are as follows:

- **Contact- and geometry-aware whole-body dexterous policy.** A unified policy, trained jointly across multiple objects and interaction clips, that tracks robot and object motion while explicitly learning hand–object contact, reproducing whole-body coordination, object transport, and finger-level grasping within one policy.
- **Interaction-preserving trajectory construction.** A pipeline that converts captured human–object interactions into executable robot–object trajectories through approach-motion completion and contact-aware retargeting, preserving object motion and hand–object contacts across the human–robot morphology gap.
- **Systematic evaluation and physically plausible dataset.** We conduct extensive experiments to assess interaction execution, quantify the benefits of joint multi-object learning, and examine how policy architecture and optimization affect learning efficiency, together with the resulting rollout dataset released for downstream policy learning and humanoid-object interaction modeling.

II. RELATED WORKS

A. Interaction-Preserving Retargeting and Motion Generation

Motion retargeting constructs reference motions from human demonstrations, conventionally by optimizing joint configurations to match Cartesian keypoints [13]–[15]. For human–object interaction, however, pose similarity alone is insufficient: embodiment differences can alter hand–object contacts, break the relative interaction geometry, or produce configurations that cannot exert the intended effect on the object. Interaction-preserving retargeting therefore incorporates contact constraints: OmniRetarget [3] preserves whole-body interaction structure for humanoids, TopoRetarget [16] and REGRIND [17] encode local hand–object topology for dexterous hands, and other works [18]–[21] further enforce force consistency of the retargeted contacts. In parallel, human motion generation has progressed from parametric motion synthesis [22] to more scalable and controllable models: Kimodo [12] handles heterogeneous spatial and temporal constraints for both human and humanoid embodiments, and ARDY [23] extends to autoregressive diffusion for streaming generation. WEAVE combines contact-aware retargeting with Kimodo-based approach-motion completion to construct robot–object reference trajectories. These kinematic references then supervise reinforcement learning, which converts them into physically executable interactions in simulation.

B. Humanoid Loco-Manipulation

Humanoid loco-manipulation requires coordinating locomotion, balance, and object interaction over extended task horizons. Modular approaches organize navigation, locomotion, reaching, and manipulation into separate components [24]–[27]. Such decomposition provides structured interfaces, while coordinating balance and hand–object contact across these interfaces remains an important consideration. Reference-tracking approaches instead learn a physics-based policy that tracks diverse humanoid motions and object interactions [1]–[5], [28]. Physics-based grasping and human–object interaction methods [6], [8]–[11] highlight the importance of coupling motion tracking with hand–object contact, motivating unified control across objects and interaction sequences. A parallel line of work introduces egocentric depth or RGB observations for visually guided loco-manipulation [29]–[33]. Data-generation approaches further diversify humanoid interactions: GRAIL [34] synthesizes 4D humanoid–object interactions

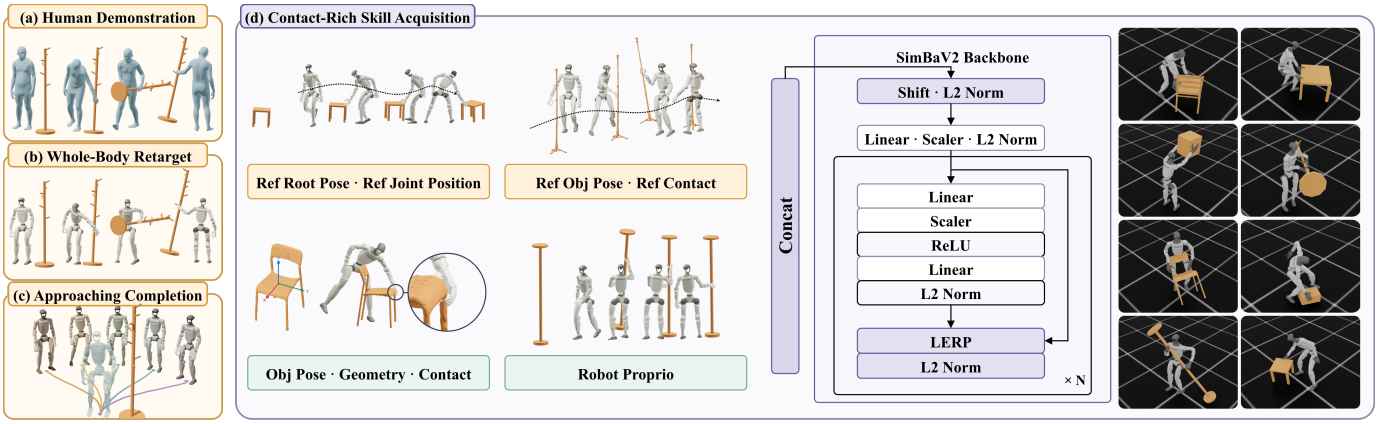

Figure 2. **Overview of WEAVE. (a)** Captured human–object interactions provide paired SMPL-X body motion and object trajectories. **(b)** Whole-body inverse kinematics retargets each interaction while preserving the hand–object contacts and the object motion. **(c)** Kimodo [12] synthesizes locomotion prefixes that approach the initial interaction pose from different azimuths, producing several complete references from one captured sequence. **(d)** A unified reference-conditioned policy is trained in simulation across all objects and interaction clips.

with a video generation model [35], HumanoidMimicGen [36] augments a few teleoperated demonstrations through whole-body planning, and VLK [37] synthesizes paired vision–language–kinematics data in reconstructed scenes. Relative to this literature, WEAVE supervises finger-level contact rather than body- and arm-level motion alone, learns a single policy across objects and interaction clips rather than per-object controllers, and releases the executed rollouts as reusable interaction data.

C. Humanoid Vision–Language–Action Models

VLA models connect semantic task descriptions and visual observations with executable robot actions. General-purpose systems [38]–[41] leverage large-scale robot datasets and pretrained VLM representations to generalize across scenes and tasks. Most of these models were initially developed for fixed-base manipulators, whose action spaces and stability requirements differ substantially from those of humanoid robots.

Extending VLA models to humanoids requires reasoning over high-dimensional whole-body motion while maintaining balance and physical contact. Recent humanoid-specific systems [42]–[45], explore unified or hierarchical architectures that connect semantic reasoning with whole-body motor control. Nevertheless, acquiring large-scale data that jointly contain locomotion, precise object interaction, and articulated finger motion remains substantially more difficult than collecting arm-level manipulation trajectories.

Therefore, WEAVE is complementary rather than competing: instead of collecting interactions through teleoperation or generation, we acquire the skills through physics-based reinforcement learning on human-object motion, producing grounded, finger-level behavior. Such physically grounded skills can serve as reusable low-level capabilities for future humanoid VLA systems, while VLA models can provide the semantic task specifications and high-level motion commands needed to select and compose them.

III. PROBLEM FORMULATION

We study whole-body dexterous loco-manipulation, in which a humanoid equipped with dexterous hands must approach an object, establish a stable grasp, and transport the object toward a desired configuration while maintaining whole-body balance. We formulate the closed-loop control problem as a partially observable Markov decision process (POMDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{O}$ denote the state and observation spaces, $\mathcal{A}$ is the action space, $\mathcal{T}$ is the state-transition function, r is the reward function, and γ is the discount factor. At each time step t, the policy outputs an action $a_t \sim \pi_\theta(\cdot \mid o_t)$, where a_t specifies desired joint positions for low-level PD controller.

Policy learning is bootstrapped from captured human–object interaction data [46]. After retargeting and motion completion, each sequence provides a robot–object reference trajectory

$$\tau^{\mathrm{ref}} = \left\{\hat{s}_t^{\mathrm{ref}}\right\}_{t=0}^{T-1}, \qquad \hat{s}_t^{\mathrm{ref}} = \left(\hat{x}_t^{\mathrm{robot}}, \hat{x}_t^{\mathrm{obj}}\right), \tag{1}$$

where $\hat{x}_t^{\mathrm{robot}}$ specifies the reference humanoid configuration and $\hat{x}_t^{\mathrm{obj}}$ specifies the corresponding object pose. The reference trajectory is used to initialize state in simulation and construct the tracking objectives during policy training.

IV. METHOD

A. Overview

WEAVE transforms captured human–object interactions into whole-body dexterous loco-manipulation skills through two main stages. First, we construct reference trajectories using whole-body retargeting, contact-aware hand refinement, and approach-motion completion. Then we train a unified contact- and geometry-aware policy through reference-tracking, jointly coordinating the humanoid body, articulated fingers, and object motion across multiple objects and interaction clips.

TABLE I

REWARD TERMS FOR REFERENCE TRACKING. HATTED QUANTITIES DENOTE REFERENCES. TRACKING TERMS CONTRIBUTE $w_i \exp(-e_{t,i}/\sigma_i^2)$.

Reward term	Computation	Definition	Weight w	Scale σ
Robot and object tracking				
Pelvis position	$\|p_t^r - \hat{p}_t^r\|_2^2$	$p_t^r, \hat{p}_t^r$: current/reference pelvis position	1.0	0.3
Pelvis orientation	$\angle(R_t^r, \hat{R}_t^r)^2$	$R_t^r, \hat{R}_t^r$: pelvis orientation; $\angle$: geodesic error	1.0	0.4
Body positions	$\frac{1}{\lvert\mathcal{B}\rvert}\sum_{b\in\mathcal{B}}\|p_{t,b} - \hat{p}_{t,b}\|_2^2$	$\mathcal{B}$: tracked robot links	1.0	0.3
Body orientations	$\frac{1}{\lvert\mathcal{B}\rvert}\sum_{b\in\mathcal{B}}\angle(R_{t,b}, \hat{R}_{t,b})^2$	$R_{t,b}, \hat{R}_{t,b}$: current/reference link orientation	1.0	0.4
Object position	$\|p_t^o - \hat{p}_t^o\|_2^2$	$p_t^o, \hat{p}_t^o$: current/reference object position	2.0	0.2
Object orientation	$\angle(R_t^o, \hat{R}_t^o)^2$	$R_t^o, \hat{R}_t^o$: current/reference object orientation	2.0	0.3
Grasp and contact				
Hand opposition	$\frac{\sum_h \eta_{t,h}\left(\frac{1}{\lvert\mathcal{K}_h\rvert}\sum_{k\in\mathcal{K}_h}\frac{1-u_{t,h,0}^\top u_{t,h,k}}{2}\right)}{\sum_h \eta_{t,h}+\epsilon}$	$u_{t,h,k}$: surface-to-fingertip unit vector; $k=0$: thumb; $\eta_{t,h}$: expected-contact gate	2.0	–
Contact matching	$\frac{\sum_{h\in\mathcal{H}} m_{t,h}(1-\lvert y_{t,h}-\tilde{c}_{t,h}\rvert)}{\sum_{h\in\mathcal{H}} m_{t,h}+\epsilon}$	$m_{t,h}=\mathbf{1}[\hat{c}_{t,h}\neq 0]$; $y_{t,h}=\mathbf{1}[\hat{c}_{t,h}>0]$; $\tilde{c}_{t,h}=\min(\|f_{t,h}\|/\bar{f},1)$	2.0	–
Regularization				
Foot sliding	$\sum_{f\in\mathcal{B}_{\text{foot}}}\chi_{t,f}^{\text{ground}}\|v_{t,f}^{xy}\|_2$	$\chi_{t,f}^{\text{ground}}$: foot–ground contact indicator	-1.0	–
Action rate	$\|a_t - a_{t-1}\|_2^2$	a_t, a_{t-1}: current/previous joint command	-0.1	–
Joint limits	$\sum_{j\in\mathcal{J}_{\text{body}}}\left([q_{t,j}-\overline{q}_j]_+ + [\underline{q}_j - q_{t,j}]_+\right)$	$\mathcal{J}_{\text{body}}$: non-finger joints; $[\underline{q}_j, \overline{q}_j]$: soft joint limits	-10.0	–

B. Interaction-Preserving Reference Construction

Before policy learning, we convert captured human–object interaction sequences into reference motions. Starting from the paired SMPL-X motion and object trajectory, our pipeline first retargets the interaction segment to the humanoid through whole-body inverse kinematics and contact-aware hand refinement. We then use Kimodo [12] to synthesize a locomotion prefix that approaches the initial interaction pose. This process produces a synchronized robot–object trajectory spanning approach, grasping, and object transport, together with contact annotations used for policy training.

Whole-body retargeting. We first obtain an initial robot trajectory $\{\tilde{x}_t^{\text{robot}}\}_{t=0}^{T-1}$ using a GMR-style whole-body inverse-kinematics objective [47], which aligns selected robot links with their SMPL-X counterparts. To account for morphological differences, we constrain the pelvis only in the horizontal plane and optimize its height based on ground contact in place of scale calibration, while joint-limit, velocity, and acceleration penalties promote feasible and temporally smooth motion.

Contact-aware hand refinement. Whole-body IK captures the arm and wrist motion but does not ensure a stable fingertip grasp. We therefore jointly refine the arm and finger configurations of both limbs, $\xi_t = (q_t^{\text{arm}}, q_t^{\text{hand}})$, initialized from $\{\tilde{x}_t^{\text{robot}}\}_{t=0}^{T-1}$ and held fixed elsewhere: the pelvis pose and the leg configuration remain at their IK solution, so the refinement does not alter the retargeted whole-body motion. For fingertip k at frame t, let $p_{t,k}(\xi_t)$ be its forward-kinematics position, $\bar{p}_{t,k}$ the object-surface point nearest to the corresponding human fingertip, $n_{t,k}$ its outward normal, and $c_{t,k} \in \{0,1\}$ the contact label. Following [19], we minimize

$$\mathcal{L}_{\text{hand}} = \underbrace{\lambda_a \sum_{t,k} c_{t,k}\|p_{t,k} - \bar{p}_{t,k}\|}_{\text{contact attraction}} - \underbrace{\lambda_q \sum_t \hat{Q}_{\text{FC}}(q_t^h)}_{\text{force-closure objective}} + \underbrace{\lambda_p \sum_{t,k}\left[-(p_{t,k}-\bar{p}_{t,k})^\top n_{t,k}\right]_+}_{\text{penetration penalty}} + \mathcal{L}_{\text{reg}}. \tag{2}$$

where $[z]_+ = \max(z, 0)$. The attraction term alone only places fingertips on the object surface, which admits configurations that touch the object without being able to hold it. We therefore additionally require the contacts to resist arbitrary disturbance wrenches, and maximize a differentiable approximation $\hat{Q}_{\text{FC}}$ of the force-closure quality [48],

$$Q_{\text{FC}}(\xi_t) = \min_{\|w\|_2=1} \max_{\substack{f\in\mathcal{F}(\xi_t)\\ \|f\|_1\le 1}} w^\top G(\xi_t) f. \tag{3}$$

Here, G maps feasible contact forces f to object-centric wrenches, and $\mathcal{F}$ is the linearized friction cone with soft-finger torsion. Intuitively, Q_{FC} measures the weakest disturbance

wrench that the grasp can resist; $Q_{\mathrm{FC}} > 0$ indicates force closure. We approximate it by a fixed set of wrench directions.

Approach motion completion. The retargeted interaction begins near the object and therefore lacks the locomotion required to approach it. We use Kimodo [12] to generate an approach prefix τ^{pre} for the refined interaction trajectory τ^{int}. The initial frames of τ^{int} provide terminal whole-body constraints, while the robot follows a path ending at the initial pose and heading. We concatenate the two segments:

$$\tau^{\mathrm{ref}} = \tau^{\mathrm{pre}} \oplus \tau^{\mathrm{int}}, \qquad T = T_{\mathrm{pre}} + T_{\mathrm{int}}. \tag{4}$$

Sampling different approach azimuths produces multiple complete references from the same interaction sequence. We sample $K = 3$ for training set and $K = 5$ for test set, which is the main source of the trajectory counts reported in V-A.

Reference annotations. After concatenation, the completed reference is re-indexed as $\tau^{\mathrm{ref}} = \{\hat{s}_t^{\mathrm{ref}}\}_{t=0}^{T-1}$, where

$$\hat{s}_t^{\mathrm{ref}} = \left(\hat{x}_t^{\mathrm{robot}}, \hat{x}_t^{\mathrm{obj}}\right) = \left(\hat{p}_t^r, \hat{R}_t^r, \hat{q}_t, \hat{p}_t^o, \hat{R}_t^o\right). \tag{5}$$

Here, $\hat{p}_t^r$, $\hat{R}_t^r$, and $\hat{q}_t$ denote the humanoid root position, root orientation, and joint configuration, while $\hat{p}_t^o$ and $\hat{R}_t^o$ denote the object pose. For each robot link ℓ, we additionally assign

$$\hat{c}_{t,\ell} = \mathbf{1}\left[d_{t,\ell}^o < \delta_{\mathrm{contact}}\right] - \mathbf{1}\left[d_{t,\ell}^o > \delta_{\mathrm{far}}\right], \tag{6}$$

where $d_{t,\ell}^o$ is the distance from the link position to the object surface. Thus, $\hat{c}_{t,\ell} \in \{-1, 0, 1\}$ denotes separated, neutral, and contact states, respectively. The completed trajectories provide initialization states and reference motion for policy learning, while the contact annotations specify the desired interaction pattern. Together, they guide a policy that learns to realize the reference interactions under the simulator dynamics.

C. Contact-Rich Skill Acquisition via Reference Tracking

We train a unified reference-conditioned policy in simulation, jointly across multiple objects and interaction clips. The policy uses robot proprioception, object state and geometry features, and reference motion to coordinate whole-body control with finger-level interaction:

$$a_t \sim \pi_{\mathrm{track}}\left(\cdot \mid o_t^{\mathrm{prop}}, o_t^{\mathrm{obj}}, \hat{x}_{t:t+H}^{\mathrm{robot}}, \hat{x}_{t:t+H}^{\mathrm{obj}},\right), \tag{7}$$

where o_t^{prop} is the robot proprioception, o_t^{obj} is the simulated object observation, and $\{\hat{x}_{t:t+H}^{\mathrm{robot}}, \hat{x}_{t:t+H}^{\mathrm{obj}}\}$ is a short-horizon command extracted from τ^{ref}. We optimize the policy with PPO [49], using reference state initialization and reference-based early termination following [50].

Observation and action. All relative poses and geometric features are expressed in robot local frame. The proprioceptive observation contains the base angular velocity, projected gravity, joint positions and velocities, and the previous action. The object observation contains its ground-truth relative pose, fingertip-to-surface vectors, binary object-contact flags, and BPS-SDF descriptor encoding object geometry. The reference command $\{\hat{x}_{t:t+H}^{\mathrm{robot}}, \hat{x}_{t:t+H}^{\mathrm{obj}}\}$ contains the joint configuration, pelvis pose, object pose, and contact labels of a short horizon.

We use an asymmetric actor–critic: the actor receives $(o_t^{\mathrm{prop}}, o_t^{\mathrm{obj}}, \hat{x}_{t:t+H}^{\mathrm{robot}}, \hat{x}_{t:t+H}^{\mathrm{obj}})$, whereas the critic replaces o_t^{prop} with a privileged robot observation that additionally contains the base linear velocity and current tracked-body poses. The action a_t specifies joint-position targets tracked by PD controllers at 50 Hz. For the underactuated Inspire hands, the policy commands only the proximal finger joints, while the intermediate and distal joints follow fixed mimic couplings.

Tracking objective. The reward combines robot and object tracking, geometry-aware grasping, contact matching, and motion regularization:

$$\begin{aligned} r_t = &\sum_{i \in \mathcal{I}} w_i \exp\left(-e_{t,i}/\sigma_i^2\right) + w_{\mathrm{opp}} r_t^{\mathrm{opp}} + w_{\mathrm{c}} r_t^{\mathrm{contact}} \\ &+ w_{\mathrm{slide}} r_t^{\mathrm{slide}} + w_{\mathrm{act}} r_t^{\mathrm{act}} + w_{\mathrm{lim}} r_t^{\mathrm{lim}}, \end{aligned} \tag{8}$$

where $\mathcal{I}$ includes pelvis pose, tracked-body pose over the body set $\mathcal{B}$, and object pose. The hand-opposition reward r_t^{opp} encourages the thumb and opposing fingers to lie on different sides of the object surface whenever the reference specifies a grasp following [10]. For hand link $h \in \mathcal{H}$, let $\hat{c}_{t,h}$ be the reference contact label and $\tilde{c}_{t,h} = \min(\|f_{t,h}\|/\bar{f}, 1)$ the normalized contact-force magnitude. Contact matching is computed only for non-neutral labels:

$$r_t^{\mathrm{contact}} = \frac{\sum_{h \in \mathcal{H}} m_{t,h}\left(1 - |y_{t,h} - \tilde{c}_{t,h}|\right)}{\sum_{h \in \mathcal{H}} m_{t,h} + \epsilon}, \tag{9}$$

where $m_{t,h} = \mathbf{1}[\hat{c}_{t,h} \neq 0]$ and $y_{t,h} = \mathbf{1}[\hat{c}_{t,h} > 0]$. The remaining terms penalize foot sliding, action variation, and non-finger joint-limit violations as showed in Table I.

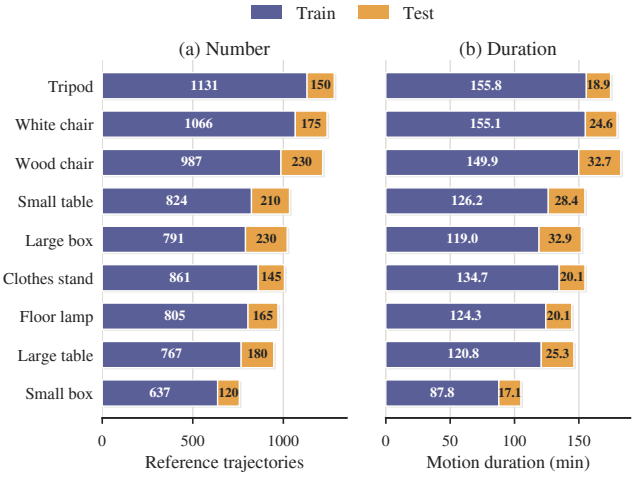

Figure 3. **Reference-motion dataset distribution.** Per-object composition of the training and test splits in terms of (a) the number of reference trajectories and (b) total motion duration.

Randomization and termination. Across parallel environments, we randomize robot and object friction and restitution, torso center of mass, and finger actuator properties. Episodes terminate when the pelvis-position error exceeds 0.25 m, the object-position error exceeds 0.30 m, the projected-gravity

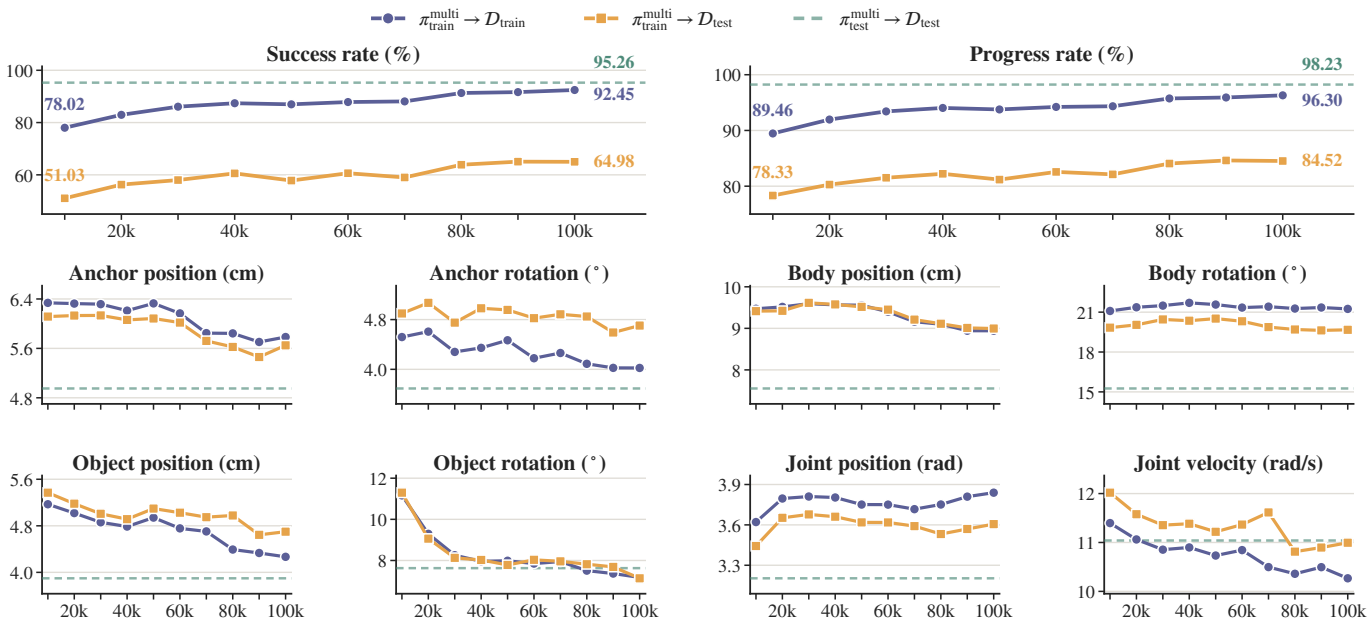

Figure 4. **Evaluation against training iterations.** Success rate (top) and tracking errors (bottom) for $\pi^{\text{multi}}_{\text{train}}$ evaluated on $\mathcal{D}_{\text{train}}$ and on $\mathcal{D}_{\text{test}}$. The dashed line is $\pi^{\text{multi}}_{\text{test}}$ evaluated on $\mathcal{D}_{\text{test}}$, which serves as an oracle for measuring generalization.

error of the pelvis or object exceeds 0.8 or 0.3, respectively, or the vertical tracking error of an ankle or wrist exceeds $0.25\,\text{m}$. We also terminate after all expected hand–object contacts are absent for ten consecutive control steps. These conditions keep the on-policy state distribution close to the reference [9], [51].

Policy optimization. Each observation group is encoded and projected to latent representation. The concatenated features are processed by SimBaV2 networks [52]. Two-dimensional weight matrices are optimized with Muon [53], while biases and other parameters use AdamW. We train a unified policy jointly over all objects and reference clips under a $200\,\text{Hz}$ simulation frequency and a $50\,\text{Hz}$ control frequency.

V. Experiments

We evaluate WEAVE to answer the following questions:

Q1: How effectively does the learned policy execute whole-body dexterous loco-manipulation across objects and interaction sequences?

Q2: How does joint multi-object training compare with single-object specialist in interaction completion and tracking accuracy?

Q3: How do the policy architecture and optimizer affect learning efficiency?

A. Experimental Setup

Embodiment and dataset. We instantiate WEAVE on Unitree G1 humanoid with 29 actuated body DoFs and two Inspire dexterous hands with 12 actuated finger DoFs. We construct two disjoint reference splits, $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, containing 7,869 (19.56h) and 1,605 (3.67h) reference trajectories respectively as showed in Figure 3.

Evaluation metrics. For N evaluation clips, let T_i denote the number of executed frames in clip i. We define the clip-balanced average $\langle z_{i,t} \rangle = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=0}^{T_i-1} z_{i,t}$. The tracking errors are then defined as

$$
\begin{aligned}
E^{p}_{\text{anchor}} &= \left\langle \|p^{a}_{i,t} - \hat{p}^{a}_{i,t}\|_2 \right\rangle, \quad E^{R}_{\text{anchor}} = \left\langle \angle(R^{a}_{i,t}, \hat{R}^{a}_{i,t}) \right\rangle, \\
E^{p}_{\text{body}} &= \left\langle \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} \|p_{i,t,b} - \hat{p}_{i,t,b}\|_2 \right\rangle, \\
E^{R}_{\text{body}} &= \left\langle \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} \angle(R_{i,t,b}, \hat{R}_{i,t,b}) \right\rangle.
\end{aligned}
\tag{10}
$$

For joint and object tracking, we report

$$
\begin{aligned}
E^{q}_{\text{joint}} &= \left\langle \|q_{i,t} - \hat{q}_{i,t}\|_2 \right\rangle, \quad E^{\dot{q}}_{\text{joint}} = \left\langle \|\dot{q}_{i,t} - \hat{\dot{q}}_{i,t}\|_2 \right\rangle, \\
E^{p}_{\text{obj}} &= \left\langle \|p^{o}_{i,t} - \hat{p}^{o}_{i,t}\|_2 \right\rangle, \quad E^{R}_{\text{obj}} = \left\langle \angle(R^{o}_{i,t}, \hat{R}^{o}_{i,t}) \right\rangle.
\end{aligned}
\tag{11}
$$

Here, a denotes the pelvis anchor, $\mathcal{B}$ is the set of tracked robot bodies, and $\angle(\cdot,\cdot)$ denotes the geodesic rotation error.

B. Whole-Body Dexterous Loco-Manipulation

Protocol. We train a multi-object tracking policy $\pi^{\text{multi}}_{\text{train}}$ on $\mathcal{D}_{\text{train}}$ and evaluate it on both $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ to measure generalization to unseen interaction. We additionally train $\pi^{\text{multi}}_{\text{test}}$ directly on $\mathcal{D}_{\text{test}}$ to serve as a comparison. The three evaluations are complementary. The first measures how well a single policy fits the references it is trained on, the second measures transfer to unseen interaction sequences of the same objects, and the third serves as an oracle.

Results. Figure 4 reports both evaluations against training iterations. Training on $\mathcal{D}_{\text{train}}$ progresses steadily. The success

rate rises from 78.0% at 10k iterations to 92.5% at 100k and the progress rate from 89.5% to 96.3%, while the tracking errors decrease over the same span, with object rotation falling from 11.17° to 7.18° and object position from 5.17 cm to 4.27 cm. A single policy thus absorbs 7,869 reference trajectories spanning nine objects, and performance is still improving at the end of training. The policy transfers to unseen interactions, reaching 65.0% SR and 84.5% PR on $\mathcal{D}_{\text{test}}$ with tracking errors close to those on $\mathcal{D}_{\text{train}}$, but the gap in completion is substantial. A likely cause is that the two splits are not identically distributed: approach-motion completion generates $K = 3$ variations per captured interaction for $\mathcal{D}_{\text{train}}$ and $K = 5$ for $\mathcal{D}_{\text{test}}$ as introduced in IV-B, so the test split spans a wider range of approach directions and generated locomotion prefixes than the policy encounters during training, leading to a mismatch of data distribution.

C. Joint Multi-Object Skill Learning

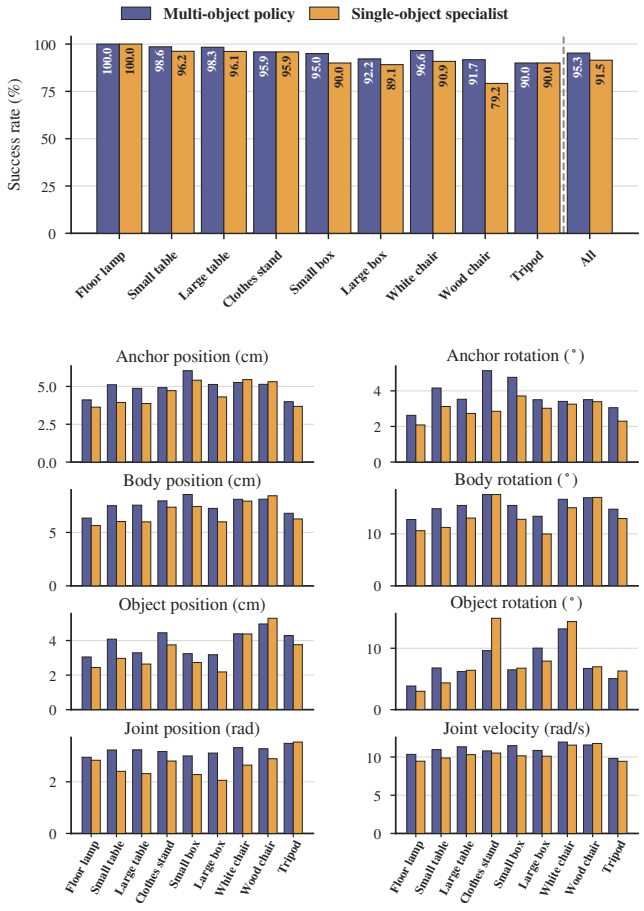

Figure 5. **Joint multi-object training versus object-specific specialists.** Top: per-object and pooled success rates on $\mathcal{D}_{\text{test}}$, where the pooled result is weighted by the number of evaluation trajectories. Bottom: tracking errors computed over rollouts.

Protocol. We compare a single multi-object policy $\pi_{\text{test}}^{\text{multi}}$, trained on all objects in $\mathcal{D}_{\text{test}}$, with nine single-object specialist $\{\pi_{\text{test}}^{j}\}_{j=1}^{9}$, each trained on the corresponding object's trajectories. Each single-object specialist is trained for 3k iterations. The unified policy is trained for $N_{\text{obj}} \times 3$k iterations, where $N_{\text{obj}} = 9$, giving 27k iterations in total. Thus, the unified policy and the collection of nine specialists receive the same aggregate number of training iterations. This comparison measures joint skill acquisition on a shared reference collection. All policies use the same observations, actions, reward, network architecture, and PPO configuration. We do not tune hyperparameters for either the specialists or the unified policy.

Results. As shown in Figure 5, the unified policy achieves a pooled success rate of 95.3%, compared with 91.5% for the specialists. Joint training matches specialist success on three objects and improves it on the other six, so the benefit is distributed across multiple objects rather than driven by a single easy category, and no object regresses. The pooled rates are weighted by the number of evaluation trajectories and therefore give greater weight to objects with more clips.

Success and imitation fidelity dissociate. The specialists often attain slightly lower tracking errors while completing fewer interactions. A specialist fits a single object's reference closely, whereas the unified policy allocates capacity across nine contact patterns and converges to a more conservative solution that trades pose fidelity for recovery margin. Lower imitation error therefore does not imply more reliable interaction completion, which argues against using tracking error alone as the primary metric for contact-rich interaction.

Why joint training helps. We offer two mechanisms consistent with the evidence. First, the approach and transport phases share whole-body structure across objects, such as walking to a pose, balancing the torso, and bearing a load, so clips from different objects act as mutual augmentation. Second, contact patterns are shared across objects of similar shape: a small table and a large table are grasped in much the same way, as are a clothes stand and a floor lamp. A specialist sees one geometry and can memorize a single grasp that fits it, whereas joint training exposes the policy to families of related shapes and pushes it toward selecting a grasp according to the object's geometry, which benefits to unseen interaction generalization.

D. Architecture and Optimizer Ablation

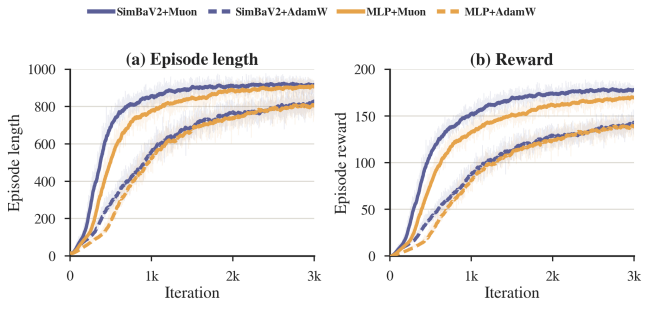

Figure 6. **Sample-efficiency ablation on the small-table task.** We compare four combinations under a common training budget of 3,000 iterations. Light curves show the raw mean episode length and reward, while dark curves show exponential moving averages with a span of 50 iterations.

Protocol. We evaluate all four combinations of MLP and SimBaV2 with AdamW and Muon on the small-table task under a common budget of 3,000 training iterations. Figure 6 reports mean episode length and reward, with raw curves and exponential moving averages.

Results. Both Muon configurations improve earlier than their AdamW counterparts, suggesting that the optimizer contributes the larger gain in sample efficiency in this comparison. A plausible explanation is that Muon's orthogonalized momentum updates, so no single direction dominates a gradient update step. This matters because the reward couples body tracking, object tracking, and finger contact, whose gradients differ substantially in scale: an update dominated by the body tracking term yields a policy that follows the reference pose without closing fingers, a local optimum that early training falls into easily. SimBaV2 provides a smaller additional improvement. Its hyperspherical normalization constrains weight and feature norms, which limits how far the policy moves in a single updates as the state distribution shifts at the onset of the contact, where the reward landscape changes most abruptly. The two effects are complementary: the optimizer governs how the gradient is distributed across the coupled objectives, while the architecture governs how much the policy is allowed to change when the state distribution moves.

VI. Conclusions and Limitations

We present WEAVE, a framework that converts captured human-object interactions into whole-body dexterous loco-manipulation skills. A reference-construction pipeline preserves task-relevant hand-object contacts across the morphology gap and completes the missing approach phases, and a unified contact- and geometry-aware policy, trained jointly over diverse object and multiple interaction clips, coordinates locomotion, whole-body balance, and finger-level grasping. The unified policy completes 92.5% success rate on trained interactions, and without any additional training, 65.0% on sequences beyond training data distribution, overall produces $\sim$23h of physically executed rollouts.

However, several limitations remain. Evaluation is carried out entirely in simulation: the policy consumes ground-truth odometry and object pose, so real-world deployment requires either onboard state estimation or distillation into a perception-based student, and the sim2real gap for contact-rich interaction is not accessed here. Concern for embodiment, the Inspire hand is underactuated and the policy commands proximal joints only, which bounds the achievable finger-level fidelity, and heavy or highly articulated objects are outside the range covered by our references and domain randomizations. Finally, the behaviors are reference-conditioned: the policy executes an interaction that must be supplied to it, and pairing it with a higher-level planner that selects and composes interaction, for instance, the humanoid VLA systems discussed in II-C, remains future work.

References

[1] F. Liu, Z. Gu, Y. Cai, Z. Zhou, H. Jung, J. Jang, S. Zhao, S. Ha, Y. Chen, D. Xu *et al.*, "Opt2skill: Imitating dynamically-feasible whole-body trajectories for versatile humanoid loco-manipulation," *IEEE Robotics and Automation Letters*, 2025.

[2] H. Weng, Y. Li, N. Sobanbabu, Z. Wang, Z. Luo, T. He, D. Ramanan, and G. Shi, "Hdmi: Learning interactive humanoid whole-body control from human videos," *arXiv preprint arXiv:2509.16757*, 2025.

[3] L. Yang, X. Huang, Z. Wu, A. Kanazawa, P. Abbeel, C. Sferrazza, C. K. Liu, R. Duan, and G. Shi, "Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction," *arXiv preprint arXiv:2509.26633*, 2025.

[4] S. Zhao, Y. Ze, Y. Wang, C. K. Liu, P. Abbeel, G. Shi, and R. Duan, "Resmimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning," *arXiv preprint arXiv:2510.05070*, 2025.

[5] S. Chen, S. Zhao, Z. Wu, J. Li, G. Shi, and C. K. Liu, "Scenebot: Contact-prompted general humanoid whole body tracking with scene-interaction," *arXiv preprint arXiv:2606.27581*, 2026.

[6] Z. Luo, J. Cao, S. Christen, A. Winkler, K. Kitani, and W. Xu, "Omnigrasp: Grasping diverse objects with simulated humanoids," *Advances in Neural Information Processing Systems*, vol. 37, pp. 2161–2184, 2024.

[7] Z. Luo, C. Tessler, T. Lin, Y. Yuan, T. He, W. Xiao, Y. Guo, G. Chechik, K. Kitani, L. Fan *et al.*, "Emergent active perception and dexterity of simulated humanoids from visual reinforcement learning," *arXiv preprint arXiv:2505.12278*, 2025.

[8] C. Tessler, Y. Jiang, E. Coumans, Z. Luo, X. B. Peng, and G. Chechik, "Maskedmanipulator: Versatile whole-body control for loco-manipulation," in *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, 2025, pp. 1–11.

[9] S. Xu, H. Y. Ling, Y.-X. Wang, and L.-Y. Gui, "Intermimic: Towards universal whole-body control for physics-based human-object interactions," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 266–12 277.

[10] S. Xu, S. Schulter, M. Ziyadi, X. He, X. Fei, Y.-X. Wang, and L. Gui, "Interprior: Scaling generative control for physics-based human-object interactions," *arXiv preprint arXiv:2602.06035*, 2026.

[11] Z. Wu, J. Li, P. Xu, and C. K. Liu, "Human-object interaction from human-level instructions," in *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2025, pp. 11 176–11 186.

[12] D. Rempe, M. Petrovich, Y. Yuan, H. Zhang, X. B. Peng, Y. Jiang, T. Wang, U. Iqbal, D. Minor, M. de Ruyter *et al.*, "Kimodo: Scaling controllable human motion generation," *arXiv preprint arXiv:2603.15546*, 2026.

[13] Z. Luo, J. Cao, K. Kitani, W. Xu *et al.*, "Perpetual humanoid control for real-time simulated avatars," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 895–10 904.

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

[15] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," *arXiv preprint arXiv:2406.08858*, 2024.

[16] J. Wu, S. Yao, G. He, X. Liu, Z. Zeng, X. Jiang, H. Yang, W. Zhang, and H. Zhao, "Toporetarget: Interaction-preserving retargeting for dexterous manipulation," *arXiv preprint arXiv:2606.16272*, 2026.

[17] Y. Feng, N. Leung, J. Wang, L. Yang, H. Qi, and P. Culbertson, "A minimalist retargeting-guided reinforcement learning recipe for dexterous manipulation," *arXiv preprint arXiv:2607.11874*, 2026.

[18] Y. Wu, L. Zeng, C. Jing, J. Ye, and X. Wang, "Reforce: Learning force-aware retargeting for dexterous manipulation," *arXiv preprint arXiv:2608.15560*, 2026.

[19] Y. Shao, Z. Chen, W. Lin, M. Zhou, T. Chen, X. Yang, Y. Chi, and Y. Mu, "Synmandex: Synthesizing human-like dexterous grasps from synthetic human pre-grasps," *arXiv preprint arXiv:2606.09798*, 2026.

[20] X. Zhu, Z. Liu, S. Jain, C. Li, M. Noori, M. A. Lin, H. Zhao, J. Welsh, M. Verghese, W. Liu *et al.*, "Learning dexterous manipulation using contact wrench guidance from human demonstration," *arXiv preprint arXiv:2607.00033*, 2026.

[21] R. Chen, F. Ruan, L. Cao, Z. Wang, B. Xu, S. Tong, J. Liu, M. Pei, C. Zhang, W. Xing *et al.*, "Dex-x: Learning visual-tactile dexterous manipulation from human videos with simulated interaction," *challenge*, vol. 1, no. 29, p. 30.

[22] T. Wang, O. Dionne, M. De Ruyter, D. Minor, D. Rempe, K. Zhao, M. Petrovich, Y. Yuan, C. Li, Z. Luo *et al.*, "Motionbricks: Scalable real-time motions with modular latent generative model and smart

primitives," *ACM Transactions on Graphics (TOG)*, vol. 45, no. 4, pp. 1–22, 2026.

[23] K. Zhao, M. Petrovich, H. Zhang, T. Wang, S. Tang, and D. Rempe, "Ardy: Autoregressive diffusion with hybrid representation for interactive human motion generation," *arXiv preprint arXiv:2607.08741*, 2026.

[24] S. Chen, Y. Ye, Z.-a. Cao, J. Lew, P. Xu, and C. K. Liu, "Hand-eye autonomous delivery: Learning humanoid navigation, locomotion and reaching," *arXiv preprint arXiv:2508.03068*, 2025.

[25] S. Arnaud, P. McVay, A. Martin, A. Majumdar, K. M. Jatavallabhula, P. Thomas, R. Partsey, D. Dugas, A. Gejji, A. Sax *et al.*, "Locate 3d: Real-world object localization via self-supervised learning in 3d," *arXiv preprint arXiv:2504.14151*, 2025.

[26] W. Liu, H. Zhao, C. Li, Y. Deng, J. Biswas, S. Pouya, and Y. Chang, "Compass: Cross-embodiment mobility policy via residual rl and skill synthesis," *arXiv preprint arXiv:2502.16372*, 2025.

[27] T. He, Z. Wang, H. Xue, Q. Ben, Z. Luo, W. Xiao, Y. Yuan, X. Da, F. Castañeda, S. Sastry *et al.*, "Viral: Visual sim-to-real at scale for humanoid loco-manipulation," *arXiv preprint arXiv:2511.15200*, 2025.

[28] I. Taouil, M. Ciebelski, S. Omar, H. Zhao, A. Dai, A. M. Johnson, and M. Khadiv, "Motiondisco: Motion discovery for extreme humanoid loco-manipulation," 2026.

[29] D. Li, Q. Wu, X. Chen, L. Li, Y. Lin, S. Wu, G. Zhang, M. Zhou, D. Xiang, Q. Zhang *et al.*, "Vaic: Vision-guided humanoid agile object interaction control via decoupled commands," *arXiv preprint arXiv:2606.09286*, 2026.

[30] H. Wang, W. Zhang, R. Yu, T. Huang, J. Ren, F. Jia, Z. Wang, X. Niu, X. Chen, J. Chen *et al.*, "Physhsi: Towards a real-world generalizable and natural humanoid-scene interaction system," *arXiv preprint arXiv:2510.11072*, 2025.

[31] X. He, S. Xu, X. Li, R. Dong, L. Bian, Y.-X. Wang, and L.-Y. Gui, "Ultra: Unified multimodal control for autonomous humanoid whole-body loco-manipulation," *arXiv preprint arXiv:2603.03279*, 2026.

[32] M. Byrd, D. Baek, K. Garg, H. Jung, D. Cho, M. Sorokin, R. Wright, and S. Ha, "Adaptmanip: Learning adaptive whole-body object lifting and delivery with online recurrent state estimation," *arXiv preprint arXiv:2602.14363*, 2026.

[33] S. Yin, Y. Ze, H.-X. Yu, C. K. Liu, and J. Wu, "Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation," *arXiv preprint arXiv:2509.20322*, 2025.

[34] T. Xie, H. Zhang, J. Park, Z. Wang, B. Wen, J. Li, X. Li, Q. Ben, H. Weng, Y. Ye, D. Minor, T. Wang, C. Jiang, S. Fidler, J. Kautz, L. Fan, Y. Zhu, Z. Luo, U. Iqbal, and Y. Yuan, "Grail: Generating humanoid loco-manipulation from 3d assets and video priors," 2026. [Online]. Available: https://arxiv.org/abs/2606.05160

[35] K. Team, J. Chen, Y. Ding, Z. Fang, K. Gai, K. He, X. He, J. Hua, M. Lao, X. Li *et al.*, "Kling-motioncontrol technical report," *arXiv preprint arXiv:2603.03160*, 2026.

[36] K. Lin, A. Mandlekar, C. R. Garrett, N. Chernyadev, Y. Fang, R. Ding, Y. Xie, J. Tran, L. Fan, and Y. Zhu, "Humanoidmimicgen: Data generation for loco-manipulation via whole-body planning," *arXiv preprint arXiv:2605.27724*, 2026.

[37] Y.-J. Wang, J. Li, S. Chen, T. E. Truong, P. Xu, P. Abbeel, R. Duan, K. Sreenath, A. Kanazawa, C. Sferrazza *et al.*, "Vlk: Learning humanoid loco-manipulation from synthetic interactions in reconstructed scenes," *arXiv preprint arXiv:2606.30645*, 2026.

[38] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[39] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[40] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[41] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[42] H. Jiang, J. Chen, Q. Bu, L. Chen, M. Shi, Y. Zhang, D. Li, C. Suo, C. Wang, Z. Peng *et al.*, "Wholebodyvla: Towards unified latent vla for whole-body loco-manipulation control," *arXiv preprint arXiv:2512.11047*, 2025.

[43] S. Bai, M. Li, X. Lv, J. Wang, X. Wang, F. Liao, C. Hou, L. Gu, W. Zhou, K. Wu *et al.*, "Hex: Humanoid-aligned experts for cross-embodiment whole-body manipulation," *arXiv preprint arXiv:2604.07993*, 2026.

[44] S. Wei, H. Jing, B. Li, Z. Zhao, J. Mao, Z. Ni, S. He, J. Liu, X. Liu, K. Kang *et al.*, "Ψ_0: An open foundation model towards universal humanoid loco-manipulation," *arXiv preprint arXiv:2603.12263*, 2026.

[45] Z. Li, Z. Zhang, Y. Wei, W. Zhang, X. Yuan, P. Zhi, G. Li, X. Guo, F. Gao, J. Yang *et al.*, "$omega$-0: A latent predictive world action model for concurrent humanoid loco-manipulation," *arXiv preprint arXiv:2608.06375*, 2026.

[46] J. Li, J. Wu, and C. K. Liu, "Object motion guided human motion synthesis," 2023. [Online]. Available: https://arxiv.org/abs/2309.16237

[47] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," *arXiv preprint arXiv:2510.02252*, 2025.

[48] C. Ferrari, J. Canny *et al.*, "Planning optimal grasps," in *Proceedings., 1992 IEEE International Conference on Robotics and Automation, 1992.*, vol. 3. IEEE, 1992, pp. 2290–2295.

[49] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[50] 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.

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

[52] H. Lee, D. Hwang, D. Kim, H. Kim, J. J. Tai, K. Subramanian, P. Wurman, J. Choo, P. Stone, and T. Seno, "Simba: Simplicity bias for scaling up parameters in deep reinforcement learning," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 46 816–46 848.

[53] J. Liu, J. Su, X. Yao, Z. Jiang, G. Lai, Y. Du, Y. Qin, W. Xu, E. Lu, J. Yan *et al.*, "Muon is scalable for llm training," *arXiv preprint arXiv:2502.16982*, 2025.