POLICY-AS-DATA: LEARNING GENERALIZABLE HOI DIFFUSION MODELS FROM SIMULATED PHYSICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthesizing realistic Human-Object Interactions (HOI) is critical for creating embodied avatars and functional virtual environments. However, current data-driven approaches primarily rely on motion capture datasets, which are expensive to scale and limited in functional diversity. Models trained with these datasets struggle to generalize to unseen objects and maintain physical consistency over long horizons. In this paper, we propose a novel framework that leverages a physics simulator to overcome the data-scarcity bottleneck in HOI generation. Specifically, we propose a scalable pipeline, called PAD-HOI, which leverages policies trained with reinforcement learning in a physics simulator for task-oriented data generation and trains a generative model on the augmented dataset for generalizable HOI generation. To seamlessly utilize the synthetic data, we introduce a coarse-to-fine retargeting process that bridges the representation gap between the simplified model used in the physics simulator and the standard parametric body models required for generative training. Validated through comprehensive experiments, our method demonstrates improved generalization to unseen objects and supports long-horizon generation, while providing broader task and object coverage and improved contact consistency.

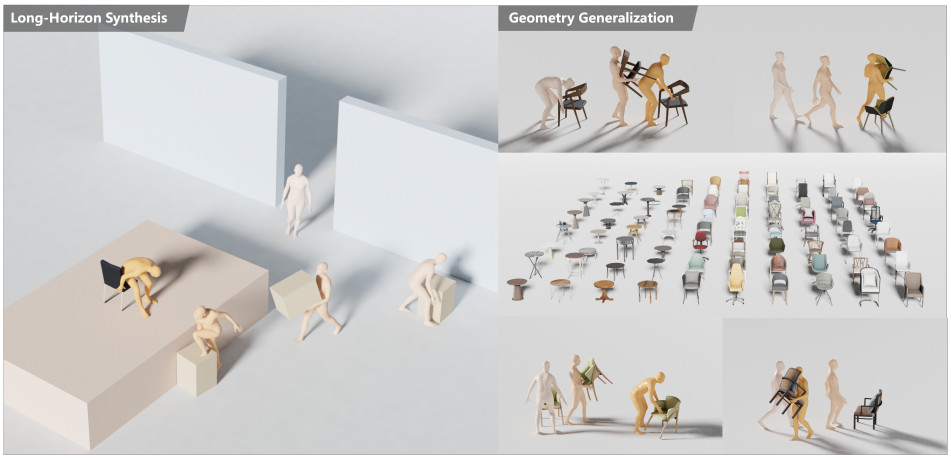


Figure 1: **Generalizable and Physically Grounded HOI Generation.** Our framework composes multiple tasks in long-horizon sequences while maintaining stable contact and generalizes to diverse object geometries. By training on simulation data covering over 100 procedurally varied object instances, the model learns robust interactions across unseen objects and extended task sequences.

1 INTRODUCTION

Humans routinely perform long-horizon interactions that require adapting body postures and hand contacts to object geometry, such as carrying a box before sitting on a chair. Synthesizing such realistic Human-Object Interactions (HOI) is important for embodied AI, high-fidelity simulation, and immersive virtual environments.

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

Recent advancements in generative models, such as diffusion models (Ho et al., 2020), have demonstrated remarkable capabilities across the image (Rombach et al., 2022; Saharia et al., 2022; Esser et al., 2024), video (Ho et al., 2022; Blattmann et al., 2023; Bar-Tal et al., 2024), and 3D (Zhang et al., 2024; Poole et al., 2022; Yi et al., 2023) domains. However, their potential in HOI synthesis remains constrained by severe data scarcity and physical inaccuracies of available motion capture (MoCap) datasets. Prohibitively expensive and sophisticated acquisition pipelines restrict the scaling of MoCap datasets to diverse object geometries and long-horizon functional interactions. Furthermore, capture hardware and post-processing inevitably introduce physical artifacts, such as penetration and floating. Consequently, models Li et al. (2023; 2024) trained exclusively on these datasets overfit to a narrow range of standardized objects and exhibit physical inconsistencies, and struggle to generalize to unseen object geometries or execute long-horizon tasks.

To bypass the limitations of MoCap acquisition, an alternative paradigm relies on physics simulation. Policies trained using reinforcement learning (RL) can generate task-oriented trajectories across varied object geometries through domain randomization. Prior works (Peng et al., 2021; Hassan et al., 2023; Gao et al., 2024; Pan et al., 2025) combine task objectives with motion priors to generate natural, task-oriented interactions. However, developing a universal RL controller that generalizes across varying object geometries is restricted by the complexities of RL training and reward engineering. In generative pipelines, physics-based control is often used downstream to track or correct generated motions (Yuan et al., 2023; Ji et al., 2025; Wu et al., 2025b). Such correction improves the physical consistency of a given motion, but does not by itself expand the interaction examples available to train the generator. We explore a complementary approach: leveraging task-specific RL policies as upstream data generators for diffusion models. Generative models excel at capturing distributions from large-scale datasets, whereas RL policies efficiently generate diverse trajectories across geometries that are expensive to capture at scale. However, these two sources do not share a common motion representation. Realizing this approach requires bridging a *representation gap*. To maintain training efficiency and feasible policy convergence, low-DoF rigid-body representations are used in RL training. These simplified skeletons fundamentally lack the high-DoF articulation and dense surface geometry required to represent rich, expressive human movements. Consequently, these physics-based data cannot be directly integrated with high-fidelity MoCap datasets, which use the SMPL format Loper et al. (2023); Pavlakos et al. (2019). To fully exploit simulator-generated data, a retargeting process is required to bridge this gap.

To address this challenge, we propose Policy-as-Data for HOI generation (PAD-HOI), a framework that leverages simulation-trained policies as a scalable source of interaction training data. Our core insight is that simulation rollouts from task-oriented policies can expand the interaction data available for training generative models along two complementary dimensions: object geometry and task execution. Through large-scale simulation across varied object instances, these rollouts provide physically simulated examples of object interactions and task trajectories that are scarce in existing MoCap datasets. To bridge the representation gap between simulation and generative training, we introduce a coarse-to-fine retargeting process that reconstructs SMPL motion from reduced rigid-body trajectories and refines contact alignment between the human mesh and object surfaces. By jointly training on retargeted synthetic data and real MoCap sequences, our model learns interactions across diverse object geometries and task executions while maintaining generation quality on the real-data distribution. We validate our approach through unseen-object generalization and long-horizon generation, as illustrated in Fig. 1. In summary, our main contributions are as follows:

- **Policy-as-Data Framework:** To our knowledge, we are the first to use task-oriented physics policies as an upstream source of training data for generative HOI synthesis. By converting policy rollouts into generative supervision, our framework expands the range of object geometries and task executions available during training.
- **Contact-Aware Retargeting Bridge:** We introduce a two-stage retargeting process that reconstructs SMPL motion from reduced rigid-body trajectories and refines mesh-level contacts, enabling joint training on simulated and real interactions.
- **Expanded Generative Capabilities:** We show that the synthetic data expands the support of generative learning in object geometry and task execution. This improves generalization to held-out object instances and enables long-horizon composition of multiple task types.

2 Related Work

Data-driven Interaction Synthesis. Recent advancements in generative modeling have enabled the synthesis of realistic human-object interactions conditioned on text or object states (Li et al., 2025; Peng et al., 2023; Wu et al., 2024; Li et al., 2023; Yang et al., 2024; Zeng et al., 2025; Cong et al., 2025; Lu et al., 2025; Lou et al., 2025; Diller & Dai, 2024; Wang et al., 2026; Bal et al., 2026; Cai et al., 2026). Early approaches focused on synthesizing human motion in static 3D scenes (Hassan et al., 2019; 2021; Huang et al., 2023; Wang et al., 2022; Zhang et al., 2022; Yi et al., 2024). With the advent of diffusion models, methods like OMOMO (Li et al., 2023) and CHOIS (Li et al., 2024) have demonstrated the ability to generate diverse interactions with dynamic objects by learning from motion capture datasets (Bhatnagar et al., 2022; Li et al., 2023; Huang et al., 2024; Fan et al., 2023; Jiang et al., 2023; Zhang et al., 2022; Liu et al., 2025; Geng et al., 2025). However, these kinematic-based approaches suffer from two fundamental limitations. First, they lack explicit physical constraints, often resulting in artifacts such as hand-object penetration, foot sliding, or floating contacts. Second, because they are trained on fixed Mocap datasets with limited object instances, they struggle to generalize to objects with unseen geometries or scales. In contrast, our approach leverages physics-based simulator to enforce contact stability and introduces domain randomization to achieve robustness to novel object shapes.

Physics-Based Character Control. Physics-based simulators offer a rigorous framework for synthesizing functionally valid motion. Reinforcement learning has been successfully applied to train characters that can perform complex tasks, ranging from locomotion (Peng et al., 2021; 2018; 2022; Tessler et al., 2023; Tevet et al., 2024; Peng, 2025) to dexterous manipulation (Tessler et al., 2025; Luo et al., 2024a; Deng et al., 2026). A prominent line of research focuses on physics-aware motion imitation, developing unified controllers capable of tracking large-scale reference motions (Wu et al., 2025a; Luo et al., 2022b;a; 2021; Wu et al., 2025b; Wang et al., 2025; Yu et al., 2025; Luo et al., 2023) or establishing versatile motion manifolds (Peng et al., 2022; Dou et al., 2023; Luo et al., 2024b) for reusable skills. While these imitation-based policies excel at accurately reproducing existing kinematic data within a simulator, they are fundamentally constrained by their reliance on dense reference trajectories. Consequently, they are ill-suited to act as generators for *unseen* object geometries or novel functional tasks where ground-truth motion capture does not exist. To overcome this limitation recent works like TokenHSI (Pan et al., 2025) and UniHSI (Xiao et al., 2024) propose task oriented goal conditioned controllers. By utilizing adversarial motion priors to encourage naturalism these policies can autonomously discover and synthesize physically plausible manipulation strategies for diverse object shapes. However while these methods guarantee physical plausibility they typically operate on simplified reduced order models to ensure simulator stability. This structural gap limits their direct applicability to modern avatar creation pipelines which require high fidelity parametric meshes (Loper et al., 2015; Romero et al., 2022; Pavlakos et al., 2019) and fine grained text controllability. Our work bypasses this architectural limitation entirely by separating the physical simulation from the final generative output.

Physics-Guided Motion Synthesis. Physics simulation plays diverse roles in human motion synthesis. One common paradigm integrates physics engines into the generation process to enforce physical consistency. PhysDiff (Yuan et al., 2023) and POMP (Ji et al., 2025) incorporate physics-based projection or dynamically consistent denoising during inference. RLPF (Yue et al., 2025) and PARC (Xu et al., 2025a) utilize reinforcement learning fine-tuning and iterative physics-based correction to refine generated motions. InterPhys (Hassan et al., 2023) and InterAct (Xu et al., 2025b) perform post-hoc optimization to resolve contact artifacts in motion capture data. We explore a complementary direction: using physics simulation as an upstream data source for training. By training task-specific RL policies across randomized object geometries, we generate large-scale synthetic interaction data that augments limited MoCap coverage. This retargeted synthetic data provides the generative model with diverse geometric and functional priors during training, enabling improved generalization to unseen objects and long-horizon composition. Our approach is orthogonal to inference-time physics-based refinement methods, which could potentially be applied to our generated outputs for further consistency improvements.

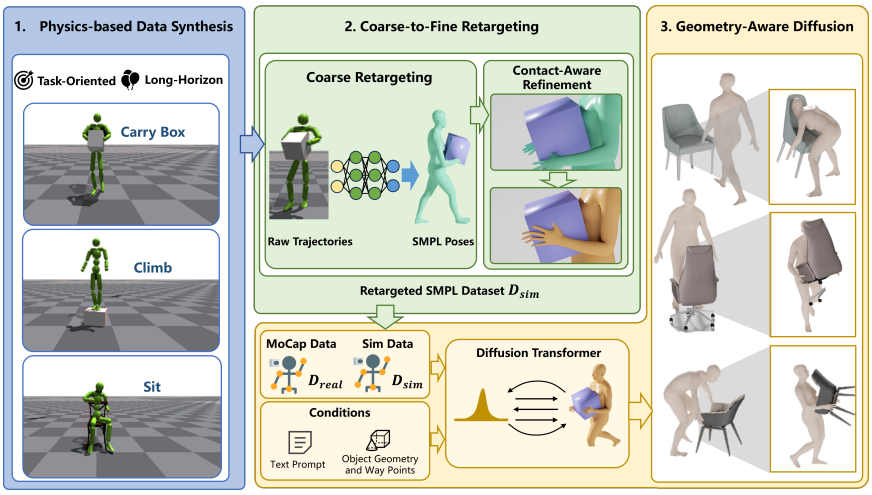


Figure 2: Overview of PAD-HOI. Our framework addresses MoCap data scarcity through physics simulation. **(a)** RL experts generate diverse trajectories across randomized object geometries. **(b)** A retargeting module converts rigid-body states to SMPL poses with contact refinement. **(c)** A diffusion model trained on augmented data generates interactions for diverse objects.

3 METHOD

To address the scarcity of MoCap data for diverse object geometries and long-horizon interactions, we use physics-based policies to generate HOI training data that augments existing MoCap datasets. As illustrated in Figure 2, PAD-HOI combines RL-based trajectory generation, coarse-to-fine retargeting to SMPL, and geometry-aware diffusion trained jointly on synthetic and real motions.

3.1 PHYSICS-BASED DATA SYNTHESIS

Physics simulation provides a mechanism for generating interaction data across varied object geometries. To facilitate efficient policy learning, we use a simplified humanoid with 15 rigid bodies and 32 actuated degrees of freedom (DoF) (Pan et al., 2025), controlled through PD target joint rotations. We roll out task-specific RL experts in procedurally randomized environments to construct a rigid-body trajectory dataset $\mathcal{D}_{\text{phys}}$, which is subsequently converted to SMPL format (Section 3.2).

To ensure that synthesized interactions are both functional and anthropomorphically natural, we adopt two rewards that decouples task specification from motion style (Peng et al., 2021; Hassan et al., 2023). Given s_t and a_t which are the character state and action at time step t, the reward is defined as:

$$r(s_t, a_t, s_{t+1}, g) = w_G\, r_G(s_t, a_t, s_{t+1}, g) + w_S\, r_S(s_t, s_{t+1}), \tag{1}$$

where r_G specifies the *task* objective, e.g., reaching a target waypoint g, and r_S specifies a *style* objective that encourages human-like transitions. To provide a task-agnostic motion prior, r_S comes from an adversarial discriminator trained to distinguish simulated transitions from real transitions from an unstructured MoCap dataset. The policy is then trained via reinforcement learning to maximize the expected cumulative reward, encouraging the agent to complete tasks successfully while simultaneously adopting the kinematic styles of reference MoCap data. The adversarial objective and reward formulation are detailed in Appendix A.

To enable generalization to novel object geometries, we leverage domain randomization during data synthesis. Instead of training the reinforcement learning policy on a single static instance, we procedurally randomize the physical dimensions and topologies of the interaction objects within the simulator (Xu et al., 2023a; Pan et al., 2025). Specifically, our experts interact with diverse categories, including randomized boxes, tables and chairs. Consequently, the generated dataset encompasses structural variations. By training on this geometrically diverse data, the diffusion model learns robust structural priors that support generalization to unseen object instances.

3.2 CONTACT-AWARE RETARGETING BRIDGE

The simplified physics character lacks the articulated pose and surface representation needed for SMPL-based generative training. We represent human motion using SMPL root translation, root orientation, and local body-joint rotations, together with object translation and orientation. Our contact-aware retargeting bridge combines coarse pose reconstruction with contact-aware refinement to convert $\mathcal{D}_{\text{phys}}$ into SMPL interaction sequences $\mathcal{D}_{\text{sim}}$ that can be used jointly with MoCap data for training. Detailed representations and dimensions are provided in Appendix D.

Coarse Retargeting. Mapping the low-dimensional physics skeleton back to the high-dimensional SMPL space is an under-determined reconstruction task, as the reduced skeleton inherently discards internal joint dimensions and surface geometry. To learn this mapping with strong kinematic priors, we construct a paired training dataset using real-world AMASS (Mahmood et al., 2019) motions. For each AMASS sequence, we extract a subset of joints from the SMPL body that corresponds to the physics character's skeleton, creating paired examples of reduced rigid-body states and full SMPL poses.

For each physics rollout, we extract per-body features $\tilde{\mathbf{S}}$ comprising local and global rotations in 6D representation, linear velocity, and angular velocity. A temporal Transformer $\mathcal{F}_\phi$ processes a sliding window $\mathbf{w}_t = \tilde{\mathbf{S}}[t - W + 1 : t]$ of length W to predict the SMPL parameters at time t:

$$\hat{\boldsymbol{x}}_t^h = \mathcal{F}_\phi(\mathbf{w}_t) = \{\hat{\boldsymbol{p}}_t^h, \hat{\mathbf{R}}_t^h, \hat{\boldsymbol{q}}_t^h\}, \tag{2}$$

where $\hat{\boldsymbol{p}}_t^h$ is the global root translation, $\hat{\mathbf{R}}_t^h$ is the global root orientation, and $\hat{\boldsymbol{q}}_t^h$ stacks the local 6D rotations of the SMPL body joints. We train $\mathcal{F}_\phi$ using rotation supervision and joint position and velocity losses computed through differentiable SMPL forward kinematics:

$$\mathcal{L}_\phi = \lambda_R \mathcal{L}_{\text{rot}}(\hat{\mathbf{R}}^h, \mathbf{R}^h) + \lambda_P \mathcal{L}_{\text{rot}}(\hat{\boldsymbol{q}}^h, \boldsymbol{q}^h) + \lambda_J \|\hat{\mathbf{J}} - \mathbf{J}\|_2^2 + \lambda_V \|\hat{\dot{\mathbf{J}}} - \dot{\mathbf{J}}\|_2^2, \tag{3}$$

where $\mathcal{L}_{\text{rot}}$ measures the distance between rotations, and $\hat{\mathbf{J}}$ and $\hat{\dot{\mathbf{J}}}$ denote SMPL joint positions and joint velocities obtained by applying forward kinematics to $\hat{\boldsymbol{x}}_{1:T}^h$. This coarse stage produces SMPL motion that maintains temporal coherence and kinematic consistency, providing an initialization for contact-aware refinement.

Contact-Aware Refinement. Coarse retargeting can still introduce hand-object contact errors. We therefore apply offline, gradient-based refinement following (Xu et al., 2025b), keeping the object trajectory fixed while optimizing human poses to improve contact and reduce penetration and jitter:

$$\mathcal{L}(\hat{\boldsymbol{x}}_t^h) = E_{\text{rec}} + \lambda_{\text{cont}} E_{\text{cont}} + \lambda_{\text{pen}} E_{\text{pen}} + \lambda_{\text{sm}} E_{\text{sm}} + \lambda_{\text{prior}} E_{\text{prior}}, \tag{4}$$

where E_{rec} keeps the refined human pose close to the coarse retargeted initialization. To address hand-object contact errors, we define E_{cont} using simulator-provided contact indicators c_t and a set of interaction vertices $\mathcal{V}_{\text{contact}}$ sampled from the two palms:

$$E_{\text{cont}} = \sum_{t=1}^{T} c_t \sum_{j \in \mathcal{V}_{\text{contact}}} d(\mathbf{v}_j(t), \mathcal{O}(t)), \tag{5}$$

where $\mathbf{v}_j(t)$ denotes the j-th SMPL vertex and $d(\cdot, \mathcal{O}(t))$ is a point-to-object distance computed via the Signed Distance Field (SDF) of the object. As for the other terms, E_{pen} penalizes human vertices that penetrate into the object, E_{sm} minimizes joint accelerations to suppress temporal jitter, and E_{prior} regularizes the SMPL parameters to maintain natural human articulations. These refinement terms and their implementation are described in Appendix B.

3.3 GEOMETRY-AWARE DIFFUSION FOR HOI SYNTHESIS

Augmented with the retargeted dataset, which contains diverse and physically consistent interactions, we can finally train a generalizable diffusion model. By learning the distribution of our hybrid dataset combining synthetic trajectories $\mathcal{D}_{\text{sim}}$ and real motion capture $\mathcal{D}_{\text{real}}$, the model absorbs the physical rules obeyed by the simulation data while retaining the natural human kinematics of the real data. This synergy enables the model to adapt to unseen object geometries and seamlessly execute complex multi-stage tasks.

We implement our generative model using a standard Transformer-based diffusion architecture conceptually building upon CHOIS (Li et al., 2024). To fully exploit the rich geometric and temporal variations in our hybrid dataset, the model relies on a unified conditioning context. We encode text prompts via CLIP (Radford et al., 2021) for semantic task guidance. To explicitly capture varying object sizes and topologies, we extract Basis Point Set (Prokudin et al., 2019) features from the object shape. We also formulate a masked motion representation containing initial kinematic states and sparse future waypoints to guide coherent generation across long horizon interactions. By utilizing this architecture, we effectively translate the physical and kinematic priors of our hybrid dataset into a highly controllable synthesis engine.

Long-Horizon Multi-Task Generation. Long-horizon interactions are generated by sequentially executing multiple tasks in a predefined scene. Each task is conditioned on a task-specific text prompt, the geometry and current pose of the interaction object, and sparse waypoints specifying the desired trajectory. At task transitions, the terminal human pose of the preceding task initializes the next, while the text and object conditions switch to those of the subsequent task. Within each task, generation proceeds autoregressively over overlapping temporal windows. The overlapping human-object states are transformed into the current local coordinate frame and enforced as hard constraints during denoising. The resulting motion segments are assembled in global coordinates.

4 EXPERIMENTS

We evaluate PAD-HOI along four axes. First, we validate that retargeting converts simulator states into usable SMPL training data (Section 4.2). Second, we demonstrate zero-shot generalization to unseen objects and long-horizon multi-task generation (Section 4.3). Third, we verify that simulator augmentation preserves competitive benchmark performance (Section 4.4). Finally, we analyze the contribution of each training data source (Section 4.5).

4.1 EXPERIMENTAL SETTINGS

Datasets. We train PAD-HOI on a hybrid dataset $\mathcal{D} = (\mathcal{D}_{\text{real}}, \mathcal{D}_{\text{sim}})$. $\mathcal{D}_{\text{real}}$ is the OMOMO dataset (Li et al., 2023), which supplies natural human kinematics. $\mathcal{D}_{\text{sim}}$ contains over 4,400 physically grounded trajectories generated by RL experts for carry, sit, and climb tasks over procedurally varied boxes, tables, chairs, and platforms. These rigid-body rollouts are converted into the SMPL representation by our coarse-to-fine retargeter (Section 3.2); dataset composition and implementation details are provided in Appendix C and D.

HOI Generation Baselines. We compare PAD-HOI with InterDiff (Xu et al., 2023b), MDM (Tevet et al., 2023), HOI-Diff (Peng et al., 2023), CHOIS (Li et al., 2024), OMOMO (Li et al., 2023), and HOI-HFLI (Wu et al., 2025b). All baselines are trained on the training split of $\mathcal{D}_{\text{real}}$ following their official implementations; HOI-HFLI additionally applies physics tracking to refine generated motions. PAD-HOI shares the generative backbone of CHOIS but is trained on $\mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{sim}}$, making this comparison a direct assessment of simulator augmentation.

Evaluation Metrics. Following prior work (Li et al., 2024), we report interaction quality metrics ($C_{\text{prec}}, C_{\text{rec}}, C_{F1}, C_{\%}, P_{\text{hand}}$) to assess human-object contacts and geometric penetrations, and ground truth difference metrics (MPJPE, $T_{\text{root}}, T_{\text{obj}}, O_{\text{obj}}$) to measure kinematic deviation from reference motions. An overview of these metrics is provided in Appendix E; multi-seed results are reported in Appendix G.

4.2 RETARGETING QUALITY

The retargeting module bridges the representation gap between reduced-DoF simulator states and full-body SMPL motion (Sec. 3.2). To generate usable training data, it must maintain interaction semantics such as accurate contacts while avoiding mesh penetrations. We evaluate retargeting quality directly on simulator rollouts that contain diverse object geometries and contact-rich behaviors.

We compare our learned retargeter (**Coarse**) and its contact-refined variant (**Coarse+Refine**) against **Mapping+IK** (Aristidou et al., 2018), which fits SMPL poses to reduced-skeleton joint targets without explicit contact objectives. This comparison evaluates the effects of learned reconstruction and contact-aware refinement on interaction quality.

Table 1: Quantitative comparison of retargeting strategies on held-out simulator rollouts.

Method	$C_{prec} \uparrow$	$C_{rec} \uparrow$	$C_{F1} \uparrow$	$C_{\%}$	$P_{\text{hand}} \downarrow$
Mapping+IK (Aristidou et al., 2018)	0.50	0.08	0.13	0.05	0.00
Coarse (Ours)	**0.72**	**0.99**	**0.83**	**0.41**	0.18
Coarse+Refine (Ours)	**0.72**	**0.99**	**0.83**	**0.41**	**0.02**

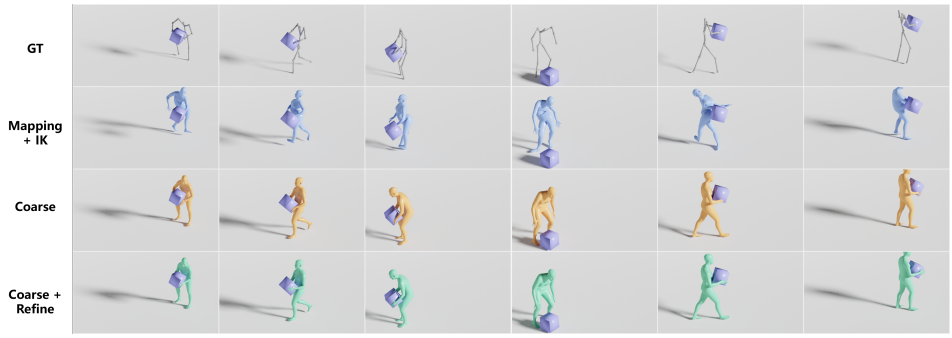


Figure 3: Qualitative comparison of retargeting strategies.

Quantitative Results. As shown in Table 1, our Coarse method and our Coarse with Refinement method achieve identical contact scores. The addition of the refinement stage maintains these exact high quality metrics while reducing the hand penetration P_{hand} from 0.18 down to 0.02. This indicates that our offline optimization successfully resolves mesh interpenetrations against the object SDF without degrading the overall interaction fidelity.

Qualitative Analysis. We visually compare the retargeted sequences against the ground truth in Figure 3. The Mapping+IK baseline struggles to maintain consistent spatial relationships and produces unnaturally distorted postures. Conversely, our Coarse retargeter resolves these artifacts, producing a globally coherent pose that respects the physical constraints of the simulator. The Coarse with Refinement demonstrates precise contact alignment by subtly adjusting the body to perfectly support the object geometry without penetration.

4.3 Generalization and Long-Horizon Generation

We evaluate PAD-HOI on unseen object geometries and long-horizon multi-task sequences to assess its generalization beyond the MoCap training data.

Zero-Shot Generalization.

Quantitative Results. We use the unseen-object test split of $\mathcal{D}_{\text{real}}$, following the evaluation protocol of Li et al. (2024) with 10 training object instances and 5 held-out instances. To ensure a strict zero-shot setting, the test objects are excluded from both $\mathcal{D}_{\text{real}}$ and $\mathcal{D}_{\text{sim}}$ during training. As shown in Table 2, compared with CHOIS, the Coarse+Refine variant of PAD-HOI improves contact F_1 from 0.64 to 0.68, precision from 0.76 to 0.83, and reduces MPJPE from 16.50 to 16.18 cm. Overall, the combination of higher recall, precision, F_1, and lower MPJPE demonstrates stronger interaction quality on unseen geometries. We additionally evaluate a category-level holdout split in Appendix H.

Qualitative Results. We examine whether the diffusion model can learn from $\mathcal{D}_{\text{sim}}$, which contains interactions across diverse box sizes and chair geometries not present in $\mathcal{D}_{\text{real}}$. As shown in Figure 4, the trained model synthesizes interactions that adapt to varying object shapes and sizes. These examples provide qualitative evidence that joint training on retargeted simulation data extends the generative model's coverage beyond $\mathcal{D}_{\text{real}}$.

Long-Horizon Multi-Task Generation. The simulated data provides task executions absent from or sparsely represented in the real MoCap data. We evaluate whether the hybrid dataset enables PAD-HOI to compose multiple task skills into coherent long-horizon sequences. Each sequence

Table 2: Quantitative comparison on five unseen OMOMO objects.

Method	Interaction Quality					Ground Truth Difference			
	$C_{rec}\uparrow$	$C_{prec}\uparrow$	$C_{F1}\uparrow$	$C_{\%}$	$P_{hand}\downarrow$	MPJPE↓	$T_{root}\downarrow$	$T_{obj}\downarrow$	$O_{obj}\downarrow$
InterDiff (Xu et al., 2023b)	0.18	0.48	0.25	0.20	0.62	30.15	68.40	96.50	1.95
MDM (Tevet et al., 2023)	0.32	0.58	0.39	0.33	0.74	21.60	42.10	36.20	2.10
HOI-Diff (Peng et al., 2023)	0.45	0.65	0.51	0.40	0.76	19.85	37.55	28.40	1.75
OMOMO (Li et al., 2023)	0.03	0.16	0.05	0.10	0.26	33.40	80.15	0.00	0.00
CHOIS (Li et al., 2024)	0.59	0.76	0.64	0.56	**0.13**	16.50	28.78	14.29	1.04
HOI-HFLI (Wu et al., 2025b)	0.59	0.78	0.63	0.55	0.16	21.8	34.5	18.9	1.36
PAD-HOI (Coarse Only)	**0.64**	0.81	**0.68**	0.58	0.19	16.21	27.80	14.27	1.03
PAD-HOI (Coarse + Refine)	**0.64**	**0.83**	**0.68**	**0.59**	0.14	**16.18**	**27.76**	**14.24**	**1.03**

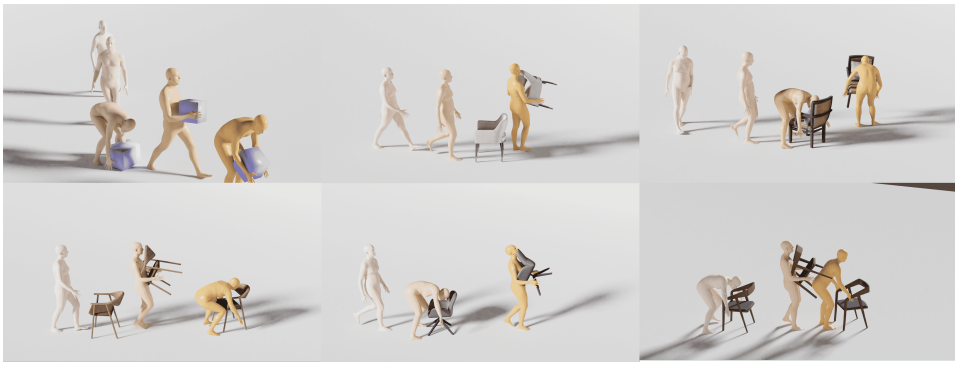


Figure 4: HOI synthesis across varying object shapes and sizes beyond the coverage of $\mathcal{D}_{\text{real}}$.

combines navigation, carrying, climbing, and sitting, with each stage specified by its text prompt, object state, and waypoints.

Qualitative Results. We qualitatively assess our framework's ability to synthesize coherent long-horizon, multi-task motions. As shown in Figure 5, guided by text prompts, object geometries, and spatial waypoints, the generated avatar successfully navigates through a narrow wall gap, approaches and carries a box to a platform, climbs onto the box and then the platform, and finally sits on a chair. The sequence demonstrates that PAD-HOI can compose multiple functional skills into a coherent continuous interaction while preserving the intended task order.

Quantitative Results. Quantitatively, we evaluate long-horizon generation over 400 rollouts with three random seeds. We report final and waypoint success (S_{final} and S_{wp}), transition MPJPE, object displacement, and hand penetration. As shown in Table 3, PAD-HOI achieves 97.5% final-target success and 94.1% waypoint success, with consistent transition and contact quality. CHOIS is included as a real-data-only capability reference under the same staged conditioning protocol. Metrics and evaluation details are provided in Appendix F.

Table 3: Quantitative analysis of long-horizon multi-task composition.

Metric	S_{final}	S_{wp}	MPJPE	$\text{MPJPE}_{\text{trans}}$	T_{obj}	$T_{\text{obj}}^{\text{trans}}$	P_{hand}
CHOIS	$49.5 \pm 7.0\%$	$43.8 \pm 7.5\%$	18.2 ± 2.4	16.1 ± 1.8	15.3 ± 3.2	5.4 ± 0.9	4.0 ± 0.8
PAD-HOI	$\mathbf{97.5 \pm 1.5\%}$	$\mathbf{94.1 \pm 2.3\%}$	$\mathbf{8.10 \pm 0.70}$	$\mathbf{8.17 \pm 0.38}$	$\mathbf{6.43 \pm 1.41}$	$\mathbf{1.95 \pm 0.13}$	$\mathbf{1.51 \pm 0.20}$

4.4 Validation on Standard Benchmarks

We evaluate on the seen-object test split of $\mathcal{D}_{\text{real}}$ to assess whether simulator augmentation preserves in-distribution generation quality. As shown in Table 4, PAD-HOI improves C_{F1} from 0.67 to 0.69 and reduces MPJPE from 15.30 to 14.97 relative to the strongest baseline, CHOIS. Comparing the two PAD-HOI variants, contact-aware refinement further reduces hand penetration from 0.61 to 0.54 cm while maintaining contact F_1 and MPJPE, confirming that refinement quality affects downstream generation. These results show that $\mathcal{D}_{\text{sim}}$ provides auxiliary supervision that improves contact and human pose accuracy while retaining generation quality on $\mathcal{D}_{\text{real}}$.

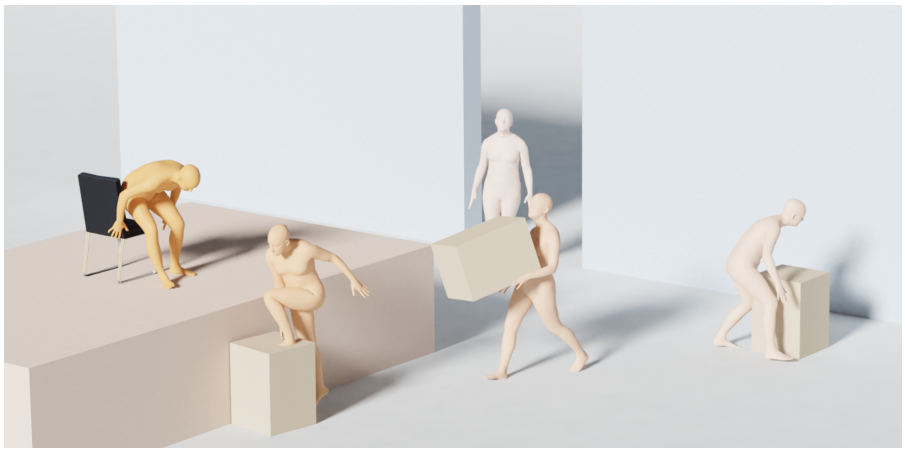


Figure 5: Long-horizon generation combining navigation, carrying, climbing, and sitting.

Table 4: Quantitative comparison on seen OMOMO objects.

Method	$C_{rec} \uparrow$	$C_{prec} \uparrow$	$C_{F1} \uparrow$	$C_{\%}$	$P_{hand} \downarrow$	MPJPE↓	$T_{root} \downarrow$	$T_{obj} \downarrow$	$O_{obj} \downarrow$
InterDiff (Xu et al., 2023b)	0.28	0.63	0.33	0.27	0.55	25.91	63.44	88.35	1.65
MDM (Tevet et al., 2023)	0.47	0.72	0.53	0.43	0.66	17.86	34.16	24.46	1.85
HOI-Diff (Peng et al., 2023)	0.62	0.77	0.64	0.49	0.69	17.44	32.28	22.75	1.51
OMOMO (Li et al., 2023)	0.66	0.77	0.67	0.56	0.55	15.82	24.75	0.00	0.00
CHOIS (Li et al., 2024)	0.64	**0.80**	0.67	0.54	0.59	15.30	24.43	**12.53**	0.99
PAD-HOI (Coarse Only)	**0.67**	0.79	**0.69**	0.57	0.61	**14.97**	**22.39**	13.89	**0.90**
PAD-HOI (Coarse + Refine)	**0.67**	0.79	**0.69**	**0.58**	**0.54**	**14.97**	**22.39**	13.89	**0.90**

4.5 Analysis of Training Data Sources

We analyze the contribution of each training data source to understand whether real and simulated interactions provide complementary supervision. Specifically, we compare models trained on $\mathcal{D}_{\text{real}}$, $\mathcal{D}_{\text{sim}}$, and their union, and evaluate them on both real and simulated test splits. The $\mathcal{D}_{\text{sim}}$ test split contains 200 trajectories excluded from all training phases.

Table 5: Cross-domain evaluation by training source.

	$\mathcal{D}_{\text{real}}$ Test			$\mathcal{D}_{\text{sim}}$ Test		
Training Data	$C_{F1} \uparrow$	MPJPE↓	$P_{hand} \downarrow$	$C_{F1} \uparrow$	MPJPE↓	$P_{hand} \downarrow$
$\mathcal{D}_{\text{real}}$-only	0.67	15.30	0.59	0.38	22.45	2.41
$\mathcal{D}_{\text{sim}}$-only	0.36	28.12	2.43	**0.74**	**11.26**	0.86
$\mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{sim}}$	**0.69**	**14.97**	**0.54**	0.71	12.03	**0.78**

As shown in Table 5, each single-source model performs better on its own domain than on the other. Joint training maintains strong performance across both test splits while achieving the lowest penetration on both, supporting the complementarity of the two data sources.

5 Conclusion

We presented PAD-HOI, a framework that leverages reinforcement learning policies in physics simulation to generate diverse interaction training data, augmenting limited motion capture datasets for generalizable HOI synthesis. Despite these advantages, scaling the framework to entirely novel task distributions remains constrained by the computational efficiency and reward engineering required to train task-specific expert policies; detailed limitations are discussed in Appendix J. A promising direction for future work is to leverage the learned diffusion model as a prior to guide the RL training for new tasks. Ultimately, this work establishes a robust foundation for generalizable and physically grounded motion synthesis.

AI USE STATEMENT

We used generative AI tools to assist with writing and literature discovery. For writing, these tools helped identify and correct grammatical errors and improve sentence fluency. For literature discovery, we used AI-assisted internet searches during the preliminary literature review to identify relevant papers in human–object interaction synthesis. The authors reviewed the suggested edits to ensure that they preserved the intended technical meaning, and checked the retrieved papers against their original publications to verify their relevance and bibliographic information. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work uses only publicly available, open-source motion-capture datasets, body models, and object assets, all of which are used in accordance with their respective licenses and usage conditions. We do not collect new human-subject data, conduct human-subject experiments, or release private personal information. Therefore, this work does not raise additional concerns regarding participant consent, privacy, or ethics approval.

REPRODUCIBILITY STATEMENT

The data synthesis, retargeting, and diffusion procedures are described in Sections 3.1, 3.2, and 3.3. Appendices A and B provide additional formulations for expert training and contact refinement. Dataset composition and model implementation details are provided in Appendices C and D. Evaluation metrics and the long-horizon evaluation protocol are described in Appendices E and F.

REFERENCES

Andreas Aristidou, Joan Lasenby, Yiorgos Chrysanthou, and Ariel Shamir. Inverse kinematics techniques in computer graphics: A survey. In *Computer Graphics Forum*, volume 37(6), pp. 35–58, 2018.

Ananya Bal, Kartik Sharma, Ethan Lai, Samyak Tiwari, Liza Dahiya, Chaitanya Chawla, and Laszlo A Jeni. Mad-hoi: Masked autoregressive diffusion for generating articulated hand object interactions from text. *arXiv preprint arXiv:2608.10162*, 2026.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15935–15946, 2022.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Songjin Cai, Linjie Zhong, Ling Guo, and Changxing Ding. Vihoi: Human-object interaction synthesis with visual priors. *arXiv preprint arXiv:2603.24383*, 2026.

Peishan Cong, Ziyi Wang, Yuexin Ma, and Xiangyu Yue. Semgeomo: Dynamic contextual human motion generation with semantic and geometric guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17561–17570, 2025.

Zekai Deng, Ye Shi, Kaiyang Ji, Lan Xu, Shaoli Huang, and Jingya Wang. Human-object interaction via automatically designed vlm-guided motion policy. In *International Conference on Learning Representations*, volume 2026, pp. 100691–100718, 2026.

Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19888–19901, 2024.

Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. C·ase: Learning conditional adversarial skill embeddings for physics-based characters. *arXiv preprint arXiv:2309.11351*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12943–12954, 2023.

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 *Advances in Neural Information Processing Systems*, pp. 79741–79763, 2024. doi: 10.52202/079017-2532. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/918b9487f8ea4661e8ba5a02b2126658-Paper-Conference.pdf.

Zichen Geng, Zeeshan Hayder, Wei Liu, and Ajmal Saeed Mian. Auto-regressive diffusion for generating 3d human-object interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3131–3139, 2025.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2282–2292, 2019.

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11374–11384, 2021.

Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701597. doi: 10.1145/3588432.3591525. URL https://doi.org/10.1145/3588432.3591525.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16750–16761, 2023.

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction from Multi-view RGB-D Images. *International Journal of Computer Vision*, 132(7):2551–2566, July 2024. ISSN 1573-1405. doi: 10.1007/s11263-024-01984-1. URL https://doi.org/10.1007/s11263-024-01984-1.

Bin Ji, Ye Pan, Zhimeng Liu, Shuai Tan, Xiaogang Jin, and Xiaokang Yang. Pomp: Physics-consistent motion generative model through phase manifolds. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22690–22701, 2025.

Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *Proceedings of European Conference on Computer Vision*. Springer, 2022.

Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *ICCV*, 2023.

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

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pp. 54–72. Springer, 2024.

Shujia Li, Haiyu Zhang, Xinyuan Chen, Yaohui Wang, and Yutong Ban. Genhoi: Generalizing text-driven 4d human-object interaction synthesis for unseen objects. *arXiv preprint arXiv:2506.15483*, 2025.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1769–1782, 2025.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. ISBN 9798400708978. URL https://doi.org/10.1145/3596711.3596800.

Yuke Lou, Yiming Wang, Zhen Wu, Rui Zhao, Wenjia Wang, Mingyi Shi, and Taku Komura. Zero-shot human-object interaction synthesis with multimodal priors. *arXiv preprint arXiv:2503.20118*, 2025.

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. *ACM Transactions on Graphics*, 45(2):1–18, 2025.

Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. In *Advances in Neural Information Processing Systems*, 2021.

Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. In *Advances in Neural Information Processing Systems*, 2022a.

Zhengyi Luo, Ye Yuan, and Kris M. Kitani. From universal humanoid control to automatic physically valid character creation. *ArXiv*, abs/2206.09286, 2022b.

Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10895–10904, 2023.

Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Omnigrasp: Grasping diverse objects with simulated humanoids. *Advances in Neural Information Processing Systems*, 37:2161–2184, 2024a.

Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris M. Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=OrOd8PxOO2.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pp. 5442–5451, October 2019.

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. *arXiv preprint arXiv:2108.10470*, 2021.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5379–5391, 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019.

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.

Xue Bin Peng. Mimickit: A reinforcement learning framework for motion imitation and control. *arXiv preprint arXiv:2510.13794*, 2025. URL `https://arxiv.org/abs/2510.13794`.

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.

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.

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.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4332–4341, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. Calm: Conditional adversarial latent models for directable virtual characters. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–9, 2023.

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

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

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. *arXiv preprint arXiv:2410.03441*, 2024.

Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, et al. Skillmimic: Learning basketball interaction skills from demonstrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17540–17549, 2025.

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022.

Ziyin Wang, Sirui Xu, Bing Zhou, Jiangshan Gong, Jian Wang, Yu-Xiong Wang, Liang-Yan Gui, et al. Unleashing guidance without classifiers for human-object interaction animation. In *International Conference on Learning Representations*, volume 2026, pp. 46121–46141, 2026.

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

Yan Wu, Korrawe Karunratanakul, Zhengyi Luo, and Siyu Tang. Uniphys: Unified planner and controller with diffusion for flexible physics-based character control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. Human-object interaction from human-level instructions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11176–11186. IEEE, 2025b.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1vCnDyQkjg.

Michael Xu, Yi Shi, KangKang Yin, and Xue Bin Peng. Parc: Physics-based augmentation with reinforcement learning for character controllers. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025a.

Pei Xu, Kaixiang Xie, Sheldon Andrews, Paul G Kry, Michael Neff, Morgan McGuire, Ioannis Karamouzas, and Victor Zordan. AdaptNet: Policy adaptation for physics-based character control. *ACM Transactions on Graphics*, 42(6), 2023a. doi: 10.1145/3618375.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14928–14940, 2023b.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, et al. Interact: Advancing large-scale versatile 3d human-object interaction generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7048–7060, 2025b.

Jie Yang, Xuesong Niu, Nan Jiang, Ruimao Zhang, and Siyuan Huang. F-hoi: Toward fine-grained semantic-aligned 3d human-object interactions, 2024. URL https://arxiv.org/abs/2407.12435.

Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*, pp. 246–263. Springer, 2024.

Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2(3):5, 2023.

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 *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16010–16021, 2023.

Junpeng Yue, Zepeng Wang, Yuxuan Wang, Weishuai Zeng, Jiangxing Wang, Xinrun Xu, Yu Zhang, Sipeng Zheng, Ziluo Ding, and Zongqing Lu. Rl from physical feedback: Aligning large motion models with humanoid control. *arXiv preprint arXiv:2506.12769*, 2025.

Ling-An Zeng, Guohong Huang, Yi-Lin Wei, Shengbo Gu, Yu-Ming Tang, Jingke Meng, and Wei-Shi Zheng. Chainhoi: Joint-based kinematic chain modeling for human-object interaction generation. *arXiv preprint arXiv:2503.13130*, 2025.

Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4diffusion: Multi-view video diffusion model for 4d generation. *Advances in Neural Information Processing Systems*, 37:15272–15295, 2024.

Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision*, pp. 518–535. Springer, 2022.

Policy-as-Data: Learning Generalizable HOI Diffusion Models from Simulated Physics
Appendix

A Expert Policy Training

This section details the comprehensive mathematical formulation and optimization strategy for the reinforcement learning experts introduced in Section 3.1 of the main text. The objective is to train a control policy π via reinforcement learning (RL) that enables the character to perform functional tasks while maintaining realistic movement. To ensure that synthesized interactions are both functional and anthropomorphically natural, we adopt a dual-component reward that decouples task specification from motion style (Peng et al., 2021; Hassan et al., 2023). Let s_t denote the character-environment state at time step t, and let a_t denote the action applied by the policy. The reward is defined as:

$$r(s_t, a_t, s_{t+1}, g) = w_G\, r_G(s_t, a_t, s_{t+1}, g) + w_S\, r_S(s_t, s_{t+1}), \tag{6}$$

where r_G specifies the *task* objective, e.g., reaching an object or transporting it to a target waypoint g, and r_S specifies a *style* objective that encourages human-like transitions.

To provide a task-agnostic motion prior, r_S employs a discriminator D_ψ trained to distinguish simulated policy transitions (s_t, s_{t+1}) from real transitions in an unstructured MoCap dataset $\mathcal{M}$. D_ψ is optimized via the standard adversarial objective:

$$\mathcal{L}_\psi = \mathbb{E}_{(s_t, s_{t+1})\sim d^{\mathcal{M}}}\left[-\log D_\psi(s_t, s_{t+1})\right] + \mathbb{E}_{(s_t, s_{t+1})\sim d^{\pi}}\left[-\log(1 - D_\psi(s_t, s_{t+1}))\right], \tag{7}$$

where $d^{\mathcal{M}}$ represents the distribution of state transitions in the MoCap dataset and d^{π} denotes the state-transition distribution induced by the policy.

The policy is then rewarded for fooling the discriminator, which encourages natural human kinematics across diverse tasks without rigidly tracking specific reference clips. The style reward for a given transition is then defined as:

$$r_S(s_t, s_{t+1}) = -\log(1 - D_\psi(s_t, s_{t+1})). \tag{8}$$

The policy $\pi(a_t|s_t, g)$ is trained to maximize the expected cumulative discounted reward over a distribution of goals $g \sim p(g)$ and trajectories $\{s_0, a_0, s_1, a_1, ...\}$ induced by the policy:

$$J(\pi) = \mathbb{E}_{g\sim p(g),(s_t,a_t,s_{t+1})\sim\pi}\left[\sum_{t=0}^{T}\gamma^t r(s_t, a_t, s_{t+1}, g)\right], \tag{9}$$

where $\gamma \in [0, 1)$ is the discount factor. This formulation encourages the agent to complete tasks successfully while simultaneously adopting the kinematic styles found in the reference MoCap data.

B Contact Aware Refinement

This section provides the detailed mathematical formulation for the contact aware refinement stage introduced in Section 3.2 of the main text.

Geometric Formulation and Signed Distance Field. We represent the object geometry at each frame by its signed distance field (SDF) in a canonical object frame. Let $\mathbf{R}_t \in SO(3)$ and $\mathbf{o}_t \in \mathbb{R}^3$ denote the object rotation and translation at frame t. The object-frame coordinate of a 3D point $\mathbf{p}$ is

$$\tilde{\mathbf{p}}_t = \mathbf{R}_t^\top(\mathbf{p} - \mathbf{o}_t), \tag{10}$$

and the world-space SDF is evaluated as

$$\phi(\mathbf{p}, t) = \mathrm{SDF}_{\mathcal{O}_0}(\tilde{\mathbf{p}}_t), \tag{11}$$

where $\mathcal{O}_0$ is the canonical object surface, $\phi(\mathbf{p}, t) > 0$ denotes a point outside the object, and $\phi(\mathbf{p}, t) < 0$ denotes penetration. The SDF can be queried for each object mesh used in simulation, including procedurally varied boxes, tables, chairs, and platforms. In our final setting $\{\mathbf{R}_t, \mathbf{o}_t\}_{t=1}^T$ correspond to the frozen object trajectory $\hat{\boldsymbol{x}}_t^o$ and are kept fixed during refinement.

Optimization Space and Temporal Masking. As stated in the main text the human state $\hat{\boldsymbol{x}}_t^h$ is refined to improve contact. The implementation achieves this palm centered refinement by optimizing a small upper body subset of the SMPL X pose parameters $\boldsymbol{\theta}_t$ extracted from $\hat{\boldsymbol{x}}_t^h$ namely the left and right collar shoulder elbow and wrist joints. Let $\hat{\boldsymbol{\theta}}_t$ be the retargeted initialization and $\Delta\boldsymbol{\theta}_t$ the optimized pose offset on these joints. Then the refined pose is

$$\boldsymbol{\theta}_t = \hat{\boldsymbol{\theta}}_t + m_t^{\mathrm{opt}} \odot \Delta\boldsymbol{\theta}_t, \tag{12}$$

where $m_t^{\mathrm{opt}} \in \{0, 1\}$ is a binary optimization mask. The body shape parameters hand pose parameters root translation and root orientation remain fixed.

Crucially we define the contact periods using the simulator provided contact indicators c_t. We derive specific active contact masks m_t^{cont} directly from c_t and spatial proximity. In practice strict contact is identified using a distance threshold of $\tau_{\mathrm{cont}} = 0.02$ m while a close range mask m_t^{close} is defined using $\tau_{\mathrm{close}} = 0.12$ m. To avoid unstable corrections from severely invalid frames we only refine frames whose initial hand object SDF values are above a lower correction bound of -0.15 m. The final optimization mask m_t^{opt} is obtained by dilating these simulator guided masks by 4 frames in time to ensure smooth transitions.

Contact Enforcement Objectives. The contact term E_{cont} from the main text is practically expanded into surface attraction close range attraction no slip regularization and palm aware face consistency. For the interaction vertices $\mathcal{V}_{\mathrm{contact}}$ sampled from the palms we define proximity weights

$$w_{j,t} = \exp\left[-\frac{|\phi(\mathbf{v}_j(t), t)|}{\sigma}\right], \tag{13}$$

where we set $\sigma = 0.01$ m. The surface attraction loss on strict contact frames is

$$E_{\mathrm{surf}} = \frac{1}{|\mathcal{T}_{\mathrm{cont}}|}\sum_{t\in\mathcal{T}_{\mathrm{cont}}}\frac{\sum_{j\in\mathcal{V}_{\mathrm{contact}}} w_{j,t}\,\phi(\mathbf{v}_j(t), t)^2}{\sum_{j\in\mathcal{V}_{\mathrm{contact}}} w_{j,t}}, \tag{14}$$

where $\mathcal{T}_{\mathrm{cont}}$ is the set of frames where c_t indicates active contact. We apply a similar weighted SDF loss E_{close} on close frames to avoid hovering palms. The no slip term enforces temporal consistency of contacted hand vertices in object local coordinates

$$E_{\mathrm{slip}} = \frac{\sum_{t=1}^{T-1}\sum_{j\in\mathcal{V}_{\mathrm{contact}}}\bar{w}_{j,t}\;\|\tilde{\mathbf{v}}_j(t+1) - \tilde{\mathbf{v}}_j(t)\|_2^2}{\sum_{t=1}^{T-1}\sum_{j\in\mathcal{V}_{\mathrm{contact}}}\bar{w}_{j,t}}, \tag{15}$$

where $\tilde{\mathbf{v}}_j(t)$ denotes the object local coordinate and $\bar{w}_{j,t} = \frac{1}{2}[w_{j,t} + w_{j,t+1}]$.

To aggregate these objectives the corresponding weights are set to $\lambda_{\text{cont}} = 3.0$ for surface attraction $\lambda_{\text{close}} = 1.0$ for close range attraction and $\lambda_{\text{noslip}} = 1.0$ for the no slip term. We further use palm face consistency weights of $\lambda_{\text{plane}} = 1.0$ $\lambda_{\text{orient}} = 0.5$ and a fallback palm attraction weight of $\lambda_{\text{palm}} = 0.5$ to encourage stable palm object attachment. The final contact objective is formulated as

$$E_{\text{cont}} = \lambda_{\text{cont}} E_{\text{surf}} + \lambda_{\text{close}} E_{\text{close}} + \lambda_{\text{noslip}} E_{\text{slip}} + \lambda_{\text{palm}} E_{\text{palm}} + \lambda_{\text{plane}} E_{\text{plane}} + \lambda_{\text{orient}} E_{\text{orient}}. \quad (16)$$

Penetration Smoothness and Prior Penalties. To ensure the refined motion remains physically plausible and free of jitter we explicitly define the smoothness term E_{sm} introduced in the main text as an ℓ_2 penalty on the joint accelerations of the optimized offsets

$$E_{\text{sm}} = \sum_{t=2}^{T-1} \|\Delta\boldsymbol{\theta}_{t+1} - 2\Delta\boldsymbol{\theta}_t + \Delta\boldsymbol{\theta}_{t-1}\|_2^2 . \quad (17)$$

The penetration term E_{pen} is

$$E_{\text{pen}} = E_{\text{hand_pen}} + E_{\text{body_pen}} + E_{\text{arm_clr}}. \quad (18)$$

Here $E_{\text{hand_pen}}$ penalizes negative SDF values on $\mathcal{V}_{\text{contact}}$ while $E_{\text{body_pen}}$ penalizes penetration of a subset of nearest-to-surface body vertices and $E_{\text{arm_clr}}$ penalizes elbow-to-wrist segments that breach a safety margin around the object surface. In practice the hand penetration depth is clamped at 0.10 m, the hand penetration weight is set to 8.0, the body penetration weight is set to 3.0, and 2000 body vertices nearest to the object surface are used for the coarse body penetration term.

Finally the reconstruction and prior terms are implemented as an ℓ_2 penalty on the optimized pose offsets keeping the refined motion close to the retargeted initialization. We set the pose regularization weight to 0.1 and the pose smoothness weight to 3.0. Optimization is performed using Adam with a learning rate of 5×10^{-3}. During the first stage of 200 iterations we optimize strictly for penetration resolution smoothness and pose regularization with all contact objectives disabled. In the second stage of 400 iterations we activate the full E_{cont} objective to securely anchor the hands to the object geometry while maintaining the collision free state achieved in the first stage.

C Dataset

C.1 Physics Simulator Configuration

Simulation Environment. We implement all physics-based policy training and trajectory generation in Isaac Gym (Makoviychuk et al., 2021), a GPU-accelerated physics simulator. The simulator runs at 60 Hz with a fixed timestep of $\Delta t = 1/60$ seconds and uses 4 substeps per frame for collision resolution. Ground contact friction is set to 0.9, and object-hand friction is set to 1.5 to enable stable grasping. We use the PhysX backend with position-based dynamics for fast and stable contact resolution.

Humanoid Model. The simulated character is a reduced rigid-body humanoid with 15 bodies and 32 degrees of freedom, designed for efficient RL policy learning. The bodies include pelvis, torso, head, left and right upper arms, forearms, hands, thighs, shins, and feet. Each joint is actuated using PD controllers with position targets provided by the policy at 30 Hz control frequency. The humanoid has a total mass of approximately 75 kg with mass distribution proportional to real human anthropometry. Joint angle limits follow physiological ranges to prevent unnatural poses during policy rollouts.

Domain Randomization. To encourage geometric generalization, we randomize object properties and initial conditions during policy training. For carry tasks with boxes, dimensions are sampled uniformly from $[0.3, 0.4]$ meters per side. For procedurally generated tables and chairs, we vary leg heights, seat widths, and tabletop dimensions within plausible ranges, resulting in over 100 distinct object instances across training. Initial object positions are randomized within a 3-4 meter radius from the humanoid spawn point, and target waypoints are sampled from a circle with radius 1.5-3 meters around the initial object position. Additionally, we apply small perturbations to humanoid joint positions at episode initialization to avoid overfitting to a single starting configuration.

C.2 Motion Capture and Synthetic Data

To train our generative model, we utilize a hybrid dataset consisting of real-world motion capture and synthetic trajectories generated via our Policy-as-Data framework.**Real-World Data ($\mathcal{D}_{\text{real}}$):** We utilize OMOMO (Li et al., 2023), a large-scale motion capture dataset containing diverse human-object interactions. Specifically, it features full-body manipulation sequences where subjects perform everyday tasks, such as pushing, pulling, carrying, and repositioning various household objects. These sequences capture complex human-object coordination, providing our framework with a robust prior for natural postural synthesis.**Simulator Data ($\mathcal{D}_{\text{sim}}$):** To provide the generative model with diverse, physically grounded functional skills, we generate raw rigid-body trajectories across three distinct interaction tasks. To support language-conditioned generation, the text prompts associated with these synthetic trajectories follow standardized templates based on the task category.**(1) Carry Tasks:** This category encompasses complete, multi-stage behaviors including approaching, picking up, carrying, and placing objects. We collect two subsets to encourage different types of generalization. First, to impart robust geometric adaptation, we generate 400 episodes (each lasting 300 frames) of carrying boxes with randomized sizes sampled uniformly from $[0.3, 0.4]$ meters. Moreover, the boxes are initialized at a random distance of 3 to 4 meters from the humanoid. The goal positions are sampled from a circle around the initial position of the object with a radius of 1.5 to 3 meters. Second, to ensure instance-level generalization across complex topologies, we replace the boxes with 30 distinct tables and 90 distinct chairs, collecting an additional 2048 episodes. The initialization distances and goal sampling strategies remain identical to those used for the boxes. All sequences in this category use the text prompt: "A person approaches a chair/box/table, picks it up, and places it in the designated location." **(2) Sit Tasks:** To further capture motions in diverse human-object interactions, we generate 1024 trajectories, each lasting 200 frames, of the humanoid interacting with and sitting on 30 different chair geometries. The associated text prompt for these episodes is: "A person approaches and sits on a chair."**(3) Climb Tasks:** To provide the model with dynamic locomotion and whole-body coordination skills, we collect 1024 trajectories, each lasting 200 frames, where the humanoid character climbs from the ground onto an elevated platform or a table. These sequences are annotated with the prompt: "A person climbs onto a platform/table."

D Implementation Details

We implement the complete framework, including the coarse-to-fine retargeter and diffusion model, in PyTorch. The physics-based trajectories are generated in Isaac Gym for parallelized data synthesis. The diffusion model is trained on eight NVIDIA RTX 4090 GPUs with mixed-precision training, while the retargeter follows the AvatarPoser training setup.

Diffusion Model. Our diffusion model follows the implementation in CHOIS (Li et al., 2024). It is a transformer-based conditional Gaussian diffusion model defined over motion windows of length 120. The predicted motion representation has dimension 216, including object translation (3), object rotation matrix (9), global human joint positions (24×3), and human joint rotations in 6D representation (22×6). When semantic contact channels are enabled, an additional 4 contact dimensions are appended. The object condition is represented using a 1024×3 basis-point-set descriptor, which is flattened and encoded by a two-layer MLP with dimensions $3072 \to 512 \to 256$. The denoiser is a transformer decoder with model dimension 512, 4 decoder layers, 4 attention heads, and key/value dimensions of 256. A sinusoidal diffusion-timestep embedding is processed by an MLP and injected as an extra conditioning token. The model uses full self-attention over the entire temporal window. When language conditioning is enabled, text is encoded by CLIP ViT-B/32 and projected to the same 512-dimensional conditioning space. We use a cosine diffusion schedule with 1000 denoising steps and train the model with the `pred_x0` objective and an ℓ_1 reconstruction loss. Training uses Adam with learning rate 2×10^{-4}, batch size 32, gradient accumulation over 2 steps, EMA decay 0.995, and a total of 8×10^5 optimization steps.

Coarse-to-Fine Retargeter. We designed our retargeter based on AvatarPoser (Jiang et al., 2022). Unlike the original AvatarPoser which only takes the head and left-right hands as input, our retargeter takes a 270-dimensional sparse-motion descriptor extracted from the full 15-joint physics humanoid. To mitigate the scale and kinematic differences between the SMPL model and the physics humanoid, we first optimize a set of SMPL shape parameters β. This pre-optimization aligns the bone lengths (e.g., arms and legs) and the pelvis height of the SMPL model to match the specific kinematic

attributes of the physics humanoid. For the network input, we concatenate local rotation in 6D, global rotation in 6D, linear velocity, and angular velocity for each joint, yielding $15 \times (6 + 6 + 3 + 3) = 270$ dimensions per frame. The network processes a temporal window of 40 frames using a linear input embedding of size $270 \rightarrow 256$, followed by a 3-layer Transformer encoder with 8 attention heads. Two lightweight MLP heads decode the final temporal features: one predicts the root orientation in 6D ($256 \rightarrow 256 \rightarrow 6$), and the other predicts the 21 body-joint rotations in 6D ($256 \rightarrow 256 \rightarrow 126$). The model supervises the last 3 frames of each window. During training, a differentiable SMPL-H body model is used for forward kinematics to recover 24 joints for positional supervision. Crucially, we apply the previously optimized β parameters rather than the default zero shape vector during these kinematic computations to ensure accurate spatial alignment. We use an ℓ_1 loss on root orientation, body-joint rotations, and joint positions, together with a velocity loss on predicted joint trajectories. The total loss is

$$\mathcal{L}_{\text{ret}} = 0.02\mathcal{L}_{\text{root}} + \mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{jpos}} + 0.5\mathcal{L}_{\text{vel}}. \tag{19}$$

Training uses Adam with learning rate 10^{-4} and batch size 256. The learning rate is decayed by a factor of 0.5 at iterations 60k, 120k, 180k, 240k, 300k, and 360k.

E Evaluation Metrics

To assess generation quality and physical interaction, we employ two sets of standard metrics following prior work (Li et al., 2024):

Interaction Quality Metric: This metric assesses the physical fidelity of human-object contact. We report contact Precision (C_{prec}), Recall (C_{rec}), F1 score (C_{F1}), and overall contact percentage ($C_{\%}$). To measure physical violations, we report a penetration score (P_{hand}), calculated as the average penetration depth of interacting human vertices into the Signed Distance Field (SDF) of the object.

Ground Truth Difference Metric: This metric measures kinematic deviation from the reference motions. We compute the Mean Per-Joint Position Error (MPJPE), root translation error (T_{root}), and object position error (T_{obj}) via Euclidean distance (in cm). We also report the object orientation error (O_{obj}), which is measured using the Frobenius norm of the rotational difference.

F Long-Horizon Evaluation Protocol

We evaluate long-horizon generation over multi-stage sequences combining navigation, carrying, climbing, and sitting. Each sequence is generated over overlapping temporal windows using stage-specific text, object, and waypoint conditions. We report results over 400 rollouts with three random seeds and an average sequence length of 600 frames.

Goal Success. S_{final} is the fraction of rollouts whose final target is reached within 10 cm, while S_{wp} measures success at the intermediate waypoints under the same threshold.

Transition Quality. $\text{MPJPE}_{\text{trans}}$ measures the mean per-joint position error on frames around skill boundaries, defined as a 20-frame window centered on each task transition point. $T_{\text{obj}}^{\text{trans}}$ measures object displacement over the same transition intervals. The overall MPJPE, object position error, and hand penetration are described in Appendix E.

Evaluation Protocol Details. Each long-horizon sequence is generated autoregressively using overlapping 120-frame windows with approximately 10-frame overlap. At each stage transition, we switch the text condition and update spatial waypoints according to the task specification. Results are aggregated over three random seeds by computing per-seed averages and then reporting the mean across seeds. Goal success requires the final human root position to be within 10 cm of the target waypoint at sequence end; intermediate waypoint success S_{wp} applies the same spatial threshold at each stage boundary.

G Multi-Seed Robustness

To assess sensitivity to random initialization, we evaluate PAD-HOI (Coarse+Refine) on the unseen-object benchmark under three independent random seeds and report the mean and standard deviation

of the key metrics in Table 6. The main-text benchmark tables report single-seed results to follow the reporting convention of prior HOI work; we provide this additional multi-seed evaluation to show the variability of our method. Performance is stable across seeds: contact F_1 is 0.68 ± 0.006, and MPJPE is 16.18 ± 0.15 cm. The other reported metrics also vary little, indicating that the main-text results are representative of our runs.

Table 6: Multi-seed evaluation of PAD-HOI (Coarse+Refine) on five unseen OMOMO objects.

Method	$C_{rec} \uparrow$	$C_{prec} \uparrow$	$C_{F1} \uparrow$	$C_{\%}$	$P_{hand} \downarrow$	MPJPE↓	$T_{root} \downarrow$	$T_{obj} \downarrow$	$O_{obj} \downarrow$
PAD-HOI (Coarse+Refine)	0.64 ± 0.005	0.83 ± 0.003	0.68 ± 0.006	0.59 ± 0.005	0.14 ± 0.001	16.18 ± 0.15	27.76 ± 0.68	14.24 ± 0.15	1.03 ± 0.01

H Category-Level Generalization

The unseen-object OMOMO split in the main text holds out object instances while retaining their categories. We additionally evaluate a category-level split that holds out one full object category. Table 7 reports the results for CHOIS and PAD-HOI. On this split, PAD-HOI improves contact F_1 from 0.53 to 0.60 and reduces MPJPE from 19.62 to 19.17 cm. The contact F_1 improvement is approximately 13% relative to CHOIS, compared with approximately 6% on the held-out instance split in the main text (0.64 to 0.68). This larger gain under a category-level holdout provides further evidence that the augmented data improves interaction quality on unseen object geometries.

Table 7: Category-level generalization with one object category held out.

Method	$C_{rec} \uparrow$	$C_{prec} \uparrow$	$C_{F1} \uparrow$	$P_{hand} \downarrow$	MPJPE↓	$T_{root} \downarrow$
CHOIS (Li et al., 2024)	0.49 ± 0.01	0.67 ± 0.01	0.53 ± 0.01	0.14 ± 0.01	19.62 ± 0.21	32.38 ± 0.45
PAD-HOI	$\mathbf{0.53 \pm 0.01}$	$\mathbf{0.72 \pm 0.01}$	$\mathbf{0.60 \pm 0.01}$	$\mathbf{0.13 \pm 0.00}$	$\mathbf{19.17 \pm 0.20}$	$\mathbf{29.08 \pm 0.23}$

I Scaling with Simulator Data

To examine how performance changes as simulator data is added, we keep the training procedure fixed and vary the proportion of $\mathcal{D}_{\text{sim}}$ in the training set. Figure 6 reports contact recall, contact precision, and contact F_1 on the unseen-object split. Adding a small amount of simulator data produces the largest improvement: at 10%, C_{rec} increases from 0.59 to 0.63, C_{prec} from 0.76 to 0.83, and C_{F1} from 0.64 to 0.67. Increasing the proportion to 100% yields smaller additional gains, reaching 0.64, 0.83, and 0.68, respectively. These results suggest that simulator data improves contact coverage efficiently, with diminishing returns as more synthetic trajectories are added.

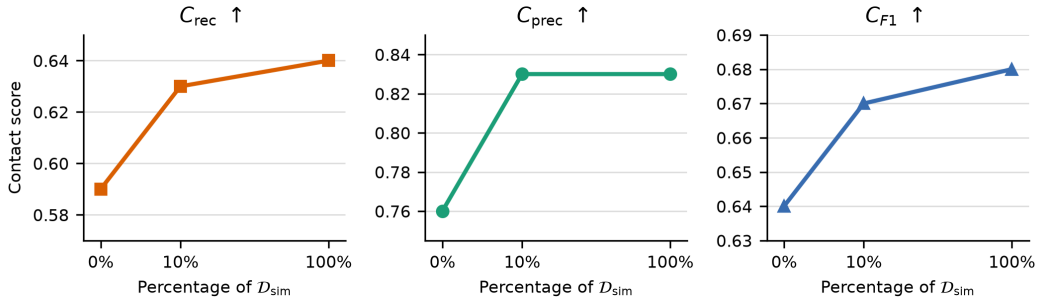


Figure 6: Scaling performance with the proportion of simulator trajectories in the training data. Results are reported on the unseen-object split.

J Limitations

Task-Specific Policy Requirements. Our framework requires training separate RL policies for each task category (carry, sit, climb). Extending to new task types necessitates reward engineering

and policy training, which demands computational resources and domain expertise. This constrains rapid adaptation to novel interaction scenarios.

Simplified Physics Representation. The simulator employs a reduced rigid-body model with 15 bodies and 32 DoF, which abstracts away fine-grained biomechanical constraints and muscle dynamics. While our retargeting bridge mitigates this gap, certain subtle motor adjustments present in real human motion may not be fully captured.

Object Category Coverage. The current implementation focuses on rigid objects (boxes, tables, chairs, platforms) with procedural geometric variation. Extension to fundamentally different object types—such as deformable materials, articulated mechanisms, or tools with complex affordances—would require designing appropriate task definitions and reward structures.