INTERMIMICGEN: SCALING HUMANOID LOCO-MANIPULATION THROUGH SELF-EVOLVING MOTION IMITATION

Anonymous authors
Paper under double-blind review



Figure 1: InterMimicGen grows a finite set of human interaction captures into an expanding repertoire of executable robot motions. **Left:** One formulation turns the captures into executable motions across diverse objects, tasks, and robots, from humanoids to a wheeled mobile manipulator. **Top right:** Self-evolving motion imitation. A single motion grows round by round into many variants. **Bottom right:** The resulting motions run on real robots. *Videos are on the supplementary website*.

ABSTRACT

Captured human-object interactions provide rich supervision for humanoid loco-manipulation, but they are sparse, heterogeneous, and not directly executable by robots. We introduce InterMimicGen, a self-evolving motion-imitation framework in which robot motion data and a tracking policy improve each other. First, we consolidate motion-captured human-object interaction datasets and retarget them into humanoid robot references while preserving whole-body coordination and dexterous hand-object relationships. This produces a large and diverse humanoid robot reference collection for dexterous whole-body loco-manipulation. Second, we train a physics-based generalist tracker that executes these references in simulation on a humanoid with dexterous hands, covering a scale and diversity beyond prior humanoid tracking systems for loco-manipulation. Third, we close a data flywheel: each round makes small, task-preserving changes to where an interaction takes place and how the body performs it, fine-tunes the tracker on them, and keeps only the variants whose simulated execution completes the task, which seed the next round. With more iterations, these small edits compound into broader coverage around the sparse original demonstrations while preserving task semantics and motion quality. Experiments show contact-preserving retargeting across robot configurations, broad tracking with a single generalist policy, executable motions that keep growing over augmentation rounds, and transfer to real robots. InterMimicGen provides a unified path from heterogeneous human demonstrations to a continually expanding motion resource for humanoid robot learning.

1 INTRODUCTION

The value of human demonstrations for humanoid loco-manipulation extends beyond reproducing the trajectories they record. Each demonstration captures one coordinated solution in a high-dimensional interaction space, but even a large motion-capture corpus provides sparse samples of this space rather

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

than dense coverage of their feasible variations. Covering these variations therefore requires either a policy that can infer the physically feasible manifold from these sparse samples, or a data-generation process that enriches the data with coordinated, physically valid variations.

A natural route is to train a tracking policy by collecting more references, but any fixed set of references still leaves these variations uncovered. When we look back, generalist trackers already generalize well in locomotion (Luo et al., 2026), where a motion close to a demonstration usually remains feasible. Loco-manipulation is less forgiving. The robot must hold contact with an object while keeping its balance, so a small change in posture or object placement can break a grasp, push the hand into the object, or make the robot fall. As a result, a tracker that executes a demonstration may still fail on the same task where the trajectory is shifted by only a few centimeters (Sec. 4.4). Covering them therefore requires data that grows with the tracker rather than a fixed collection. We build this as a *data flywheel*, in which executable data and the tracker grow together.

To instantiate this flywheel, we introduce **InterMimicGen**, a self-evolving motion-imitation data engine that couples human-object interaction retargeting, large-scale motion imitation, and iterative data expansion. We first consolidate heterogeneous human-object motion-capture data (Figure 3) and retarget the interactions into humanoid robot references. Our retargeter keeps both long-horizon body-object relations and the finger contact that holds small objects: a whole-body solver shapes the grasp to follow the object's geometry rather than captured finger poses alone, and a fine-grained cleanup then refines the contact at the millimeter scale. The result is a large and diverse robot reference collection for dexterous whole-body loco-manipulation, as shown in Figure 1. Physics-based trackers then execute these references in simulation on a humanoid robot with dexterous hands, and the same trackers later grow together with their reference data through iterative data expansion.

The goal of this expansion is not to invent new tasks; instead, each human interaction becomes a seed for physically distinct realizations of the same semantic behavior (Mandlekar et al., 2023). We vary object and body motion through small edits around previously validated interactions, such as placing a box farther to the left or crouching deeper, so the generated candidates stay close to interactions that already execute. Each round fine-tunes the tracker on the current references and these candidates, executes them in simulation, and keeps only those that reach the end without failure, which then seed the next round. Over more rounds, these small edits accumulate into large variations. We call this reciprocal growth *self-evolving motion imitation*: a broader set of executable references supports a more capable tracker, which in turn validates a wider set of coordinated variations. The generated references keep the comparable quality of the original data, and the fine-tuned tracker executes them, whereas the original tracker fails on about half of them, even though it was trained with scale.

Our contributions are threefold. First, we consolidate heterogeneous human-object interaction (HOI) data and retarget them into, to our knowledge, the most diverse humanoid robot reference collection (See Sec. 4.1) for dexterous whole-body loco-manipulation. Second, using these references, we train a physics-based generalist tracker that realizes this repertoire on a target embodiment at a scale beyond prior humanoid tracking systems (Weng et al., 2025; Yang et al., 2026; He et al., 2026; Cao et al., 2026b) for loco-manipulation. Third, we introduce a self-evolving augmentation-learning loop that expands this repertoire through coupled variations in object and whole-body motions, turning validated interactions into a growing resource for subsequent learning of whole-body controllers.

2 Related Work

Human-object interaction data. Human-object and human-scene capture datasets provide complementary interaction references across whole-body grasping, object tracking, bimanual manipulation, and scene-level coordination (Taheri et al., 2020; Bhatnagar et al., 2022; Huang et al., 2024; Lv et al., 2024; Fan et al., 2023; Lu et al., 2025; Zhang et al., 2025; Liu et al., 2025b; Zhang et al., 2023a; Zhao et al., 2024; Li et al., 2023). These resources differ in object diversity, contact structure, and body and hand annotation, shaping the supervision available for imitation. Among them, InterAct (Xu et al., 2025c) consolidates heterogeneous interaction sources, while HiPHI (Ji et al., 2026) adds extensive body-object capture. We combine their motion into a robot reference collection and expand physically executable coverage through retargeting and humanoid robot execution in simulation.

Whole-body and dexterous motion retargeting. Motion retargeting adapts demonstrations across morphologies using learned mappings or optimization objectives built from skeleton structure, contact

cues, and geometric correspondences (Villegas et al., 2018; Aberman et al., 2020; Villegas et al., 2021; Reda et al., 2023; Araújo et al., 2025; Cao et al., 2026a). For interaction, the relationships between body, object, hand, and scene must survive the change in embodiment rather than only the joint trajectory (Kim et al., 2016; Zhang et al., 2023b; Yang et al., 2026). Dexterous teleoperation and synthesis study high-quality hand or hand-arm transfer (Handa et al., 2020; Qin et al., 2023; Zhao et al., 2026; Pan et al., 2025; Xu et al., 2025b), but whole-body interaction also requires those contacts to remain consistent with locomotion and balance. Our retargeter therefore separates the problem into a coarse whole-body interaction solve and a local dense contact refinement, so hand motion is improved without disturbing the body-object motion.

Physics-based imitation and whole-body manipulation. Physics-based imitation began with example-guided control of individual skills (Peng et al., 2018) and later scaled to shared controllers over diverse motion libraries (Won et al., 2020; Luo et al., 2023). On physical humanoids, teleoperation and imitation enable whole-body control (He et al., 2024; Ze et al., 2025), while universal tracking policies and visuomotor extensions broaden the motion and sensing interfaces (Yin et al., 2026; Luo et al., 2026; Ding et al., 2025). Object-centric extensions couple body and object dynamics for grasping and dynamic skills (Wang et al., 2023; Luo et al., 2024; Wang et al., 2025b; Tessler et al., 2025; Xu et al., 2025d; 2026a), and then move to real-world humanoid loco-manipulation (Weng et al., 2025; Zhao et al., 2025; Fu et al., 2026; Liu et al., 2025a; Dao et al., 2024; Sun et al., 2025; Kalaria et al., 2025; Yin et al., 2025; Wang et al., 2025a; He et al., 2026). Our work starts from reference tracking and extends it into a self-evolving data engine, with real deployment.

Demonstration augmentation and iterative generation. A demonstration can be reused beyond its recorded trajectory (Mandlekar et al., 2023): collected segments preserve task intent, while execution under adapted conditions becomes valid training data. Later systems extend this execution-validated reuse to bimanual dexterity (Jiang et al., 2025), skill composition (Garrett et al., 2024), intervention-driven data collection (Hoque et al., 2024; Luo et al., 2025), policy self-improvement (Bousmalis et al., 2024), and humanoid loco-manipulation (Lin et al., 2026). A related line in character and humanoid control scales policies through generated experience or closed-loop motion generation (Xu et al., 2025a; Debbad et al., 2026; Xu et al., 2026b; Wang et al., 2026). Our setting differs in both domain and mechanism: we expand dexterous and agile humanoid loco-manipulation from motion-capture references, and we self-evolve through motion imitation itself instead of a separate kinematic motion synthesizer, which is often brittle for object interaction.

3 METHOD

InterMimicGen turns captured human-object interactions into an expanding collection of robot references that a physics-based tracker can execute (Figure 2). A reference $\boldsymbol{\xi}$ stores the target robot motion, object motion, and contact intent of one clip. We first consolidate heterogeneous captures (Sec. 3.1) and retarget them to robot embodiments (Sec. 3.2). A generalist tracker then realizes these references under robot dynamics (Sec. 3.3). Finally, self-evolving motion imitation edits how verified interactions are realized and keeps only the edits that the tracker executes in physics. Repeating this over rounds, with a tracker that keeps learning from the new motions, lets the verified references and the tracker grow together (Sec. 3.4).

3.1 A CONSOLIDATED INTERACTION REFERENCE COLLECTION

Because our learning loop expands skills already represented in the source motions, the seed collection should cover interaction semantics as broadly as possible. We combine InterAct (Xu et al., 2025c) and the object-interaction portion of HiPHI (Ji et al., 2026) into a shared collection. InterAct contributes diverse tasks and object categories by consolidating public motion-capture datasets (Taheri et al., 2020; Bhatnagar et al., 2022; Huang et al., 2024; Zhang et al., 2023a; Zhao et al., 2024; Li et al., 2023), while HiPHI contributes longer synchronized human-object motions. Figure 3 summarizes this complementarity: the sources occupy different distributions of human motion, object geometry, and interaction space, and their union improves coverage on external motion probes. We convert both sources to a common body-object-hand representation and treat absent hand annotations as unknown rather than as evidence of no contact.

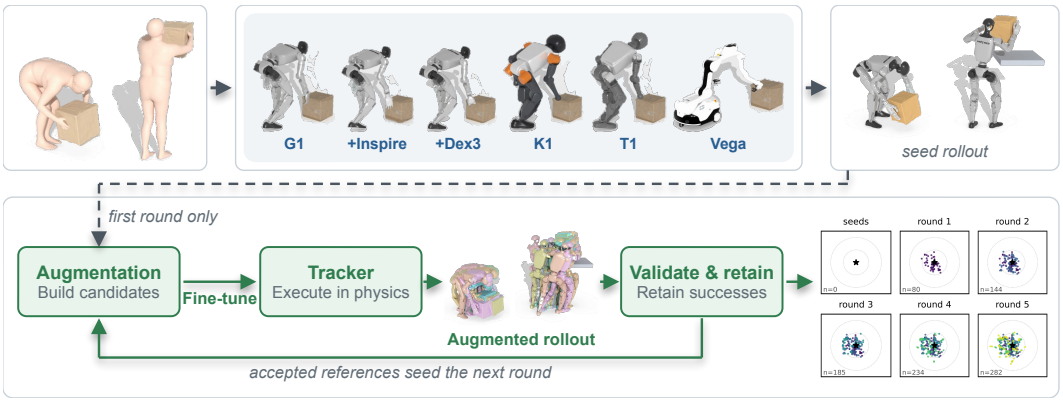

Figure 2: **Top:** Reference construction. Human-object motion capture is retargeted into whole-body dexterous references, which a physics-based tracker imitates to produce seed rollouts. **Bottom:** Self-evolving motion imitation. Each round augments verified references into candidates, fine-tunes the tracker on them, executes the candidates in physics, and retains only successful rollouts as seeds for the next round; **Bottom right:** the archive of accepted augmentation expanding.

We then keep only the interactions each target robot can physically support. Sequences that simulation cannot represent, such as bag hanging with an unmodeled strap, are removed, and platforms without dexterous hands exclude interactions whose success depends on articulated finger contact. The resulting references share one format for downstream training (Sec. A).

3.2 Retargeting for dexterous whole-body interaction

Retargeting must satisfy constraints at two scales. The whole body must preserve support and body-object geometry over long horizons, while the hand must resolve millimeter-scale contact around small object regions. Optimizing both with one objective tends to loosen contact or over-constrain the body. We therefore solve the whole body first, with a finger-opposition term that shapes the grasp, and then remove local artifacts in a cleanup that preserves contact (Algorithm 1).

Interaction-preserving whole-body retargeting. The coarse stage transfers body-object relationships through an interaction mesh (Yang et al., 2026) that connects sparse body landmarks with sampled object geometry, so matching the mesh preserves spatial interaction structure rather than only joint poses. It solves for the floating base and articulated body under the target robot's constraints, with sparse hand landmarks placing the wrists and palms. Because different motion clips might fail under different local constraints, we try several constrained solvers in turn and keep the first feasible solution. The solve keeps the source object trajectory fixed and already includes the finger-opposition term below, so the hands close around the object while the body is solved.

Finger opposition. Surface proximity alone might promote several fingertips settle on the same side of an object, which does not hold it. We therefore add an opposition term to the whole-body solve. For fingertip f, let $\boldsymbol{n}_{t,f}$ be the normalized displacement from its selected object-surface point to the fingertip at time t, and let $\boldsymbol{n}_{t,\mathrm{thumb}}$ be the same direction for the thumb:

$$E_{\mathrm{cover}}(\boldsymbol{q}_t) = \sum_{f \in \mathcal{F}_e} g_{t,f} \left[\boldsymbol{n}_{t,\mathrm{thumb}}^{\top} \boldsymbol{n}_{t,f} + m \right]_+^2, \qquad [x]_+ = \max(x, 0), \tag{1}$$

where $\mathcal{F}_e$ contains the supported thumb-finger pairs, $g_{t,f}$ activates a pair once the thumb and finger approach the object, and m sets the opposition margin. The term penalizes pairs whose directions are not sufficiently opposed, so fingers approach from the side opposite the thumb.

Geometry cleanup with contact preservation. The whole-body solve can still leave small penetrations, jitter, and contact drift. We therefore refine the robot and object trajectories with two successive Adam passes (Kingma & Ba, 2015). The first reduces robot-object, floor, and hand self-collision while regularizing deviation from the coarse motion, foot support, joint limits, and temporal smoothness. The second smooths the repaired trajectory under collision constraints and preserves hand-object

relationships during contact. Each pass is retained only if its output satisfies deviation and quality checks relative to its input. Robots without articulated fingers omit finger-specific terms. Sec. B gives the objectives, acceptance checks, and solver settings.

3.3 A GENERALIST FOR INTERACTION MOTION IMITATION

The generalist tracker turns kinematic references into physically executable motions, and later serves as the validator of augmented references. It observes the robot state, upcoming reference frames, and object-relative geometry, including distance-decayed vectors from each body link to its nearest point on the object surface. It outputs joint targets for proportional-derivative (PD) control, and on the Inspire hands, coupled finger joints follow their driving joints through mimic-joint constraints instead of acting as independent actuators. We train a single policy π_θ with proximal policy optimization (Schulman et al., 2017) across parallel environments, sharing parameters across motions and objects. The reward balances body motion, object motion, interaction geometry, and contact fidelity. Reference-state initialization exposes different phases of long sequences, and adaptive sampling emphasizes difficult motions. More details are presented in Sec. C.

3.4 SELF-EVOLVING MOTION IMITATION

A human capture shows what an interaction achieves but only one way of performing it. We therefore generate new ways of performing the same interaction. This follows MimicGen (Mandlekar et al., 2023), which turns one demonstration into many training examples by replaying it under new conditions and keeping the successful replays, and here the replay is motion imitation by our tracker. Each variant is derived from a verified parent reference $\boldsymbol{\xi}$ and must keep four properties of the parent, which form the invariant set $\mathcal{I}$: the interaction type, the intended contacts, the object, and the task outcome. Variants thus add new motions around each capture without changing its meaning.

What an edit changes. An edit changes either where the interaction takes place or how the body performs it. We describe it by coordinates $\boldsymbol{z} = (\boldsymbol{z}^{\text{task}}, \boldsymbol{z}^{\text{body}})$ and construct a proposal $\tilde{\boldsymbol{\xi}}$ that keeps every invariant of its parent:

$$\tilde{\boldsymbol{\xi}} = \text{Augment}(\boldsymbol{\xi}; \boldsymbol{z}, \mathcal{I}), \qquad r_j(\tilde{\boldsymbol{\xi}}, \boldsymbol{\xi}) = 0 \quad \forall j \in \mathcal{I}, \tag{2}$$

where r_j measures how much the proposal violates invariant j. *Spatial edits* $\boldsymbol{z}^{\text{task}}$ move or turn the object's trajectory, for example placing a box further to the left, and leave the whole-body response to the tracker, which must reach, grasp, carry, and place the object accordingly. *Body edits* $\boldsymbol{z}^{\text{body}}$ keep the object path and change the posture that realizes it: a deeper crouch or a shifted and turned pelvis, a wider stance or a different toe angle, or a different elbow posture. In every case the hands stay on their original targets, so the contacts with the object are preserved. After each edit, we recompute the kinematics, velocities, and body-object interaction features of the new reference. Sec. D define the coordinates, their bounds, and the checks that each invariant of the parent still holds.

Fine-tune, then keep only what executes. Each round starts from a set of verified parents. Let $\mathcal{D}_k$ be the verified references after round k, where $\mathcal{D}_0$ holds the seed rollouts of the base tracker on the retargeted references, and let $\mathcal{P}_k \subseteq \mathcal{D}_k$ be the parents of round k. The round edits the parents into proposals $\mathcal{C}_k$, fine-tunes the tracker π_θ on the verified references together with the proposals, since the proposals lie partly outside what it has learned, and then executes each proposal in physics and keeps the successful ones:

$$\begin{aligned} \mathcal{C}_k &= \text{Augment}(\mathcal{P}_k), \\ \theta_{k+1} &= \text{Train}(\theta_k, \mathcal{D}_k \cup \mathcal{C}_k), \\ \mathcal{B}_k &= \big\{ \text{Store}(\tilde{\boldsymbol{\xi}}, \boldsymbol{\tau}) : \tilde{\boldsymbol{\xi}} \in \mathcal{C}_k,\ \boldsymbol{\tau} \sim \text{Rollout}(\pi_{\theta_{k+1}}, \tilde{\boldsymbol{\xi}}),\ V(\tilde{\boldsymbol{\xi}}, \boldsymbol{\tau}; \mathcal{I}) = 1 \big\}, \\ \mathcal{D}_{k+1} &= \mathcal{D}_k \cup \text{Select}(\mathcal{B}_k). \end{aligned} \tag{3}$$

Here, Train continues the tracker's reinforcement learning on the proposed references. The validation function V accepts a proposal only if its rollout reaches the end of the reference without early termination or a fall, passes motion-smoothness checks, and preserves the task outcome, such as holding or placing the object at the end (Sec. D.4). Satisfying Eq. 2 alone does not make a proposal executable under robot dynamics. Fine-tuning expands the set of variations the tracker can execute: without it, only a small fraction of second-round proposals pass validation (Sec. F.2). Thus, the loop

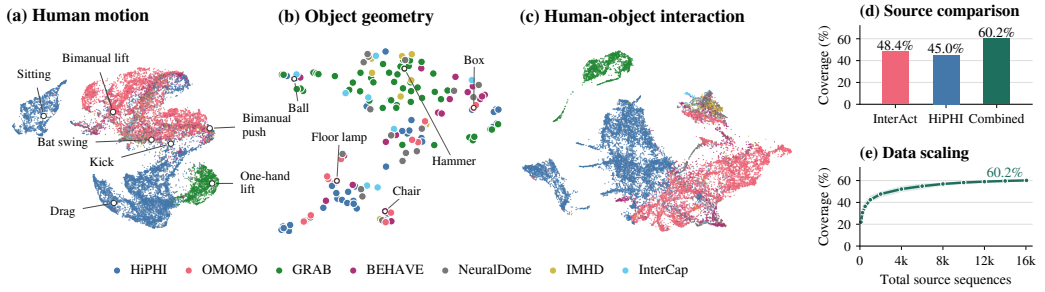

Figure 3: **Our data has broad coverage.** (a-c) UMAP embeddings of human motion, object geometry, and human-object interaction, colored by source dataset by our collection. (d) How well our collection cover external motions from GRAIL (Xie et al., 2026), MOVIN (Jang et al., 2023), and ELMO (Jang et al., 2024); a motion is covered when one of our sequences is close to it (Sec. E.1). (e) The same coverage as the number of our source sequences grows.

improves the policy's capability as it expands the motion collection, rather than merely collecting variations the initial generalist could already track.

Growing coverage over rounds. Accepted proposals become the parents of the next round, so the verified references and the tracker's skills grow together across more iterations. A proposal that fails is never edited again, so errors cannot accumulate across rounds. Each round picks a diverse set of the successful motions that reached farthest from their seed and moves each a fixed step further (Sec. D.5). Coverage is measured independently of these edit coordinates (Sec. 4.1).

4 Experiments

We evaluate four questions. (**i**) Can one retargeter preserve both long-horizon body-object relations and fine finger contact, and does this yield cleaner references than prior retargeters (Sec. 4.2)? (**ii**) How broadly can trackers execute these references across objects and embodiments, and which interactions remain hard (Sec. 4.3)? (**iii**) Does growing the data together with the tracker expand what it can execute beyond the original references, and does this growth come from learning rather than from filtering more rollouts (Sec. 4.4)? (**iv**) Do the resulting motions carry over from simulation to real humanoid and wheeled robots (Sec. 4.5)?

4.1 Experimental setup

Data and embodiments. We consolidate InterAct (Xu et al., 2025c) and HiPHI (Ji et al., 2026) into a collection of 16,059 motions, 140.68 reference hours, and 157 object entries (Table 4), which covers a wide range of everyday interactions (Figure 3). We retarget it to six robot configurations: G1 (Unitree Robotics, b) with non-dexterous, Inspire (Inspire Robots), or Dex3 (Unitree Robotics, a) hands, Booster T1 and K1 (Booster Robotics, b;a), and Dexmate Vega (Dexmate). Sec. A.1 details which experiments run on each embodiment.

Metrics. Each experiment uses its own group of metrics, and Sec. E defines each of them once. Retargeted references are scored on physical plausibility and fidelity to the human motion: hand-contact preservation, in-hand slip, penetration, foot sliding, MPJPE, and joint acceleration. Tracking is scored by the success rate, the fraction of clips executed to the end without a fall or a large body, object, or interaction error, and by body and object errors on successful clips. For the augmentation loop, we report the growth of the verified reference set, the success of the tracker before and after self-evolution on the generated and original references, and the quality of their executions.

Baselines. For dexterous retargeting, we compare with Weave (Cao et al., 2026b), whose released retargeted references use the same G1 with Inspire hands. For retargeting without dexterous hands, we compare with OmniRetarget (Yang et al., 2026), UMR (Cao et al., 2026a), and ULTRA (He et al., 2026) on G1 with non-dexterous hands. For tracking, no publicly available tracker is trained on an interaction collection of comparable scale, so we compare our generalist with per-object specialists, each trained only on the references of one object, which serve as a practical upper bound. For the augmentation loop, we compare with the scene-scaling augmentation of ULTRA (He et al., 2026),

Table 1: **Quantitative comparison on retargeting** on G1 with Inspire hands, on the 3,059 OMOMO clips. Ours show better motion quality. Metrics are defined in Sec. E.2.

	Hand contact			Penetration			Motion			Jnt. acc. (rad/s^2)↓			Jnt. limit (%)↓	
	Frame (%)↑	Hand (%)↑	Slip (m/s)↓	Obj. (%)↓	Depth (cm)↓	Gnd. (%)↓	Foot sl. (%)↓	MPJPE (cm)↓	EE (cm)↓	All	Body	Fing.	Body	Fing.
Weave	96.1	97.1	**0.225**	67.4	2.62	31.9	62.0	9.54	12.6	12.2	6.3	19.3	0.6	33.4
Ours	**98.2**	**98.8**	0.239	**37.2**	**2.39**	**0.0**	**12.3**	**5.74**	**5.3**	**2.8**	**3.2**	**2.3**	**0.0**	**0.0**

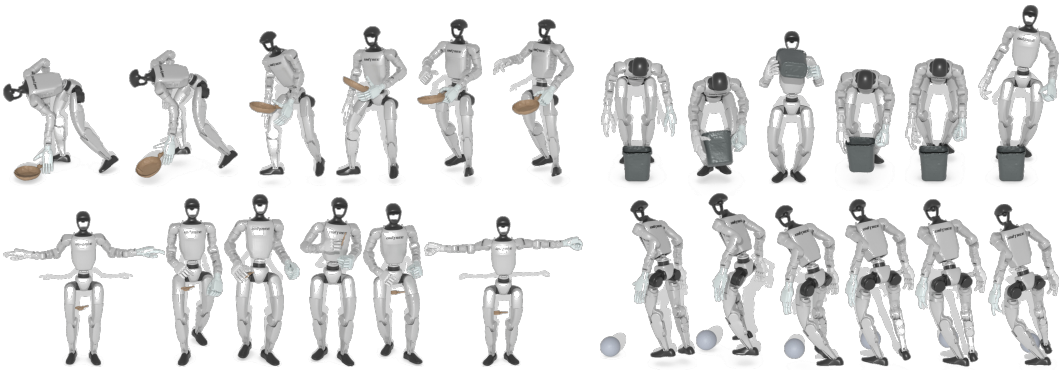

Figure 4: **Qualitative evaluation on retargeting.** High-quality results on representative tasks.

which also expands the training references of a humanoid tracker (Sec. F.4), and we isolate the effectiveness of augmentation with two controls: the base tracker before any augmentation and a frozen tracker that only filters new candidates without finetuning.

4.2 Retargeting quality and dexterous contact

Our retargeter produces cleaner dexterous references. Table 1 compares our retargeting with Weave on G1 with Inspire hands. Our references keep hand contact slightly better, greatly reduce object penetration, foot sliding, and joint acceleration, and stay closer to the raw motion, at the cost of slightly higher in-hand slip. We attribute this to explicitly shaping the grasp and then additionally refining contact in a second cleanup, so better hands do not come at the cost of the body.

Dexterous hands keep grasps that rigid hands can only approximate. Figures 10 and 11 compare against OmniRetarget (Yang et al., 2026) and UMR (Cao et al., 2026a) on shared source motions. With articulated fingers, our references preserve the grasp and hand-object relationship of the demonstration by explicit grasp modeling. When all methods use the same non-dexterous G1 (Sec. F.1), each favors a different trade-off: ours is the smoothest and most accurate at the end effectors, OmniRetarget penetrates least, and UMR preserves the most contact.

One formulation retargets across embodiments and objects. Figures 8, 4, and 9 show a grasp retargeted to two dexterous hands, interactions with objects of different shapes, and one interaction retargeted to all six embodiments. The formulation is shared across robots: each configuration only declares its own landmarks and hand mappings, so a new robot needs only its configuration.

4.3 Physics-based interaction tracking

Table 2 reports tracking on our constructed benchmark over 15 objects, grouped into bimanual, sitting, and grasping interactions. Bimanual interactions are tracked most reliably on every G1 configuration. Sitting and grasping are harder, since sitting shifts the body's weight onto the object and grasping requires finger contact on small object regions. The Booster K1 specialist trails the G1 specialists on Bimanual interactions. A single policy trained on the whole collection tracks all 15 benchmark objects (Table 2, bottom). Its body error stays close to that of the specialists, whereas its success rate and object rotation error fall behind, most on grasping. Sharing one policy across objects costs mainly object control rather than body tracking, and fine grasps are where per-object capacity matters.

Table 2: **Quantitative results on motion tracking.** One generalist nearly matches the per-object specialists in body and object position error, while trailing them in success rate and object rotation. Metrics are defined in Sec. E.3; – marks tasks the embodiment cannot perform.

	Bimanual				Sitting				Grasping			
Embodiment	SR↑	MPJPE↓	Obj.↓	Rot.↓	SR↑	MPJPE↓	Obj.↓	Rot.↓	SR↑	MPJPE↓	Obj.↓	Rot.↓
Specialists (one policy per object)												
G1 + Inspire	93.8	5.19	5.79	10.4	75.4	5.94	4.76	7.4	84.9	5.30	7.17	14.1
G1 + Dex3-1	93.1	4.92	7.02	11.0	82.8	5.36	5.01	6.5	83.9	5.14	7.32	12.7
G1	94.4	4.52	5.88	8.5	82.8	4.93	5.26	6.2	–	–	–	–
K1	75.7	6.21	7.35	14.3	75.4	6.41	5.08	8.3	–	–	–	–
Generalist (one policy for all objects)												
G1 + Inspire	73.2	5.86	6.53	23.4	64.9	6.14	5.94	7.8	58.3	5.95	7.63	35.5

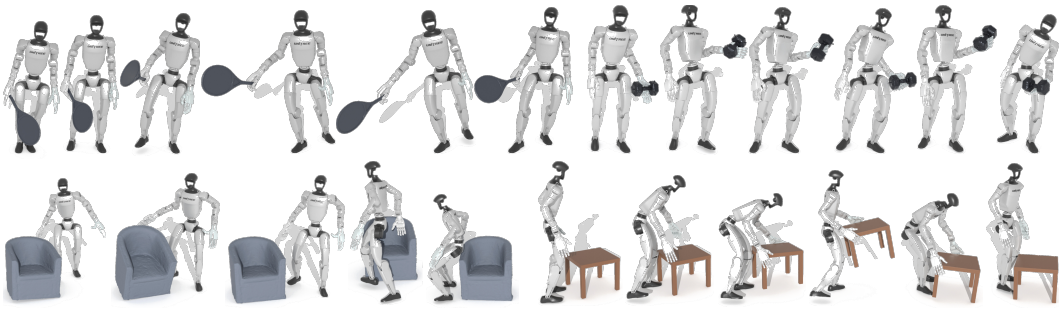

Figure 5: **Physics-based tracking across objects.** Executed rollouts on G1 with Inspire hands for four objects: a racket, a dumbbell, an armchair, and a table.

Figures 5, 12, and 13 show executed rollouts across objects, across embodiments, and with two dexterous hands.

4.4 Does iterative generation expand learned competence?

We run five rounds of self-evolving motion imitation on bimanual and grasping interactions with G1 and Inspire hands, and on bimanual interactions with G1 and Dex3 hands, grouped as in the tracking benchmark (Sec. 4.3). Each round applies spatial edits, which move and turn where the object is placed relative to the seed's own execution, and body edits, which change the posture that realizes the interaction; later rounds edit only references verified in earlier rounds. We ask three questions: are the augmented references of good quality, does each round expand their range, and does training on them improve the tracker?

Augmented references stay close to the originals in quality. Averaged over all rounds, executions of augmented references stay close to those of the original references in body joint acceleration and hand jitter, while foot sliding rises from about 1% to about 2% of contact frames (quality columns of Table 3). All three measures rise moderately in later rounds, which by design accept edits farther from the captures (Sec. F.3).

Each round expands the set of executable references. Every setting keeps accepting new spatial and body references: the verified reference set grows about **twentyfold** in the first round and more than a **hundredfold** by the fifth (growth column of Table 3 and Figure 6; per-round values in Sec. F.3).



Figure 6: **A single capture grows into many executable variants.** Each group shows the original reference followed by the verified augmentations accumulated by round 1 and by round 5.

Table 3: Over five rounds, the verified references grow **more than a hundredfold**, and the evolved tracker executes nearly all of them without losing the original ones. Growth: verified references relative to the original references. Success: rate at which the tracker before (Orig.) and after (Evol.) self-evolution. Quality: executions of original (Orig.) and augmented (Aug.) references.

	Growth (×)	Success (%)				Quality, Orig. → Aug.		
		Generated		Original		Body acc.	Foot sl.	Hand jit.
Setting	R1 → R5	Orig.	Evol.	Orig.	Evol.	(rad/s^2)↓	(%)↓	(mm)↓
Inspire, Bimanual	21.8 → 150.5	52.1	**98.4**	100	100	11.4 → 12.4	1.18 → 1.87	1.00 → 1.17
Dex3, Bimanual	21.3 → 142.0	59.3	**98.7**	100	100	13.9 → 16.0	1.14 → 2.21	1.12 → 1.30
Inspire, grasping	19.6 → 146.4	64.2	**98.9**	100	100	13.6 → 15.3	0.70 → 2.05	0.96 → 1.14

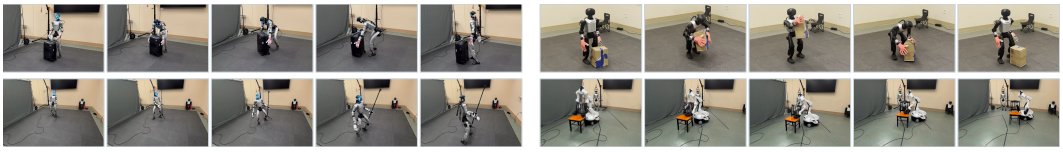

Figure 7: **Real-world execution.** We show the motion from our paradigm transfer to different embodiments, from Unitree G1 pulling a suitcase and G1 with Inspire hands carrying a tripod (left); Booster K1 lifting and placing a box and Dexmate Vega relocating a chair (right).

Body edits also widen the range along another axis, posture: their references spread over a region of body motion several times wider than the scene-scaling variants of ULTRA (He et al., 2026) (Sec. F.4). This expansion depends on learning. A frozen tracker that only filters candidates accepts almost none of the new candidates after the first round, because the new candidates start from first-round successes that already lie beyond its reach (Sec. F.2).

Self-evolution makes the tracker better. We replay every accepted reference and every original reference with the tracker before and after self-evolution, counting a reference as executed when at least one of three rollouts reaches its end (success columns of Table 3). The original tracker fails on more than a third of the accepted references even though it is trained with large-scale data, whereas the evolved tracker executes nearly all of them, and both execute every original reference. Self-evolution therefore extends what the tracker can execute without losing what it could already do.

4.5 Cross-embodiment execution and real-world transfer

To verify that the trajectories produced by InterMimicGen work beyond simulation, we execute them on real robots. Each embodiment uses the parts of the framework it supports (Sec. A.1), and each platform runs its own low-level controller, trained by us or publicly available (Sec. G). These controllers are blind to the object and track only the robot motion, yet our trajectories execute well on every platform, suggesting that they are physically consistent enough to transfer without object feedback. Figure 7 shows box lifting and placement, suitcase pulling, tripod carrying, and chair relocation, with more videos on the *supplementary website*.

5 Discussion and Conclusion

InterMimicGen treats a humanoid tracker as both a consumer and a producer of interaction data. A consolidated human-object motion collection supplies diverse task semantics, hand-aware retargeting transfers them to six robot configurations, and physics-based tracking grounds them in executable motion. The augmentation-learning loop then turns validated variations into new references. Every round enlarges the archive, and the fine-tuned tracker learns interactions beyond the base tracker's reach while keeping its seed skills. Learning is what keeps the loop going, since a frozen tracker stops accepting new variations after the first round.

Limitations. The loop varies how an interaction is realized, so it densifies coverage around each demonstration without creating new task semantics. Iterative selection may also favor easier objects or compound reference errors.

AI USE STATEMENT

Generative AI assisted with refining this manuscript and preparing the LaTeX source and placeholder layouts. The authors verify the final technical descriptions, citations, implementation correspondence, and empirical claims.

REPRODUCIBILITY STATEMENT

The appendix specifies the reference format, retargeting objectives, tracker configuration, augmentation protocol, and evaluation metrics. We will release source-provenance manifests, robot mappings, model configurations, the lineage of accepted trajectories, and scripts to reproduce all comparisons, following the redistribution terms of each source dataset.

REFERENCES

Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics*, 39 (4):62, 2020.

Joel A E Andersson, Joris Gillis, Greg Horn, James B Rawlings, and Moritz Diehl. CasADi – A software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation*, 11(1):1–36, 2019. doi: 10.1007/s12532-018-0139-4.

João Pedro Araújo, 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.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *CVPR*, 2022.

Booster Robotics. Booster K1. Official product page, a. URL https://www.booster.tech/booster-k1/.

Booster Robotics. Booster T1. Official product page, b. URL https://www.booster.tech/booster-t1/.

Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Manon Devin, Alex X. Lee, Maria Bauzá Villalonga, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, Antoine Laurens, Claudio Fantacci, Valentin Dalibard, Martina Zambelli, Murilo Fernandes Martins, Rugile Pevceviciute, Michiel Blokzijl, Misha Denil, Nathan Batchelor, Thomas Lampe, Emilio Parisotto, Konrad Zolna, Scott E. Reed, Sergio Gómez Colmenarejo, Jon Scholz, Abbas Abdolmaleki, Oliver Groth, Jean-Baptiste Regli, Oleg Sushkov, Thomas Rothörl, José Enrique Chen, Yusuf Aytar, Dave Barker, Joy Ortiz, Martin A. Riedmiller, Jost Tobias Springenberg, Raia Hadsell, Francesco Nori, and Nicolas Heess. RoboCat: A self-improving generalist agent for robotic manipulation. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=vsCpILiWHu.

Hanyang Cao, Yuetong Fang, Taesoo Kwon, Runyi Yu, Ji Ma, Jing Tan, Yangchen Zhou, Baoze Du, Yi Gu, Yukang Gao, Ruoli Dai, Lei Han, and Renjing Xu. Unified motion retargeting for humanoids with learned point cloud correspondence. *arXiv preprint arXiv:2609.02134*, 2026a.

Liu Cao, Xingze Wu, Jingzhi Cui, Botian Xu, Mingzhi Pei, Ruoqu Chen, and Mengdi Xu. Weave: Learning whole-body dexterous loco-manipulation from human-object interactions. *arXiv preprint arXiv:2609.16683*, 2026b.

Jeremy Dao, Helei Duan, and Alan Fern. Sim-to-real learning for humanoid box loco-manipulation. In *ICRA*, 2024.

Pranav Debbad, Kanish Thiagarajan, Victor Dhédin, Shafeef Omar, and Majid Khadiv. Humanoid-DART: Humanoid loco-manipulation using diffusion-guided augmentation through relabeling and tracking. *arXiv preprint arXiv:2606.26855*, 2026.

Dexmate. Vega. Official product page. URL https://www.dexmate.ai/product/vega.

Pengxiang Ding, Jianfei Ma, Xinyang Tong, Binghong Zou, Xinxin Luo, Yiguo Fan, Ting Wang, Hongchao Lu, Panzhong Mo, Jinxin Liu, Yuefan Wang, Huaicheng Zhou, Wenshuo Feng, Jiacheng Liu, Siteng Huang, and Donglin Wang. Humanoid-VLA: Towards universal humanoid control with visual integration. *arXiv preprint arXiv:2502.14795*, 2025.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023.

Yuhui Fu, Feiyang Xie, Chaoyi Xu, Jing Xiong, Haoqi Yuan, and Zongqing Lu. DemoHLM: From one demonstration to generalizable humanoid loco-manipulation. *IEEE Robotics and Automation Letters*, 11(4):4393–4400, 2026. doi: 10.1109/LRA.2026.3666395.

Caelan Reed Garrett, Ajay Mandlekar, Bowen Wen, and Dieter Fox. SkillMimicGen: Automated demonstration generation for efficient skill learning and deployment. In *CoRL*, 2024.

Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan D. Ratliff, and Dieter Fox. DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *ICRA*, 2020.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris M. Kitani, Changliu Liu, and Guanya Shi. OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *CoRL*, 2024.

Xialin He, Sirui Xu, Xinyao Li, Runpei Dong, Liuyu Bian, Yu-Xiong Wang, and Liang-Yan Gui. ULTRA: Unified multimodal control for autonomous humanoid whole-body loco-manipulation. In *IROS*, 2026.

Ryan Hoque, Ajay Mandlekar, Caelan Reed Garrett, Ken Goldberg, and Dieter Fox. IntervenGen: Interventional data generation for robust and data-efficient robot imitation learning. In *IROS*, 2024.

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction from multi-view RGB-D images. *International Journal of Computer Vision*, 132(7):2551–2566, 2024.

Inspire Robots. The Dexterous Hands. Official product catalog. URL https://en.inspire-robots.com/product-category/the-dexterous-hands/.

Deok-Kyeong Jang, Dongseok Yang, Deok-Yun Jang, Byeoli Choi, Taeil Jin, and Sung-Hee Lee. MOVIN: Real-time motion capture using a single LiDAR. *Computer Graphics Forum*, 42(7): e14961, 2023.

Deok-Kyeong Jang, Dongseok Yang, Deok-Yun Jang, Byeoli Choi, Sung-Hee Lee, and Donghoon Shin. ELMO: Enhanced real-time LiDAR motion capture through upsampling. *ACM Transactions on Graphics*, 2024.

Jiahao Ji, Ji Ma, Runhan Zhang, Runyi Yu, Wenjia Wang, Weiheng Chi, Qianqian Peng, Weichao Yan, Yongfei Gu, Ye Tian, Ting Wu, Longwei Li, Chun Yuan, Ruoli Dai, and Lei Han. HiPHI: A large-scale benchmark for high-precision human motion and object-interaction. In *CoRL*, 2026.

Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Jim Fan, and Yuke Zhu. DexMimicGen: Automated data generation for bimanual dexterous manipulation via imitation learning. In *ICRA*, 2025.

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.

Yeonjoon Kim, Hangil Park, Seungbae Bang, and Sung-Hee Lee. Retargeting human-object interaction to virtual avatars. *IEEE Transactions on Visualization and Computer Graphics*, 2016.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics*, 2023.

Kevin Lin, Ajay Mandlekar, Caelan Reed Garrett, Nikita Chernyadev, Yu Fang, Runyu Ding, Yuqi Xie, Justin Tran, Linxi Fan, and Yuke Zhu. HumanoidMimicGen: Data generation for loco-manipulation via whole-body planning. *arXiv preprint arXiv:2605.27724*, 2026.

Fukang Liu, Zhaoyuan Gu, Yilin Cai, Ziyi Zhou, Hyunyoung Jung, Jaehwi Jang, Shijie Zhao, Sehoon Ha, Yue Chen, Danfei Xu, and Ye Zhao. Opt2Skill: Imitating dynamically-feasible whole-body trajectories for versatile humanoid loco-manipulation. *IEEE Robotics and Automation Letters*, 2025a.

Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. Core4D: A 4D human-object-human interaction dataset for collaborative object rearrangement. In *CVPR*, 2025b.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. HUMOTO: A 4D dataset of mocap human object interactions. In *ICCV*, 2025.

Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. *Science Robotics*, 10(105):eads5033, 2025. doi: 10.1126/scirobotics.ads5033.

Zhengyi Luo, Jinkun Cao, Alexander Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *ICCV*, 2023.

Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Omnigrasp: Grasping diverse objects with simulated humanoids. In *NeurIPS*, 2024.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, Jinhyung Park, David Sami, Zi Wang, Xingye Da, Runyu Ding, Cyrus Hogg, Lina Song, Edy Lim, Eugene Jeong, Tairan He, Haoru Xue, Wenli Xiao, Simon Yuen, Jan Kautz, Yan Chang, Umar Iqbal, Linxi Jim Fan, and Yuke Zhu. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026. doi: 10.1126/scirobotics.aed4592.

Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, and Xiaokang Yang. HIMO: A new benchmark for full-body human interacting with multiple objects. In *ECCV*, 2024.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. MimicGen: A data generation system for scalable robot learning using human demonstrations. In *CoRL*, 2023.

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.

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*, 37(4):1–14, 2018.

Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. AnyTeleop: A general vision-based dexterous robot arm-hand teleoperation system. In *RSS*, 2023.

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

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

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.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020.

Chen Tessler, Yifeng Jiang, Erwin Coumans, Zhengyi Luo, Gal Chechik, and Xue Bin Peng. Masked-Manipulator: Versatile whole-body manipulation. In *SIGGRAPH Asia*, 2025.

Unitree Robotics. Unitree Dex3-1. Official product page, a. URL https://www.unitree.com/Dex3-1/.

Unitree Robotics. Unitree G1. Official product page, b. URL https://www.unitree.com/g1/.

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

Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, and Jun Saito. Contact-aware retargeting of skinned motion. In *ICCV*, 2021.

Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

Huayi Wang, Wentao Zhang, Runyi Yu, Tao Huang, Junli Ren, Feiyu Jia, Zirui Wang, Xiaojie Niu, Xiao Chen, Jiahe Chen, Qifeng Chen, Jingbo Wang, and Jiangmiao Pang. PhysHSI: Towards a real-world generalizable and natural humanoid-scene interaction system. *arXiv preprint arXiv:2510.11072*, 2025a.

Jiashun Wang, Yifeng Jiang, Haotian Zhang, Chen Tessler, Davis Rempe, Jessica Hodgins, and Xue Bin Peng. HIL: Hybrid imitation learning for dynamic athletic control. *ACM Transactions on Graphics*, 2026. doi: 10.1145/3829364.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. PhysHOI: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023.

Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, Jian Zhang, Lei Zhang, and Ping Tan. SkillMimic: Learning basketball interaction skills from demonstrations. In *CVPR*, 2025b.

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.

Jungdam Won, Deepak Gopinath, and Jessica Hodgins. A scalable approach to control diverse behaviors for physically simulated characters. *ACM Transactions on Graphics*, 39(4):33, 2020. doi: 10.1145/3386569.3392381.

Tianyi Xie, Haotian Zhang, Jinhyung Park, Zi Wang, Bowen Wen, Jiefeng Li, Xueting Li, Qingwei Ben, Haoyang Weng, Yufei Ye, David Minor, Tingwu Wang, Chenfanfu Jiang, Sanja Fidler, Jan Kautz, Linxi Fan, Yuke Zhu, Zhengyi Luo, Umar Iqbal, and Ye Yuan. GRAIL: Generating humanoid loco-manipulation from 3D assets and video priors. *arXiv preprint arXiv:2606.05160*, 2026.

Michael Xu, Yi Shi, KangKang Yin, and Xue Bin Peng. PARC: Physics-based augmentation with reinforcement learning for character controllers. In *SIGGRAPH*, 2025a.

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*, 2025b.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. InterAct: Advancing large-scale versatile 3D human-object interaction generation. In *CVPR*, 2025c.

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*, 2025d.

Sirui Xu, Samuel Schulter, Morteza Ziyadi, Xialin He, Xiaohan Fei, Yu-Xiong Wang, and Liang-Yan Gui. InterPrior: Scaling generative control for physics-based human-object interactions. In *CVPR*, 2026a.

Weisheng Xu, Qiwei Wu, Jiaxi Zhang, Jing Tan, Yangfan Li, Yuetong Fang, Jiaqi Xiong, Kai Wu, Rong Ou, and Renjing Xu. Iterative closed-loop motion synthesis for scaling the capabilities of humanoid control. In *CVPR*, 2026b.

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. In *ICRA*, 2026.

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. *IEEE Robotics and Automation Letters*, 11(7):8124–8131, 2026. doi: 10.1109/LRA.2026.3692091.

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.

Yanjie Ze, Zixuan Chen, João Pedro Araújo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, and C. Karen Liu. TWIST: Teleoperated whole-body imitation system. In *CoRL*, 2025.

Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. NeuralDome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023a.

Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez-Pellitero, and Gerard Pons-Moll. FORCE: Physics-aware human-object interaction. In *3DV*, 2025.

Yunbo Zhang, Deepak Gopinath, Yuting Ye, Jessica Hodgins, Greg Turk, and Jungdam Won. Simulation and retargeting of complex multi-character interactions. In *SIGGRAPH*, 2023b.

Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'M HOI: Inertia-aware monocular capture of 3D human-object interactions. In *CVPR*, 2024.

Mandi Zhao, Yifan Hou, Dieter Fox, Yashraj Narang, Ajay Mandlekar, and Shuran Song. DexMachina: Functional retargeting for bimanual dexterous manipulation. In *ICML*, 2026.

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.

The supplementary material provides implementation details and extended evaluations that complement the main paper. Sec. A describes the data and embodiment scope, together with additional retargeting details in Sec. B. Sec. C specifies the tracker's inputs and outputs, training setup, reference sampling, and rollout termination rules used to train and evaluate the generalist policy. Sec. D details the augmentation protocol, including spatial edits, body edits, and how the tracker is fine-tuned over rounds. Sec. E defines the evaluation metrics. Sec. F reports additional retargeting and augmentation experiments, and Sec. G documents the hardware protocol and details on real-world deployment.

We also provide an anonymous supplementary site in the supplementary material which is self-contained and follows the structure of the paper: embodiment transfer and retargeting comparisons, generalist tracking, iterative augmentation with spatial and body edits, and hardware execution. Each video is labeled with its embodiment, method, and generation round, and marks simulation or hardware. Comparison videos are synchronized by interaction phase, and augmentation videos show the edit next to the achieved motion.

A Data and embodiments

All embodiments draw on one consolidated collection of 16,059 sequences, 9,030 from InterAct and 7,029 from HiPHI. Every retargeted reference $\boldsymbol{\xi}$ provides the robot state, the object pose and twist, and the contact intent, so tracking and augmentation use the same format on every robot.

Table 4: **Reference collection** for G1 with Inspire hands. Object entries are dataset-object pairs, so an object that appears in two source datasets counts twice. Reference hours sum the frames of the selected robot references at 50 fps and may include overlapping source segments.

Source	Object entries	Motions	Reference h
InterAct (Xu et al., 2025c)	117	9,030	17.82
HiPHI (Ji et al., 2026)	40	7,029	122.86
Consolidated	157	16,059	140.68

A.1 Embodiment scope

Table 5 lists the studies run on each platform after the capability filtering of Sec. 3.1. Generalist training and iterative augmentation are evaluated mainly on G1 with dexterous hands, and the other platforms use the subsets they can perform. Motions a platform cannot support are excluded from its evaluation.

Table 5: Studies run on each platform. Configurations with dexterous hands use all data, including the dexterous benchmark; the generalist is trained on G1 with Inspire hands.

Platform	Retargeting	Specialist tracking	Generalist	Augmentation loop	Sim-to-real	Excluded tasks
G1	✓	✓	–	–	✓	dexterous
G1 + Inspire	✓	✓	✓	✓	✓	none
G1 + Dex3-1	✓	✓	–	✓	–	none
Booster T1	✓	–	–	–	–	dexterous
Booster K1	✓	✓	–	–	✓	dexterous
Dexmate Vega	✓	–	–	–	✓	dexterous, seated

B Retargeting

The retargeter of Sec. 3.2 runs two stages (Algorithm 1): an interaction-preserving whole-body solve that includes finger opposition, followed by a two-pass geometry cleanup that preserves contact.

Variables. The output is the robot trajectory $\boldsymbol{q}_{1:T}$, where $\boldsymbol{q}_t = (\boldsymbol{q}_t^{\text{root}}, \boldsymbol{q}_t^{\text{body}}, \boldsymbol{q}_t^{\text{hand}})$ holds the floating-base pose, the body joints, and the independent finger joints at frame t. The whole-body

Algorithm 1 Retargeting one interaction clip to a target robot.

Require: scaled human motion and object trajectory $\boldsymbol{x}^o_{1:T}$; robot configuration (landmarks, thumb-finger pairs $\mathcal{F}_e$, coupling maps)
Ensure: robot trajectory $\boldsymbol{q}_{1:T}$ and object trajectory $\boldsymbol{x}^o_{1:T} + \Delta\boldsymbol{x}^o_{1:T}$
Stage 1: interaction-preserving whole-body solve with finger opposition
for $t = 1, \ldots, T$ **do**
  update the opposition switches $g_{t,f}$ from finger-object distances and reference contact labels
  $\boldsymbol{q}_t \leftarrow \arg\min E_{\text{coarse}}$ (Eq. 4) under joint, support, and collision constraints, with $\boldsymbol{x}^o_t$ fixed and warm-started from $\boldsymbol{q}_{t-1}$
    try the solvers in turn and keep the first solution that passes the feasibility checks
end for
Stage 2: geometry cleanup with contact preservation
for pass $\in$ (penetration repair, collision-aware smoothing) **do**
  repeat
    Adam on $(\boldsymbol{q}_{1:T}, \Delta\boldsymbol{x}^o_{1:T})$ with the pass objective (Table 6), then projection onto joint limits
    apply the acceptance check of the pass to the complete candidate against the pass input (Table 7)
  until the candidate passes, or all retry settings have been tried
end for
return $\boldsymbol{q}_{1:T}$, $\boldsymbol{x}^o_{1:T} + \Delta\boldsymbol{x}^o_{1:T}$

solve updates $\boldsymbol{q}_t$ frame by frame and keeps the source object trajectory $\boldsymbol{x}^o_{1:T}$ fixed. The cleanup updates the whole trajectory $\boldsymbol{q}_{1:T}$ together with small object-pose corrections $\Delta\boldsymbol{x}^o_{1:T}$.

Interaction-preserving whole-body solve. The whole-body stage transfers body-object structure, rather than joint poses alone, from the scaled human reference to the robot. It builds an interaction mesh over body landmarks and sampled object geometry and minimizes

$$E_{\text{coarse}} = \lambda_I E_{\text{interaction}} + \lambda_C E_{\text{cover}} + E_{\text{temporal}} + E_{\text{posture}}, \tag{4}$$

subject to joint, support, and collision constraints, with the object trajectory fixed. The interaction term $E_{\text{interaction}}$ matches the Laplacian coordinates of the robot-object mesh to the scaled source interaction (Yang et al., 2026), while the temporal and posture terms regularize frame-to-frame motion and keep the posture plausible. We set $\lambda_I = 10$, uniform over the mesh vertices. E_{temporal} penalizes the squared change of each coordinate of $\boldsymbol{q}_t$ from the previous frame's solution, with weight 2 for the floating base, legs, and waist and 1 for the arm and finger joints. E_{posture} penalizes the waist yaw and roll angles with weight 0.2 each. Sparse hand landmarks place the wrist and palm, and the opposition term E_{cover} of Eq. 1 shapes the fingers, with weight $\lambda_C = 3$ and margin $m = 1$. Because the cosine between two directions is at least -1, this margin keeps the term active for every active pair, so the term pulls each pair toward full opposition.

When opposition is active. The switch $g_{t,f}$ turns the opposition term on only when a grasp is forming. It activates when the thumb and at least one other finger are within 10 cm of the object, holds for 15 frames, and resets only after the fingers move more than 15 cm away. Contact labels in the reference can also trigger it.

Solvers. Different clips fail under different local constraints and solvers, so we solve the same interaction-mesh objective with several constrained solvers, tried in turn: MOSEK, Clarabel, ECOS, CVXOPT, OSQP, and SCS, plus box-bound variants of Clarabel, OSQP, and SCS that replace the local Euclidean step bound with componentwise bounds. The coarse stage is solved frame by frame, warm-started from the previous frame. The first solution that passes the feasibility checks initializes the cleanup stage.

Surface correspondences. The opposition term needs an object-surface point for every active fingertip. For fingertip landmark $\boldsymbol{p}_{t,f}$ and selected object-surface point $\boldsymbol{s}_{t,f}$ on the object's collision geometry, we define $\boldsymbol{v}_{t,f} = \boldsymbol{p}_{t,f} - \boldsymbol{s}_{t,f}$ and the direction $\boldsymbol{n}_{t,f} = \boldsymbol{v}_{t,f}/\|\boldsymbol{v}_{t,f}\|$ of Eq. 1. The term depends on the robot configuration through the wrist and hand forward kinematics and is purely geometric: it compares the thumb direction with each active opposing fingertip direction, which pushes the fingers to contact the object from complementary sides.

Actuation mappings. One hand formulation serves different robots because each configuration declares its own kinematics to the shared objective. A dependent joint, such as a coupled finger joint of the Inspire hands, is written as $q_j = \alpha_j q_{i(j)} + \beta_j$ and reconstructed inside the differentiable hand

Table 6: **Cleanup objective weights** (default settings). "acc/jerk": the weight applies to both. ✓: term used, weight unlisted.

Term	Repair	Smoothing
Collision		
Robot-object penetration	3000	3000 (new)
Robot-floor penetration	3000	3000 (new)
Object-floor penetration	5000	5000
Hand self-collision (new)	300	300
Fidelity and contact		
Trajectory deviation	✓	1
Body-landmark drift	2	2
Planted-foot drift	30	10
Joint limits	1	1
Contact keeping	✓	20
Hand pose rel. object	–	0.5
Smoothness		
Body joints acc/jerk	20	40
Hand joints acc/jerk	20	80
Root translation / rotation	4000 (acc)	4000 / 300
Object translation / rotation	8000 (acc)	16000 / 300
Body landmarks	2000 (acc)	1000
Edit roughness	800	–

Table 7: **Cleanup acceptance checks.** Repair is measured against the whole-body solution, smoothing against the repaired trajectory.

Maximum change	Repair	Smoothing
Root position	5 cm	5 cm
Body joint	0.5 rad	0.5 rad[a]
Hand joint	1.5 rad	1.5 rad
Object position	6 cm	4 cm
Object rotation	15°	25°
Non-hand body position	15 cm	12 cm
Hand body / root endpoint	–	20 / 2 cm
Object penetration	+3 mm	+2 mm
Non-foot floor penetration	+3 mm[b]	+2 mm
Jerk metrics	3×	1.5×
Fingertip-aperture acc.	–	+0.02
Retained contact frames	≥50%	≥60%

[a] RMS ≤ 0.08 rad. [b] Above $\max(\text{input}, 3\,\text{mm})$.

model before forward kinematics. Its derivative is folded into the Jacobian of the independent joint, $J_{i(j)} \leftarrow J_{i(j)} + \alpha_j J_j$, for the landmark, collision, and opposition terms. Each configuration also declares its body and fingertip landmarks and its supported thumb-finger pairs $\mathcal{F}_e$. Hands with fewer fingers use the pairs they have, and robots without articulated fingers skip the finger terms.

Geometry cleanup with contact preservation. The whole-body solution can still leave millimeter-scale penetration and jitter. The cleanup optimizes the robot joints, the floating base, and small object-pose corrections with two successive Adam passes, each followed by projection onto joint limits. Both passes are strongly regularized toward their input. Table 6 lists the default weights. Penetration repair runs 300 Adam and 60 projection steps. It penalizes robot-object penetration with a 2 mm clearance margin, robot-floor and object-floor penetration, and hand self-collision beyond the overlap already present in its input, while fidelity and smoothness terms keep the edit small and free of new roughness. Collision-aware smoothing then runs 400 Adam and 60 projection steps. It penalizes acceleration and jerk of the body, hands, root, object, and body landmarks, keeps the collision penalties so that smoothing cannot reintroduce penetration, and weakens data fidelity to 0.02 at frames with sudden hand jumps so that these jumps can be smoothed away. A pass that fails is retried with adjusted weights and iteration counts.

How contact is preserved. Two terms protect contact. In both passes, a contact-keeping term penalizes robot points that were near contact in the input and move more than 2 cm away from the object. Smoothing also keeps the hand-base pose fixed relative to the object at frames in contact, detected as an object distance below 1.5 cm, with a relative-pose scale of 2 cm. These terms preserve and stabilize existing contact but do not attract fingertips to new surface points. Grasp shape comes from the opposition term of the whole-body solve.

Acceptance checks. Each pass accepts a candidate only after the complete trajectory passes its check against that pass's input: the repair check compares with the whole-body solution, and the smoothing check compares with the repaired trajectory (Table 7). The checks bound how far the root, joints, object, and body may move, how much penetration may grow, and how many contact frames must survive, which applies when the input has at least five. Repair must also improve penetration: if the input's maximum penetration exceeds 3 mm, the output must reach at most 3 mm or cut that maximum by at least 25%. Smoothing must not worsen the aggregate jerk score. Jerk limits allow each metric to grow by the listed factor plus a small fixed margin. Both checks reject nonfinite results,

quaternion-norm errors of at least 10^{-6}, new joint-limit violations, and objects that sink more than 5 mm below their starting height or the floor.

Retargeting to wheeled robots. For wheeled embodiments such as Dexmate Vega we use, we keep the interaction-mesh objective and adapt the configuration space and landmark correspondences. The planar base translation and yaw are optimized jointly with the upper body, while the chassis height stays at its nominal value and the base roll and pitch are fixed at zero. Human leg and foot landmarks are removed from the interaction mesh and the foot-contact constraints are disabled, so the base, torso, and arms together preserve the demonstrated manipulation without passing changes in human pelvis height to the chassis. The three torso joints are constrained only through the mapped shoulder, elbow, wrist, and hand landmarks rather than by matching human spine angles. Temporal regularization and bounds on frame-to-frame changes of the upper-body joints and the planar base limit abrupt motion, and the wheel joints stay fixed during this kinematic stage. We discuss how the trajectories can be used by the real controller in Sec. G.

C GENERALIST INTERACTION TRACKER

The tracker π_θ of Sec. 3.3 is a privileged tracker, trained with PPO on G1 with Inspire hands using the settings in Table 8. Its observation, action, and reward define each control step, and its training and evaluation protocol decides which references it sees and when a rollout ends.

C.1 OBSERVATION, ACTION, AND REWARD

Observation. The policy observes robot proprioception, heading-local differences to upcoming reference frames, object-relative state, and body-object geometry. The geometric part encodes, for each body link i, the heading-local displacement v_i to its nearest object-surface point as a distance-decayed vector:

$$\psi(v_i) = \frac{v_i}{\|v_i\|_2 + \epsilon} \exp(-\kappa \|v_i\|_2), \qquad \epsilon = 10^{-6}, \quad \kappa = 5\,\mathrm{m}^{-1}. \tag{5}$$

The feature is computed for both the current and the reference geometry. It makes the policy sensitive to nearby interaction structure and applies unchanged to hands with different joint spaces.

Action and PD control. The policy outputs joint-position offsets a_t, which a PD controller converts to torques:

$$\tau_t = \mathrm{clip}_{\tau_{\max}} \left[K_p \big(q_{\mathrm{rest}} + s_a a_t - q_t \big) - K_d \dot{q}_t \right], \tag{6}$$

where q_t and $\dot{q}_t$ are the joint positions and velocities, q_{rest} is the rest pose, and the gains K_p and K_d, action scale s_a, and torque limit $\tau_{\max}$ are embodiment-specific. On the Inspire hands, coupled finger joints follow their driving joints through the mimic-joint mapping, which is reconstructed before simulation, instead of acting as independent actuators.

Reward. The reward multiplies body, object, interaction, and contact tracking terms, so success in one term cannot offset failure in another, and adds regularization:

$$r_t^{\mathrm{track}} = s_r r_t^{\mathrm{body}} r_t^{\mathrm{object}} r_t^{\mathrm{interaction}} r_t^{\mathrm{contact}}, \qquad r_t = r_t^{\mathrm{track}} + \lambda_t r_t^{\mathrm{regularization}}, \tag{7}$$

where s_r scales the tracking reward and λ_t weights the regularization. The body and object terms track pose and velocity, and the interaction and contact terms maintain the intended body-object relation. The interaction error compares body-to-surface vectors directly and weights close pairs most:

$$w_{ij} = \frac{\max(\|d_{ij}\|_2^2, \eta)^{-1}}{\sum_{a,b} \max(\|d_{ab}\|_2^2, \eta)^{-1}}, \qquad E_{\mathrm{int}}^{\mathrm{track}} = \frac{1}{2} \sum_{i,j} (w_{ij} + \hat{w}_{ij}) \|d_{ij} - \hat{d}_{ij}\|_2^2, \tag{8}$$

where d_{ij} and $\hat{d}_{ij}$ are the current and reference body-surface vectors, $\hat{w}_{ij}$ is the same weight computed from $\hat{d}_{ij}$, and η floors the squared distance. The regularization includes an object-energy factor that penalizes the simulated linear and angular acceleration of the object, obtained by differencing velocities at the control interval.

C.2 Training and evaluation protocol

Object-compatible adaptive sampling. Each simulated environment holds one object asset, so it samples only the references compatible with that object. Within this compatible set $\mathcal{R}_o$, reference i is drawn with probability

$$p(i \mid o) = \frac{\rho}{|\mathcal{R}_o|} + (1-\rho)\frac{[\mathrm{clip}(\bar{c}_i, c_{\min}, 1)]^{-\beta}}{\sum_{j\in\mathcal{R}_o}[\mathrm{clip}(\bar{c}_j, c_{\min}, 1)]^{-\beta}}, \quad \rho = 0.25,\ c_{\min} = 0.05,\ \beta = 0.5, \tag{9}$$

where $\bar{c}_i$ is the moving-average completion of reference i. The uniform share ρ keeps every reference in training, and the second term shifts the remaining mass toward clips the policy rarely completes.

Start frames and termination. Reference-state initialization starts training episodes at sampled frames before the terminal hold, exposing different phases of long sequences. A rollout terminates on a fall, a non-hand body-position error above 0.5 m, an object-position error above 0.5 m, a normalized interaction error above one, or a nonfinite state. Physical tests use a short initialization grace interval, and terminal-hold frames are excluded from source-duration and coverage accounting.

Contact-prioritized hand reward. The generalist scores grasp contact during reference-active hand phases. For distal pad f, let $\boldsymbol{b}_f = \boldsymbol{p}_f - \boldsymbol{u}_f$ point from its nearest object-surface sample to the pad center, and let $\Delta_f = \max(\|\boldsymbol{b}_f\|_2 - r_f, 0)$ be the gap beyond its collision-mesh-derived effective radius r_f. We compute proximity $P_f = \exp[-(\Delta_f/\sigma)^2]$, contact proxy $C_f = S_f[1 - \Delta_f/d_c]_0^1$, and thumb-finger opposition $O_f = [(1 - \hat{\boldsymbol{b}}_h^\top \hat{\boldsymbol{b}}_f)/2]_0^1$, where h denotes the thumb, $\sigma = 0.05\,\mathrm{m}$, and $d_c = 0.025\,\mathrm{m}$. Here S_f thresholds the sum of absolute simulator contact-force components on the distal link at 0.1; proximity filtering reduces credit for contact away from the target object but does not identify the contacting body. For non-thumb fingers f, define

$$A = \max_f \sqrt{P_h P_f}, \quad E = \max_f(\sqrt{P_h P_f}\, O_f), \quad C_{\mathrm{opp}} = C_h \max_f(C_f O_f), \quad C_{\mathrm{any}} = \max_{f\in\mathrm{pads}} C_f. \tag{10}$$

The palm proxy $C_{\mathrm{palm}} = S_{\mathrm{palm}}[1 - \Delta_{\mathrm{palm}}/d_c]_0^1$ uses the same proximity filter, where Δ_{palm} is the palm-to-object-surface gap. The per-hand grasp score is

$$G = 0.05A + 0.10E + 0.65C_{\mathrm{opp}} + 0.15C_{\mathrm{any}} + 0.05C_{\mathrm{palm}}. \tag{11}$$

The active-hand scores are combined with a limited support-contact fallback, then converted to a bounded reward factor with strength 0.25. This factor replaces the earlier hand-tracking factor rather than multiplying on top of it. The score favors the same thumb-finger opposition encouraged by E_{cover} in Eq. 1.

Table 8: Generalist tracker settings.

Setting	Value
Reference roster	16,059 motions
Parallelization	4 GPUs, 4,096 environments per GPU
Control / physics	50 Hz control, 4 physics substeps per action
PPO rollout / minibatch / epochs	32 steps / 16,384 samples / 6 epochs
Actor / critic	Separate ReLU MLPs, $1024 \to 1024 \to 512$
Privileged input / action outputs	6,596 / 53, incl. coupled-finger reconstruction
Learning rate / discount / GAE	2×10^{-5} / 0.99 / 0.95
PPO clip / critic / action-bound coefficients	0.2 / 5 / 10

Per-object specialists use the same PPO configuration as the generalist in Table 8, but each trains on the references compatible with its object.

Evaluation. Evaluation starts at frame zero from the reference's prescribed initial state, runs to the reference endpoint, and freezes the normalization statistics. Appendix E.3 defines the success rate and tracking errors computed from these rollouts.

D Self-evolving motion imitation

Augmentation edits a verified reference in interaction coordinates, and the loop keeps only edits that pass physical execution (Sec. 3.4). Each edit type has its own coordinates, its own way of blending

the edit in over time, and its own set of quantities held fixed, while building, execution, and storage are shared by all edits.

D.1 Edit space and invariants

Edits broaden how each task is physically realized, while its manipulation semantics remain those of the source capture. The edit $\boldsymbol{z} = (\boldsymbol{z}^{\text{task}}, \boldsymbol{z}^{\text{body}})$ of Eq. 2 combines spatial coordinates, the object translation and yaw in the seed's interaction frame, with body coordinates that change the pelvis, the stance and toe angle, or the elbow posture. Every edited reference keeps the invariant set $\mathcal{I}$ of its parent: the interaction type, the intended contacts, the object, and the task outcome. Body edits also keep the object path, palm targets, and relevant support constraints of the parent, and a zero edit reproduces the parent. Each edit is blended in gradually over the interaction through a weight α_t, after which the edited reference recomputes its forward kinematics and velocities and its body-object interaction features from the edited geometry (Figure 6).

D.2 Spatial edits

Spatial edits move where the interaction takes place, and the whole body must respond. They act in the seed's interaction frame, whose origin is where a noise-free execution of the seed releases the object, and every edit derived from that seed uses the same frame. With g_t the intended hand contact inherited from the parent, the edit is blended in over the contact phase:

$$\alpha_t = \begin{cases} \dfrac{\sum_{u \le t} g_u}{\sum_u g_u}, & \sum_u g_u > 0, \\ 0, & \text{otherwise}, \end{cases} \qquad x'_o(t) = x_o^{\text{parent}}(t) + \alpha_t H_{\text{rel}} \delta, \tag{12}$$

where $x_o^{\text{parent}}(t)$ is the object position of the verified parent being edited, δ is the spatial edit $\boldsymbol{z}^{\text{task}}$ in the release frame, and H_{rel} maps it to the world frame. The weight is zero before first contact and one after release, so the edit leaves the approach untouched and accumulates during contact. Round 1 requests planar translations within ± 5 cm and yaw within $\pm 10^\circ$. Later rounds push outward from the successful edits that reached farthest, each receiving 16 new edits at radial offsets of 5–20 cm and tangential offsets of up to ± 7.5 cm, and keep the parent's yaw. Achieved positions and headings are recorded as measured, on a 2.5 cm and 5° grid. Moving the goal preserves the type of outcome, such as placement on a support, rather than the exact terminal pose.

In the five-round experiments, only the object reference is edited, and the tracker supplies the whole-body response during execution. An alternative re-solves the waist and arms with inverse kinematics so that the hand follows the moved object. It reaches larger displacements but is not used in those experiments.

D.3 Body edits

Body edits vary the robot's posture around a fixed task, and each of their three types keeps its own set of quantities fixed. Pelvis edits change the pelvis and torso configuration, namely crouch, translation, and orientation, and keep the object path, palms, ankle targets, and contact intent. Stance edits change foot separation and toe orientation and keep the root, arms, fingers, object path, and contact intent. Arm edits use the extra freedom of each seven-DoF arm and keep the palms, object, root, legs, fingers, and contact intent. Each edit is blended in and out smoothly around the interaction before the validity checks of Sec. D.4.

Fixed-palm arm edits. Arm edits move each arm in the nullspace of its palm pose constraint, so the palm stays on its source target. With palm Jacobian $J_h(q_t) \in \mathbb{R}^{6\times 7}$,

$$J_h(q_t) n_t = 0, \qquad q_t^{\text{trial}} = q_t + \alpha_t \gamma s_h n_t, \tag{13}$$

where n_t is a nullspace direction, consecutive n_t are sign-aligned, $s_h \in \{-1, 1\}$ selects the arm-specific direction, γ sets the edit amplitude, and α_t is the edit weight. The constructed pose is projected back to the original palm target before rollout.

Edit magnitudes. Round 1 applies 16 strong variants to each source: eight pelvis edits that crouch by 6–10 cm, combined with a planar shift of up to 5 cm or a pitch, roll, or yaw of up to 12°; four

stance edits that widen the stance by up to 10 cm or narrow it by 5 cm, two of them also turning the toes by $\pm 18°$; and four arm edits of $25°$ covering the four combinations of arm directions. Later rounds push outward from verified parents in steps of 5 cm or $5°$ per axis, with arm angles capped at $30°$. These are requested magnitudes, and an infeasible pose is rejected before execution.

D.4 Building, executing, and keeping edits

Spatial and body edits share the acceptance principle of Sec. 3.4, but differ in when the tracker learns them (Table 9). Before physics, a candidate is checked only for basic validity. Spatial candidates, with near-duplicates removed, then join fine-tuning before they are judged. Body candidates are judged with the current tracker and join training in later rounds once verified. Each candidate is executed with three rollouts under different random seeds.

Table 9: **Acceptance rules for the two edit types.** ✓: checked; –: not checked.

	Spatial edits	Body edits
Checked before physics	finite values	finite, object unchanged
Tracker used	fine-tuned first	current
Rollouts needed to pass	2 of 3	1 of 3
Completes the reference	✓	✓
Terminal hand contact	✓	✓ (held tasks)
No sustained fall	✓	✓
Smooth motion	✓	✓
Object placed and at rest	–	✓ (placed tasks)
Object path matches its source	–	✓
Stored as	executed motion	verified proposal

Criteria. Terminal hand contact holds when the simulated hand contacts match the reference contact labels in six of the final ten frames. A sustained fall is a collapsed or strongly misoriented pose that lasts ten consecutive frames or six of the final ten. Smoothness compares the high-frequency motion of the hands and body, above 6 Hz, with that of the original source and of the other candidates, and rejects clear outliers. A placed object must rest stably on its support in six of the final ten frames. These checks preserve the type of outcome rather than the exact target pose.

Storage. An accepted spatial candidate is stored as its executed motion, relabeled with the simulated body and object states. An accepted body candidate keeps its constructed reference, and its rollout is stored as evidence that the tracker can execute it.

D.5 Growth over rounds

New edits are proposed only around verified parents, so each round starts where the previous one succeeded. For spatial edits, each task selects up to 16 parents among the successful edits that reached farthest from their seed, favoring distinct sources and reliable execution, and gives each 16 edits that move further outward. For body edits, each source's successful rollouts are described by their achieved posture, and the most extreme successful postures receive the outward edits of Sec. D.3. Each round fine-tunes the tracker for a fixed wall-clock budget, about 1,050–1,420 PPO epochs or 87–105 M transitions, sampling original and augmented references 1:1, and original, verified body, and spatial references 2:1:1 once body references exist.

E Evaluation metrics

This section defines each metric of Sec. 4.1 once, grouped in the order of the experiments: data coverage (Figure 3), retargeted references (Sec. 4.2), tracking (Sec. 4.3), and the augmentation loop (Sec. 4.4). All compared methods share the same features, thresholds, and frame rates.

E.1 Data coverage

The panels of Figure 3 are computed from the following features.

Human motion. Each sequence is described by 21 non-finger joint positions relative to the pelvis, normalized by skeletal height, plus root displacement, height, and heading. Each channel is resampled to 128 frames and truncated to its first 16 DCT coefficients (1,088 dimensions), and PCA-64 then UMAP-2 gives one point per sequence (16,059). The feature uses no object information.

Object geometry. Each mesh is sampled with 10,000 surface points, centered, scaled to the unit sphere, and oriented by its first source pose, then embedded with a frozen OpenShape PointBERT encoder; embeddings are averaged over four yaw rotations. PCA and UMAP give one point per distinct geometry (155). Unit-sphere scaling removes physical size.

Human-object interaction. The feature concatenates the object position relative to the pelvis, the wrist positions and surface distances relative to the object, the wrist separation, and the object height, translation, and rotation, encoded with the same DCT scheme. Each block is normalized separately before PCA-64 and UMAP-2.

External coverage. Probes from GRAIL, MOVIN, and ELMO are matched to the source library in a 73-dimensional body-motion space: PCA-64 of the pelvis-relative joint DCT features plus nine per-limb repetition descriptors. A probe is covered when its nearest source sequence lies within the 75th percentile of the source leave-one-out nearest-neighbor distance, and coverage averages the three probe sets equally. We compare the full InterAct, HiPHI, and their union without subsampling.

Scaling. Coverage is computed on nested random subsets of the union and averaged over 20 orderings; shading marks the 5th–95th percentiles.

E.2 Retargeted references

These metrics score kinematic references in Table 1 and Table 10. Each method is evaluated on its own robot model at 30 fps, on the source clips shared by the compared methods, against the human reference scaled by the subject's body scale.

Contact preservation. During human-labeled hand contact, a hand preserves contact when the minimum distance between its wrist, palm, and finger meshes and the object mesh is at most 2 cm, with interpenetration counted as zero. The per-frame variant requires all contacting hands to be preserved; the per-hand variant counts each hand–frame separately.

In-hand slip. Mean speed of eleven hand keypoints in the object frame over consecutive contact frames.

Penetration. A frame penetrates when any robot collision geometry is deeper than 1 cm inside the object (a shared 20-part convex decomposition). We report their mean maximum depth. Ground penetration (Gnd.) is the fraction of frames in which any robot collision geometry is more than 1 cm below the floor.

Foot sliding. A foot is in stance when the human toe moves horizontally at most 0.15 m/s, and slides when the robot toe then moves faster than 0.15 m/s. We report sliding frames over stance frames. Sliding-frame speed is the mean robot toe speed over sliding frames.

Fidelity. MPJPE is the pelvis-relative error of twelve landmarks (hips, knees, ankles, shoulders, elbows, wrists); end-effector error restricts it to the wrists and ankles.

Smoothness. Absolute second finite difference of joint angles times the squared frame rate, averaged over frames and joints. Body (29) and finger (24) joints are reported separately, and the all-joint value averages over all 53. P95 is the 95th percentile of the same quantity over all frames and body joints of a clip, averaged over clips.

Joint limits. Fraction of frames with any joint outside its own range (10^{-4} rad tolerance). Joint-frame violations count violating joint–frame pairs instead of frames.

E.3 Tracking

These metrics score the trackers in Table 2. Each benchmark clip is executed once under the evaluation protocol of Sec. C.2. The benchmark groups its 15 objects into three categories. Box-like: OMOMO largebox and smallbox, BEHAVE boxlarge and suitcase. Sitting: BEHAVE stool, IMHD chair, and

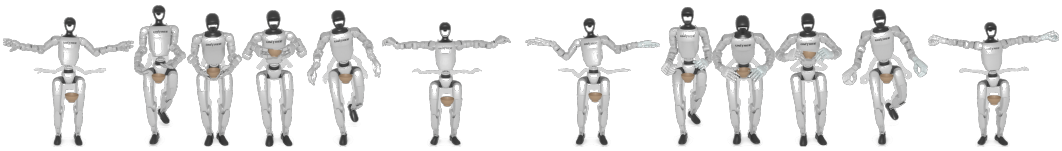

Figure 8: **Retargeting with dexterous hands.** A bowl-holding interaction retargeted to G1 with Dex3-1 (left) and Inspire (right) hands.

NeuralDome smallsofa. Grasping: OMOMO tripod and largetable, IMHD dumbbell, InterCap cup, racket, and umbrella, GRAB mug, and NeuralDome table.

Success rate. A rollout succeeds when it reaches the last reference frame without triggering a termination condition of Sec. C.2. The success rate is the fraction of successful clips.

Tracking errors. Errors are computed on successful clips by comparing each simulated frame with the reference frame of the same index, excluding the terminal hold frames. MPJPE is the mean world-frame position error of the 30 body links shared by all three G1 hand configurations (pelvis, legs, waist, torso, and arms). Object position error is the error of the object origin, and object rotation error is the geodesic angle between the simulated and reference orientations. Per-frame values are averaged within a clip and then across clips.

E.4 Self-evolution

These metrics score self-evolution in Tables 3, 12, and 11. Spatial and body edits are pooled throughout.

Growth. The number of distinct verified references after a given round, spatial and body together, divided by the number of original references of the setting.

Motion quality. Quality is computed on the executed trajectory of each original reference and of each reference newly accepted in a round, at 50 fps and without the 20 terminal-hold frames, and each motion counts equally. Body joint acceleration is the mean absolute second finite difference of the 29 body-joint angles times the squared frame rate. Foot sliding is the fraction of adjacent frames with foot contact in which the ankle link moves horizontally faster than 0.15 m/s; unlike Sec. E.2, contact is taken from the simulation rather than from the human motion. Hand jitter applies a fourth-order Butterworth high-pass filter above 6 Hz to pelvis-relative hand positions and reports the 75th percentile of the root-mean-square over 50-frame windows.

Success before and after self-evolution. Each accepted reference and each original reference is replayed with the tracker before round 1 and after round 5, with three rollouts under fixed random seeds and deterministic actions. A reference counts as executed when at least one rollout reaches its last frame without triggering a termination condition of Sec. C.2, and the success rate is the fraction of references executed.

Yield. In the frozen-tracker ablation, the yield is the fraction of a round's rollouts that pass the acceptance checks.

F Additional results

Additional tables and figures extend the retargeting, tracking, and loop comparisons of Sec. 4, and a frozen-tracker ablation isolates the role of fine-tuning in the loop.

F.1 Retargeting comparison on G1 with non-dexterous hands

All four retargeters are compared on G1 with non-dexterous hands, using the 537 OMOMO clips they share (largebox, smallbox, plasticbox, and suitcase) and the metrics of Sec. E.2. Our entries use the final cleaned references for every object. Contact preservation uses the supplied left/right human-hand contact labels, which are positive in 479 clips. ULTRA produces several variants per source sequence, which we average before averaging all metrics over source sequences. ULTRA's variants have no matching human reference, so their foot-sliding and fidelity metrics are unavailable.



Figure 9: **Retargeting across embodiments.** One suitcase interaction retargeted to G1 with Dex3-1, Inspire, and non-dexterous hands (left, top to bottom) and to Booster K1, Booster T1, and Dexmate Vega (right, top to bottom).

Table 10: **Retargeting on G1 with non-dexterous hands**, over the 537 OMOMO clips shared by all four methods. A dash denotes no qualifying event and N/A an unavailable metric; bold marks the unique best.

Metric	OmniRetarget	UMR	ULTRA	Ours
Contact preservation				
Hand contact preservation (%) ↑	79.09	**95.13**	79.06	83.45
Penetration (> 1 cm)				
Robot–object depth (cm) ↓	**1.27**	1.60	1.49	1.97
Foot sliding (0.15 m/s)				
Sliding frames (%) ↓	**0.00**	7.02	N/A	7.33
Sliding-frame speed (m/s) ↓	–	0.22	N/A	0.26
Fidelity to human				
MPJPE (cm) ↓	**5.55**	6.70	N/A	5.62
End-effector error (cm) ↓	6.36	7.92	N/A	**5.86**
Smoothness				
Body joint acc. (rad/s^2) ↓	6.65	6.11	10.54	**2.54**
Body joint acc., P95 (rad/s^2) ↓	23.19	21.97	37.29	**8.96**
Joint limits				
Body violation frames (%) ↓	0.00	0.00	1.04	0.00
Body joint-frame violations (%) ↓	0.00	0.00	0.04	0.00

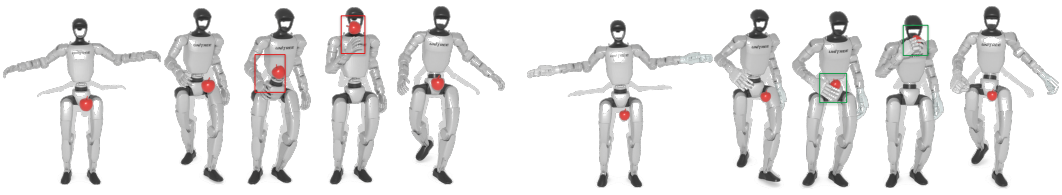

Figure 10: **Retargeting comparison with OmniRetarget.** OmniRetarget to G1 with rigid hands (left) and our retargeter for G1 with Inspire hands (right), at corresponding interaction phases. The dexterous hand keeps the grasp and the hand-object relationship of the demonstration.

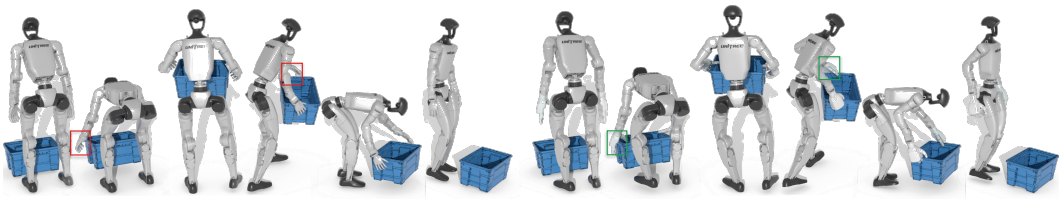

Figure 11: **Retargeting comparison with UMR.** UMR to G1 (left) and our retargeter to G1 with Inspire hands (right), at corresponding interaction phases.

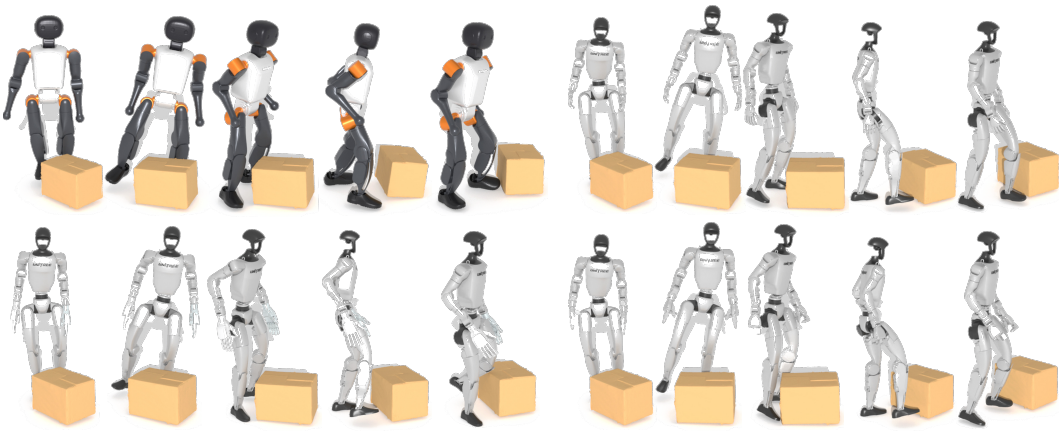

Figure 12: **Physics-based tracking across embodiments.** A box-lifting reference executed by specialist trackers on Booster K1 (top left), G1 with non-dexterous hands (top right), G1 with Inspire hands (bottom left), and G1 with Dex3-1 hands (bottom right).

With rigid hands, each method favors a different trade-off. Table 10 shows no single winner between fidelity, geometric constraints, and smoothness. Our references are by far the smoothest and the most accurate at the end effectors, OmniRetarget penetrates the object least, and UMR preserves the most hand contact. Rigid hands cannot close around an object, so contact and penetration trade off directly. This is the gap our dexterous retargeting addresses: with articulated fingers, our references keep the grasp and hand–object relationship of the demonstration (Figures 10 and 11).

F.2 Ablation on learning in the loop

More sampling cannot replace learning. A fixed tracker can only filter candidates it already executes, and after the first round it has nothing left to filter. Because the base tracker fails on many accepted references (Table 3), a simpler alternative to fine-tuning would keep the tracker fixed, sample more rollouts, and keep those that pass. Table 11 tests this alternative in a separate run on a subset

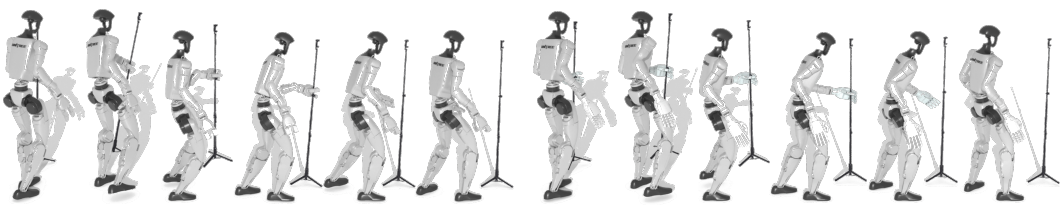

Figure 13: **Physics-based tracking with dexterous hands.** A tripod grasp executed on G1 with Dex3-1 (left) and Inspire (right) hands.

Table 11: **Frozen versus fine-tuned tracker** in a separate run on a subset of the box-like seeds with translation edits, whose counts are not directly comparable with Table 3. Yield and growth follow Sec. E.4. The frozen run stops after round 3.

	Yield (%)		Growth (×)	
Round	Frozen	Tuned	Frozen	Tuned
1	29.4	25.3	163.0	188.5
2	0.4	72.9	170.5	426.0
3	0.4	68.9	184.9	573.0

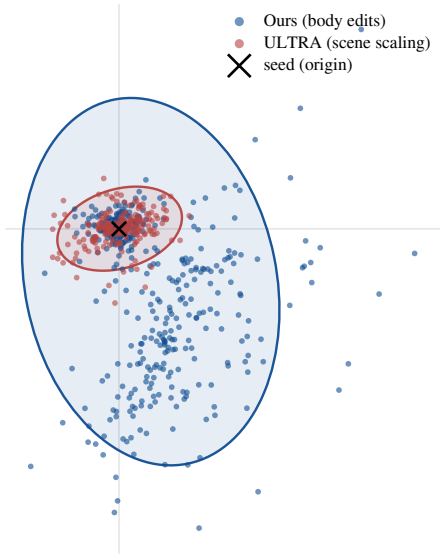

Figure 14: **Body-motion variation around shared seeds.** Each reference is embedded by the DCT coefficients of its body joint angles, taken relative to its seed and projected with PCA, so all seeds sit at the origin. Ellipses enclose 95% of a Gaussian fit.

of the box-like seeds. In the first round, both trackers execute the same candidates with the same checkpoint, so their yields are similar. From the second round on, the frozen tracker accepts almost nothing, because the new candidates start from first-round successes that already lie beyond its reach, and its growth stalls. The fine-tuned tracker keeps a high yield in round 2 and 3, so its verified references keep growing.

F.3 Self-evolution by round

The reference set grows every round, while quality measure rises moderately. Table 12 extends Table 3 to every round. Growth is measured on the cumulative set of verified references, and quality on the references newly accepted in each round. The verified set grows in every round in all three settings. All three quality measures rise in later rounds, most clearly foot sliding; later rounds, however, accept edits farther from the captures and a different mix of spatial and body edits, so the trend reflects harder motions as well as execution quality. Success rates are measured only before and after self-evolution and appear in Table 3.

F.4 Body-motion range of augmented references

Body edits move references farther from their seed than scene scaling does. Figure 14 compares how far augmented references depart from their seed in body motion, on the box-like seeds shared by both methods. Our body edits spread over a region several times wider than ULTRA's scene-scaling

Table 12: **Self-evolution by round**, extending Table 3. Growth (×): cumulative verified references relative to the original references. Quality of the original references (Orig.) and of the references newly accepted in each round: body joint acceleration (Acc., rad/s^2), foot sliding (Foot, %), and hand jitter (Hand, mm).

	Inspire, box-like				Dex3, box-like				Inspire, grasping			
Round	Growth	Acc.	Foot	Hand	Growth	Acc.	Foot	Hand	Growth	Acc.	Foot	Hand
Orig.	1.0	11.4	1.18	1.00	1.0	13.9	1.14	1.12	1.0	13.6	0.70	0.96
1	21.8	11.8	1.39	1.07	21.3	15.6	1.59	1.25	19.6	15.3	0.99	1.07
2	34.8	11.3	1.39	1.09	31.6	14.8	1.77	1.23	33.3	14.4	1.42	1.09
3	80.4	12.1	1.63	1.16	73.3	14.9	1.87	1.24	77.8	14.6	1.78	1.08
4	118.1	12.9	2.14	1.22	110.8	16.4	2.41	1.33	112.8	15.7	2.42	1.16
5	150.5	13.2	2.37	1.24	142.0	17.4	2.99	1.38	146.4	16.4	2.87	1.26

variants, and mostly along directions that scaling does not vary, since the edits change posture while the object path is kept.

G Hardware and real-world deployment

Deployment for Unitree G1. Experiments are performed on a Unitree G1 humanoid with 29 DOFs (six per leg, three at the waist, and seven per arm), running robot firmware version 1.5.4. The policy is executed on a laptop with an NVIDIA RTX 5090 Laptop GPU and communicates over Ethernet through Unitree's SDK. We use SONIC (Luo et al., 2026) as the low-level whole-body controller: it takes the reference motion at 50 Hz, which is tracked by the onboard joint-level PD controllers at 500 Hz, using SONIC's default per-joint PD gains. On top of these, stiffness and damping scales of 1.5 are applied to the two ankle pitch motors; a constant 1 degree pitch offset is further added to the measured pelvis IMU orientation to account for the calibration bias of the robot's own IMU.

Deployment for Unitree G1 with Inspire dexterous hands. For tasks require dexterous manipulation, we equip Unitree G1 with two Inspire RH56DFTP dexterous hands, each with six actuated DOFs, plus six mimic joints. The humanoid body is controlled by SONIC as described above, while the hands directly replay the finger trajectory of the rollout motion. For the synchronization, the body deployment publishes its current playback frame, and a separate process sends the hand pose for that frame to the robot at about 12 Hz. Notably, we initiate the thumb motion first, followed by the motion of the other four fingers.

Deployment for Booster K1. Experiments are performed on a Booster K1 humanoid with 22 DOFs (six per leg, four per arm, and two at the head). For policy training, we build on Booster Train (`https://github.com/BoosterRobotics/booster_train`), extending it with object-interaction tracking and targeted domain randomization. We use an asymmetric actor-critic formulation, with privileged object and contact information provided only to the critic during training. Robot and object dynamics are randomized, including ground friction, object mass, center of mass, object placement, actuator properties, and external perturbations, while preserving close tracking of the reference body and object motion. For real-world execution, policies are deployed using Booster Deploy (`https://github.com/BoosterRobotics/booster_deploy`).

Deployment for Dexmate Vega. Experiments are performed on a Dexmate Vega-1 Pro, a bimanual mobile manipulator with 23 DOFs (three at the torso, seven per arm, three at the head, and three DOFs at the omnidirectional wheeled base). Since the wheeled base is stable and dexterous, no whole-body controller is needed. Retargeted upper-body joint trajectories are sent directly to the onboard joint controllers through Dexmate's dexcontrol SDK (v0.5.0). Notably, due to the torso joint speed limit, all trajectories are uniformly time-scaled by $0.3\times$ speed during deployment. Base motion is converted to wheel commands by solving the chassis inverse kinematics using trajectory optimization, formulated in CasADi (Andersson et al., 2019) and solved with IPOPT (Wächter & Biegler, 2006). The solver penalizes position, yaw, terminal-pose tracking errors, and regularizes body acceleration, steering rate and acceleration, and wheel acceleration. All commands are streamed at 100 Hz.