InterPrior: Scaling Generative Control for Physics-Based Human-Object Interactions

Sirui Xu[1] Samuel Schulter[2] Morteza Ziyadi[2] Xialin He[1]
Xiaohan Fei[2] Yu-Xiong Wang[1†] Liang-Yan Gui[1†]
[1]University of Illinois Urbana-Champaign [2]Amazon
[†]Equal Advising

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

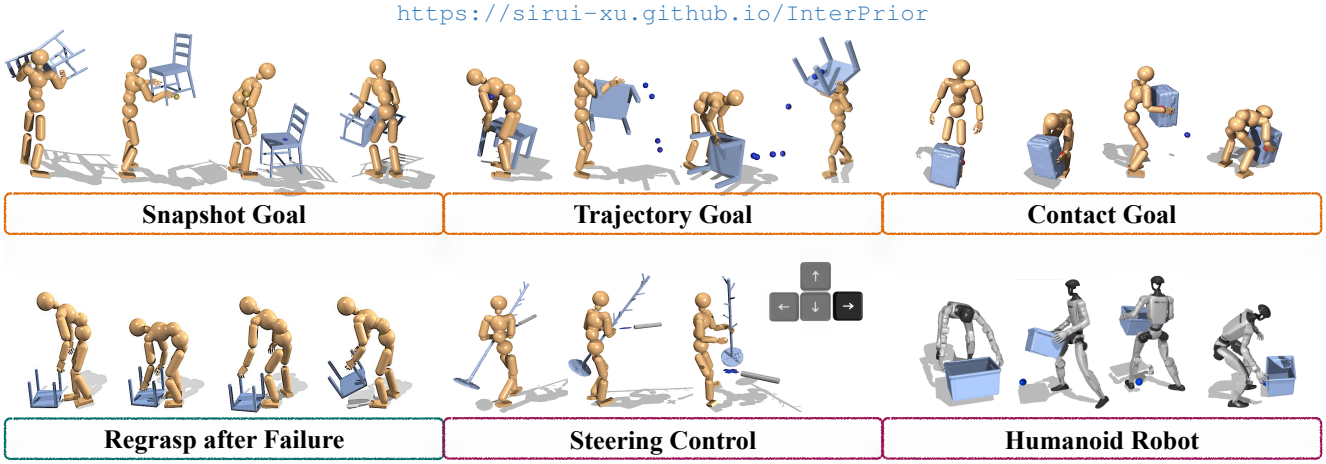

Figure 1. InterPrior is a *versatile generative controller* instantiated as a goal-conditioned policy that controls a simulated humanoid to follow goal guidance and interact with objects in a physics-based simulator. Three core, composable capabilities enable pursuing (**I**) long-horizon snapshot goals, (**II**) trajectory goals, and (**III**) contact goals (*Top*). *Yellow*, *blue*, and *red* dots respectively denote *human*, *object*, and *contact* goals. It demonstrates failure recovery (*Bottom Left*) from unsuccessful grasps. InterPrior enables steering control from a human operator and can be applied to humanoid robot embodiments (*Bottom Right*). More demo videos are provided in the webpage.

Abstract

Humans rarely plan whole-body interactions with objects at the level of explicit whole-body movements. High-level intentions, such as affordance, define the goal, while coordinated balance, contact, and manipulation can emerge naturally from underlying physical and motor priors. Scaling such priors is key to enabling humanoids to compose and generalize loco-manipulation skills across diverse contexts while maintaining physically coherent whole-body coordination. To this end, we introduce InterPrior, a scalable framework that learns a unified generative controller through large-scale imitation pretraining and post-training by reinforcement learning. InterPrior first distills a full-reference imitation expert into a versatile, goal-conditioned variational policy that reconstructs motion from multimodal observations and high-level intent. While the distilled policy reconstructs training behaviors, it does not generalize reliably due to the vast configuration space of large-scale human-object interactions. To address this, we apply data augmentation with physical perturbations, and then perform reinforcement learning finetuning to improve competence on unseen goals and initializations. Together, these steps consolidate the reconstructed latent skills into a valid manifold, yielding a motion prior that generalizes beyond the training data, e.g., it can incorporate new behaviors such as interactions with unseen objects. We further demonstrate its effectiveness for user-interactive control and its potential for real robot deployment.

1. Introduction

Human-object interaction (HOI) is inherently hierarchical: humans plan at a high level with sparse intentions, while detailed limb coordination, balance, and contact emerge through *fast*, *intuitive* motor responses [62]. For instance, when reaching for a bottle, we plan the hand's target and object motion, while the rest of the body follows through *subconscious* coordination. Motion imitation policies [88] have scaled to large HOI skills but rely on explicit planners for dense full-body and object references. In contrast,

an *interaction motor prior* should sample feasible loco-manipulation behaviors from a distribution conditioned on sparse goals, *e.g.*, next-second hand contact, rather than simply mimicking deterministic, fully specified trajectories.

To model a distribution over feasible loco-manipulation behaviors, early work [15, 45] learns a generative controller via adversarial distributional matching and then uses reinforcement learning (RL) to promote task achievement under it. These methods can expand motion coverage beyond demonstrations, but are hard to scale due to unstable optimization, discriminator mode collapse, and handcrafted task objectives. An alternative is to distill reference imitation policies [38], with goal conditioning [60] achieved without task-specific design. While these approaches can absorb large-scale data, they can be brittle when reference coverage lags far behind the configuration space—as in loco-manipulation, where even a few object degrees of freedom can induce a combinatorial explosion of contact modes and relative poses with different geometries.

To address these limitations, we introduce *InterPrior*, a physics-based HOI controller that is *scalable* along four axes (Figure 1). (**I**) *task coverage*: a single policy supports multiple goal formulations, *e.g.*, sparse targets and their compositions; (**II**) *skill coverage*: the same training recipe scales to large HOI data and enables affordance-rich interactions beyond simple grasping; (**III**) *motion coverage*: it generates expressive trajectories instead of merely reconstructing demonstrations; and (**IV**) *dynamics coverage*: it maintains task success under varied physical properties.

Our key insight is that *RL finetuning* is essential for turning distillation from data reconstruction into a robust, generalizable policy. Distillation alone cannot cover the full HOI configuration space, yet RL applied in isolation often drifts toward unnatural reward-hacking behaviors. We therefore use distillation to provide a strong, natural initialization, and apply RL as a *local optimizer* that improves robustness while remaining anchored to the pretrained model. Concretely, we leverage *distillation* to inherit broad skills from large-scale HOI demonstrations, by training a masked conditional variational policy to reconstruct motor control from sparse, multimodal goals, distilled from a reference imitation expert. We then *RL finetune* this policy to consolidate its latent skills into a *valid interaction manifold*. The finetuning optimizes two objectives: improving success on unseen goals and initializations, and preserving pretrained knowledge through regularization. It leverages the pretrained base policy to synthesize natural in-between motions, with failure states to acquire recovery behaviors, *e.g.* re-approach and re-grasp. Together, these steps transform reconstructed latent skills into a stable, continuous manifold that generalizes beyond the training trajectories.

Our contributions are fourfold. (**I**) We present *InterPrior*, a generalizable generative controller for physics-based human-object interaction, encompassing diverse skills rather than fixed procedural routines (*e.g.*, approach, grasp, place) typical of prior work. (**II**) We develop an RL finetuning strategy that enables robust failure recovery and goal execution across varied configurations while maintaining human-like coordination. The resulting controller supports mid-trajectory command switching, re-grasps after failures, and remains stable under perturbations. (**III**) We show that our finetuning strategy naturally extends to *novel objects and interactions*, functioning as a reusable prior. (**IV**) We demonstrate embodiment flexibility by training on the G1 humanoid [65] with sim-to-sim evaluation and enabling real-time control via keyboard interfaces.

2. Related Work

Data-driven human interaction animation has progressed from kinematic models assuming simplified object dynamics [67, 100, 104] to methods generating whole-body motions with dynamic objects [5, 8, 13, 14, 16, 20, 22, 23, 26, 34, 46, 50, 51, 75, 77, 84, 85, 89, 93, 98]. However, these kinematic approaches often exhibit implausible contact drift and interpenetration. Such limitations partly arise from existing HOI datasets [3, 19, 21, 27, 29, 32, 35, 41, 79, 82, 96, 97, 99, 103], which contain spatial or physical inconsistencies that impede the learning of realistic interactions. Physics-based methods seek to address this gap but often rely on early curated datasets [57] focusing on limited yet high-fidelity hand-centric manipulations [38, 60, 72]. Recent advances in humanoid hardware [2, 10, 25, 56, 105] have begun to bridge the virtual and physical domains, though typically without too much agility. Together, these developments highlight the need for *scalable HOI priors*, models capable of generalizing across tasks, remaining robust to imperfect data, and synthesizing realistic HOIs.

2.1. Physics-based Character Animation

Physics-based character animation learns simulated controllers via RL, *e.g.*, tracking reference motions [47, 102]. Scalability has been improved through multi-clip trackers with reference planners [24, 70, 73] without or with closed-loop schemes [61, 83]. Nevertheless, such controllers remain constrained by their reference motion planners, making them fragile when the planned motions are dynamically unstable, a very common issue in HOI, where kinematic planners often neglect physical feasibility. *Learned generative priors* address this limitation by encoding physically plausible motor memory encoded into policies. One line of research employs adversarial imitation with discriminators [48] to learn the motor prior, and later extends to skill embeddings [49] and conditional control [9, 58]. These approaches promote motion diversity but remain sample-inefficient and challenging to scale. A complementary line distills motor skills into compact latent codes. Earlier work

adopts model learning to train a variational autoencoder (VAE) [28] based controller [11, 74, 90, 91], while recent studies pretrain universal trackers [36] and distill them into latent priors [37], masked policies [59], or offline training with diffusion models [18, 64, 76]. Yet, these methods are often limited by the expert coverage. Our InterPrior synergizes the strength of both lines: it first distills large-scale motion imitators and finetunes it via RL, bridging a generative controller with versatile conditions while enhancing the control by alleviating out-of-distribution brittleness.

2.2. Physics-based Human-Object Interaction

Advances in physics-based character control have progressively expandeded the scope of HOI animation. Early approaches primarily focus on simple object dynamics, such as striking or sitting [4, 6, 44, 49, 78], whereas recent developments have extended to complex, scenario-specific sports and games [1, 31, 39, 68, 69, 72, 80, 94]. Progress has also been observed in generalizable tasks, such as object carrying and rearrangement [7, 12, 15, 30, 43, 45, 55, 71, 95, 101], predominantly enabled by adversarial imitation learning, while most systems remain skill-specific, relying on fixed procedural routines (*e.g.* approach, grasp, place with regular-shaped objects). They struggle to adapt to objects that require careful affordances and fine-grained interaction skills (*e.g.*, grasping a chair bar with one hand). To address these limitations, HOI motion imitation [77, 81, 88, 92] has emerged as a promising paradigm for scaling skill repertoires and capturing fine-grained interactions, as it directly emphasizes precision and stability. Distilling such imitation policies therefore represents a crucial step toward establishing a *versatile HOI controller*. However, existing efforts often exhibit narrow task coverage, emphasizing single-object proficiency [92] or relying on curated dataset with low-dynamic and hand-centric skills [38, 40, 60]. Our InterPrior provides a principled solution for generalizing a generative controller for agile whole-body loco-manipulation.

3. Methodology

Task Formulation. We aim to learn a policy π that operates in a physics simulator and produces human-object interaction motion from high-level goals rather than full reference. Such goals can be extracted from a human user (*e.g.*, steering control), a HOI kinematic motion generator (see Sec. F), or keypoints from Motion Captured (MoCap) data. The policy π conditions on the current human-object state and recent history together with these goals, and samples control signals from its learned distribution to drive the simulated human or humanoid to interact with the object. The outcome is a rollout motion sequence that is physically simulated, follows the provided goals where available, and remains diverse and natural in aspects that are not specified.
Overview. Figure 2 illustrates our three-stage paradigm.

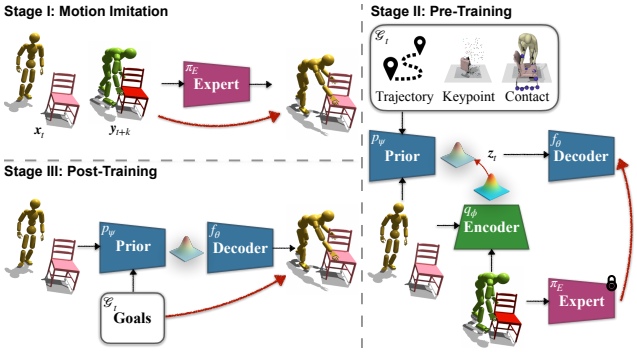

Figure 2. Overview of the proposed InterPrior framework. It consists of: (**I**) full-reference imitation expert training on large-scale human-object interaction data; (**II**) distillation of the expert into a variational policy with a structured latent space for skill embeddings; and (**III**) post-training of the variational policy to enhance generalization. Blue modules denote the final policy used at inference; green and red modules are training-only components, and red arrows denote supervision signals (rewards/losses).

First, we train an expert policy π_E for *large-scale* HOI motion imitation, incorporating data augmentation, physical perturbations, and shaped rewards to promote stable whole-body coordination and precise grasping across diverse configurations (Sec. 3.2). Second, we distill the expert into a masked conditional variational policy π that maps sparse goal inputs to a multi-modal distribution (Sec. 3.3). Third, we finetune this policy π using RL to enhance robustness under unseen configurations, employing failure-state resets to encourage recovery behaviors (Sec. 3.4). Each stage is modeled as a Markov Decision Process (MDP), which shares a consistent input formulation comprising observations and goal conditioning, as well as an output action corresponding to low-level actuation commands (Sec. 3.1).

3.1. Policy States and Actions

Observation. The policy input at time t includes an observation that aggregates human kinematics, object kinematics, and their interaction and contact states, $\boldsymbol{x}_t = [\underbrace{\boldsymbol{r}_t^h, \boldsymbol{\theta}_t^h, \dot{\boldsymbol{r}}_t^h, \dot{\boldsymbol{\theta}}_t^h}_{\text{human}}, \underbrace{\boldsymbol{r}_t^o, \boldsymbol{\theta}_t^o, \dot{\boldsymbol{r}}_t^o, \dot{\boldsymbol{\theta}}_t^o}_{\text{object}}, \underbrace{\boldsymbol{D}_t, \boldsymbol{C}_t}_{\text{interaction}}]$. Here, the superscripts h and o denote human and object quantities, respectively. $\boldsymbol{r}$ and $\boldsymbol{\theta}$ denote positions and orientations, respectively; the dotted terms indicate linear and angular velocities. The interaction terms include signed distances from body segments to object surfaces $\boldsymbol{D}_t$ and binary contacts $\boldsymbol{C}_t$ derived from simulator contact forces, following [88]. All continuous quantities are normalized in a human root-centric and local heading frame for invariance to global placement. The human-related terms contain 52 components for the SMPL humanoid [36] and 39 for the Unitree G1 robot [65]. Each rigid body contributes one element to human-related variables in $\boldsymbol{x}_t$, including $\boldsymbol{D}_t$ and $\boldsymbol{C}_t$, *e.g.*, $\boldsymbol{D}_t \in \mathbb{R}^{39\times 3}$ for G1. Objects are all rigid.

Goal Conditioning. The policy is also conditioned on a set of future *goals* that specify desired human-object configurations at different horizons. During training, we extract goals from reference, where each reference $\boldsymbol{y}_t$ shares the same state space as observation $\boldsymbol{x}_t$, including human, object, and contact components. A corresponding binary mask $\boldsymbol{m}_t$ indicates which components of the reference are provided to the policy [59]. To capture both near-term and distant intentions, we employ two types of goal conditioning: (**I**) a short-horizon preview sequence and (**II**) a long-horizon snapshot. Let H denote the maximum prediction horizon, $K \subset \{1, \dots, H\}$ a set of short-horizon offsets, and L a long-horizon offset. The long-horizon offset L is initialized randomly, decremented by one at each timestep, and re-sampled when it reaches zero. For each $k \in K \cup \{L\}$, we retrieve $(\boldsymbol{y}_{t+k}, \boldsymbol{m}_{t+k})$, where the mask $\boldsymbol{m}_{t+k}$ is sampled to cover every possible condition *e.g.*, end-effector pose, object pose, human-object contacts, their combination, *etc*. (see Sec. C for details of the sampling). Each goal is represented using a *masked residual encoding*: $\tilde{\boldsymbol{y}}_{t+k} = \boldsymbol{m}_{t+k} \odot \Delta\big(\boldsymbol{y}_{t+k}, \boldsymbol{x}_t\big)$, $\mathcal{G}_t = \{\, (\tilde{\boldsymbol{y}}_{t+k}, \boldsymbol{m}_{t+k}) \mid k \in K \cup \{L\} \,\}$, where $\odot$ denotes elementwise masking and Δ applies a log-map to rotational components and subtraction to Euclidean quantities. During inference, user-specified or model generated sparse targets can be supplied by filling only the informed components, setting the corresponding mask to one, and zeroing the rest.

Action. The policy outputs an action vector $\boldsymbol{a}_t$, defining the actuation as $\boldsymbol{a}_t \in \mathbb{R}^{51\times 3}$ for SMPL [33, 52] and $\boldsymbol{a}_t \in \mathbb{R}^{29}$ for the G1 humanoid [65]. Each action represents a joint position target expressed in the exponential map, which is subsequently converted into joint torques via proportional-derivative (PD) control. The resulting torques are applied to the corresponding joints in the physics simulator, driving the human-object interactions and generating the next state $\boldsymbol{x}_{t+1}$ according to the simulator's dynamics.

3.2. InterMimic+: Full-Reference Imitation Expert

Serving as the teacher for the final policy π, we formulate large-scale co-tracking of human and object motions following InterMimic [88]. At each timestep t, the expert policy π_E receives the observation along with future references, which contain complete information without masking. The policy outputs low-level actuation commands $\boldsymbol{a}_t$ and is trained using Proximal Policy Optimization (PPO) [54] to maximize a composite reward function: $r = r_{\text{track}} \times r_{\text{energy}}$, where r_{track} promotes alignment between the reference $\boldsymbol{y}_t$ and simulation state $\boldsymbol{x}_t$, and r_{energy} encourages physically plausible and efficient behaviors. This formulation enforces *strict adherence to the reference*.

The policy from the original InterMimic achieves high-fidelity imitation and broad loco-manipulation coverage. However, in practice, we observe key issues due to the policy's strong reliance on references, which we address with our advanced version. (**I**) The policy shows a degradation of precision when interacting with thin or small objects, as it tends to rigidly follow reference trajectories (See Figure 3) without utilizing fine-grained hand-object relations. (**II**) This limitation is more severe if the rollout deviates from reference trajectories. To mitigate these issues, we expand reference scope and introduce reference-free rewards.

Expanding Reference Scope. To reduce reliance on reference trajectories, we apply *randomization*, *perturbation*, and *augmentation*. We initialize each episode from reference frames with random variations in human-object poses. During rollouts, we apply sparse impulses, *i.e.*, random velocity perturbations to the pelvis and object, to induce deviations from the references. We augment object shapes and randomize physical properties such as mass density, center-of-mass offsets, inertia, and friction, with details presented in Sec. E. This exposes the policy to diverse dynamics, without alternating the reference. Unlike common sim-to-real practices, we do not randomize actuation parameters or add observation noise, as these do not directly enhance state or dynamics coverage. However, perturbations alone are insufficient; it is necessary to introduce a termination penalty that discourages the policy from entering failure under perturbation. We define $r_{\text{ter}} = -w_{\text{ter}} \times c_{\text{ter}}$, where c_{ter} is triggered by a human fall or large deviations in states from references, following [88], and w_{ter} is a scaling coefficient.

Reference-Free Reward. A key challenge in precise hand grasping under randomization and perturbation is that strict reference-based tracking becomes unreliable. To address this, we introduce a hand reward r_{h} that encourages the hand to *target* and *wrap* around the object based on the current simulation state, rather than relying on reference trajectories. Details of the formulation can be found in Sec. D. When combined with the reference imitation reward, it serves as a corrective term that guides the hand to orient, align, and close around the actual object, potentially deviated from the reference due to perturbations, rather than strictly following the reference trajectory. The full reward is defined as $r_t = (r_{\text{track}} \times r_{\text{energy}} \times r_{\text{h}}) + r_{\text{ter}}$.

3.3. InterPrior: Variational Distillation

Given an imitation expert policy π_E (Sec. 3.2) trained to master motor skills for HOI, our objective is to distill it into a *variational policy* π. Unlike the expert policy π_E, which operates under densely supervised and fully observed reference trajectories, the variational policy π must preserve naturalness and diversity with sparse cues. This is achieved by sampling from a latent skill distribution, which endows π with the capacity to generate plausible variations in action space. Our framework builds upon [59, 86] with two new designs: (**I**) *multi-modal conditioning*, including contact for versatile human-object conditioning, and (**II**) *prior shaping*

and bounding regularization for robustness.

Model. We model the policy π with a latent $\boldsymbol{z}_t \in \mathbb{R}^{d_z}$ to for multi-modality. As shown in Fig. 2, π consists of:

$$\begin{aligned} &\textbf{Prior:} && p_\psi(\boldsymbol{z}_t \mid \boldsymbol{x}_{t-\ell:t}, \mathcal{G}_t), \\ &\textbf{Encoder:} && q_\phi(\boldsymbol{z}_t \mid \boldsymbol{x}_t, \mathcal{G}_t, \boldsymbol{y}_{t:t+H}, \boldsymbol{y}_{t+L}), \\ &\textbf{Decoder:} && f_\theta(\boldsymbol{a}_t \mid \boldsymbol{x}_{t-\ell:t}, \boldsymbol{z}_t). \end{aligned}$$

The encoder is an MLP used only during training; given the full future reference, it outputs a Gaussian $\mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$. In parallel, a prior Transformer encodes recent history, with history length ℓ, and a sparse goal, producing a Gaussian $\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$. Following [59], we form a residual posterior $\mathcal{N}(\boldsymbol{\mu}_p + \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$. During training we sample the latent skill via reparameterization: $\boldsymbol{z}_t = (\boldsymbol{\mu}_p + \boldsymbol{\mu}_q) + \boldsymbol{\Sigma}_q^{1/2}\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and hold $\boldsymbol{\epsilon}$ fixed within an episode to promote temporally consistency [90]. During inference, only the prior is used to sample $\boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$. The decoder MLP maps the latent and observation to the action. The decoder also includes an auxiliary head during training that reconstructs the *masked* entries of the goal, encouraging a meaningful latent space by learning to *complete* intent from context.

Bounding the Latent. To improve robustness and prevent unnatural behaviors induced by out-of-distribution latents, after sampling we project $\boldsymbol{z}_t \leftarrow \boldsymbol{z}_t / \|\boldsymbol{z}_t\|$ so that the policy operates on a hypersphere, following [49]. This normalization stabilizes skill learning by limiting the rare latent draws while preserving directional variability for multi-modal behaviors. Note that we apply the projection after sampling, thus KL regularization can still be computed on the Gaussian p_ψ and q_ϕ before projection.

Online Distillation and Regularization. We utilize an online distillation framework following DAgger [53], where the student policy π learns from a mixture of expert π_E and self-generated rollouts. Training begins with trajectories fully controlled by the expert π_E, and the ratio of student-driven states is gradually increased as learning progresses. At each step, the expert provides its action output as supervision for the student. The policy is optimized using a composite objective consisting of multiple loss terms: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ELBO}} + \lambda_{\text{scale}}\mathcal{L}_{\text{scale}} + \lambda_{\text{tc}}\mathcal{L}_{\text{tc}}$. The primary objective, $\mathcal{L}_{\text{ELBO}}$, is a weighted evidence lower bound [28] that combines three components: (**I**) an *imitation loss* encouraging the student to reproduce expert actions, (**II**) a *goal reconstruction loss* promoting accurate completion of masked goal entries to align with the ground truth, and (**III**) a *KL regularization loss* that penalizes divergence between the posterior $\mathcal{N}(\boldsymbol{\mu}_p + \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ and the prior distribution $\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$. We introduce two auxiliary losses to further shape the latent. $\mathcal{L}_{\text{scale}}$ constrains the prior mean $\boldsymbol{\mu}_p$ to maintain unit magnitude, preventing degeneracy given hypersphere normalization. $\mathcal{L}_{\text{tc}}$ encourage consecutive prior distributions to remain similar across time steps. Details of these losses are provided in Sec. D.

3.4. InterPrior: Post-Training Beyond Reference

The distilled policy π (Sec. 3.3) exhibits goal following, yet it is brittle when the goal or human-object state drifts off the dataset distribution, *e.g.*, during transitions between skills. Unlike human-only motion [37] or small-object grasping [60], loco-manipulation tasks with coupled affordances span a far larger configuration space that references alone cannot cover. This follows from the learning dynamics of distillation: training proceeds by replaying dataset trajectories. Our key observation is that the pretrained π provides a strong and natural initialization for RL finetuning as a local optimizer that expands its scope along three axes: (**I**) recover from near-failure or failure states, (**II**) explore unseen yet plausible configurations without trajectory replay, and at the same time (**III**) preserve the naturalness of behaviors encoded by the pretrained policy. A natural alternative is to sample novel multi-frame trajectories that combine diverse human, object, and contact configurations and then train the policy to track them [38], but this requires a strong trajectory sampler, which is particularly challenging at loco-manipulation scale. Instead, we target *single-frame* goals: composing goals observed in data can induce unseen configurations, and we further combine such goals with randomized initializations and offsets to systematically broaden the state distribution encountered during RL.

In-Betweening for Finetuning. To mitigate the cost of exhaustive trajectory sampling, we formulate finetuning as an *in-betweening* task, where the policy tracks from a randomly sampled initial configuration toward a single-frame goal randomly drawn from the dataset. The policy is rewarded for progressing toward this sampled goal. The reward is defined as,

$$\begin{aligned} r_t^{\text{PT}} &= \left(r_{\text{energy}} \times r_{\text{h}}\right) \;+\; r_{\text{goal}} \;+\; r_{\text{ter}}, \\ r_{\text{goal}} &= \begin{cases} r_{\text{succ}}, & \text{if } \left\|\boldsymbol{m}_{t+L} \odot \Delta(\tilde{\boldsymbol{y}}_{t+L}, \boldsymbol{x}_t)\right\|_1 < \tau, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

where the terms r_{energy}, r_{ter}, and r_{h} are defined in Sec. 3.2. Since the goal is arbitrary by the random masking, we do not use a dense distance-based reward. The goal reward r_{goal} provides a sparse success signal that activates when the masked feature distance between the current state $\boldsymbol{x}_t$ and target $\tilde{\boldsymbol{y}}_{t+L}$ falls below a threshold τ. r_{succ} is a constant.

Learning New Skills. As shown in Figure 1, our RL finetuning can expand the distilled policy by handling two common regimes. (**I**) *In-distribution extensions* reuse and compose behaviors already supported by the demonstrations. A representative example is *regrasping*, which arises naturally from goal-conditioned in-betweening: training the policy to reach goals from diverse initializations and perturbed states encourages self-correction from near-failure outcomes without additional supervision. (**II**) *Out-of-distribution skills* must be learned explicitly when the re-

quired behavior is absent from the dataset. A representative example is *getting up*. Following prior practice [45, 66], we append a learnable *token* to the (Sec. 3.3) to indicate this new subtask and add an auxiliary reward that encourages upright posture and center-of-mass elevation (Sec. D).

Prior Preservation. During finetuning, rather than freezing network components to mitigate catastrophic forgetting as in prior work [45, 66], we adopt a simple multi-objective schedule. Specifically, we maintain a subset of environments that continue optimizing the original distillation objective (Sec. 3.3), while the remaining environments perform RL finetuning (Sec. D). This anchors the policy to the pretrained prior during adaptation without restricting model capacity. Given the environment mixtures and the joint execution of RL and distillation, we distribute tasks across multiple GPUs and aggregate gradients via a map-reduce scheme. Further details are provided in Sec. D.

4. Experiments

We evaluate InterPrior on two tasks: (**I**) *full-reference tracking* and (**II**) *sparse goal following*. The evaluation covers snapshot, trajectory, and contact specification, as well as their *compositions*. Since our goal representation is formed by masking arbitrary subsets of targets, these settings subsume a *broad family of task formulations*, ranging from single-frame constraints to multi-step trajectories over different joints and contacts. We further study InterPrior as a reusable prior for novel objects, and for tracking trajectories generated by kinematic models (Sec. F).

Datasets. We employ the InterAct [87] dataset with its preprocessing, which features diverse daily interactions encompassing a wide range of subjects and objects. Following [88], we use the OMOMO subset [29] repaired by their teacher rollout. To assess generalizability, we apply InterPrior to other InterAct subsets including selected data from BEHAVE [3] and HODome [96]. We exclude interactions dominated by soft-body dynamics (*e.g.*, backpack shoulder straps) when choosing evaluation examples.

Baselines and Tasks. We focus on baselines that cover diverse objects and skills and therefore omit methods that are for single object or task-specific proficiency [45, 72, 92]. (**I**) *Full-reference tracking.* We compare against the original InterMimic [88], with InterPrior, which supports full-reference imitation by removing masks. Evaluations target challenging regimes involving *thin-object interactions* and *initialization noise*. (**II**) *Sparse goal following.* We evaluate the complete InterPrior framework against adapted MaskedMimic [59, 60], to our task under identical goals, following Figure 1: (a) *Snapshot goals:* a ground truth frame specifies a few human joints or object position in the long term; (b) *Trajectory goals:* a sequence of ground-truth keyframes defines the a few joints or object trajectories; (c) *Contact goals:* a contact schedule specifies the desired active contact regions on objects, which will be converted to goals for human joints; (d) *Multi-goal chaining:* To evaluate long-horizon robustness, we concatenate three randomly sampled ground-truth subgoals, each canonicalized with respect to the preceding one. The concatenated sequence may include a mixture of snapshot, trajectory, and contact-following segments, with randomized goal transitions. For consistency, the same goals are used across all baselines; (e) *Random initialization:* To test motion coverage, we initialize the humanoid within five meters of the object and define the task as lifting the object by 0.5 meters from its initial position.

Metrics. (**I**) *Full-reference tracking.* Following [88], we report the following metrics: (a) *Success Rate (SR)*: the proportion of rollouts completed without violating the early-termination criteria; (b) *Human Position Error* E_{h} (m): the mean per-joint positional deviation between the simulated and reference humans, excluding hands due to the missing ground truth from the dataset; and (c) *Object Position Error* E_{o} (m): the mean positional deviation between the simulated and reference objects. (**II**) *Sparse goal following.* The evaluation metrics include: (a) *Success Rate (SR)*; (b) *Human* and *Object Errors* (E_{h}, E_{o}): the deviation from the target goal state, computed over the unmasked region; and (c) *Failure Rate (Fail)*: proportion of rollouts that directly fail *e.g.*, fall. More details are presented in Sec. F.

Implementation Details. All control policies operate at 30 Hz in IsaacGym [42]. The imitation expert policy, along with the encoder and decoder used during distillation, are implemented as MLPs with hidden layers of (1024, 1024, 512). The prior network is a four-layer Transformer encoder, and the critics use the same MLP architecture for expert training and RL finetuning. We retrain InterPrior on the G1 embodiment using our three-stage paradigm. During the first stage, we incorporate additional rewards and domain randomization to enhance stability on G1 and facilitate robust sim-to-sim transfer. All auxiliary rewards are multiplied with the imitation reward in exponential form $\exp(-\cdot)$, except for the termination term, which is added directly. The formulation of each G1-specific reward term is provided in Table C, and the dynamics randomization ranges used during training are summarized in Table D. We exclude thin-geometry objects for G1 because we do not include dexterous hands supporting single-hand grasps.

4.1. Quantitative Results

(**I**) *Full-reference tracking.* Table 2 shows that InterPrior achieves higher success rates under thin-geometry interactions and initialization noise. While InterMimic attains lower position error by strictly tracking the reference, InterPrior sometimes yields slightly higher human position error because it intentionally deviates when needed to re-align contact, trading strict tracking for interaction completion. (**II**) *Goal-conditioned tasks.* Under identical goal specifi-

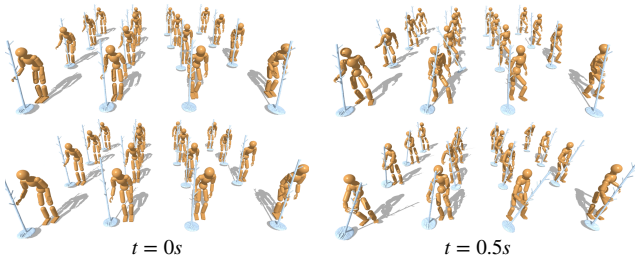

Figure 3. **Qualitative comparison** of same reference imitation between InterMimic [88] (**top**) and our InterMimic+ (**bottom**). InterMimic strictly follows the reference humanoid motion but fails to grasp the thin cloth stand when initialized with perturbations.

Figure 4. **Qualitative results** on a multi-object task. The model input is shifted to the second object once the first object is released.

cations (Table 1), InterPrior consistently improves success and reduces errors, with the largest gains on long-horizon multi-goal chaining and random-initialization stress tests. Distillation-based policies (including InterPrior pre-RL) fit the demonstration-induced state distribution; long rollouts with goal switching can enter under-covered intermediate states, causing drift and failure. RL finetuning directly trains the policy to reach sparse targets from diverse initializations, improving interpolation across goal sequences and recovery from off-distribution states. The position error trends follow a goal-sparsity continuum: broader state coverage benefits sparse goals more, and the gap narrows as goals densify. With full-reference tracking (Table 2), InterMimic for strict tracking achieves the lowest errors.

4.2. Qualitative Results

(**I**) *Full-reference tracking.* Figure 3 shows that InterMimic rigidly follows the reference but often fails to acquire or maintain contact on thin geometries under perturbations. In contrast, our tracking policy allows small, targeted deviations to correct hand-object alignment, producing stable grasps and more reliable completion. (**II**) *Long-horizon tasks.* Figures 4 and 1 show that InterPrior sustains minute-long whole-body interaction with multiple objects and smooth transitions across skills (*e.g.*, approach, grasp, lift, reposition). When drift begins (contact or balance), InterPrior self-corrects instead of compounding errors, consistent with the robustness induced by RL finetuning. (**III**) *Novel objects and interactions.* Figures 5 and 7 demonstrate zero-shot generalization to unseen objects and interaction styles. Guided only by sparse snapshot goals, InterPrior

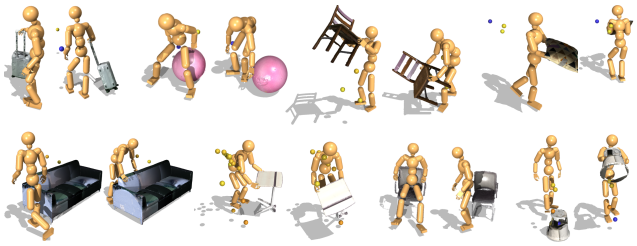

Figure 5. **Zero-shot qualitative results.** A single InterPrior model trained from OMOMO [29] demonstrates generalization to *unseen* objects and interactions from BEHAVE [3] and HODome [96].

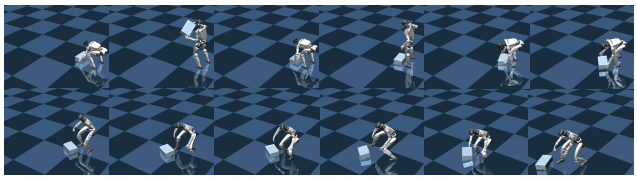

Figure 6. **Qualitative results** on sim-to-sim from IsaacGym [42] to MuJoCo [63] with object trajectory as condition, showing a sustained interaction involving box pickup, pushing, and kicking.

complete unspecified degrees of freedom and converge to feasible contact, even the original data in BEHAVE [3] and HODome [96] is for different human shape. (**IV**) *Sim-to-sim transfer.* Figure 6 illustrates transfer from IsaacGym [42] to MuJoCo [63]: InterPrior maintains coherent long-horizon interactions under object-conditioned goals, showing the potential to transfer to the real world.

4.3. Ablation Study

We conduct a cumulative ablation study reported in Table 1. Starting from a MaskedMimic baseline with an InterMimic expert, we progressively enable the components of InterPrior: upgrading to an InterMimic+ expert, incorporating the latent shaping loss, bounding both latent and observation spaces, and finally applying RL finetuning.

Impact of Latent Shaping and Bounding. Introducing the latent shaping loss yields modest improvements on in-distribution tasks but provides clear gains for long-horizon behavior and under random initialization. This indicates that a well-shaped and properly bounded latent is essential for mitigating drift in challenging, contact-rich interactions.

Effectiveness of Finetuning. Comparing the full InterPrior model with the variant before finetuning shows that RL finetuning chiefly enhances robustness. The improvement is also more pronounced on stress tests, suggesting that finetuning helps the policy exploring the feasible motion space and recover from distributional shift, while maintaining the policy with similar precision on standard tasks.

Impact of Finetuning on Trajectory Following. As discussed in Sec. 3.4, our *in-betweening* finetuning is applied only on snapshot goals rather than full trajectories, which may raise concerns about degrading trajectory-following performance. However, as shown in Table 1, trajectory following is well preserved for two reasons: (**I**) the fine-

Figure 7. **Qualitative comparison** between InterMimic [88] (**left**, full reference), MaskedMimic [59] (**middle**), and our InterPrior (**right**) on unseen and imperfect interactions from the BEHAVE [3] dataset. InterPrior can recover from data imperfection and continue the rollout.

Table 1. **Quantitative evaluation** and **ablation study** on in-distribution goal-conditioned tasks, including snapshot, trajectory, contact (Figure 1), plus out-of-distribution **stress tests** on challenging scenerio, such as long-horizon multi-goal chains and object lifting under random human initialization. For the random initialization, only the object is assigned a goal, thus the human error is omitted.

Method		Snapshot				Trajectory				Contact				**Chain**			**Rand Init**	
Variant	Additions (cumulative)	Succ ↑	E_h ↓	E_o ↓	Fail ↓	Succ ↑	E_h ↓	E_o ↓	Fail ↓	Succ ↑	E_c ↓	E_o ↓	Fail ↓	Succ ↑	E_h ↓	E_o ↓	Succ ↑	E_o ↓
MaskedMimic [59]	InterMimic [88] as Expert	64.2	29.3	22.1	12.6	88.0	9.0	8.1	8.5	52.2	49.2	25.7	13.9	29.1	40.2	43.9	31.7	26.8
InterPrior (**Ours**)	InterMimic+ as Expert	71.4	18.6	11.7	11.0	92.7	8.2	7.7	5.2	69.3	25.6	18.2	9.7	33.9	37.1	39.6	30.1	22.1
	+ Latent Shaping Loss	74.9	20.4	15.5	10.6	92.4	**7.9**	**6.6**	5.3	71.9	26.7	15.3	11.9	40.0	37.0	40.8	30.9	13.9
	+ Bounded Latent & Observations	89.1	**11.7**	**8.9**	6.0	93.6	8.1	**6.6**	4.6	88.5	17.0	**8.1**	5.4	45.1	31.5	37.2	41.1	19.6
	+ RL Finetuning (= **full**)	**90.0**	13.6	9.5	**3.7**	**94.6**	**7.9**	6.9	**2.5**	**90.7**	**15.9**	9.9	**2.9**	**68.8**	**30.2**	**35.7**	**88.6**	**11.9**

Table 2. **Quantitative evaluation** of full-reference imitation on OMOMO with *thin objects* and *initialization perturbations*, and adaptation to *novel object* and *interaction* skills, evaluated before and after finetuning on new data. For novel interactions, E_h and E_o not directly comparable since InterPrior now uses random sparse goals. Results show that InterPrior functions as a **reusable prior** with stronger adaptation capability than the full-reference imitator.

	OMOMO [29] select			BEHAVE [3]	HODome [96]
Method	SR ↑	E_h ↓	E_o ↓	SR ↑	SR ↑
InterMimic [88]	63.9	**7.1**	**11.4**	10.7	27.8
InterMimic + finetuning	/	/	/	38.9	55.5
InterPrior	**83.2**	8.9	11.7	27.4	40.1
InterPrior + finetuning	/	/	/	**52.0**	**72.4**

tuning procedure does not alter the model under trajectory-conditioned inputs, which are explicitly protected by a concurrent distillation loss; and (**II**) we redefine a snapshot goal if deviations from the target trajectory appears, and thus trajectory-following can implicitly benefit from the RL finetuning on snapshot goal following.

Scalable Prior. Beyond the generalization results in Figure 5, Table 2 and Figure 7 further demonstrate that InterPrior scales more robustly to novel objects and interactions, with or without finetuning, compared to the full-reference InterMimic baseline. A key factor is the prevalent dataset imperfections. For example, in Figure 7, baselines fail as contact artifacts cause failure initialization, whereas InterPrior can re-establish contact and continue the task. This flexibility allows the learned model to better absorb additional interaction data, even when such data are imperfect.

Failure Cases. Despite its improved robustness over the baselines, InterPrior still exhibits failure modes, as shown in Figure A. The human loses contact and moves without the object, whereas the baseline demonstrates a significantly higher failure rate, often resulting in human fall. We find typical failure scenarios include: (**I**) challenges with extremely thin or elongated objects that were unseen during training; and (**II**) partial goal completion in multi-goal chaining, where canonicalization introduces large alignment discrepancies, leading the policy to favor maintaining balance over achieving precise goal configurations.

Real-World Deployment. InterPrior encodes motor skills into a compact latent space through variational distillation and RL finetuning, yielding a modular prior that decouples skill acquisition from downstream deployment. This modularity enables adaptation to *new embodiments* and *sensing modalities* without retraining the prior itself. In [17], we validate this property by deploying a unified controller built on InterPrior to a real Unitree G1 humanoid, where it achieves autonomous loco-manipulation from egocentric depth and sparse task goals, suggesting that scaling interaction priors in simulation is a viable path toward versatile real-world humanoid behavior.

5. Conclusion

We present InterPrior, a physics-based generative motion controller that scales human-object interaction by combining large-scale imitation distillation with reinforcement finetuning. Using a distilled, goal-conditioned latent policy and optimizing it with RL yields a controller that maintains natural whole-body coordination while substantially improving robustness and competence. It composes loco-manipulation skills, transitions smoothly, and recovers from failures across diverse contact and dynamic conditions. This decoupled recipe broadens task, skill, and dynamics coverage while enabling interactive control and can be applied to different embodiments. We hope this scalable paradigm to provide a practical recipe for humanoid loco-manipulation. Future directions include integrating perception, language-conditioned goals, and richer affordances to advance InterPrior toward robust sim-to-real assistive manipulation and teleoperation.

Acknowledgments. This work was supported in part by the Amazon-Illinois Center on AI for Interactive Conversational Experiences, NSF under Grants 2106825 and 2519216, the DARPA Young Faculty Award, the ONR Grant N00014-26-1-2099, and the NIFA Award 2020-67021-32799. This work used computational resources, including the NCSA Delta and DeltaAI and the PTI Jetstream2 supercomputers through allocations CIS230012, CIS230013, CIS240311, and CIS240428 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, as well as the TACC Frontera supercomputer, Amazon Web Services (AWS), and OpenAI API 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] Donghoon Baek, Amartya Purushottam, Jason J Choi, and Joao Ramos. Whole-body bilateral teleoperation with multi-stage object parameter estimation for wheeled humanoid locomanipulation. *arXiv preprint arXiv:2508.09846*, 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, 6, 7, 8, 1, 4

[4] Yu-Wei Chao, Jimei Yang, Weifeng Chen, and Jia Deng. Learning to sit: Synthesizing human-chair interactions via hierarchical control. In *AAAI*, 2021. 3

[5] Peishan Cong, Ziyi Wang, Yuexin Ma, and Xiangyu Yue. Semgeomo: Dynamic contextual human motion generation with semantic and geometric guidance. In *CVPR*, 2025. 2

[6] 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

[7] Zekai Deng, Ye Shi, Kaiyang Ji, Lan Xu, Shaoli Huang, and Jingya Wang. Human-object interaction via automatically designed vlm-guided motion policy. *arXiv preprint arXiv:2503.18349*, 2025. 3

[8] Christian Diller and Angela Dai. CG-HOI: Contact-guided 3d human-object interaction generation. In *CVPR*, 2024. 2

[9] Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. C· ase: Learning conditional adversarial skill embeddings for physics-based characters. In *SIGGRAPH Asia*, 2023. 2

[10] Yuhui Fu, Feiyang Xie, Chaoyi Xu, Jing Xiong, Haoqi Yuan, and Zongqing Lu. DemoHLM: From one demonstration to generalizable humanoid loco-manipulation. *arXiv preprint arXiv:2510.11258*, 2025. 2

[11] Levi Fussell, Kevin Bergamin, and Daniel Holden. Supertrack: Motion tracking for physically simulated characters using supervised learning. *ACM Transactions on Graphics (TOG)*, 40(6):1–13, 2021. 3

[12] 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

[13] Zichen Geng, Zeeshan Hayder, Wei Liu, and Ajmal Saeed Mian. Auto-regressive diffusion for generating 3d human-object interactions. In *AAAI*, 2025. 2

[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. 2

[15] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *SIGGRAPH*, 2023. 2, 3

[16] Wenkun He, Yun Liu, Ruitao Liu, and Li Yi. Syncdiff: Synchronized motion diffusion for multi-body human-object interaction synthesis. In *ICCV*, 2025. 2

[17] 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. *arXiv preprint arXiv:2603.03279*, 2026. 8, 1

[18] Xiaoyu Huang, Takara Truong, Yunbo Zhang, Fangzhou Yu, Jean Pierre Sleiman, Jessica Hodgins, Koushil Sreenath, and Farbod Farshidian. Diffuse-cloc: Guided diffusion for physics-based character look-ahead control. *ACM Transactions on Graphics (TOG)*, 44(4):1–12, 2025. 3

[19] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *GCPR*, 2022. 2

[20] Kai Jia, Tengyu Liu, Mingtao Pei, Yixin Zhu, and Siyuan Huang. PrimHOI: Compositional human-object interaction via reusable primitives. In *ICCV*, 2025. 2

[21] 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. 2

[22] 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. 2

[23] 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. 2

[24] Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. SuperPADL: Scaling language-directed physics-based control with progressive supervised distillation. In *SIGGRAPH*, 2024. 2

[25] 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. 2

[26] 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. 2

[27] 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. 2

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3, 5

[29] 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, 6, 7, 8

[30] Yitang Li, Mingxian Lin, Zhuo Lin, Yipeng Deng, Yue Cao, and Li Yi. Learning physics-based full-body human reaching and grasping from brief walking references. In *CVPR*, 2025. 3

[31] 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

[32] 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. 2

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics*, 2015. 4, 1

[34] 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. 2

[35] 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. 2

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

[37] Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. *arXiv preprint arXiv:2310.04582*, 2023. 3, 5

[38] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Grasping diverse objects with simulated humanoids. In *NeurIPS*, 2024. 2, 3, 5

[39] 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

[40] Zhengyi Luo, Chen Tessler, Toru Lin, Ye Yuan, Tairan He, Wenli Xiao, Yunrong Guo, Gal Chechik, Kris Kitani, Linxi Fan, et al. Emergent active perception and dexterity of simulated humanoids from visual reinforcement learning. *arXiv preprint arXiv:2505.12278*, 2025. 3

[41] 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. 2

[42] 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, 1, 2

[43] 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

[44] 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. 3

[45] 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. 2, 3, 6

[46] 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. 2

[47] 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, 4

[48] 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. 2

[49] 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. 2, 3, 5

[50] Ilya A Petrov, Vladimir Guzov, Riccardo Marin, Emre Aksan, Xu Chen, Daniel Cremers, Thabo Beeler, and Gerard Pons-Moll. ECHO: Ego-centric modeling of human-object interactions. *arXiv preprint arXiv:2508.21556*, 2025. 2

[51] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Tridi: Trilateral diffusion of 3d humans, objects, and interactions. In *ICCV*, 2025. 2

[52] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017. 4, 1

[53] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction

to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011. 5, 4

[54] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 4

[55] Yutong Shen, Hangxu Liu, Lei Zhang, Penghui Liu, Ruizhe Xia, Tianyi Yao, and Tongtong Feng. Detach: Cross-domain learning for long-horizon tasks via mixture of disentangled experts. *arXiv preprint arXiv:2508.07842*, 2025. 3

[56] 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. 2

[57] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 2

[58] 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. 2

[59] 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. 3, 4, 5, 6, 8, 1, 2

[60] Chen Tessler, Yifeng Jiang, Erwin Coumans, Zhengyi Luo, Gal Chechik, and Xue Bin Peng. MaskedManipulator: Versatile whole-body control for loco-manipulation. *arXiv preprint arXiv:2505.19086*, 2025. 2, 3, 5, 6, 1

[61] 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. 2

[62] Emanuel Todorov and Michael I Jordan. Optimal feedback control as a theory of motor coordination. *Nature neuroscience*, 5(11):1226–1235, 2002. 1

[63] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IROS*, 2012. 7, 1

[64] Takara Everest Truong, Michael Piseno, Zhaoming Xie, and Karen Liu. Pdp: Physics-based character animation via diffusion policy. In *SIGGRAPH Asia*, 2024. 3

[65] Unitree. Unitree g1 humanoid agent ai avatar. https://www.unitree.com/g1/. 2, 3, 4, 1

[66] Ron Vainshtein, Zohar Rimon, Shie Mannor, and Chen Tessler. Task Tokens: A flexible approach to adapting behavior foundation models. *arXiv preprint arXiv:2503.22886*, 2025. 6

[67] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *CVPR*, 2021. 2

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

[69] Jiashun Wang, Yifeng Jiang, Haotian Zhang, Chen Tessler, Davis Rempe, Jessica Hodgins, and Xue Bin Peng. Hil: Hybrid imitation learning of diverse parkour skills from videos. *arXiv preprint arXiv:2505.12619*, 2025. 3

[70] Tingwu Wang, Yunrong Guo, Maria Shugrina, and Sanja Fidler. Unicon: Universal neural controller for physics-based character motion. *arXiv preprint arXiv:2011.15119*, 2020. 2

[71] 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. In *ICCV*, 2025. 3

[72] 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. 2, 3, 6

[73] 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

[74] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. Physics-based character controllers using conditional vaes. *ACM Transactions on Graphics (TOG)*, 41(4):1–12, 2022. 3

[75] Lin Wu, Zhixiang Chen, and Jianglin Lan. HOI-Dyn: Learning interaction dynamics for human-object motion diffusion. *arXiv preprint arXiv:2507.01737*, 2025. 2

[76] Yan Wu, Korrawe Karunratanakul, Zhengyi Luo, and Siyu Tang. UniPhys: Unified planner and controller with diffusion for flexible physics-based character control. In *ICCV*, 2025. 3

[77] Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. Human-object interaction from human-level instructions. In *ICCV*, 2025. 2, 3

[78] 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. 3

[79] Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. InterTrack: Tracking human object interaction without object templates. In *3DV*, 2024. 2

[80] 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

[81] 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

[82] Liang Xu, Chengqun Yang, Zili Lin, Fei Xu, Yifan Liu, Congsheng Xu, Yiyi Zhang, Jie Qin, Xingdong Sheng, Yunhui Liu, et al. Perceiving and acting in first-person: A dataset and benchmark for egocentric human-object-human interactions. In *ICCV*, 2025. 2

[83] Michael Xu, Yi Shi, KangKang Yin, and Xue Bin Peng. Parc: Physics-based augmentation with reinforcement learning for character controllers. In *SIGGRAPH*, 2025. 2

[84] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 2, 5

[85] Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. In *NeurIPS*, 2024. 2

[86] Sirui Xu, Yu-Wei Chao, Liuyu Bian, Arsalan Mousavian, Yu-Xiong Wang, Liangyan Gui, and Wei Yang. Dexplore: Scalable neural control for dexterous manipulation from reference scoped exploration. In *CoRL*, 2025. 4

[87] 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. 6

[88] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *CVPR*, 2025. 1, 3, 4, 6, 7, 8, 2

[89] Mengqing Xue, Yifei Liu, Ling Guo, Shaoli Huang, and Changxing Ding. Guiding human-object interactions with rich geometry and relations. In *CVPR*, 2025. 2

[90] Heyuan Yao, Zhenhua Song, Baoquan Chen, and Libin Liu. Controlvae: Model-based learning of generative controllers for physics-based characters. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. 3, 5

[91] Heyuan Yao, Zhenhua Song, Yuyang Zhou, Tenglong Ao, Baoquan Chen, and Libin Liu. MoConVQ: Unified physics-based motion control via scalable discrete representations. *arXiv preprint arXiv:2310.10198*, 2023. 3

[92] Runyi Yu, Yinhuai Wang, Qihan Zhao, Hok Wai Tsui, Jingbo Wang, Ping Tan, and Qifeng Chen. Skillmimic-v2: Learning robust and generalizable interaction skills from sparse and noisy demonstrations. In *SIGGRAPH*, 2025. 3, 6

[93] Ling-An Zeng, Guohong Huang, Yi-Lin Wei, Shengbo Gu, Yu-Ming Tang, Jingke Meng, and Wei-Shi Zheng. Chain-HOI: Joint-based kinematic chain modeling for human-object interaction generation. In *CVPR*, 2025. 2

[94] 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

[95] Haozhuo Zhang, Jingkai Sun, Michele Caprio, Jian Tang, Shanghang Zhang, Qiang Zhang, and Wei Pan. HumanoidVerse: A versatile humanoid for vision-language guided multi-object rearrangement. *arXiv preprint arXiv:2508.16943*, 2025. 3

[96] 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. 2, 6, 7, 8, 1, 4

[97] 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. 2

[98] Jinlu Zhang, Yixin Chen, Zan Wang, Jie Yang, Yizhou Wang, and Siyuan Huang. InteractAnything: Zero-shot human object interaction synthesis via llm feedback and object affordance parsing. In *CVPR*, 2025. 2

[99] 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. 2

[100] Xiaohan Zhang, Sebastian Starke, Vladimir Guzov, Zhensong Zhang, Eduardo Pérez Pellitero, and Gerard Pons-Moll. SCENIC: Scene-aware semantic navigation with instruction-guided control. *arXiv preprint arXiv:2412.15664*, 2024. 2

[101] 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

[102] Ziyu Zhang, Sergey Bashkirov, Dun Yang, Michael Taylor, and Xue Bin Peng. ADD: Physics-based motion imitation with adversarial differential discriminators. *arXiv preprint arXiv:2505.04961*, 2025. 2

[103] 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. 2

[104] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *ICCV*, 2023. 2

[105] 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. 2

InterPrior: Scaling Generative Control for Physics-Based Human-Object Interactions

Supplementary Material

In this supplementary, we provide additional details of our InterPrior framework with extended experiments:

- **(i)** Sec. A describes the organization of the demo video.
- **(ii)** Sec. B details the overall simulation configuration.
- **(iii)** Sec. C provides additional information on our goal representation, *e.g.*, how snapshot, trajectory, and contact goals are constructed at training and evaluation time with the masks.
- **(iv)** Sec. D gives a comprehensive explanation on: (**I**) the detailed formulation of the reference-free hand reward; (**II**) the losses used for variational distillation and latent shaping, and (**III**) RL finetuning.
- **(v)** Sec. E specifies additional implementation details, including network architectures, training schedules, and how we apply data augmentation to expert training, as well as additional techniques we use during G1 training for sim-to-sim experiments.
- **(vi)** Sec. F presents further qualitative results, *e.g.*, the integration of InterPrior with kinematic HOI generators, additional details of metrics, and failure cases.
- **(vii)** Sec. G examines the limitations of our current system and its potential societal implications.

Contents

A. Demo Video

The demo video on the webpage visualizes behaviors produced by InterPrior across settings detailed in the following. All sequences are rendered from the physics simulator [42, 63] using the same SMPL [33, 52] and G1 [65] model as for training. No post-processing is applied other than camera selection and cropping for visualization.

Core Capability. We show examples of snapshot, trajectory, and contact-conditioned control corresponding to the scenarios illustrated in Figure 1 of the main paper, for objects with diverse shapes.

Failure Recovery and Regrasping. We visualize rollouts perturbed or initialized from failure states. The video highlights re-approaching, re-grasping, and recovery from falls as described in Sec. 3.4.

Long-Horizon Multi-Goal Chains. We include long sequences where three canonicalized sub-goals are chained (Sec. 4, "Chain" tasks) and the policy must transition smoothly between different interaction while maintaining task success.

Diverse Task Execution from the Same Goal. We show that our model is able to control the simulated human achieving the same task with different execution.

Baseline Comparison. We demonstrate that InterPrior achieves superior performance compared to existing baseline methods [59, 60, 88].

Novel Interaction Generalization. We visualize qualitative results on BEHAVE [3] and HODome [96], as a complementary to Figure 5 and Figure 7 in the main paper.

Interaction with multiple objects. We showcase that InterPrior supports human interactions with multiple objects, without requiring any task-specific training.

Sim-to-Sim for G1. We include more examples of the G1 humanoid with sim-to-sim transfer, as a complementary to Figure 6, for controlling a humanoid only based on object future snapshot goal.

Sim-to-Real for G1. Our work, ULTRA [17], extends InterPrior with additional perception modules and enables perception-in-the-loop loco-manipulation on the Unitree G1. More details are available on the webpage.

Interactive Steering Control. Finally, we show real-time keyboard control where a user steers high-level goals and InterPrior produces coherent whole-body motion online.

B. Simulation

All experiments are performed in IsaacGym [42] with the GPU PhysX backend. Control policies run at 30Hz, while the simulator is stepped at 60Hz with two internal substeps per control step. The main simulation hyperparameters are summarized in Table A.

Table A. Simulation hyperparameters used in IsaacGym [42]. We largely follow the settings from prior work [72, 88].

Hyperparameter	Value
Simulation step Δt	$1/60$ s
Control step Δt	$1/30$ s
Physics substeps per control step	2
Position solver iterations	4
Velocity solver iterations	1
Contact offset	0.02
Rest offset	0.0
Max depenetration velocity	100
Object & ground restitution	0.7
Object & ground friction	0.9
Object density	200
Max convex hulls per object	64
Object rest offset	0.01

We introduce a small object rest offset to reduce human-object interpenetration, especially for thin geometries. Although this slightly enlarges the effective collision boundary, it avoids the substantial cost associated with increasing solver accuracy to compensate for collision handling.

C. Goal Formulation

This section details the construction of snapshot, trajectory, and contact goals and the associated masks used. Specifically, a goal state $\boldsymbol{y}_t$ shares the same structure as the observation $\boldsymbol{x}_t$, and a binary mask $\boldsymbol{m}_t$ indicates which components of $\boldsymbol{y}_t$ are provided to the policy.

C.1. Horizon for Goals

Short-Horizon Preview. We use a small set of offsets $K = \{1, 2, 4, 16\}$ to provide short-horizon previews relative to the current timestep t. For each offset $k \in K$, we construct a goal pair $(\boldsymbol{y}_{t+k}, \boldsymbol{m}_{t+k})$.

Long-Horizon Snapshot. A long-horizon offset sampled by $L \in [1, 128]$ defines a single far-future goal $(\boldsymbol{y}_{t+L}, \boldsymbol{m}_{t+L})$. During training, L is initialized randomly at the start of each episode and then decremented each timestep, being resampled once it reaches zero. Although termed a long-horizon snapshot, its value naturally decreases at each step and may temporarily fall below the short-horizon offsets.

C.2. Stochastic Mask Sampling during Training

During training, masks are not tied to specific tasks (snapshot; trajectory; contact). Instead, we randomly decide which parts of the future state are revealed to the policy, so that the policy is exposed to a *wide variety* of partial and sparse goals, following [59]. We operate at the level of rigid bodies, including objects with following three rules:

Body-Wise Masking. Visibility is enforced at the body level. For each rigid body, we maintain a single binary variable. If it is *false*, all all state features associated with that body at time $t+k$ are masked out, positions, orientations, and linear and angular velocities. The same rule applies to the entries in the interaction vectors D_{t+k} and the contact state C_{t+k}, defined in Sect. 3.1, which are masked or revealed together.

Independent Sampling in Rigid Bodies. At each horizon offset k, each body is sampled independently according to a fixed Bernoulli distribution: human-state and interaction components are revealed with probability 0.1, and object components with probability 0.5. This procedure produces diverse, randomly constructed combinations of visible and masked human, object, and contact features, rather than relying on any task-specific mask templates.

Temporal Consistency of Masks. To avoid flickering visibility, masks evolve over time with a high probability of staying the same and a small probability of being resampled. Concretely, for $k > 1$ we define a first-order Markov process:

$$\boldsymbol{m}_{t+k} = \begin{cases} \boldsymbol{m}_{t+k-1}, & \text{with probability } 1 - p_{\text{reset}}, \\ \text{Bernoulli}(\boldsymbol{p}_{\text{vis}}), & \text{with probability } p_{\text{reset}}. \end{cases}$$

Here $p_{\text{reset}} = 0.01$ ensures that once a body is masked or unmasked, it tends to remain in that state for multiple steps, while occasional resets still diversify the masks. The visibility probabilities $\boldsymbol{p}_{\text{vis}}$ follow the design above.

C.3. Task Definition for Inference

During inference, masks are constructed according to the target task. For a given task, the visibility pattern remains fixed throughout the rollout. The only exception is the multi-goal chaining setting, where we resample a new mask whenever the controller transitions to the next sub-goal.

Snapshot-Conditioned Control. We unmask the long-horizon snapshot. We still apply the consistent per-body sampling to determine which body or object components are revealed. All short-horizon preview are fully masked.

Trajectory-Conditioned Control. We unmask the short-horizon preview. Following the same per-body sampling, we reveal only a subset of the joint or object components. The long-horizon snapshot goal is retained.

Contact-Conditioned Control. Contact goals are implemented as a special case of snapshot conditioning in which we reveal only contact-related information. Specifically, we unmask the contact entries of $\boldsymbol{C}_t$, the associated signed-distance fields $\boldsymbol{D}_t$ (defined in Sec. 3.1), and the relevant human body parts. To avoid ambiguity in the target, we additionally unmask the object pose in the snapshot frame.

Multi-Goal Chaining. For multi-goal chains, we extract data by concatenating different data sequences. Specifically, we canonicalize each subsequent first frame with respect to the previous last frame. Canonicalization is performed by aligning the human root position (excluding height), and heading, *i.e.*, rotation around the vertical z–axis only, rather than the full $SO(3)$ orientation. Because this transformation is applied with respect to the human frame only, the object frame may become partially misaligned after canonicalization. As a result, we do not expect the policy to perfectly satisfy all chained goals, especially when object-relative alignment becomes extremely inconsistent. Nevertheless, the presence of a long horizon makes the policy possibly compensate for canonicalization artifacts.

D. Additional Details on Methodology

This section expands the reward and loss formulations, as well as additional details for the three stages of our framework: (**I**) InterMimic+ expert training (extending Sec. 3.2), (**II**) variational distillation (extending Sec. 3.3), and (**III**) RL post-training (extending Sec. 3.4).

D.1. InterMimic+: Full-Reference Imitation Expert

Reference-Free Reward for Expert. Here we introduce the detailed formulation of the hand reward r_{h}. Let $\boldsymbol{p}_T$ denote the position of the thumb fingertip and $\{\boldsymbol{p}_j\}_{j\in S}$ the positions of the other fingertips, with $\boldsymbol{q}_T$ and $\{\boldsymbol{q}_j\}_{j\in S}$ being their respective nearest surface points on the object. We define unit bearing vectors from the object surface toward the fingertips as $\boldsymbol{u}_T = (\boldsymbol{p}_T - \boldsymbol{q}_T)/\|\boldsymbol{p}_T - \boldsymbol{q}_T\|$ and $\boldsymbol{u}_j = (\boldsymbol{p}_j - \boldsymbol{q}_j)/\|\boldsymbol{p}_j - \boldsymbol{q}_j\|$, $j \in S$. The reward is defined as $r_{\text{h}} = \exp(-w_{\text{h}} e_{\text{h}})$, where $e_{\text{h}} = 1 - \frac{1}{|S|}\sum_{j\in S}\frac{1-\boldsymbol{u}_T^\top \boldsymbol{u}_j}{2}$, and w_{h} increases as the hand-object distance decreases, activating only when the reference indicates an upcoming interaction. This reward encourages all five fingers to maximize upcoming surface contact with the object.

D.2. InterPrior: Variational Distillation

Here we introduce the formulation for our proposed losses for variational Distillation. Let $\boldsymbol{\mu}_{p,t}$ and $\boldsymbol{\Sigma}_{p,t}$ denote the prior's mean and covariance at time t, *i.e.*, $\mathcal{N}(\boldsymbol{\mu}_{p,t}, \boldsymbol{\Sigma}_{p,t}) \equiv p_\psi(\boldsymbol{z}_t \mid \boldsymbol{x}_{t-\ell:t}, \mathcal{G}_t)$.

(**I**) *Scale loss.* We regularize the prior mean to lie on the unit hypersphere. This is to prevent the output mean from collapsing or exploding, with the use of latent normalization:

$$\mathcal{L}_{\text{scale}} = \mathbb{E}_t\big[\big(\|\boldsymbol{\mu}_{p,t}\|_2 - 1\big)^2\big].$$

(**II**) *Temporal consistency loss.* To obtain a smooth latent prior over time, we use $\mathcal{L}_{\text{tc}}$ to penalize changes in the prior distribution across consecutive timesteps using the squared 2-Wasserstein distance between Gaussians.

(**III**) *Goal reconstruction loss.* The decoder includes an additional head that predicts future goal features conditioned on the latent. Let $\widehat{\boldsymbol{y}}_{t+k}$ denote the predicted goal at offset k and $\boldsymbol{m}_{t+k}$ the input mask used to construct the masked residual goal. We train this head to complete the *masked* entries of the goal, *i.e.*, those that were hidden from the policy input. Formally, the goal reconstruction loss is

$$\mathcal{L}_{\text{goal}} = \mathbb{E}_{t,k}\big[\big\|\big(\mathbf{1} - \boldsymbol{m}_{t+k}\big) \odot \big(\widehat{\boldsymbol{y}}_{t+k} - \boldsymbol{y}_{t+k}\big)\big\|_2^2\big],$$

where $\odot$ denotes element-wise multiplication and $\mathbf{1}$ is an all-ones vector. This loss encourages the latent $\boldsymbol{z}_t$ to capture intent and context sufficient to reconstruct the missing parts of the goal, given only the visible subset provided by the mask. In practice, we reconstruct short future with $k = 1$.

D.3. InterPrior: Post-Training Beyond Reference

Get-Up Training. To learn the get-up behavior, in addition to the new learnable token as discussed in Sec. 3.4, we introduce an auxiliary reward that becomes active, with episodes initialized from a fallen state. The reward encourages both elevation of the pelvis and reorientation of the torso toward an upright configuration:

$$r^{\text{getup}} = w_{\text{height}}\,\sigma\big(h_t - h_{\text{target}}\big) \;+\; w_{\text{upright}}\,\sigma\big(\mathbf{n}_t \cdot \mathbf{n}_{\text{up}}\big), \quad (1)$$

where h_t is the pelvis height, h_{target} is set as 0.7, $\mathbf{n}_t$ is the torso's up vector, $\mathbf{n}_{\text{up}}$ is the world up direction, and $\sigma(\cdot)$ denotes a clipped linear shaping function.

Distributed Training. To mitigate catastrophic forgetting, we divide the parallel simulation environments into three groups: (**I**) *RL environments*, optimized solely with the post-training reward r_t^{PT}; (**II**) *Distillation environments*, optimized using the ELBO objective and supervised by the expert policy, as described in Sec. 3.3. The policy parameters are shared across all environments. Gradients are aggregated synchronously to update the shared policy.

Mask Prompt Engineering during Inference. To further enhance robustness during inference without additional learning, we apply lightweight *mask-based prompting* over the goal specification $\mathcal{G}_t$ (Sec. 3.1): (**I**) When following a trajectory and the state lags behind, we remove the trajectory goal but redefine the nearest waypoint as the snapshot goal. (**II**) For snapshot goals with distant target joints ($>$1 m), we retain only the root translation goal while masking out all other components, prompting locomotion before fine manipulation. (**III**) When human-object targets are contradictory, *e.g.*, both are moving but no grasp is established, we set the human root goal to the current object position while maintaining root height, masking all other joints. This encourages natural re-approach and regrasping behaviors. These inference-time edits operate solely on the goal $\mathcal{G}_t$, while the policy parameters remain fixed.

Finetuning on Additional HOI Datasets. The same finetuning mechanism naturally extends to absorbing new interaction datasets. Given any additional HOI corpus (*e.g.*, BEHAVE [3] or HODome [96] in Sec. 4), states from such new dataset are treated as additional sources of long-horizon goals and initializations for RL rollouts, while the distillation group continues to regularize the policy toward the original prior. This allows InterPrior to incrementally acquire new object categories and interaction styles without retraining from scratch.

E. Implementation Details

This section summarizes key implementation details, including network configurations, hyperparameters, randomization settings used for expert training, and additional techniques used during G1 training for sim-to-sim experiments.

PPO Setup. For both the expert and RL finetuning stages, we use PPO with generalized advantage estimation (GAE) and a clipped surrogate objective, and train with Adam. Following common practice [47], we keep the PPO discount factor γ, GAE parameter λ, clip ratio, and entropy regularization as shown in Table B, and apply gradient clipping.

InterMimic+: Full-Reference Imitation Expert. The InterMimic+ expert policy and critic are MLPs with three hidden layers of sizes $(1024, 1024, 512)$, using ReLU activations. Actor and critic are parameterized separately, and the critic outputs a scalar value with full observation and reference as input. Please refer to [88] for more details.

InterPrior: Variational Distillation. The encoder and decoder used for variational distillation share the same MLP backbone with hidden sizes $(1024, 1024, 512)$. The prior p_ψ is implemented as a 4-layer Transformer encoder with 4 attention heads, a latent dimension of 512, and a feedforward width of 1024. For the distillation objective (Sec. D), we use unit weight for the action reconstruction loss, and assign a weight of 10^{-3} to all auxiliary terms (goal reconstruction, scale loss, and temporal consistency loss). The KL regularizer follows a β-VAE style schedule: the KL weight β is annealed from 10^{-3} to 1.0 over the course of training. We first perform 500 epochs of warm-up using only teacher-controlled rollouts, and then gradually increase the fraction of student-controlled rollouts [53] until epoch $10,000$, at which point 95% of environments are driven by the student policy while the remaining 5% always use the teacher for fresh expert trajectories.

InterPrior: Post-Training Beyond Reference. For the post-training stage, we retain the same loss weights used for the distillation branch, and combine with the PPO loss weights specified in Table B for the RL branches.

Inference Efficiency. The runtime breakdown is: observation 20.16,ms, physics 19.02,ms, policy inference 0.43,ms, *SDF* 0.134,ms, and other overheads 0.057,ms, highlighting the policy's potential for real-world deployment.

Table B. Hyperparamters for training teacher and student policies.

Hyperparameters	value
Discount factor γ	0.99
Generalized advantage estimation λ	0.95
Learning rate	2e-5
Action loss weight	1
Critic loss weight	5
Action bounds loss weight	10
Minibatch size	16384
Horizon length H	32
Maximum episode length	300

Table C. **Additional reward terms for G1** used in Stage I expert training. Here, $\boldsymbol{\tau}$ denotes the vector of joint torques with elementwise limits $[\boldsymbol{\tau}_{\min}, \boldsymbol{\tau}_{\max}]$; $\boldsymbol{q}$ and $\dot{\boldsymbol{q}}$ are joint degrees and velocities with limits $[\boldsymbol{q}_{\min}, \boldsymbol{q}_{\max}]$; $\boldsymbol{a}_t$ is the control action at time t; $\boldsymbol{\omega}$ and $\boldsymbol{v}$ are the base (root) angular and linear velocities; F_z^{feet} is the vertical ground-reaction force at the feet; $\boldsymbol{v}^{\text{feet}}$ is the tangential (ground-plane) velocity of the feet; d_{feet} is the horizontal distance between the two feet, with desired bounds $[d_{\min}, d_{\max}]$; $\boldsymbol{g}_{xy}^{\text{feet}}$ is the projection of the gravity direction onto the foot frame's ground plane; $\mathbb{1}(\cdot)$ and $\mathbb{1}_{\text{termination}}$ are indicator functions. All norms $\|\cdot\|$ and $\|\cdot\|_2$ are Euclidean.

TERM	EXPRESSION	WEIGHT
Penalty:		
Torque limits	$\mathbb{1}(\boldsymbol{\tau} \notin [\boldsymbol{\tau}_{\min}, \boldsymbol{\tau}_{\max}])$	2
DoF position limits	$\mathbb{1}(\boldsymbol{q} \notin [\boldsymbol{q}_{\min}, \boldsymbol{q}_{\max}])$	5
Energy	$\|\boldsymbol{\tau} \odot \dot{\boldsymbol{q}}\|$	10^{-4}
Termination	$\mathbb{1}_{\text{termination}}$	-30
Regularization:		
DoF velocity	$\|\dot{\boldsymbol{q}}\|_2^2$	4×10^{-4}
Action rate	$\|\boldsymbol{a}_t\|_2^2$	0.1
Torque	$\|\boldsymbol{\tau}\|$	2×10^{-3}
Angular velocity	$\|\boldsymbol{\omega}\|^2$	0.01
Base velocity	$\|\boldsymbol{v}\|^2$	0.1
Foot slip	$\mathbb{1}(F_z^{\text{feet}} > 5.0) \cdot \sqrt{\|\boldsymbol{v}^{\text{feet}}\|}$	0.03
Feet distance reward	$\frac{1}{2}\exp\left(-100\,\lvert\max(d_{\text{feet}} - d_{\min}, -0.5)\rvert\right) + \frac{1}{2}\exp\left(-100\,\lvert\max(d_{\text{feet}} - d_{\max}, 0)\rvert\right)$	0.5
Feet orientation	$\sqrt{\|\boldsymbol{g}_{xy}^{\text{feet}}\|}$	1

F. Additional Experimental Results

In this section, we introduce metric details, provide supplementary qualitative results, and discuss failure cases.

Additional Details on Evaluation Metrics. For *trajectory-following* tasks, we evaluate the policy at each timestep by comparing the rollout state with the corresponding reference, and compute pose and object errors only over the unmasked components. For *snapshot goal-following* tasks, there is no time-aligned reference trajectory. Instead, we compute the error between the rollout state and the snapshot goal at every timestep and report the *minimum* of this

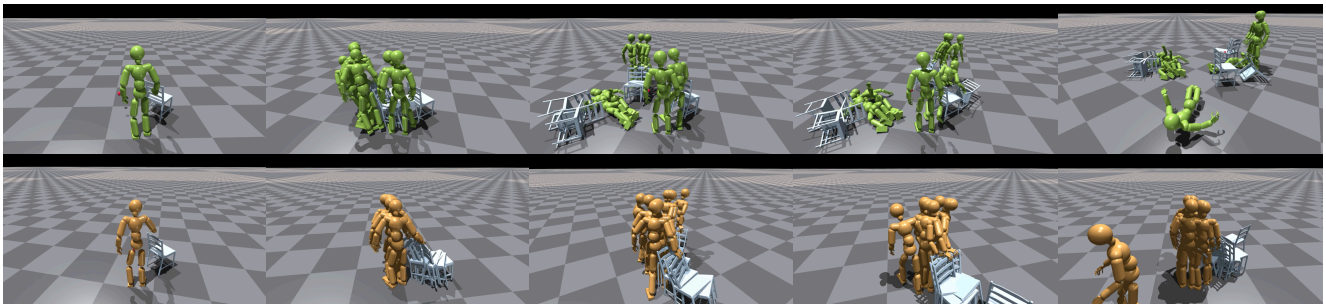

Figure A. **Additional qualitative comparisons** with baseline method [59, 60] (**Top**). Our InterPrior shows higher success rate under the same task goal.

Table D. **Range of dynamics randomization.** "default" refers to the parameter value from the unitree G1 official 29DoF model. v_{xy} is the planar (horizontal) push velocity.

Term	Range / Value
Dynamics randomization	
Friction coefficient	$\mathcal{U}(1.0,\ 3.0)$
Base CoM offset	$\mathcal{U}(-0.05,\ 0.05)$ m
Base mass offset	$\mathcal{U}(-3.0,\ 3.0)$ kg
P gain scaling	$\mathcal{U}(0.8,\ 1.2)\ \times$ default
D gain scaling	$\mathcal{U}(0.8,\ 1.2)\ \times$ default
External perturbation	
Push robot	interval $= 4$ s, $\quad v_{xy} = 1$ m/s

distance over the rollout. This reflects whether the policy is capable of reaching the target configuration.

Diverse Behaviors Under the Same Goal. Beyond the examples shown in the main paper, Figure B illustrates how InterPrior behaves diversely given the same goal, showing that our learned latent space is meaningful and is able to capture diverse behaviors.

Integration with Kinematic HOI Generators. To demonstrate that InterPrior's generalization, we integrate it with InterDiff [84] that produces physically unconstrained interaction trajectories. The integration proceeds as follows: (**I**) the kinematic generator produces a 25 frames of human-object poses given the past 15 frames following [84]; (**II**) we convert these sequences into our goal representation by extracting snapshot and trajectory goals; and (**III**) we feed these goals into InterPrior. The result is shown in Figure C.

G. Discussion

Limitations and Future Work. InterPrior is still bounded by the coverage and quality of its training data: highly corrupted or unseen interaction patterns are not reliably recovered, and in such cases the policy often defaults to conservative strategies, maintaining balance without fully solving the task. Our model is tailored to rigid object, and we still observe occasional artifacts such as shallow interpenetrations, foot skating, or failure cases such as object drop over long rollouts. The current hand and contact representation is also not designed for fine-grained finger dexterity or in-hand manipulation. Finally, our three-stage training introduces additional complexity and hyperparameters. Future work includes expanding dataset diversity, incorporating richer hand models, and simplifying or unifying the training scheme.

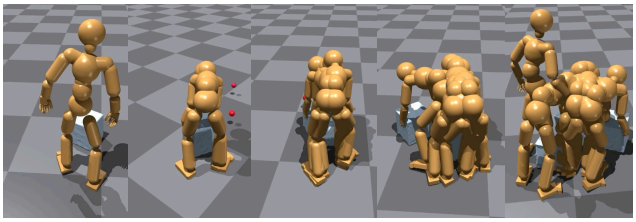

Figure B. **Qualitative results** given the same goal. Our framework produces multiple valid yet distinct interaction trajectories.

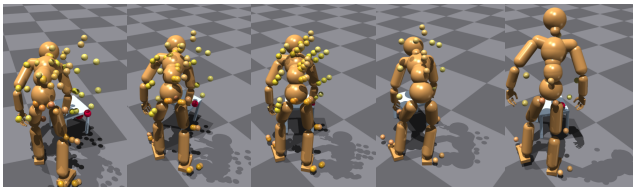

Figure C. **Qualitative results** of InterPrior following the targets generated by InterDiff [84] (*yellow* and *red* dots). InterPrior adaptively completes the task without strictly adhering to the targets, using only sparse inputs of wrist, feet, and object target.

Societal and Ethical Considerations. InterPrior enables more general-purpose, physically grounded humanoid controller, which can be beneficial for animation, simulation, and robotics, but also raises potential risks. More capable humanoid controllers could be deployed in unsafe settings or for applications that conflict with societal norms (*e.g.*, surveillance or coercive scenarios). We therefore encourage careful consideration of safety mechanisms, usage policies, and ethical guidelines when applying this type of model beyond controlled research environments.