InterMimic: Towards Universal Whole-Body Control for Physics-Based Human-Object Interactions

Sirui Xu[1] Hung Yu Ling[2] Yu-Xiong Wang[1†] Liang-Yan Gui[1†]
[1] University of Illinois Urbana-Champaign [2] Electronic Arts
[†] Equal Advising

https://sirui-xu.github.io/InterMimic

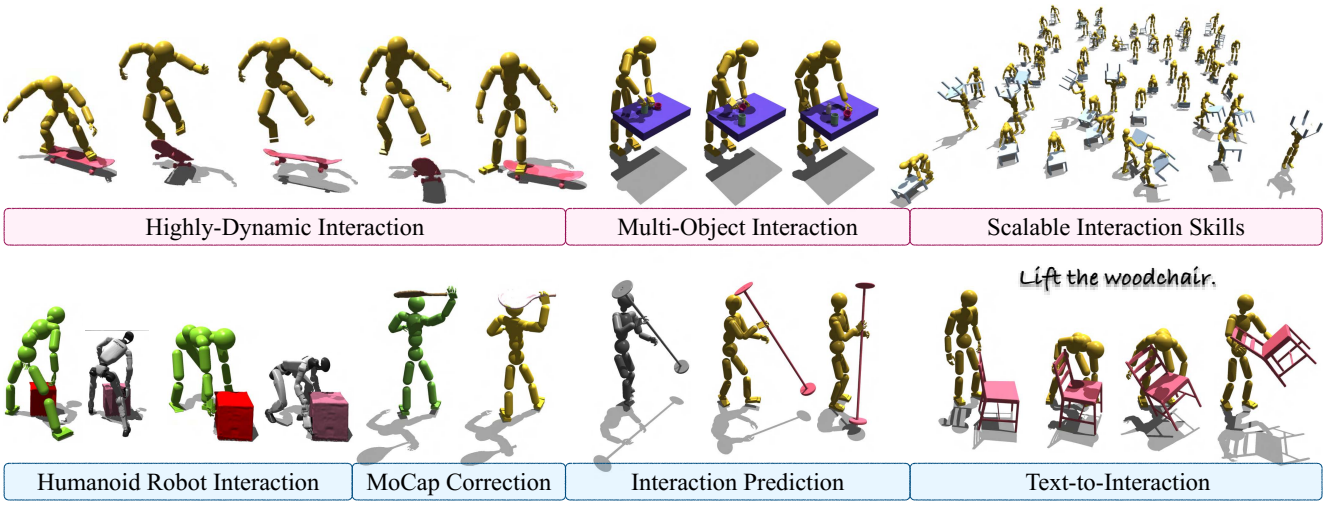


Figure 1. InterMimic enables physically simulated humans to perform interactions with dynamic and diverse objects. It supports highly-dynamic, multi-object interactions and scalable skill learning (**Top**), making it adaptable for versatile downstream applications (**Bottom**): it can translate whole-body loco-manipulation skills to a humanoid robot [24, 81], perfect interaction MoCap data, and bridge kinematic generation, *e.g.*, predicting future interactions from past (InterDiff [101]) or generating interactions given text prompts (InterDreamer [102]).

Abstract

Achieving realistic simulations of humans interacting with a wide range of objects has long been a fundamental goal. Extending physics-based motion imitation to complex human-object interactions (HOIs) is challenging due to intricate human-object coupling, variability in object geometries, and artifacts in motion capture data, such as inaccurate contacts and limited hand detail. We introduce InterMimic, a framework that enables a single policy to robustly learn from hours of imperfect MoCap data covering diverse full-body interactions with dynamic and varied objects. Our key insight is to employ a curriculum strategy – perfect first, then scale up. We first train subject-specific teacher policies to mimic, retarget, and refine motion capture data. Next, we distill these teachers into a student policy, with the teachers acting as online experts providing direct supervision, as well as high-quality references. Notably, we incorporate RL fine-tuning on the student policy to surpass mere demonstration replication and achieve higher-quality solutions. Our experiments demonstrate that InterMimic produces realistic and diverse interactions across multiple HOI datasets. The learned policy generalizes in a zero-shot manner and seamlessly integrates with kinematic generators, elevating the framework from mere imitation to generative modeling of complex human-object interactions.

1. Introduction

Animating human-object interactions is a challenging and time-consuming task even for skilled animators. It requires a deep understanding of physics and meticulous attention to detail to create natural and convincing interactions. While

Motion Capture (MoCap) data provides references, animators often need to correct contact errors caused by sensor limitations and occlusions between humans and objects. However, this process remains unscalable, as refining a single motion demands a delicate balance between preserving the captured data and ensuring its physical plausibility.

Physics-based human motion imitation [38, 63] offers an alternative approach to improving motion fidelity, by training control policies to mimic reference MoCap data within a physics simulator. However, scaling up human-object interaction imitation presents significant challenges: (**i**) *MoCap Imperfection*: Contact artifacts are common, causing expected contacts to fluctuate instead of maintaining consistent zero distance, often due to MoCap limitations or missing hand capture [3, 39]. Accurately imitating MoCap kinematics can result in unrealistic dynamics in simulation. Moreover, HOI datasets often include diverse human shapes, requiring motion retargeting to adapt movements across different human models while preserving interaction dynamics. This retargeting process is imperfect and can introduce new contact artifacts or exacerbate existing ones. (**ii**) *Scaling-up*: Although large-scale motion imitation has been explored in previous works [50, 79, 92, 106], it remains largely underexplored for whole-body interactions involving dynamic and diverse objects.

In this paper, we aim to utilize rich yet imperfect motion capture interaction datasets to train a control policy capable of learning diverse motor skills while enhancing the plausibility of these actions by correcting errors, such as inaccurate hand motions and faulty contacts. Our approach is grounded on the key insight of tackling the challenges of *skill perfection* and *skill integration* progressively. We implement a curriculum-based teacher-student distillation framework, where multiple teacher policies focus on imitating and refining small subsets of interactions, and a student policy integrates these skills from the teachers.

Instead of relying on curated data that covers a limited range of actions [4, 51], we employ multiple teacher policies trained on a diverse set of imperfect interaction data and address two key challenges: *retargeting* and *recovering*. First, we unify all training policies to a canonical human model, by embedding HOI retargeting directly into the imitation. This is achieved by reframing the policy learning to optimize both imitation and retargeting objectives. Second, our teacher policies refine interaction motion through learning from it, as accurate contact dynamics enforced by a physics simulator inherently correct inaccuracies in the reference kinematics. To support this, we introduce tailored contact-guided reward and optimize trajectory collection, enabling effective skill imitation despite MoCap errors.

Introducing teacher policies offers several key benefits. By leveraging teacher rollouts, we effectively distill raw MoCap data into refined HOI references with a unified embodiment and enhanced physical fidelity. These refined references guide the subsequent student policy training, reducing the negative impact of errors in the original MoCap data. A major hurdle in scaling motion imitation is the sample inefficiency of Reinforcement Learning (RL), which can lead to prohibitively long training times. Our teacher-student approach mitigates this through a *space-time trade-off*: multiple teacher policies are trained in parallel on smaller, more manageable data subsets, and their expertise is then *distilled* into a single student policy. We begin with demonstration-based distillation to bootstrap PPO [71] updates, reducing reliance on pure trial and error and enabling more effective scaling. As training progresses, the student gradually shifts from heavy demonstration guidance to increased RL updates, ultimately surpassing simple demonstration memorization. This mirrors alignment strategies in Large Language Models (LLMs), where demonstration-based pre-training is refined through RL fine-tuning [56].

To summarize, our contributions are as follows: (**i**) We introduce *InterMimic*, which, to the best of our knowledge, is the *first* framework designed to train physically simulated humans to develop *a wide range of whole-body* motor skills for interacting with *diverse* and *dynamic* objects, extending beyond traditional grasping tasks. (**ii**) We develop a teacher-student training strategy, where teacher policies provide a unified solution to address the challenges of retargeting and refining in HOI imitation. The student distillation introduces a scalable solution by leveraging a space-time trade-off. (**iii**) We demonstrate that our unified framework, *InterMimic*, as illustrated in Figure 1, effectively handles versatile physics-based interaction animation, recovering motions with realistic and physically plausible details. Notably, by combining kinematic generators with *InterMimic*, we enable a physics-based agent to achieve tasks such as interaction prediction and text-to-interaction generation.

2. Related Work

Significant progress has been made in physics-based human interaction animation and control, with advancements in areas such as human-human interactions [45, 90], hand-object interactions [60, 86, 100, 104], human interactions with static scenes [5, 36, 57, 79, 96], and real-world humanoid control for object manipulation [2, 11, 16, 19, 20, 28, 41, 43, 72, 107]. Among these areas, we are the first to achieve universal whole-body loco-manipulation simulation with diverse dynamic objects, beyond pick-and-place and grasping actions – a novel achievement in animation and an idealized reference for real-world humanoid control. Below, we elaborate on recent studies on *whole-body* interaction animation, particularly involving *dynamic objects*.

2.1. Kinematic Interaction Animation

Generating human interactions has been a long-standing topic in animation and computer graphics [15, 37]. Significant advances in character animation have emerged with the advent of deep learning, *e.g.*, including phase-function-based methods [22] that enable object interactions like carrying a box [74] or playing basketball [75]. This is extended to more diverse but static objects approaching [35, 77, 94, 111]. Subsequent efforts [14, 26, 27, 40, 42, 49, 66, 95] integrate object motion into interactions but remain constrained by assuming that interactions occur primarily through the hands. To address this, recent developments [7, 10, 12, 21, 30, 62, 73, 84, 93, 101, 102] introduce interactions in a fashion of whole-body loco-manipulation that engages multiple body parts in contact. However, these methods often suffer from physical inaccuracies, such as floating contacts and penetrations, while they generate only body motion without considering hand dexterity. In this work, we address physical inaccuracies by refining imperfect kinematic generation through physics simulation, with InterDiff [101] and HOI-Diff [62] serving as motion planning for loco-manipulation that bridges high-level decision-making (*e.g.*, text instruction) with low-level execution.

2.2. Physics-based Interaction Animation

Physics-based methods generate motion through motor control policies within a physics simulator, *e.g.*, achieved via deep reinforcement learning to track reference motions [63]. These policies are directly applicable for executing simple interactions, such as punching or striking an object [8, 65, 78, 80]. To achieve more complex interactions, early studies focus on specific scenarios, including notable sports-related [52] examples such as basketball [88], skating [44], soccer [98], tennis [108], table tennis [85], and more proposed in [1]. Research also demonstrates flexibility in more general but simpler box carrying tasks [58, 87, 113]. These advancements are achieved through the integration of multiple control policies [55], the use of adversarial motion priors [13, 18, 64], and imitating diverse kinematic generations [95, 99]. However, these methods train their policies in a *non-scalable* manner, with each policy handling only specific object types or actions. In pursuit of a single, scalable policy to enable multiple interaction skills, existing methods either rely on fixed interaction patterns, such as approaching and grasping objects [4, 51], or extend single-object skills, *e.g.*, interactions involving a basketball [89]. Additionally, they mostly depend on highly curated data from the GRAB dataset [76], which, despite its high quality, primarily features low-dynamic full-body motion and only small-sized objects. More recent datasets [3, 23, 25, 31, 39, 46–48, 53, 97, 103, 109, 110, 112, 114] offer richer full-body interactions with diverse objects but contain noticeable artifacts that challenge existing motion imitation approaches. We process data from OMOMO [39], BEHAVE [3], HODome [109], and IMHD [114] collected in the InterAct [103] dataset, and the multi-object dataset HIMO [53], highlighting InterMimic's *scalability* to diverse interactions and its *robustness* against MoCap artifacts.

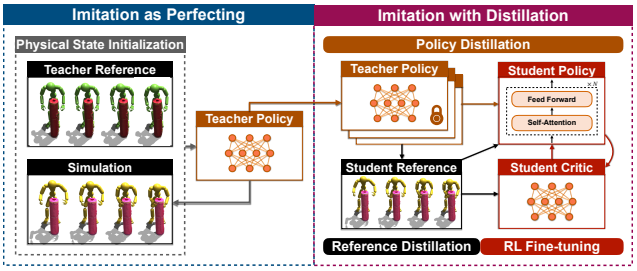


Figure 2. Our two-stage pipeline: (**i**) training each teacher policy (MLP) on a small data subset with initialization corrected via Physical State Initialization (PSI), and (**ii**) freezing the teacher policies to provide refined references for training a student policy (Transformer). The student leverages teacher supervision for effective scaling and is fine-tuned through RL.

3. Methodology

Task Formulation. The goal of human-object interaction (HOI) imitation is to learn a policy π that produces simulated human-object motion $\{\boldsymbol{q}_t\}_{t=1}^{T}$ closely matching a ground-truth reference $\{\hat{\boldsymbol{q}}_t\}_{t=1}^{T}$ derived from large-scale MoCap data. Given the geometries of the human and objects, the policy should also compensate for missing or inaccurate details in the dataset. Each pose $\boldsymbol{q}_t$ has two components: the human pose $\boldsymbol{q}_t^h$ and the object pose $\boldsymbol{q}_t^o$. The human pose is defined as $\boldsymbol{q}_t^h = \{\boldsymbol{\theta}_t^h, \boldsymbol{p}_t^h\}$, where $\boldsymbol{\theta}_t^h \in \mathbb{R}^{52\times3}$ represents the joint rotations, and $\boldsymbol{p}_t^h \in \mathbb{R}^{52\times3}$ specifies the joint positions. Specifically, our human model includes 30 hand joints and 22 joints for the rest of the body, with one root joint's rotation and position specified in global coordinates, whereas the rotations and positions of all other joints are defined relative to their respective parent joints' coordinate systems. The object pose $\boldsymbol{q}_t^o$ is represented as $\{\boldsymbol{\theta}_t^o, \boldsymbol{p}_t^o\}$, where $\boldsymbol{\theta}_t^o \in \mathbb{R}^3$ denotes the object's orientation and $\boldsymbol{p}_t^o \in \mathbb{R}^3$ the position. All simulation states have corresponding ground-truth values, denoted by the hat symbol. For instance, the reference object rotation is $\{\hat{\boldsymbol{\theta}}_t^o\}_{t=1}^{T}$. The environmental setup for the simulation is detailed in Sec. B.

Overview. We formulate interaction imitation as a Markov Decision Process (MDP), defined by states, actions, simulator-provided transition dynamics, and a reward function. Figure 2 illustrates our two-stage framework: (i) training teacher policies $\pi^{(T)}$ on small skill subsets, and (ii) distilling these teachers into a scalable student policy $\pi^{(S)}$ for large-scale skill learning. In Sec. 3.1, we define the states $\boldsymbol{s}_t$

and actions $\boldsymbol{a}_t$, applicable to both teacher $\pi^{(T)}$ and student $\pi^{(S)}$ policies. In Sec. 3.2, we describe how teacher policies are trained via RL, focusing on reward designs that facilitate retargeting, as well as techniques that mitigate the impact of imperfections in the reference data. Sec. 3.3 details the subsequent distillation of teachers into a scalable student policy, leveraging both RL and learning from demonstration.

3.1. Policy Representation

State. The state $\boldsymbol{s}_t$, which serves as input to the policy, comprises two components $\boldsymbol{s}_t = \{\boldsymbol{s}_t^s, \boldsymbol{s}_t^g\}$. The first part, $\boldsymbol{s}_t^s$, contains human proprioception and object observations, expressed as, $\{\{\boldsymbol{\theta}_t^h, \boldsymbol{p}_t^h, \boldsymbol{\omega}_t^h, \boldsymbol{v}_t^h\}, \{\boldsymbol{\theta}_t^o, \boldsymbol{p}_t^o, \boldsymbol{\omega}_t^o, \boldsymbol{v}_t^o\}, \{\boldsymbol{d}_t, \boldsymbol{c}_t\}\}$, where $\{\boldsymbol{\theta}_t^h, \boldsymbol{p}_t^h, \boldsymbol{\omega}_t^h, \boldsymbol{v}_t^h\}$ represent the rotation, position, angular velocity, and velocity of all joints, respectively, while $\{\boldsymbol{\theta}_t^o, \boldsymbol{p}_t^o, \boldsymbol{\omega}_t^o, \boldsymbol{v}_t^o\}$ represent the orientation, location, velocity, and angular velocity of the object, respectively. Motivated by [6], we include object geometry and whole-body haptic sensing from two elements: (**i**) $\boldsymbol{d}_t$, vectors from human joints to their nearest points on each object surface; and (**ii**) $\boldsymbol{c}_t$, contact markers indicating whether the human's rigid body parts experience applied forces; this serves as simplified tactile or force sensing – an important multimodal input in robot manipulation tasks [9]. The goal state $\boldsymbol{s}_t^g = \{\boldsymbol{s}_{t,t+k}^g\}_{k \in K}$ integrates reference poses from the ground truth motion, where $\boldsymbol{s}_{t,t+k}^g$ is defined as,

$$\begin{aligned} &\{\{\hat{\boldsymbol{\theta}}_{t+k}^h \ominus \boldsymbol{\theta}_t^h, \hat{\boldsymbol{p}}_{t+k}^h - \boldsymbol{p}_t^h\}, \{\hat{\boldsymbol{\theta}}_{t+k}^o \ominus \boldsymbol{\theta}_t^o, \hat{\boldsymbol{p}}_{t+k}^o - \boldsymbol{p}_t^o\}, \\ &\{\hat{\boldsymbol{d}}_{t+k} - \boldsymbol{d}_t, \hat{\boldsymbol{c}}_{t+k} - \boldsymbol{c}_t\}, \{\hat{\boldsymbol{\theta}}_{t+k}^h, \hat{\boldsymbol{p}}_{t+k}^h, \hat{\boldsymbol{\theta}}_{t+k}^o, \hat{\boldsymbol{p}}_{t+k}^o\}\}, \end{aligned} \tag{1}$$

where $\hat{\boldsymbol{\theta}}_{t+k}^h, \hat{\boldsymbol{p}}_{t+k}^h, \hat{\boldsymbol{d}}_{t+k}, \hat{\boldsymbol{c}}_{t+k}$ represent the reference information at time step $t+k$, $\ominus$ denotes the calculation of rotation difference. All continuous elements of $\boldsymbol{s}_t$ are normalized relative to the current direction of view of the human and the position of the root [63].

Given that most MoCap data does not provide reference contact or tactile information, we extract reference contact markers $\hat{\boldsymbol{c}}_{t+k}$ by inferring dynamic information, beyond relying solely on inaccurate contact distances, specifically by analyzing the object's acceleration to detect human-induced forces. To accommodate the variability in contact distances observed in reference motion, we discretize reference contact markers using varying distance thresholds, as illustrated in Fig. 3(**i**). The neutral areas serve as buffer zones, avoiding the penalization or enforcement of strict contact. See Sec. C of supplementary for details.

Action. Our human model has 51 actuated joints, defining an action space of $\boldsymbol{a}_t \in \mathbb{R}^{51\times 3}$. These actions are specified as joint PD targets using the exponential map and are converted into torques applied to each of the human joints.

3.2. Imitation as Perfecting

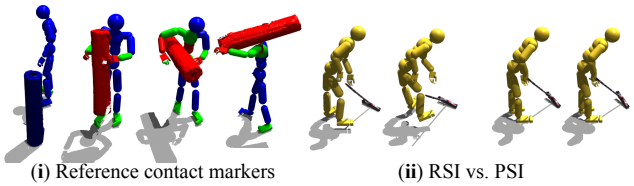
The teacher policy $\pi^{(\mathrm{T})}$ is trained via RL to maximize the expected discounted reward by comparing simulated states against potentially erroneous reference states. The training involves: (**i**) trajectory collection, where we explain how trajectories are initialized and terminated. (**ii**) policy updating, where collected trajectories and their associated rewards are used to refine the policy. In this section, we elaborate on our reward design and how we optimize trajectory collection to mitigate the impact of reference inaccuracies.

Imitation as Retargeting. We tailor teacher policies to each human subject, while all policies share the same base human model. This serves the retargeting purpose by converting HOIs from different human shapes into a unified base shape. Although motion imitation does not necessarily require a unified human model [50, 91], our approach offers two benefits: (i) It enhances integration with kinematic generation methods, which generally perform better on a single, unified shape [17]. (ii) It demonstrates possible integration with real-world humanoid deployment, which requires retargeting to a consistent physical embodiment. In Figure 1, our method translates MoCap data into motor skills on a Unitree G1 [81] with two Inspire hands [24], all without external retargeting in complex contact-rich scenarios. See Sec. F of the supplementary for additional details.

Human [83] or HOI [32] retargeting can be formulated as an optimization problem. Inverse Kinematics (IK) methods, such as those based on quadratic programming [34], demonstrate effectiveness in simplified scenarios but remain underexplored for motions featuring intricate object interactions. RL, by contrast, solves the optimization by maximizing an expected cumulative reward, prompting us to investigate whether RL-driven HOI imitation can be used for HOI retargeting. This extends existing physics-based retargeting approaches, which either omit object interactions [67] or are non-scalable with a single reference [113].

While the kinematics should differ due to the embodiment gap, we argue that the *dynamics* between human and object should remain *invariant*. Thus, we define rewards to include an embodiment-aware component that loosely aligns the simulated kinematics with the reference interaction, and an embodiment-agnostic reward component that encourages dynamics to be close to the reference.

Embodiment-Aware Reward. When the human and object are far apart, retargeting should prioritize capturing rotational motion, whereas when they are close, accurate position tracking becomes crucial for achieving contact. To reflect this, we define the weights $\boldsymbol{w}_d$ that are inversely proportional to the distances between joints and the object [113]. The reward thus includes cost functions for joint position $E_p^h = \langle \boldsymbol{\Delta}_p^h, \boldsymbol{w}_d \rangle$, rotation $E_\theta^h = \langle \boldsymbol{\Delta}_\theta^h, \mathbf{1} - \boldsymbol{w}_d \rangle$, and interaction tracking $E_d = \langle \boldsymbol{\Delta}_d, \boldsymbol{w}_d \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product, $\boldsymbol{\Delta}_p^h[i] = \|\hat{\boldsymbol{p}}^h[i] - \boldsymbol{p}^h[i]\|$, $\boldsymbol{\Delta}_\theta^h[i] =$



(i) Reference contact markers (ii) RSI vs. PSI

Figure 3. (i) Visualization of reference contact markers that accommodate varied contact distances: red to promote contact, green for neutral areas where contact is neither promoted nor penalized, and blue to penalize contact. (ii) Initializing the rollout with reference (RSI) or reference corrected via simulation (PSI).

$\|\hat{\boldsymbol{\theta}}^h[i] \ominus \boldsymbol{\theta}^h[i]\|$, and $\boldsymbol{\Delta}_d[i] = \|\hat{\boldsymbol{d}}[i] - \boldsymbol{d}[i]\|$ represent the displacement for the variables defined in Sec. 3.1 with timestep t omitted. The formulation of $\boldsymbol{w}_d$ is provided in Sec. D of supplementary. The reward to be maximized can be formulated as $\exp(-\lambda E)$ for each cost function E with a specific hyperparameter λ. Details can be found in Sec. D.

Embodiment-Agnostic Reward. The reward includes components for object tracking and contact tracking. The object tracking cost is defined for position $E_p^o = \|\hat{\boldsymbol{p}}^o - \boldsymbol{p}^o\|$ and rotation $E_\theta^o = \|\hat{\boldsymbol{\theta}}^o - \boldsymbol{\theta}^o\|$, with all values normalized to the human's current position and direction. The contact tracking reward comprises two cost functions: body contact promotion E_b^c and penalty E_p^c, both aligning the simulated contact $\boldsymbol{c}$ with reference markers $\hat{\boldsymbol{c}}$, as shown in Figure 3. We define three contact levels – promotion, penalty, and neutral – to accommodate potential inaccuracies in reference contact distances. The detailed formulation can be found in Sec. D of the supplementary. Since the physics engine does not differentiate between object, ground, and self-contact, we adopt two strategies: **(i)** we model foot-ground contact promotion and penalty. This ensures proper foot lifting during cyclic walking and mitigates foot hobbling. **(ii)** We allow self-collision to avoid self-contact promotion but to promote object interaction. This poses minimal risk as the policy is guided by MoCap reference, which, although lacking perfect contact accuracy, rarely shows self-penetration. For humanoid robots with embodiments that differ from the MoCap reference and require real-world applicability, we disable self-collision, as discussed in Sec. F.

We introduce energy consumption rewards [105] to penalize large human or object jitters, with a proposed contact energy penalizing abrupt contact to promote compliant interactions. See Sec. D of supplementary for more details.

Hand Interaction Discovery. We use data with average or flattened hand poses [3, 39], which makes accurate object manipulation imitation challenging. To address this, we activate a reference contact marker for any hand part when a *fingertip* or *palm* is near an object. Given tasks that do not demand high dexterity, employing a contact-promoting reward with this marker enables policies to develop effective hand interaction strategies, leveraging the exploratory nature of RL. Additionally, we constrain the range of motion (RoM) of the hands to ensure natural movement. See Sec. D and Sec. B of the supplementary for further details.

Policy Learning. Following [63], the control policy π is trained using PPO [71] with the policy gradient $L(\psi) = \mathbb{E}_t[\min(r_t(\psi)A_t, \text{clip}(r_t(\psi), 1-\epsilon, 1+\epsilon)A_t)$. ψ are the parameters of π and $r_t(\psi)$ quantifies the difference in action likelihoods between updated and old policies. ϵ is a small constant, and A_t is the advantage estimate given by the generalized advantage estimator GAE(λ) [70].

Physical State Initialization. Learning later-phase motion can be essential for policies to achieve high rewards during earlier phases, compared to incrementally learning from the starting phase. Thus, Reference State Initialization (RSI) [63] sets the current pose $\boldsymbol{q}_t$ to a reference pose $\hat{\boldsymbol{q}}_t$ at a random timestep t, for initializing the rollout. However, initializing with the imperfect reference can introduce *critical artifacts*, such as contact floating or incorrect hand motion, leading to unrecoverable failures, *e.g.*, object falling, as depicted in Figure 3(ii). These issues render many initializations ineffective, limiting training on certain interaction phases since successful rollouts may not reach them before the maximum length. The problem is exacerbated by the use of prioritized sampling [79, 91], which favors high-failure-rate initializations.

To address the need for higher-quality reference initialization, we propose *Physical State Initialization* (PSI). As illustrated in Figure 2, PSI begins by creating an initialization buffer that stores reference states from MoCap and simulation states from prior rollouts. For each new rollout, an initial state is randomly selected from this buffer, which increases the likelihood of starting from advantageous positions. Once a rollout is completed, trajectories are evaluated based on their expected discounted rewards; those above a certain threshold are added to the buffer using a first-in-first-out (FIFO) strategy, while older or lower-quality trajectories are discarded. This selective reintroduction of high-value states for initialization helps maintain stable policy updates. We apply PSI in a sparse manner to ensure training efficiency. As shown in Figure 3(ii), PSI can collect trajectories for policy update that RSI does not effectively utilize. Further details are provided in Sec. E of the supplementary.

Interaction Early Termination. Early Termination (ET) [63] is commonly used in motion imitation, ending an episode when a body part makes unplanned ground contact or when the character deviates significantly from the reference [50], thus stopping the policy from overvaluing invalid transitions. However, additional conditions should be considered for human-object interactions. We propose *Interaction Early Termination* (IET), which supplements ET with three extra checks: **(i)** Object points deviate from their references by more than 0.5 m on average. **(ii)** Weighted av-

erage distances between the character's joints and the object surface exceed 0.5 m from the reference. (**iii**) Any required body-object contact is lost for over 10 consecutive frames. Full conditions are detailed in Sec. E of the supplementary.

3.3. Imitation with Distillation

As shown in Figure 2, after training the teacher policies on data from each subject (Sec. 3.2), we aggregate them to train a student policy $\pi^{(\mathrm{S})}$ to master all skills. As outlined in Algorithm 1, the combined teacher policies, denoted by $\pi^{(\mathrm{T})}$ for brevity, serves dual roles by providing state-action trajectories $(\boldsymbol{s}^{(\mathrm{T})}, \boldsymbol{a}^{(\mathrm{T})})$: (**i**) the state $\boldsymbol{s}^{(\mathrm{T})}$ for reference distillation, and (**ii**) the action $\boldsymbol{a}^{(\mathrm{T})}$ for policy distillation.

Reference Distillation. Noisy MoCap data can hinder policy learning, especially at larger scales. In contrast, teacher policies trained on smaller-scale data effectively address these issues by correcting contact artifacts, refining hand placements, and recovering missing details (see Figures 1 and 5). To fully leverage teacher policies, we use their rollouts as references for defining the student policy's goal state and reward functions, distinguishing our approach from distillation based on only action output.

Policy Distillation. We apply distillation on action outputs, which we view as crucial for scaling policies to large datasets. In essence, we trade space for time: teacher policies are trained in parallel on smaller data subsets, allowing the student policy to scale through distillation. Following Algorithm 1, we begin with Behavior Cloning (BC) [29] and use RL fine-tuning to go beyond demonstration memorization, an approach common in LLM alignment [56]. We integrate BC into online policy updates with a staged schedule: we start with DAgger [69] and gradually transition to PPO. Throughout, the critic is continuously trained with the reward from Sec. 3.2. This RL fine-tuning phase is crucial as teacher policies may behave differently when performing similar skills, and simple BC can lead to suboptimal "averaging" behavior, where RL fine-tuning helps the student converge on optimal solutions.

Algorithm 1 Distillation with RL Fine-tuning

Input: A composite policy $\pi^{(\mathrm{T})}$ integrated from individual teacher policies, student policy parameters $\boldsymbol{\psi}$, student value function parameters $\boldsymbol{\phi}$, schedule hyperparameter β for DAgger, horizon length H for PPO
for $t = 0, 1, 2, \ldots$ **do**
  for h from 1 to H **do**
    Sample a variable $u \sim \mathrm{Uniform}(0, 1)$
    Collect $\boldsymbol{s}^{(\mathrm{T})}, \boldsymbol{a}^{(\mathrm{T})}$ from teacher $\pi^{(\mathrm{T})}$
    Obtain the refined reference from $\boldsymbol{s}^{(\mathrm{T})}$ to define $\boldsymbol{s}^{(\mathrm{S})}$ and $r(\cdot)$, obtain $\boldsymbol{a}^{(\mathrm{S})}$ from $\pi^{(\mathrm{S})}_{\boldsymbol{\phi}}(\boldsymbol{a}^{(\mathrm{S})}|\boldsymbol{s}^{(\mathrm{S})})$.
    if $u \leq \frac{t}{\beta}$ **then** ▷ Use the teacher
      Given $\boldsymbol{s}^{(\mathrm{S})}$, execute $\boldsymbol{a}^{(\mathrm{S})}$, observe $\boldsymbol{s}'^{(\mathrm{S})}, r$
    else ▷ Use the student
      Given $\boldsymbol{s}^{(\mathrm{S})}$, execute $\boldsymbol{a}^{(\mathrm{T})}$, observe $\boldsymbol{s}'^{(\mathrm{S})}, r$
    end if
    Store the transition $(\boldsymbol{s}^{(\mathrm{S})}, \boldsymbol{s}'^{(\mathrm{S})}, \boldsymbol{a}^{(\mathrm{S})}, \boldsymbol{a}^{(\mathrm{T})}, r)$
  end for
  Update $\boldsymbol{\phi}$ with TD(λ)
  Compute PPO objective: $L(\psi)$
  Compute $J(\psi) = \|\boldsymbol{a}^{(\mathrm{S})} - \boldsymbol{a}^{(\mathrm{T})}\|$
  Compute the weight: $w = \min(\frac{t}{\beta}, 1)$
  Update ψ by gradient: $\nabla_{\psi}(wL(\psi) + (1-w)J(\psi))$
end for

3.4. Architecture

We set the keyframe indices K (Sec. 3.1, Eq. 1) to $\{1, 16\}$ for the teacher policies and $\{1, 2, 4, 16\}$ for the student policy. The broader observation window for the student policy helps it better distinguish different skills with larger-scale data. Teacher policies employ MLPs, common in physics-based animation [63], while the student policy handles higher-dimensional observations, for which MLPs are less effective. Thus, we use a transformer [82] architecture for sequential modeling [79], as shown in Figure 2.

4. Experiments

We evaluate teacher policies on their ability to imitate imperfect HOI references, and assess the entire teacher-student framework for scalability to large-scale data and zero-shot generalization across various scenarios. Additional experiments are provided in Sec. G of supplementary.

Datasets. We use the following datasets from InterAct [103]: OMOMO [39], BEHAVE [3], HODome [109], IMHD [114], and HIMO [53]. OMOMO, containing 15 objects and approximately 10 hours of data, is our primary dataset for evaluating the full teacher-student distillation framework for its scale. We train 17 teacher policies, one per subject, with subject 14 reserved as the test set and the remaining data used for training the student policy. A small portion of data is discarded after teacher imitation due to severe MoCap errors that could not be corrected (see Sec. F and Sec. H of the supplementary). Additional datasets are used to evaluate teacher policies in various MoCap scenarios with different error levels and interaction types. We focus on highly dynamic motions (Figure 1) and interactions involving multiple body parts (Figure 4), while excluding scenarios such as carrying a bag with a strap, since the simulator [54] used lacks full support for soft bodies.

Metrics. We use the following metrics: (**i**) *Success Rate* is defined as the proportion of references that the policy successfully imitates at least once, averaged across all references, while (**ii**) *Duration* is the time (in seconds) that the imitation is maintained without triggering the interaction early termination conditions introduced in Sec. 3.2. (**iii**)

Figure 4. Qualitative comparison between PhysHOI [88] (top), the reference motion (middle) from the BEHAVE [3] dataset, and the interaction refined by our teacher trained on it (bottom). InterMimic faithfully imitates the interactions involving multiple body parts while correcting errors in the original reference.

Human Tracking Error (E_h) measures the per-joint position error (cm) between the simulated and reference human (excluding hand joints for BEHAVE [3] and OMOMO [39] due to inaccuracy). (**iv**) *Object Tracking Error* (E_o) measures the per-point position error (cm) between the simulated and reference object. Both errors are averaged over the duration of the imitation in the best-performing trial.

Baselines. To facilitate fair comparisons, we downgrade our method for teacher policy evaluation to imitate either a single MoCap clip (Figure 4) or multiple clips with a single object (Table 1), enabling direct comparison with PhysHOI [88] and SkillMimic [89] (Sec. 4.1 and 4.2). Due to the lack of established baselines for large-scale HOI imitation, we adapt the following variants for comparison with our student policy (Sec. 4.3): (**i**) **PPO** [71] trains an imitation policy from scratch, following [63]. We experiment with both versions, with and without *reference distillation*; (**ii**) **DAgger** [69] distills the student without RL fine-tuning, a process we refer to as *policy distillation*.

Implementation Details. The control policies operate at 30 Hz and are trained using the *Isaac Gym* simulator [54]. Teacher policies are implemented as MLPs with hidden layers of sizes 1024, 1024, and 512. The student policy is implemented as a three-layer Transformer encoder with 4 heads, a hidden size of 256, and a feed-forward layer of 512. The critics are also modeled as MLPs with the same architecture as the teacher policies. To integrate the student policy with kinematic generators, including text-to-HOI [62] and future interaction prediction [101], we train these models using reference data distilled by the teacher policies from the OMOMO [39] dataset, following the same train-test split as the student policy training. For the text-to-HOI model, we train it to generate 10 seconds of motion and use 24 generated samples for evaluation, while for future interaction prediction, the model generates 25 future frames given 10 past frames and we use 60 generated samples for evaluation. See Sec. F of the supplementary.

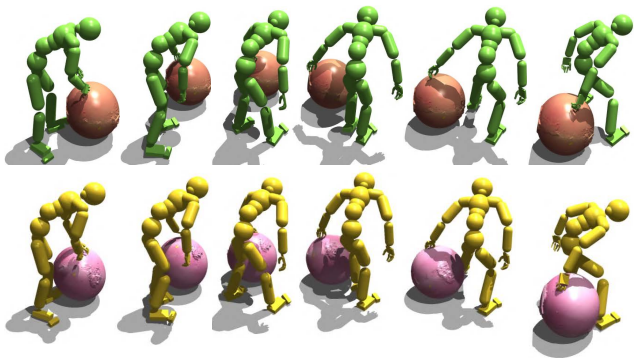


Figure 5. We recover plausible object rotations (bottom) that are challenging for motion capture due to the equivariant geometries of objects, which result in the object sliding on the ground (top).

Method	Time$^{\uparrow}$	$E_h{}^{\downarrow}$	$E_o{}^{\downarrow}$
SkillMimic [89]	12.2	7.2	13.4
InterMimic (**Ours**) w/o IET	40.3	6.7	9.9
InterMimic (**Ours**) w/o PSI	36.1	6.6	10.2
InterMimic (**Ours**)	**42.6**	**6.4**	**9.2**

Table 1. Quantitative comparison between the teacher policy from InterMimic and SkillMimic [89] to imitate data extracted from the BEHAVE [3] dataset involving a single subject interacting with yogamat. We ablate our proposed approach by removing interaction early termination and physical state initialization.

4.1. Quantitative Evaluation

Table 1 shows that the baseline struggles with MoCap imperfections, *e.g.*, incorrect hand positioning, and thus results in clearly shorter tracking durations. In contrast, our method maintains reference tracking for longer durations and produces interactions that closely match the reference. Table 2 shows that our method consistently outperforms baselines in both training data imitation and out-of-distribution generalization, including interactions from the test set and from kinematic generations. We discuss the effectiveness of specific design choices in Sec. 4.3.

4.2. Qualitative Evaluation

Figure 4 shows a representative sequence from the experiment in Table 1, illustrating how our teacher policy corrects interactions that PhysHOI [88] fails to track robustly – our method effectively withstands and corrects incorrect hand positioning and floating contacts in the reference. Beyond obvious errors, our method also rectifies the rotation of symmetric objects that MoCap inaccurately depicts as slid-

PPO	Reference Distillation	Policy Distillation	Architecture	OMOMO-Train [39]				OMOMO [39]-Test				OMOMO [39]-Test (w ×10)				HOI-Diff [62]				InterDiff [101]			
				Succ.↑	Time↑	E_h↓	E_o↓	Succ.↑	Time↑	E_h↓	E_o↓	Succ.↑	Time↑	E_h↓	E_o↓	Succ.↑	Time↑	E_h↓	E_o↓	Succ.↑	Time↑	E_h↓	E_o↓
✓	×	×	MLP	23.9	101.6	7.2	15.6	9.6	85.3	7.5	16.2	3.9	71.2	7.5	17.9	0.0	0.0	-	-	6.7	11.7	**6.2**	16.4
×	✓	✓		54.5	139.9	7.1	11.0	54.3	140.2	7.1	**11.2**	15.5	91.7	9.3	13.7	4.2	84.8	10.1	**9.7**	65.0	27.4	7.5	13.4
✓	✓	×		71.7	152.8	8.9	12.7	91.6	173.7	8.5	13.2	45.8	127.6	9.1	14.9	8.3	**130.9**	10.1	13.8	73.3	28.9	6.9	14.4
✓	✓	✓		**90.7**	**168.0**	**5.5**	**9.7**	95.5	173.9	**5.4**	11.9	**62.6**	**140.9**	**6.6**	14.5	**12.5**	121.4	8.6	12.1	75.0	29.1	**6.2**	13.5
✓	✓	✓	Transformer	88.8	167.0	6.0	10.2	**98.1**	**176.5**	5.9	11.3	56.8	134.7	**6.6**	**13.2**	**12.5**	119.0	**8.5**	12.6	**76.7**	**29.3**	6.4	**13.3**

Table 2. Quantitative evaluation of large-scale interaction imitation using OMOMO [39], kinematic generations from HOI-Diff [62], and InterDiff [101]. Additionally, we evaluate on test set when objects with weights ten times greater than those used during training.

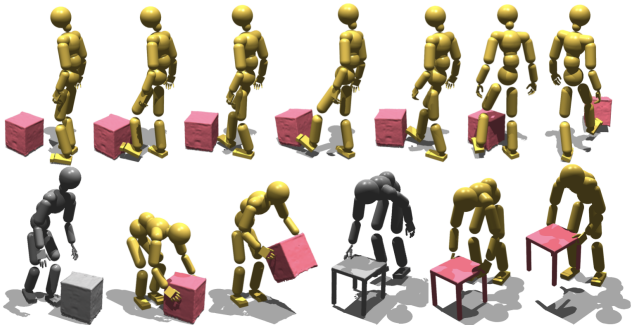


Figure 6. Zero-shot integration with a text-to-HOI model HOI-Diff [62] (**Top**), using 'Kick the large box' as the prompt, and an interaction prediction model InterDiff [101] (**Bottom**), where gray meshes are past states and colored illustrate future generations.

ing along the ground, shown in Figure 5. Figure 6 presents additional examples that complement Figure 1, demonstrating how our approach integrates with kinematic generators for future interaction prediction and text-to-interaction synthesis. This zero-shot generalization extends to novel objects unseen during training (Figure 7), highlighting the effectiveness of our object geometry and contact-encoded representation, as well as the large-scale training.

4.3. Ablation Study

Effectiveness of PSI and IET. We conduct an ablation study, as demonstrated in Table 1, comparing the full approach to "Ours w/o PSI". The results validate that Physical State Initialization (PSI) is effective by mitigating inaccuracies in the motion capture data. We also observe reduced effectiveness without our interaction early termination, as training often spends rollouts on irrelevant periods.

Effectiveness of Reference Distillation. Compared to directly scaling imitation from MoCap with potential imperfections (line 1 in Table 2), using references refined by the teacher policy (line 3) achieves consistently better performance on all metrics. The improvement is even more pronounced on the test set, where, without reference distillation, the policy struggles with unseen shapes, while retargeting by reference distillation eliminates the difficulty.

Effectiveness of Joint PPO and DAgger Updates. As shown in Table 2, training a policy from scratch (line 3) or relying solely on policy distillation (DAgger, line 2) fails to

Figure 7. Zero-shot generalization of our student policy on novel objects from BEHAVE [3] and HODome [109].

achieve optimal performance. While supervised skill learning lays the groundwork, additional PPO fine-tuning is crucial for resolving conflicts among teacher policies. This is important because our subject-based clustering may not effectively distinguish between different interaction patterns, and ambiguity arises when multiple teachers produce different actions for similar motions.

Effectiveness of Transformer for Policy Learning. From Table 2, we see that using a Transformer policy (line 5) outperforms MLP-based approaches, particularly on the test set and out-of-distribution cases generated by the kinematic model. We attribute this to the Transformer's inductive bias in sequential modeling and its capacity to incorporate longer-term observations, enabling it to handle complex spatio-temporal dependencies more effectively.

5. Conclusion

In this work, we introduce a framework for synthesizing realistic human-object interactions that are both physically grounded and generalizable. Unlike previous methods, our approach leverages a rich repository of imperfect MoCap data to facilitate the learning of various interaction skills across a wide variety of objects. To address inaccuracies in the MoCap data, we propose contact-guided rewards and optimize trajectory collection, enabling teacher policies to recover missing physical details in the original data. These teacher policies are used to train student policies within a distillation framework that combines policy distillation and reference distillation, thus enabling efficient skill scaling. Our approach shows zero-shot generalizability, which effectively bridges the gap between imitation and generative capabilities by integrating with kinematic generation. We believe that this framework can be adapted for whole-body loco-manipulation for real-world robots, enabling them to handle objects with human-like dexterity and nuance.

Acknowledgments. We thank Wei Yang, Yu-Wei Chao, Arsalan Mousavian, Ankur Handa, Samuel Schulter, Morteza Ziyadi, and Xiaohan Fei for valuable discussions. This work was supported in part by NSF Grant 2106825, NIFA Award 2020-67021-32799, the Amazon-Illinois Center on AI for Interactive Conversational Experiences, the Toyota Research Institute, the IBM-Illinois Discovery Accelerator Institute, and Snap Inc. This work used computational resources, including the NCSA Delta and DeltaAI and the PTI Jetstream2 supercomputers through allocations CIS230012, CIS230013, and CIS240311 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, as well as the TACC Frontera supercomputer and Amazon Web Services (AWS) through the National Artificial Intelligence Research Resource (NAIRR) Pilot.

References

[1] Jinseok Bae, Jungdam Won, Donggeun Lim, Cheol-Hui Min, and Young Min Kim. Pmp: Learning to physically interact with environments using part-wise motion priors. In *SIGGRAPH*, 2023. 3

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

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *CVPR*, 2022. 2, 3, 5, 6, 7, 8

[4] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *3DV*, 2024. 2, 3

[5] Yu-Wei Chao, Jimei Yang, Weifeng Chen, and Jia Deng. Learning to sit: Synthesizing human-chair interactions via hierarchical control. In *AAAI*, 2021. 2

[6] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *CVPR*, 2022. 4

[7] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *CVPR*, 2020. 3

[8] Jieming Cui, Tengyu Liu, Nian Liu, Yaodong Yang, Yixin Zhu, and Siyuan Huang. AnySkill: Learning open-vocabulary physical skill for interactive agents. In *CVPR*, 2024. 3

[9] Ravinder S Dahiya, Giorgio Metta, Maurizio Valle, and Giulio Sandini. Tactile sensing—from humans to humanoids. *IEEE transactions on robotics*, 26(1):1–20, 2009. 4

[10] Divyanshu Daiya, Damon Conover, and Aniket Bera. COLLAGE: Collaborative human-agent interaction generation using hierarchical latent diffusion and language models. *arXiv preprint arXiv:2409.20502*, 2024. 3

[11] Jeremy Dao, Helei Duan, and Alan Fern. Sim-to-real learning for humanoid box loco-manipulation. In *ICRA*, 2024. 2

[12] Christian Diller and Angela Dai. CG-HOI: Contact-guided 3d human-object interaction generation. In *CVPR*, 2024. 3

[13] Jiawei Gao, Ziqin Wang, Zeqi Xiao, Jingbo Wang, Tai Wang, Jinkun Cao, Xiaolin Hu, Si Liu, Jifeng Dai, and Jiangmiao Pang. CooHOI: Learning cooperative human-object interaction with manipulated object dynamics. In *NeurIPS*, 2024. 3

[14] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. IMoS: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, 2023. 3

[15] Michael Gleicher. Motion editing with spacetime constraints. In *Proceedings of the 1997 symposium on Interactive 3D graphics*, 1997. 3

[16] Zhaoyuan Gu, Junheng Li, Wenlan Shen, Wenhao Yu, Zhaoming Xie, Stephen McCrory, Xianyi Cheng, Abdulaziz Shamsah, Robert Griffin, C Karen Liu, et al. Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning. *arXiv preprint arXiv:2501.02116*, 2025. 2

[17] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 4

[18] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *SIGGRAPH*, 2023. 3

[19] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024. 2

[20] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *IROS*, 2024. 2

[21] Wenkun He, Yun Liu, Ruitao Liu, and Li Yi. SyncDiff: Synchronized motion diffusion for multi-body human-object interaction synthesis. *arXiv preprint arXiv:2412.20104*, 2024. 3

[22] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 3

[23] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *GCPR*, 2022. 3

[24] Inspire-robots. Smaller and higher-precision motion control experts. https://inspire-robots.store/. 1, 4, 5

[25] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. CHAIRS: Towards full-body articulated human-object interaction. In *ICCV*, 2023. 3

[26] Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-

scene interaction synthesis from text instruction. In *SIGGRAPH Asia*, 2024. 3

[27] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *CVPR*, 2024. 3

[28] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. DexMimicGen: Automated data generation for bimanual dexterous manipulation via imitation learning. *arXiv preprint arXiv:2410.24185*, 2024. 2

[29] Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. SuperPADL: Scaling language-directed physics-based control with progressive supervised distillation. In *SIGGRAPH*, 2024. 6

[30] Hyeonwoo Kim, Sangwon Beak, and Hanbyul Joo. DAViD: Modeling dynamic affordance of 3d objects using pre-trained video diffusion models. *arXiv preprint arXiv:2501.08333*, 2025. 3

[31] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. ParaHome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. *arXiv preprint arXiv:2401.10232*, 2024. 3

[32] Yeonjoon Kim, Hangil Park, Seungbae Bang, and Sung-Hee Lee. Retargeting human-object interaction to virtual avatars. *IEEE transactions on visualization and computer graphics*, 22(11):2405–2412, 2016. 4

[33] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 5

[34] Dieter Kraft. Algorithm 733: Tomp–fortran modules for optimal control calculations. *ACM Transactions on Mathematical Software (TOMS)*, 20(3):262–281, 1994. 4

[35] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. NIFTY: Neural object interaction fields for guided human motion synthesis. In *CVPR*, 2024. 3

[36] Jiye Lee and Hanbyul Joo. Locomotion-Action-Manipulation: Synthesizing human-scene interactions in complex 3d environments. In *ICCV*, 2023. 2

[37] Kang Hoon Lee, Myung Geol Choi, and Jehee Lee. Motion patches: building blocks for virtual environments annotated with motion data. In *SIGGRAPH*. 2006. 3

[38] Yoonsang Lee, Sungeun Kim, and Jehee Lee. Data-driven biped control. In *SIGGRAPH*. 2010. 2

[39] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 2, 3, 5, 6, 7, 8

[40] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *ECCV*, 2024. 3

[41] Jinhan Li, Yifeng Zhu, Yuqi Xie, Zhenyu Jiang, Mingyo Seo, Georgios Pavlakos, and Yuke Zhu. OKAMI: Teaching humanoid robots manipulation skills through single video imitation. In *CoRL*, 2024. 2

[42] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. In *WACV*, 2024. 3

[43] Fukang Liu, Zhaoyuan Gu, Yilin Cai, Ziyi Zhou, Shijie Zhao, Hyunyoung Jung, Sehoon Ha, Yue Chen, Danfei Xu, and Ye Zhao. Opt2skill: Imitating dynamically-feasible whole-body trajectories for versatile humanoid loco-manipulation. *arXiv preprint arXiv:2409.20514*, 2024. 2

[44] Libin Liu and Jessica Hodgins. Learning to schedule control fragments for physics-based characters using deep q-learning. *ACM Transactions on Graphics (TOG)*, 36(3): 1–14, 2017. 3

[45] Yunze Liu, Changxi Chen, Chenjing Ding, and Li Yi. PhysReaction: Physically plausible real-time humanoid reaction synthesis via forward dynamics guided 4d imitation. In *ACMMM*, 2024. 2

[46] Yun Liu, Bowen Yang, Licheng Zhong, He Wang, and Li Yi. Mimicking-bench: A benchmark for generalizable humanoid-scene interaction learning via human mimicking. *arXiv preprint arXiv:2412.17730*, 2024. 3

[47] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *CVPR*, 2024.

[48] 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*, 2025. 3

[49] Jintao Lu, He Zhang, Yuting Ye, Takaaki Shiratori, Sebastian Starke, and Taku Komura. CHOICE: Coordinated human-object interaction in cluttered environments for pick-and-place actions. *arXiv preprint arXiv:2412.06702*, 2024. 3

[50] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *ICCV*, 2023. 2, 4, 5, 3

[51] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Grasping diverse objects with simulated humanoids. In *NeurIPS*, 2024. 2, 3

[52] Zhengyi Luo, Jiashun Wang, Kangni Liu, Haotian Zhang, Chen Tessler, Jingbo Wang, Ye Yuan, Jinkun Cao, Zihui Lin, Fengyi Wang, et al. SMPLOlympics: Sports environments for physically simulated humanoids. *arXiv preprint arXiv:2407.00187*, 2024. 3

[53] Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, et al. HIMO: A new benchmark for full-body human interacting with multiple objects. In *ECCV*, 2024. 3, 6

[54] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. In *NeurIPS*, 2021. 6, 7, 2, 5

[55] Josh Merel, Saran Tunyasuvunakool, Arun Ahuja, Yuval Tassa, Leonard Hasenclever, Vu Pham, Tom Erez, Greg Wayne, and Nicolas Heess. Catch & carry: reusable neural controllers for vision-guided whole-body tasks. *ACM Transactions on Graphics (TOG)*, 39(4):39–1, 2020. 3

[56] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini

Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. 2, 6

[57] Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes. In *3DV*, 2024. 2

[58] Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. TokenHSI: Unified synthesis of physical human-scene interactions through task tokenization. In *CVPR*, 2025. 3

[59] Soohwan Park, Hoseok Ryu, Seyoung Lee, Sunmin Lee, and Jehee Lee. Learning predict-and-simulate policies from unorganized human motion data. *ACM Transactions on Graphics (TOG)*, 38(6):1–11, 2019. 4

[60] Sungjae Park, Seungho Lee, Mingi Choi, Jiye Lee, Jeonghwan Kim, Jisoo Kim, and Hanbyul Joo. Learning to transfer human hand skills for robot manipulations. *arXiv preprint arXiv:2501.04169*, 2025. 2

[61] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 2, 5

[62] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. HOI-Diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023. 3, 7, 8, 5

[63] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018. 2, 3, 4, 5, 6, 7

[64] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021. 3

[65] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)*, 41(4):1–17, 2022. 3

[66] Haziq Razali and Yiannis Demiris. Action-conditioned generation of bimanual object manipulation sequences. In *AAAI*, 2023. 3

[67] Daniele Reda, Jungdam Won, Yuting Ye, Michiel van de Panne, and Alexander Winkler. Physics-based motion retargeting from sparse inputs. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–19, 2023. 4

[68] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017. 2

[69] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011. 6, 7

[70] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *ICLR*, 2016. 5

[71] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2, 5, 7

[72] Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation. In *RSS*, 2024. 2

[73] Wenfeng Song, Xinyu Zhang, Shuai Li, Yang Gao, Aimin Hao, Xia Hou, Chenglizhao Chen, Ning Li, and Hong Qin. HOIAnimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In *CVPR*, 2024. 3

[74] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209–1, 2019. 3

[75] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics (TOG)*, 39(4):54–1, 2020. 3

[76] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 3

[77] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. GOAL: Generating 4d whole-body motion for hand-object grasping. In *CVPR*, 2022. 3

[78] Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. Calm: Conditional adversarial latent models for directable virtual characters. In *SIGGRAPH*, 2023. 3

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

[80] Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit H Bermano, and Michiel van de Panne. CLoSD: Closing the loop between simulation and diffusion for multi-task character control. In *ICLR*, 2025. 3

[81] Unitree. Unitree g1 humanoid agent ai avatar. https://www.unitree.com/g1/. 1, 4, 5

[82] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 6

[83] Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, and Jun Saito. Contact-aware retargeting of skinned motion. In *ICCV*, 2021. 4

[84] Weilin Wan, Lei Yang, Lingjie Liu, Zhuoying Zhang, Ruixing Jia, Yi-King Choi, Jia Pan, Christian Theobalt, Taku Komura, and Wenping Wang. Learn to predict how humans manipulate large-sized objects from interactive motions. *IEEE Robotics and Automation Letters*, 2022. 3

[85] Jiashun Wang, Jessica Hodgins, and Jungdam Won. Strategy and skill learning for physics-based table tennis animation. In *SIGGRAPH*, 2024. 3

[86] Ruocheng Wang, Pei Xu, Haochen Shi, Elizabeth Schumann, and C Karen Liu. Fürelise: Capturing and physically synthesizing hand motion of piano performance. In *SIGGRAPH Asia*, 2024. 2

[87] Wenjia Wang, Liang Pan, Zhiyang Dou, Zhouyingcheng Liao, Yuke Lou, Lei Yang, Jingbo Wang, and Taku Komura. SIMS: Simulating human-scene interactions with real world script planning. *arXiv preprint arXiv:2411.19921*, 2024. 3

[88] 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. 3, 7, 2

[89] Yinhuai Wang, Qihan Zhao, Runyi Yu, Ailing Zeng, Jing Lin, Zhengyi Luo, Hok Wai Tsui, Jiwen Yu, Xiu Li, Qifeng Chen, et al. SkillMimic: Learning reusable basketball skills from demonstrations. In *CVPR*, 2025. 3, 7, 2

[90] Zhenzhi Wang, Jingbo Wang, Dahua Lin, and Bo Dai. InterControl: Generate human motion interactions by controlling every joint. In *NeurIPS*, 2024. 2

[91] Jungdam Won and Jehee Lee. Learning body shape variation in physics-based characters. *ACM Transactions on Graphics (TOG)*, 38(6):1–12, 2019. 4, 5

[92] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. A scalable approach to control diverse behaviors for physically simulated characters. *ACM Transactions on Graphics (TOG)*, 39(4):33–1, 2020. 2, 4

[93] Qianyang Wu, Ye Shi, Xiaoshui Huang, Jingyi Yu, Lan Xu, and Jingya Wang. THOR: Text to human-object interaction diffusion via relation intervention. *arXiv preprint arXiv:2403.11208*, 2024. 3

[94] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. SAGA: Stochastic whole-body grasping with contact. In *ECCV*, 2022. 3

[95] Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. Human-object interaction from human-level instructions. *arXiv preprint arXiv:2406.17840*, 2024. 3

[96] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *ICLR*, 2024. 2

[97] Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. InterTrack: Tracking human object interaction without object templates. In *3DV*, 2024. 3, 6

[98] Zhaoming Xie, Sebastian Starke, Hung Yu Ling, and Michiel van de Panne. Learning soccer juggling skills with layer-wise mixture-of-experts. In *SIGGRAPH*, 2022. 3

[99] Zhaoming Xie, Jonathan Tseng, Sebastian Starke, Michiel van de Panne, and C Karen Liu. Hierarchical planning and control for box loco-manipulation. *arXiv preprint arXiv:2306.09532*, 2023. 3

[100] Pei Xu and Ruocheng Wang. Synchronize dual hands for physics-based dexterous guitar playing. In *SIGGRAPH Asia*, 2024. 2

[101] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 1, 3, 7, 8, 5

[102] Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. In *NeurIPS*, 2024. 1, 3

[103] 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*, 2025. 3, 6

[104] Zeshi Yang, Kangkang Yin, and Libin Liu. Learning to use chopsticks in diverse gripping styles. *ACM Transactions on Graphics (TOG)*, 41(4):1–17, 2022. 2

[105] Wenhao Yu, Greg Turk, and C Karen Liu. Learning symmetric and low-energy locomotion. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 5

[106] Ye Yuan and Kris Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. In *NeurIPS*, 2020. 2

[107] Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, and Jiajun Wu. Generalizable humanoid manipulation with improved 3d diffusion policies. *arXiv preprint arXiv:2410.10803*, 2024. 2

[108] Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 3

[109] 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*, 2023. 3, 6, 8

[110] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In *CVPR*, 2024. 3

[111] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. COUCH: Towards controllable human-chair interactions. In *ECCV*, 2022. 3

[112] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez-Pellitero, and Gerard Pons-Moll. FORCE: Dataset and method for intuitive physics guided human-object interaction. In *3DV*, 2024. 3

[113] Yunbo Zhang, Deepak Gopinath, Yuting Ye, Jessica Hodgins, Greg Turk, and Jungdam Won. Simulation and retargeting of complex multi-character interactions. In *SIGGRAPH*, 2023. 3, 4

[114] 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. 3, 6