ULTRA: Unified Multimodal Control for Autonomous Humanoid Whole-Body Loco-Manipulation

Xialin He[†] Sirui Xu[†] Xinyao Li Runpei Dong
Liuyu Bian Yu-Xiong Wang[‡] Liang-Yan Gui[‡]
University of Illinois Urbana-Champaign
[†]Equal Contribution [‡]Equal Advising

https://ultra-humanoid.github.io/

Figure 1. ULTRA is an **all-in-one** controller for humanoid loco-manipulation that supports: **Top**. dense motion tracking from *arbitrary* reference. **Middle**. *fine-grained control* from operator commands. **Bottom**. *long-horizon goal* following with *egocentric perception*. It demonstrates autonomous whole-body behavior without relying on test-time motion references under real sensing.

Abstract

Achieving autonomous and versatile whole-body loco-manipulation remains a central barrier to making humanoids practically useful. Yet existing approaches are fundamentally constrained: retargeted data are often scarce or low-quality; methods struggle to scale to large skill repertoires; and, most importantly, they rely on tracking predefined motion references rather than generating behavior from perception and high-level task specifications. To address these limitations, we propose ULTRA, a unified framework with two key components. First, we introduce a physics-driven neural retargeting algorithm that translates large-scale motion capture to humanoid embodiments while preserving physical plausibil-

ity for contact-rich interactions. Second, we learn a unified multimodal controller that supports both dense references and sparse task specifications, under sensing ranging from accurate motion-capture state to noisy egocentric visual inputs. We distill a universal tracking policy into this controller, compress motor skills into a compact latent space, and apply reinforcement learning finetuning to expand coverage and improve robustness under out-of-distribution scenarios. This enables coordinated whole-body behavior from sparse intent without test-time reference motions. We evaluate ULTRA in simulation and on a Unitree G1 humanoid. Results show that ULTRA generalizes to autonomous whole-body loco-manipulation from egocentric perception, consistently outperforming tracking-only baselines with limited skills.

1. Introduction

Real-world loco-manipulation requires autonomy beyond replaying fixed reference motions. In unstructured environments, a humanoid must span a continuum: from dense motion references to sparse task goals, and from accurate state estimation to purely onboard sensing. Yet many controllers treat these as separate regimes and focus mainly on reference tracking [36, 41]. This fragmentation creates a precision-flexibility trade-off: dense-tracking policies break down when references are missing or infeasible, while purely goal-conditioned policies often lack the fine-grained coordination needed for complex tasks. We therefore seek a unified controller that produces whole-body loco-manipulation and smoothly transitions between dense plans and sparse intent as information changes.

Despite progress in co-tracking humanoid and object dynamics [38], two bottlenecks hinder unified autonomy. First, kinematic retargeting can yield physically inconsistent demonstrations that fail in contact-rich tasks. Second, existing architectures typically assume a fixed conditioning structure tailored to one input type, and cannot interpret diverse or partial supervision within a consistent framework. Under shifting observability and goals at deployment, this rigidity leads to systemic instability. We address both barriers: limited, physically implausible demonstrations and policies designed mainly for tracking predefined trajectories rather than operating with subsets of conditioning signals.

To overcome the demonstration bottleneck, we introduce a physics-driven, *neural* retargeting algorithm that transfers large-scale motion capture (MoCap) to humanoid embodiments at scale. Unlike kinematic retargeting [1, 41], which struggles to maintain physical consistency in contact-rich tasks, our retargeting is dynamics- and contact-aware by construction. We cast retargeting as simulation-constrained optimization with kinematic, dynamic, and contact constraints, and solve it with reinforcement learning (RL) at scale. Once trained, the policy generates large-scale physically feasible trajectories and generalizes to arbitrary data, enabling augmentation by scaling both objects and motions.

Building on this expanded corpus, we learn a *U*nified mu*LT*imodal cont*R*oller for *A*utonomous humanoid control (ULTRA) that shifts from reference replay to perception-driven, goal-conditioned control. We first train a privileged universal tracker, then distill it into a student that follows diverse goal specifications, from dense references to sparse long-horizon targets (Fig. 1). This is enabled by (**i**) unified tokenization with availability masking [29], which keeps a single policy stable when references or modalities are missing; and (**ii**) a variational skill bottleneck plus RL finetuning [39] geared toward deployment with realistic perception and sensor noise. The bottleneck resolves ambiguity under sparse goals by maintaining coherent motion, while RL finetuning shifts control from reference-conditioned tracking to closed-loop goal stabilization under partial observability and distribution shift. Together, ULTRA yields one policy that tracks references when available and executes from egocentric perception and sparse intent when they are not.

In summary, ULTRA presents a unified system for practical whole-body loco-manipulation with three components: (**i**) a physics-driven neural retargeting pipeline that scales MoCap to humanoid embodiments and supports zero-shot augmentation; (**ii**) a versatile *multimodal* controller distilled from a privileged tracker that supports reference tracking and goal following across sensing modalities, including blind, MoCap-based, and depth-perception settings; and (**iii**) simulation and real-world evaluation on Unitree G1, showing a single unified model can outperform tracking-only baselines when references exist while enabling broader goal-conditioned behaviors as shown in Fig. 1.

2. Related Work

2.1. Motion Retargeting

Retargeting transfers motion across embodiments with different morphologies. It originated in animation, where inverse-kinematics optimization adapted motions under kinematic constraints [10], and later evolved into learning-based mappings that amortized transfer for better generalization [33]. Humanoid retargeting requires stronger constraints because executability is contact-dependent and further limited by joint limits and dynamics. As a result, existing robot retargeting methods trade off efficiency and physical fidelity: kinematic approaches are fast but often under-model dynamics and degrade in contact-rich settings [1, 15, 21, 29, 41], while physics-based retargeting enforces contact and dynamics for physically plausible motions, but relies on non-convex, expensive optimization, typically per-trajectory RL [24, 37] or costly sampling-based methods [20]. We target the missing regime: *physics-driven yet scalable* retargeting that preserves interaction semantics without per-trajectory RL. We perform dataset-scale retargeting with a single unified policy

in one pass, and enable augmentation to expand coverage.

2.2. Humanoid Whole-body Locomotion

Leveraging human motion data to teach humanoid robots complex skills has been widely studied. Early methods often use model-based control (*e.g.*, trajectory optimization and MPC) to bridge embodiment and dynamics, while recent learning-based systems achieve precise tracking and agile motion replay [3, 5–8, 36, 43, 44]. Beyond pure tracking, recent work moves toward foundation-style control by distilling large motion corpora into reusable priors, where a single model tracks diverse motions and supports multiple control modes [14, 16, 42, 45]. Others shape latent priors with adversarial RL [13, 17, 27, 40], but have not shown reliable scaling to large, heterogeneous loco-manipulation corpora. ULTRA follows the scalable teacher-student distillation paradigm but addresses a key bottleneck: offline distillation is limited by the state coverage of teacher rollouts. While less severe for humanoid-only control with more structured spaces, it becomes acute in high-dimensional robot-object interaction. To address this, we draw inspiration from animation practice [39], but focus on real-world deployment: we perform large-scale distillation followed by RL fine-tuning that *expands* interaction-state coverage and improves robustness to out-of-distribution goals and executions.

2.3. Humanoid Whole-body Loco-Manipulation

Most humanoid motion tracking emphasizes reproducing human motion on the robot and treats environmental dynamics as secondary [2, 5, 12, 28], which is brittle for contact-rich loco-manipulation. Recent work couples humanoid motion and object interaction via co-tracking and shows strong agility [4, 36, 41, 46], but often assumes limited data replay or relies on external object state estimation (*e.g.*, motion capture), limiting autonomy under onboard egocentric sensing. Other approaches use hierarchical designs that generate trajectories/keypoints and track them with a universal controller [9, 43]; however, stacking a high-level planner on a low-level controller can accumulate error and violate physical constraints. Adversarial motion priors broaden coverage but are typically task-specific, requiring careful objective engineering and scaling poorly to large, heterogeneous loco-manipulation corpora [34]. ULTRA addresses these issues by learning a goal-conditioned policy that unifies dense tracking and sparse task specifications in a *shared* latent space, and by using RL finetuning to induce *closed-loop* behaviors that expand interaction-state coverage. This yields a *versatile* single-policy controller under real-world perception and a *scalable* paradigm that leverages broad motion corpora.

3. Problem Formulation and Preliminaries

3.1. Task Interface

We study whole-body loco-manipulation tasks where a humanoid interacts with a manipulated object, specified by a *goal* signal $\boldsymbol{c} \in \mathcal{C}$ that defines the task objective. A rollout succeeds if the terminal outcome satisfies $\boldsymbol{c}$, *e.g.*, the humanoid root and/or the object reaches target transformations within a tolerance. At each time step t, the policy receives (**i**) an observation $\boldsymbol{o}_t \in \mathcal{O}$ and (**ii**) task conditioning $\boldsymbol{c}_t$, and outputs an action $\boldsymbol{a}_t \in \mathcal{A}$. Here $\boldsymbol{a}_t$ specifies target joint positions executed by a PD controller.

Goal specification. We consider two forms of $\boldsymbol{c}$: (**i**) *dense reference conditioning*, which provides a time-indexed motion reference and thus specifies intermediate motions; and (**ii**) *sparse goal conditioning*, which specifies long-horizon target transformations for the humanoid root and/or object while leaving intermediate motions underdetermined.

Perception. Beyond proprioception, we consider two regimes for object sensing: (**i**) *MoCap-based sensing*, where $\boldsymbol{o}_t$ includes accurate object pose (*e.g.,* from motion capture); and (**ii**) *egocentric depth perception*, where $\boldsymbol{o}_t$ includes an egocentric point cloud from a depth sensor (*e.g.,* head-mounted), from which object state must be inferred.

3.2. Preliminaries

Since the controller may rely on partial onboard sensing, we model loco-manipulation as a goal-conditioned *Partially Observable Markov Decision Process* (POMDP). Let $\boldsymbol{s}_t \in \mathcal{S}$ be the underlying system state (humanoid and scene, including the object), with dynamics $\boldsymbol{s}_{t+1} \sim \mathcal{T}(\boldsymbol{s}_t, \boldsymbol{a}_t)$. The policy acts from $\boldsymbol{o}_t = \Omega(\boldsymbol{s}_t)$ and conditioning $\boldsymbol{c}_t$, producing $\boldsymbol{a}_t \in \mathcal{A}$. We optimize $\pi(\boldsymbol{a}_t \mid \boldsymbol{o}_t, \boldsymbol{c}_t)$ to maximize expected discounted return: $\max_\pi \mathbb{E}\big[\sum_{t\geq 0} \gamma^t r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{c}_t)\big]$. where γ is the discount factor that exponentially down-weights future rewards, The following sections describe how we use PPO [26] and imitation to learn policies, including the observation/reward design and key techniques for our tasks.

4. Method

As shown in Fig. 2, ULTRA follows a four-stage training paradigm that couples physics-driven motion retargeting with teacher-student learning. In Stage 1, we learn a *retargeting policy* that maps human MoCap motions to physically feasible humanoid loco-manipulation rollouts. In Stage 2, we train a privileged *teacher policy*, leveraging full state and dense reference trajectories from the retargeted rollouts. In Stage 3, we distill the teacher into a *multimodal student* that operates under perception and sparse goal specifications. Finally, we deploy the student with separated control mode.

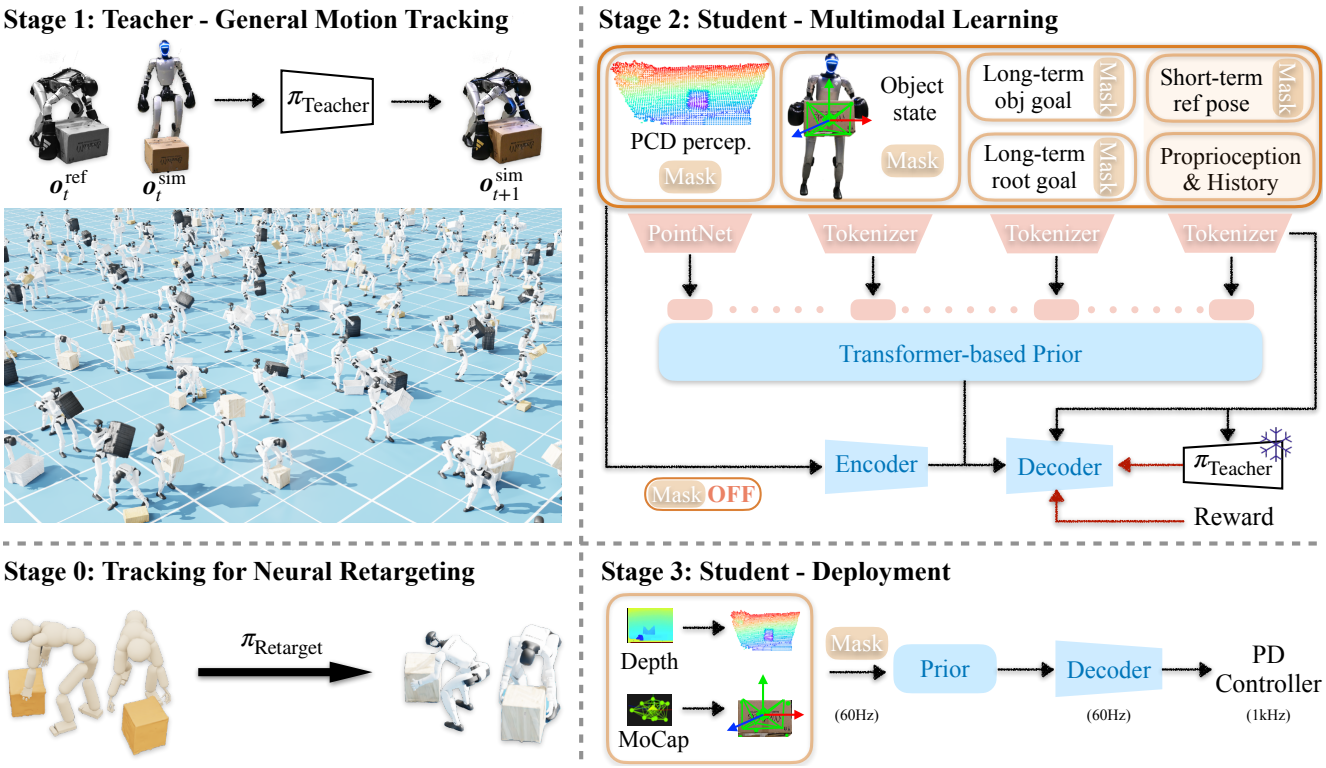


Figure 2. ULTRA follows four stages: (**i**) **Neural Retargeting:** an RL policy converts MoCap data into physically feasible G1 rollouts with augmentation; (**ii**) **Tracking:** a privileged teacher tracks these rollouts using full state and references; (**iii**) **Distillation:** we distill the teacher into a multimodal student for realistic sensing and sparse goals, with additional RL finetuning; (**iv**) **Deployment:** the student runs under real sensing, supporting depth input or MoCap-based state estimation.

4.1. Motion Tracking for Neural Retargeting

Given a human-object demonstration represented by an SMPL-X [22] motion sequence and an object pose trajectory, our goal is to generate a physically feasible rollout on the target humanoid (*e.g.,* Unitree G1) that preserves the overall motion and intended interaction. Traditional retargeting solves inverse kinematics under kinematic constraints. We instead cast retargeting as RL-based trajectory optimization: rewards encode tracking, while simulator transitions enforce kinematics, dynamics, and contacts. Following [38], this is well suited for contact-rich loco-manipulation, where contacts are hard to express as kinematic constraints. As preprocessing, we scale the human-object trajectory to match G1 and define a fixed correspondence from human key links to humanoid counterparts. We then train a *unified* retargeting policy across all motions, producing physically consistent rollouts without per-motion optimization or retraining. Dense, full-body tracking is brittle under embodiment mismatch and becomes especially fragile during object interaction, where exact link-wise targets may be infeasible and contact often requires deliberate deviations. Our key insight is to combine (**i**) relaxed tracking that prioritizes end effectors critical for loco-manipulation with (**ii**) interaction and contact rewards that correct mismatch-induced errors.

Reward. We define $r_{\text{track}} = r_p \cdot r_r \cdot r_{\text{obj}} \cdot r_{\text{int}} \cdot r_{\text{ct}} \cdot r_{\text{eng}}$, with all terms computed in a heading-aligned humanoid frame. Let $\mathcal{F}$ include only feet and palms. r_p tracks end-effector positions as sparse anchors; r_r matches normalized link directions over a fixed key edge set; and r_{eng} regularizes joint effort and foot placement. To reduce ambiguity, r_{obj} tracks object pose/velocities and r_{int} matches palm-to-surface offsets over sampled object points. We also align contact events by mapping contacts on human links to corresponding humanoid links, yielding r_{ct}. Full definitions are in Sec. .1.

Observation. The policy uses a privileged, reference-aware observation containing simulator state and its deviation from the SMPL-X reference. Since preprocessing establishes a fixed correspondence after scaling/alignment, residuals are well-defined: $\boldsymbol{o}_t = [\boldsymbol{o}_t^{\text{sim}}, \boldsymbol{o}_t^{\text{ref}}, \boldsymbol{o}_t^{\Delta}]$. $\boldsymbol{o}_t^{\text{sim}}$ includes proprioception and contact signals; $\boldsymbol{o}_t^{\text{ref}}$ provides selected correspondence-defined reference quantities (including object state); and $\boldsymbol{o}_t^{\Delta}$ encodes heading-aligned simulation-reference differences. All quantities are expressed in a heading-aligned frame to remove global yaw. See Sec. .1.

State initialization and early termination. Because we cannot reliably initialize the humanoid from an SMPL-X pose, we do not use reference-state initialization [23]. Each episode starts from a default standing pose, initially tracking

the first reference frame to stabilize the humanoid before transitioning to full tracking with smoothly varying weights. We terminate on falls, excessive deviation, or contact mismatch for 20 frames [38] to improve sample efficiency.

Simplified actuation. Since retargeting is used only to *generate reference rollouts*, we prioritize motion quality and throughput over hardware-faithful control. We use an *idealized* low-level controller in simulation (I., control frequency equal to simulation frequency), enabling stronger, more responsive tracking than onboard PD control. We train without domain randomization or perturbations, and address robustness later in Sec. 4.2.

Trajectory and object augmentation. RL-based retargeting also enables *flexible augmentation* (Fig. 4). Since preprocessing already scales positions, we can (**i**) apply anisotropic scaling along coordinate axes and (**ii**) scale the manipulated object with different coefficients, while interaction/contact rewards correct imperfections and the simulator enforces physical feasibility. Crucially, these augmentations are handled by a *single retargeting policy without retraining*.

4.2. Dense Motion Tracking for Teacher Policy

Sec. 4.1 converts human-object demonstrations into physically feasible G1 rollouts. For downstream imitation, we train a separate privileged teacher π_{teacher} to track these rollouts. The teacher uses the deployment control interface and actuation limits, but trains with privileged state and dense reference residuals to accelerate learning. We randomize physics and inject perturbations to broaden state visitation and teach recovery, producing stable behaviors that provide high-quality supervision for the student.

Observation. The teacher uses the same reference-aware observation as retargeting, but does not require cross-embodiment correspondence since the reference is already in the humanoid embodiment. (Table D).

Dense tracking objective. The teacher uses the same reward template, but replaces sparse anchoring with *full* link tracking, together with object, interaction, and contact reward. (Tables B and C).

Reference initialization and robustness training. We initialize from randomly sampled reference frames and include occasional stand still episodes that track standing references, reflecting deployment from a stable standing pose. To improve robustness, we randomize humanoid/object physical properties and inject perturbations, with a short grace period to allow recovery. We use the same early-termination criteria as retargeting and add no observation noise at this stage. See Tables I and J.

4.3. Multimodal Student Policy

We distill the privileged teacher into a multimodal student policy π_{student}. Unlike the teacher, the student observes only partial state and conditions on whatever modalities are available at test time via an availability mask randomly sampled during training. This retains teacher behavior as a prior while enabling goal-reaching under missing observations.

Multimodal observation with availability mask. The student consumes heterogeneous inputs: $\boldsymbol{o}_t^{\text{student}} = \left[\boldsymbol{o}_t^{\text{proprio}}, \boldsymbol{o}_t^{\text{goal}}, \boldsymbol{o}_t^{\text{object}}, \boldsymbol{o}_t^{\text{pcd}}, \boldsymbol{m}_t\right]$. $\boldsymbol{o}_t^{\text{proprio}}$ contains proprioception (*e.g.,* joint states, IMU), $\boldsymbol{o}_t^{\text{object}}$ provides object state (*e.g.,* MoCap), and $\boldsymbol{o}_t^{\text{pcd}}$ is an egocentric point cloud (*e.g.,* egocentric camera). $\boldsymbol{o}_t^{\text{goal}}$ encodes task objectives and commands, including (**i**) long-horizon object transforms, (**ii**) long-horizon humanoid root transforms, and (**iii**) next-frame humanoid local state changes for tracking. We also include discretized commands (*e.g.*, stand still) for deployment. $\boldsymbol{m}_t$ indicates which modalities are present. (Table G).

Distillation. We collect data with a DAgger-style loop [25]: we roll out with the teacher initially, gradually shift to the student, and query the teacher on visited states to obtain $\boldsymbol{a}_t^{\text{teacher}}$. During training, an encoder $q_\phi(\boldsymbol{z}_t^{\text{res}} \mid \boldsymbol{o}_t^{\text{student}}, \boldsymbol{o}_t^{\text{teacher}})$ infers a latent residual [29] using privileged teacher inputs, while a prior $p_\theta(\boldsymbol{z}_t^{\text{prior}} \mid m(\boldsymbol{o}_t^{\text{student}}))$ predicts a latent from masked student observations ($m(\cdot)$ applies $\boldsymbol{m}_t$). We combine them as $\boldsymbol{z}_t = \boldsymbol{z}_t^{\text{prior}} + \boldsymbol{z}_t^{\text{res}}$ and sample actions $\boldsymbol{a}_t^{\text{student}} \sim \pi_{\text{student}}(\boldsymbol{a}_t \mid \boldsymbol{o}_t^{\text{student}}, \boldsymbol{z}_t)$. We implement π_{student} with a transformer-based encoder [32] that projects each modality into shared tokens; $\boldsymbol{m}_t$ gates tokens and modulates cross-modal attention to ignore missing inputs. At deployment, we sample $\boldsymbol{z}_t$ from the prior only.

Training objective. We match teacher actions while aligning the prior with the privileged posterior:

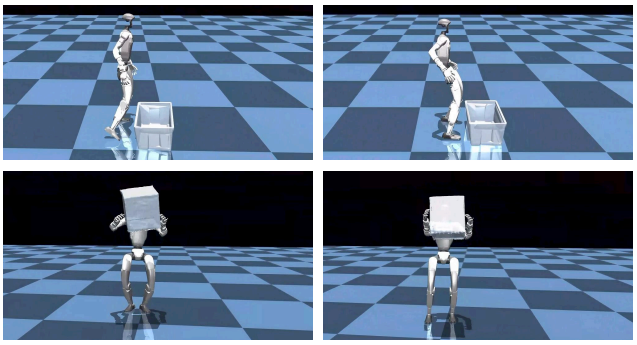
$$\mathcal{L} = \|\boldsymbol{a}_t^{\text{student}} - \boldsymbol{a}_t^{\text{teacher}}\|_2^2 + \mathcal{L}_{\text{aux}} + \lambda_{\text{KL}} D_{\text{KL}}\big(q_\phi(\boldsymbol{z}_t \mid \boldsymbol{o}_t^{\text{student}}, \boldsymbol{o}_t^{\text{teacher}}) \,\|\, p_\theta(\boldsymbol{z}_t \mid \boldsymbol{o}_t^{\text{student}})\big). \tag{1}$$

$\mathcal{L}_{\text{aux}}$ uses reconstruction heads (recovering masked modalities) to encourage $\boldsymbol{z}_t$ to retain task-relevant information.

Curriculum learning. Beyond DAgger, we use two curricula to keep the prior effective under partial observability: we progressively increase modality-masking probability, and anneal λ_{KL} and auxiliary weights to avoid posterior collapse while preserving latent skill diversity.

Shortcut for tracking. For local-goal tracking, behavior is largely deterministic, so a stochastic latent helps less. We add a residual shortcut (with the mask) from the full-body goal directly to the decoder, preserving low-level reference information and stabilizing decoding (Fig. 2).

RL finetuning. We perform RL finetuning on top of the distilled student by switching a subset of parallel environments to a goal-reaching objective while continuing distillation updates. Following [39], we partition simulators into (**i**) distillation environments replaying reference motions with imitation losses, and (**ii**) RL environments optimizing task success under state/goal perturbations. We sample random offsets for the object goal, humanoid root goal, and their



OmniRetarget ULTRA (**Ours**)

Figure 3. Qualitative comparison of our retargeting and OmniRetarget [41] at the same frame/sequence. **Top:** final frame; the baseline shows undesired standing foot placement. **Bottom:** a contact frame; ours yields more stable contacts.

initializations. Reward details are in Table H.

Deployment versatility. At test time, the student receives only o_t^{student} and samples z_t from the prior. With the same parameters, modality masking enables: (**i**) high-fidelity tracking by unmasking local reference (Fig. 1 **Top**), (**ii**) goal-conditioned control by masking local reference and unmasking long-horizon goals (Fig. 1 **Middle**), and (**iii**) vision-based manipulation by masking MoCap object state while unmasking point clouds (Fig. 1 **Bottom**).

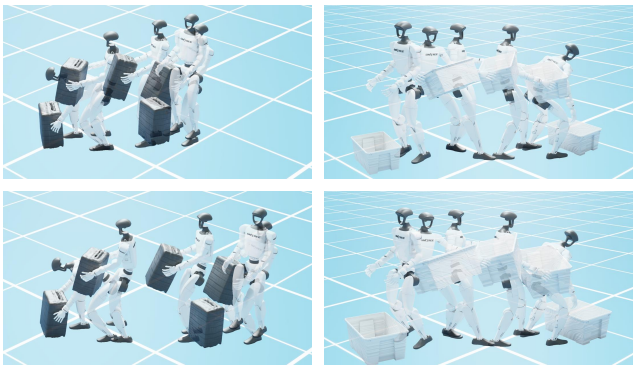
5. Experimental Results

We evaluate ULTRA end-to-end for autonomous whole-body loco-manipulation, from data generation to real-world transfer. We ask: (**i**) Can we retarget human-object MoCap into physically consistent rollouts with stable contacts and minimal sliding/penetration? (**ii**) Under dense references, can the student match a privileged teacher and specialized trackers? (**iii**) Under sparse goals, does RL finetuning improve robustness and yield a semantically organized latent skill space? (**iv**) Can one policy transfer to a real humanoid without test-time references? We evaluate four axes: retargeting, tracking, goal execution, and deployment on Unitree G1.

5.1. Experimental Setup

Simulation. We train in IsaacGym [19] with GPU-parallel environments and validate key results in MuJoCo [30]. Real trials use a physical Unitree G1 [31].

Dataset. We use OMOMO [11] human-object MoCap, using the corrected subset from [38] for a fair comparison with [41]. We focus on 4 box-shaped objects (others require dexterous hands). We retarget all sequences with our RL-based pipeline (Sec. 4.1) and augment via anisotropic trajectory scaling and object resizing, yielding a $\sim 6\times$ larger corpus (Fig. 4). We use the same train/test split for in-distributional (**ID**) evaluation and define out-of-distribution (**OOD**) by held-out motions and novel object scales from our zero-shot augmentation.



Motion Augmentation Object Augmentation

Figure 4. Zero-shot augmentation with the retargeting policy. **Left:** trajectory scaling. **Right:** object scaling. Motions remain plausible, enabling scalable data augmentation.

5.2. Motion Retargeting

Baselines. We compare against: (**i**) PHC [15] (kinematics-based retargeting for G1), (**ii**) GMR [1] (humanoid motion retargeting without objects), and (**iii**) OmniRetarget [41] (interaction-preserving kinematic engine with interaction-mesh augmentation). All retarget from the same OMOMO subset processed by [38].

Metrics. We measure: (**i**) *penetration* (duration, max depth) between humanoid/object/environment [41]; (**ii**) *foot skating* (sliding duration, max tangential stance velocity), with stance defined geometrically (foot within 2 cm of ground) to avoid noisy MoCap stance labels; and (**iii**) *contact floating*, the duration of lost hand-object contact during transport, detected via MuJoCo contact queries.

Quantitative evaluation. Table 2 shows ULTRA outperforms baselines across nearly all metrics/categories: lowest foot-skating duration/velocity and much less contact floating (near-zero on Largebox/Suitcase), while also reducing penetration. We attribute this to physics-aware retargeting that enforces contact/dynamics, keeping stance feet planted and preserving hand-object contact when lifting.

Qualitative evaluation. Fig. 3 shows more accurate hand/foot placement than OmniRetarget, whose kinematic formulation often breaks contact consistency and yields unnatural configurations relative to the object and ground.

Effectiveness of data augmentation. Our augmentation diversifies motions without retraining the retargeter and applies along the full trajectory (not only the initial frame), producing temporally consistent variations (Fig. 4). This improves downstream generalization: in Table 1, OmniRetarget retrained on our augmented data attains substantially higher OOD tracking success than when trained on its original dataset, confirming broader state/skill coverage.

5.3. General Motion Tracking

Baselines. We evaluate dense tracking (full reference provided) against: (**i**) OmniRetarget† (original data), (**ii**)

Table 1. Motion-tracking evaluation in IsaacGym. All methods are trained/evaluated on our data unless noted. **Green** highlights our primary tracking controller.

	In-Distribution (ID)							Out-of-Distribution (OOD)						
	Succ ↑	Succ ↑	Humanoid			Object		Succ ↑	Succ ↑	Humanoid			Object	
Method	(Humanoid)	(+Object)	$E_{\text{g-mpjpe}}$ ↓	E_{mpjpe} ↓	E_{jitter} ↓	E_{pos} ↓	E_{rot} ↓	(Humanoid)	(+Object)	$E_{\text{g-mpjpe}}$ ↓	E_{mpjpe} ↓	E_{jitter} ↓	E_{pos} ↓	E_{rot} ↓
(a) Unified Multimodal Controller														
ULTRA (**Ours**)	67.30±0.12	57.44±0.40	13.49±0.14	5.89±0.02	6.27±0.00	53.42±0.21	65.44±0.37	70.57±0.54	52.00±0.44	35.55±0.23	14.67±0.08	6.81±0.01	56.52±0.35	67.60±0.58
(b) Privileged Teacher														
ULTRA Teacher	97.57±0.05	89.79±0.11	12.98±0.30	5.64±0.05	14.81±0.08	17.15±0.03	23.28±0.33	97.12±0.43	81.33±0.78	19.14±0.48	7.94±0.11	15.91±0.08	25.57±0.28	33.49±0.37
(c) General Motion Tracking														
ULTRA (RL)	54.47±0.43	41.78±0.31	49.30±0.32	16.23±0.11	20.04±0.15	47.48±0.31	60.53±0.09	53.38±0.98	23.54±0.24	68.11±0.26	22.22±0.01	17.46±0.09	66.17±0.54	59.44±0.73
ULTRA (Distillation)	**85.03**±3.00	**77.15**±0.57	**15.45**±0.08	**6.84**±0.04	**8.12**±0.01	**25.48**±0.48	**33.97**±0.58	**86.63**±0.50	**52.74**±0.04	**35.01**±0.31	**13.48**±0.10	**9.35**±0.01	**36.18**±0.30	**38.18**±0.29
HDMI [36]	13.07±0.20	9.94±0.38	92.77±0.56	26.90±0.10	26.13±0.60	78.93±0.42	70.23±0.62	13.92±0.78	12.95±0.30	87.07±0.44	27.54±0.06	29.19±0.38	77.33±2.27	71.16±0.48
OmniRetarget† [41]	41.27±1.17	21.90±0.29	62.96±1.43	15.37±0.17	39.35±0.57	77.94±3.52	66.47±1.15	33.36±0.39	20.78±0.13	74.80±0.34	16.23±0.15	49.52±0.52	55.11±2.32	62.44±0.77
OmniRetarget [41]	51.34±0.67	20.91±0.52	67.12±0.86	7.43±0.07	39.92±1.44	60.67±0.54	67.03±0.19	46.71±0.74	25.82±0.52	68.34±0.82	8.98±0.19	40.08±1.77	58.57±2.37	66.70±0.84

† Trained/evaluated on original OmniRetarget dataset.

Table 2. Physical interaction quality for retargeting. Ours is better.

Method	**Penetration**		**Foot Skating**		**Contact Floating**
	Duration ↓	Max Depth (cm) ↓	Duration ↓	Max Vel. (cm/s) ↓	Duration ↓
Largebox					
PHC [15]	0.908 ± 0.125	0.073 ± 0.048	0.303 ± 0.145	0.032 ± 0.022	0.025 ± 0.054
GMR [1]	0.522 ± 0.259	0.086 ± 0.053	0.366 ± 0.317	0.029 ± 0.020	0.111 ± 0.171
OmniRetarget [41]	**0.000 ± 0.002**	0.013 ± 0.002	0.205 ± 0.106	0.035 ± 0.019	0.231 ± 0.224
ULTRA (**Ours**)	0.008 ± 0.030	**0.012 ± 0.002**	**0.061 ± 0.031**	**0.018 ± 0.010**	**0.015 ± 0.063**
Suitcase					
PHC [15]	0.914 ± 0.119	0.077 ± 0.051	0.286 ± 0.147	0.035 ± 0.024	0.032 ± 0.065
GMR [1]	0.571 ± 0.265	0.105 ± 0.050	0.399 ± 0.368	0.028 ± 0.018	0.142 ± 0.175
OmniRetarget [41]	0.003 ± 0.016	**0.012 ± 0.002**	0.264 ± 0.141	0.040 ± 0.021	0.404 ± 0.279
ULTRA (**Ours**)	**0.002 ± 0.013**	0.017 ± 0.019	**0.062 ± 0.045**	**0.017 ± 0.008**	**0.008 ± 0.040**

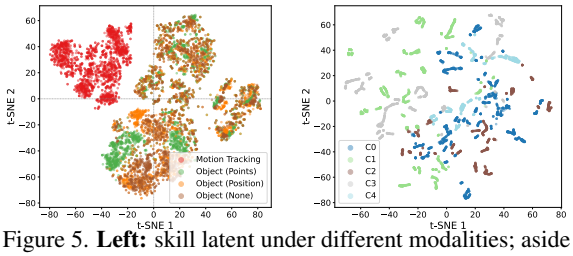


Figure 5. **Left:** skill latent under different modalities; aside from tracking, embeddings largely mix, indicating a shared skill space. **Right:** skill latent cluster by text labels (C0–C4), showing semantic structure.

OmniRetarget retrained on our augmented set, and (**iii**) HDMI [36] adapted to our setting. We also report ULTRA ablations: (**i**) direct RL under student observations (tracking only), (**ii**) tracking-only distillation, and (**iii**) all-task unified training. The privileged teacher is an upper bound.

Metrics. We report success (Succ): no fall and per-frame $E_{\text{g-mpjpe}} < 0.3\,\text{m}$ and $E_{\text{pos}} < 0.3\,\text{m}$; we also report humanoid-only success. Tracking errors include $E_{\text{g-mpjpe}}$, E_{mpjpe}, E_{jitter}, and object errors E_{pos}, E_{rot}.

Results. Table 1 shows ULTRA strongly outperforms baselines for humanoid-object tracking, especially under OOD motions/object scales. HDMI often becomes unstable at our scale and fails to converge. OmniRetarget trains smoothly but frequently fails manipulation: humanoid success is reasonable, but drops when object tracking is required, likely due to missing explicit object observations and a default-to-locomotion failure mode. ULTRA closes this gap via a privileged teacher with object signals and distillation that preserves closed-loop tracking under partial observability.

Distillation vs. direct RL under partial observation. Table 1 shows a clear gap between ULTRA trained with direct RL under student observations and the distilled student, in both ID and OOD tracking. In contact-rich loco-manipulation, direct RL must simultaneously learn whole-body stabilization and sustained object contact from partial observations, so early failures dominate rollouts and training often collapses. In contrast, the privileged teacher leverages full simulator state and dense references to learn contact-aware corrections with stable optimization, and distillation transfers this behavior to the student under realistic sensing, yielding higher success and lower object errors.

Distillation regularizes control. Although the teacher has access to more information, Table 1 shows the student can achieve *lower jitter* than the privileged teacher, this is significant for both all task student or student specialized for tracking. We attribute this to distillation acting as an implicit regularizer: matching teacher actions suppresses high-frequency, overly reactive RL corrections that reduce instantaneous error but introduce jitter and contact chattering. The student therefore learns a smoother, more contact-stable approximation that preserves the teacher's dominant strategy while discarding brittle micro-corrections.

All-task training induces a motion prior. Comparing ULTRA (Distillation) to ULTRA (**Ours**) in Table 1, unified training reduces ID tracking success while *largely preserving OOD performance*. We hypothesize that jointly optimizing dense tracking and sparse goal completion encourages the policy to learn a more trajectory-invariant motion prior that remains stabilizable under partial observability. This can reduce ID tracking fidelity, since the unified controller is not trained exclusively for reference replay, but it does not harm OOD performance, where success depends more on stable primitives than on exact replay.

5.4. Goal-Conditioned Following

Metric. Success (Succ): no fall and terminal state within 0.3 m of the goal.

Comparisons. Tracking-only baselines (*e.g.,* HDMI [36], OmniRetarget [41]) require dense references and are inapplicable; we compare ULTRA to ablations.

Tasks. We deploy ULTRA on a physical Unitree G1. The student runs onboard at the control frequency with proprioception and, when available, OptiTrack object pose (Fig. A). For dense tracking, we test OMOMO subsets (bimanual box lift/carry, suitcase transport) with household objects (Fig. B). For goal-conditioned control, we provide no motion references and specify future object transforms via simple keyboard commands.

RL finetuning expands OOD coverage. Table 3 and Fig. C show finetuning yields modest ID gains but large OOD gains under random goal offsets (nearly doubling under point clouds and tripling under position-only). This suggests finetuning expands interaction-state coverage and reinforces closed-loop recovery beyond the demonstration manifold.

Latent space shows control modes and motion semantics. We visualize the learned motion embeddings with t-SNE [18] to interpret what the motor latent capture. Fig. 5 (left) shows that the latent space cleanly separates dense reference tracking from sparse goal following across input modalities, while remaining within a shared manifold. Motion tracking stays distinct because we do not force it through the stochastic latent: when a local tracking goal is given, we pass a residual shortcut from the full-body goal directly to the decoder (Sec. 4.3). This leaves the latent to capture mainly ambiguity and multimodality under sparse goals. Fig. 5 (right) further shows *semantic structure*: we encode each motion's text description with MiniLM [35], cluster the resulting text embeddings into 5 classes with K-Means, and then plot the corresponding latents. The latent projections align with these semantic clusters, suggesting that the transformer encoder organizes motor skills by both control regime and high-level motion intent, reducing ambiguity under sparse goals by mapping them to appropriate regions of the skill manifold.

5.5. Real-World Deployment

Tasks. We deploy ULTRA on a physical Unitree G1. The student runs onboard at the control frequency with proprioception and, when available, OptiTrack object pose. For dense tracking, we test OMOMO subsets (bimanual box lift/carry, suitcase transport) with household objects. For goal-conditioned control, we provide no motion references and specify object transforms via keyboard commands.

Point cloud extraction. For egocentric perception, we extract object point clouds from depth only: back-project depth pixels using calibrated intrinsics, crop a forward ROI, remove the ground plane, take the dominant cluster as the box, and downsample to a fixed size for policy input.

Table 3. Sim-to-sim success rate on Mujoco across goal type with in-distributional (ID) goals from training and out-of-distributional (OOD) goals with random offsets, and across perception with egocentric point clouds or object position with no shape. Policies are trained in IsaacGym and evaluated in MuJoCo with 20 selected motion per setting.

RL fine-tuning	ID Goals		OOD Goals	
	Points	Position	Points	Position
✗	16 / 20	14 / 20	5 / 20	4 / 20
✓	19 / 20	16 / 20	**9 / 20**	**12 / 20**
Δ (RL gain)	+18.8%	+14.3%	**+80.0%**	**+200.0%**

Table 4. Real-world success rates on the OMOMO subset using a Unitree G1 humanoid. Each task is evaluated over two trials. MoCap provides object pose tracking for non-egocentric control modes, while the egocentric setting relies only on onboard sensing. MoCap is used for success evaluation in all settings. Dense reference tracking is direction-agnostic and thus reported as a single success rate.

Setting	Vertical	Lateral
Dense Reference Tracking	**73%** (19/26)	
Sparse Goal Following (MoCap)	**80%** (8/10)	**90%** (9/10)
Sparse Goal Following (Egocentric)	**50%** (5/10)	**60%** (6/10)

Quantitative evaluation. Table 4 reports success rates: the policy reliably grasps/transports on hardware and achieves reasonable sparse-goal success under out-of-distribution operator commands, including composed motions.

Failure analysis. Failures mainly arise from (**i**) friction gaps causing occasional grasp slip, (**ii**) depth noise/occlusion breaking point-cloud extraction, and (**iii**) disturbances beyond the recovery margin learned with domain randomization, motivating future tactile integration.

6. Conclusion

ULTRA is a unified framework for practical humanoid whole-body loco-manipulation that moves beyond reference replay toward perception- and goal-driven autonomy. It combines an RL-formulated, physics-driven retargeting policy that scales human-object MoCap into physically consistent humanoid rollouts with a distilled multimodal controller that unifies dense tracking and sparse goal specification. Experiments show improved interaction fidelity from retargeting, a student that matches tracking performance while remaining robust under distribution shift, and RL finetuning that boosts success on out-of-distribution goals. We further validate sim-to-real transfer on Unitree G1, demonstrating reliable dense tracking and sparse goal following. Overall, ULTRA points to a scalable path for versatile loco-manipulation that adapts online from realistic sensing without test-time references.

References

[1] Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025. 2, 6, 7

[2] Qingwei Ben, Feiyu Jia, Jia Zeng, Junting Dong, Dahua Lin, and Jiangmiao Pang. HOMIE: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*, 2025. 3

[3] Zixuan Chen, Mazeyu Ji, Xuxin Cheng, Xuanbin Peng, Xue Bin Peng, and Xiaolong Wang. Gmt: General motion tracking for humanoid whole-body control. *arXiv preprint arXiv:2506.14770*, 2025. 3

[4] Yuhui Fu, Feiyang Xie, Chaoyi Xu, Jing Xiong, Haoqi Yuan, and Zongqing Lu. DemoHLM: From one demonstration to generalizable humanoid loco-manipulation. *arXiv preprint arXiv:2510.11258*, 2025. 3

[5] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024. 3

[6] Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, et al. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. *arXiv preprint arXiv:2502.01143*, 2025.

[7] Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, et al. Hover: Versatile neural whole-body controller for humanoid robots. In *ICRA*, 2025.

[8] Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. Exbody2: Advanced expressive humanoid whole-body control. *arXiv preprint arXiv:2412.13196*, 2024. 3

[9] Dvij Kalaria, Sudarshan S Harithas, Pushkal Katara, Sangkyung Kwak, Sarthak Bhagat, Shankar Sastry, Srinath Sridhar, Sai Vemprala, Ashish Kapoor, and Jonathan Chung-Kuan Huang. Dreamcontrol: Human-inspired whole-body humanoid control for scene interaction via guided diffusion. *arXiv preprint arXiv:2509.14353*, 2025. 3

[10] Jehee Lee and Sung Yong Shin. A hierarchical approach to interactive motion editing for human-like figures. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 39–48, 1999. 2

[11] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 6

[12] Jinhan Li, Yifeng Zhu, Yuqi Xie, Zhenyu Jiang, Mingyo Seo, Georgios Pavlakos, and Yuke Zhu. OKAMI: Teaching humanoid robots manipulation skills through single video imitation. In *CoRL*, 2024. 3

[13] Yitang Li, Zhengyi Luo, Tonghe Zhang, Cunxi Dai, Anssi Kanervisto, Andrea Tirinzoni, Haoyang Weng, Kris Kitani, Mateusz Guzek, Ahmed Touati, et al. Bfm-zero: A promptable behavioral foundation model for humanoid control using unsupervised reinforcement learning. *arXiv preprint arXiv:2511.04131*, 2025. 3

[14] Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025. 3

[15] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *ICCV*, 2023. 2, 6, 7

[16] Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Sirui Chen, Fernando Castañeda, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025. 3

[17] Le Ma, Ziyu Meng, Tengyu Liu, Yuhan Li, Ran Song, Wei Zhang, and Siyuan Huang. Styleloco: Generative adversarial distillation for natural humanoid robot locomotion. *arXiv preprint arXiv:2503.15082*, 2025. 3

[18] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 8

[19] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. In *NeurIPS*, 2021. 6

[20] Chaoyi Pan, Changhao Wang, Haozhi Qi, Zixi Liu, Homanga Bharadhwaj, Akash Sharma, Tingfan Wu, Guanya Shi, Jitendra Malik, and Francois Hogan. Spider: Scalable physics-informed dexterous retargeting. *arXiv preprint arXiv:2511.09484*, 2025. 2

[21] Sungjae Park, Homanga Bharadhwaj, and Shubham Tulsiani. Demodiffusion: One-shot human imitation using pre-trained diffusion policy. *arXiv preprint arXiv:2506.20668*, 2025. 2

[22] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 4

[23] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018. 4

[24] Daniele Reda, Jungdam Won, Yuting Ye, Michiel van de Panne, and Alexander Winkler. Physics-based motion retargeting from sparse inputs. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–19, 2023. 2

[25] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011. 5

[26] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[27] Jiyuan Shi, Xinzhe Liu, Dewei Wang, Ouyang Lu, Sören Schwertfeger, Chi Zhang, Fuchun Sun, Chenjia Bai, and Xuelong Li. Adversarial locomotion and motion imitation for humanoid policy learning. *arXiv preprint arXiv:2504.14305*, 2025. 3

[28] Wandong Sun, Luying Feng, Baoshi Cao, Yang Liu, Yaochu Jin, and Zongwu Xie. Ulc: A unified and fine-grained controller for humanoid loco-manipulation. *arXiv preprint arXiv:2507.06905*, 2025. 3

[29] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 43(6):1–21, 2024. 2, 5

[30] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IROS*, 2012. 6

[31] Unitree. Unitree g1 humanoid agent ai avatar. https://www.unitree.com/g1/. 6

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 5

[33] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. *CVPR*, 2018. 2

[34] Huayi Wang, Wentao Zhang, Runyi Yu, Tao Huang, Junli Ren, Feiyu Jia, Zirui Wang, Xiaojie Niu, Xiao Chen, Jiahe Chen, et al. Physhsi: Towards a real-world generalizable and natural humanoid-scene interaction system. *arXiv preprint arXiv:2510.11072*, 2025. 3

[35] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*, 2020. 8

[36] Haoyang Weng, Yitang Li, Nikhil Sobanbabu, Zihan Wang, Zhengyi Luo, Tairan He, Deva Ramanan, and Guanya Shi. Hdmi: Learning interactive humanoid whole-body control from human videos. *arXiv preprint arXiv:2509.16757*, 2025. 2, 3, 7, 8

[37] Sirui Xu, Yu-Wei Chao, Liuyu Bian, Arsalan Mousavian, Yu-Xiong Wang, Liang-Yan Gui, and Wei Yang. Dexplore: Scalable neural control for dexterous manipulation from reference-scoped exploration. In *CoRL*, 2025. 2

[38] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *CVPR*, 2025. 2, 4, 5, 6

[39] Sirui Xu, Samuel Schulter, Morteza Ziyadi, Xialin He, Xiaohan Fei, Yu-Xiong Wang, and Liangyan Gui. InterPrior: Scaling generative control for physics-based human-object interactions. *arXiv preprint arXiv:2602.06035*, 2026. 2, 3, 5

[40] Haoru Xue, Xiaoyu Huang, Dantong Niu, Qiayuan Liao, Thomas Kragerud, Jan Tommy Gravdahl, Xue Bin Peng, Guanya Shi, Trevor Darrell, Koushil Sreenath, et al. Leverb: Humanoid whole-body control with latent vision-language instruction. *arXiv preprint arXiv:2506.13751*, 2025. 3

[41] Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025. 2, 3, 6, 7, 8

[42] Kangning Yin, Weishuai Zeng, Ke Fan, Minyue Dai, Zirui Wang, Qiang Zhang, Zheng Tian, Jingbo Wang, Jiangmiao Pang, and Weinan Zhang. Unitracker: Learning universal whole-body motion tracker for humanoid robots. *arXiv preprint arXiv:2507.07356*, 2025. 3

[43] Shaofeng Yin, Yanjie Ze, Hong-Xing Yu, C Karen Liu, and Jiajun Wu. Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation. *arXiv preprint arXiv:2509.20322*, 2025. 3

[44] Yanjie Ze, Zixuan Chen, Joao Pedro Araújo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, and C Karen Liu. Twist: Teleoperated whole-body imitation system. *arXiv preprint arXiv:2505.02833*, 2025. 3

[45] Weishuai Zeng, Shunlin Lu, Kangning Yin, Xiaojie Niu, Minyue Dai, Jingbo Wang, and Jiangmiao Pang. Behavior foundation model for humanoid robots. *arXiv preprint arXiv:2509.13780*, 2025. 3

[46] 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. 3