Synthesizing Whole-Body Loco-Manipulation without Human Demonstration

Yufei Zhu[1,*] **Runyi Yu**[2,*] **Yinhuai Wang**[2] **Aoru Xue**[1] **Qin Sun**[3,4]
Qingqiu Huang[5] **Xinge Zhu**[6] **Zhiyang Dou**[7] **Yuexin Ma**[1,†]

[1]ShanghaiTech University [2]The Hong Kong University of Science and Technology
[3]Guangzhou Institute of Energy Conversion, CAS
[4]University of Science and Technology of China [5]Morphi
[6]The Chinese University of Hong Kong [7]The University of Hong Kong
[*]Equal Contribution [†]Corresponding Author

Project page: https://manipsynth-tracker.github.io/

Abstract

Acquiring diverse whole-body dexterous manipulation data for humanoids remains a fundamental challenge in character animation and robotics. Existing pipelines rely on expensive mocap data collected from many subjects and require retargeting to a unified humanoid shape, which limits scalable data construction and often yields physically inconsistent interactions (e.g., unstable contacts). We present **ManipSynth**, a unified framework for whole-body loco-manipulation synthesis, with compositional generation of sparse whole-body motion, object-centric grasp anchors, and contact-aware interaction transitions. Then, we scale data generation through randomized pre-grasp motions, grasp poses, and object trajectories. ManipSynth achieves approximately 5× higher contact coverage and lower average penetration than OMOMO across representative objects. To test whether better contact-consistent data improve scalable control learning, we train **ManipTracker**, a general-purpose Humanoid-Object Interaction (HOI) tracking policy, on purely synthetic demonstrations. Under the same object set, number of training trajectories, the same ManipTracker trained on ManipSynth demonstrates stronger generalization than when trained on mocap data, while achieving over 90% training success.

1 Introduction

Humans perform loco-manipulation with the *whole body*, not the hands alone. Modeling these whole-body behaviors is fundamental to physics-based character animation and robotics, yet most existing manipulation pipelines are primarily built for robotic-arm grippers or dexterous hands Liu et al. [2024], Ye et al. [2023a], Liu et al. [2025a,b], Ye et al. [2023b], Zhang et al. [2025], Wang et al. [2025a], Physical Intelligence et al. [2025], Black et al. [2024], Zhang et al.. This motivates us modeling manipulation from a whole-body perspective.

Recent approaches Wang et al. [2023, 2025b], Yu et al. [2025], Xu et al. [2025], Tessler et al. [2025], Kim et al. [2026], Xu et al. [2026], Wen et al. [2025a], Lin et al. [2025], Wen et al. [2025b], Lou et al. [2025] have made significant progress in learning physically plausible human object interactions (HOI). These methods mainly fall into two paradigms: (1) whole-body prior-based control with task-specific rewards and (2) human demonstration-based learning with imitation rewards. The first paradigm can learn robust whole-body grasping skills Bae et al. [2023], Braun et al. [2024], Luo et al. [2024], but each new interaction task typically requires redesigning new rewards. The second

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

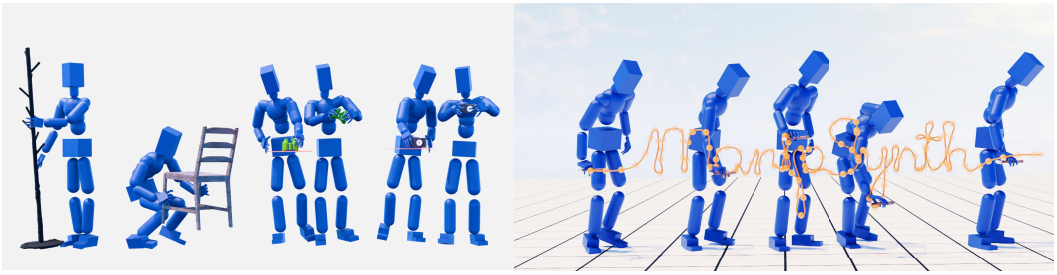


Figure 1: Our method synthesizes loco-manipulation behaviors between humanoids and diverse objects, including (**Left**) interactions with large and small objects, single-handed grasping, bimanual handovers, as well as (**Right**) long-horizon object transportation and orientation.

paradigm relies on human demonstrations Wang et al. [2023], Xu et al. [2025], Yu et al. [2025], Tessler et al. [2025], Kim et al. [2026], Wen et al. [2025b], either captured via motion capture systems or reconstructed from videos. However, motion capture typically requires specialized hardware and extensive manual effort Taheri et al. [2020], Li et al. [2023], Lu et al. [2025] to obtain high-quality data, which limits scalability across diverse object categories, interaction modes, and environments. Although video-reconstructed demonstrations Bhatnagar et al. [2022], Huang et al. [2022], Wen et al. [2025a,b] could scale more easily, they are often prone to noises, particularly in hand-object contact and object pose estimation.

Different from existing approaches, synthetic data generation provides a much more scalable alternative Tian et al. [2025], Chen et al. [2025c], Garrett et al. [2024], Wang et al. [2025a], Zhu et al. [2025], Deng et al. [2025], while enabling stronger quality through physically grounded generation constraints. Although existing synthetic data pipelines are numerous, they all target single-arm or dual-arm grippers Tian et al. [2025], Chen et al. [2025c], Garrett et al. [2024], Deng et al. [2025] or dexterous hands Wang et al. [2025a], Zhu et al. [2025], and solely consider table-top manipulation scenarios. Therefore, how to synthesize physically plausible whole-body loco-manipulation data remains largely unexplored.

In fact, while existing data synthesis methods have achieved significant progress in dexterous hand and robotic arm manipulation, extending such synthesis paradigms to whole-body dexterous interaction poses two challenges: First, the human whole body contains over 20 joints, making direct trajectory planning for each joint inefficient and impractical in such a high DOF space. Second, as shown in Table 2, tracking wrist-only trajectories yields substantially larger errors, therefore, sparse tracking alone is insufficient for precise hand-object coordination.

To address the above challenges, we propose ManipSynth, a unified framework for whole-body loco-manipulation data synthesis. Specifically, ManipSynth combines compositional contact-mode representation, object-centric grasp anchor synthesis via force-closure optimization Wang et al. [2022], and sparse whole-body trajectory synthesis from root-wrist controls using a partially conditioned controller Luo et al. [2023]. Together, these components enable us to instantiate continuous interactions by composing and stitching contact-aware key states, yielding temporally aligned humanoid, object, and contact trajectories. ManipSynth achieves $\mathbf{5}\times$ **higher** average contact coverage than OMOMO and lower average penetration, indicating broader interaction diversity and better contact quality. More importantly, these high-quality, high-coverage trajectories are generated in the same humanoid shape used by downstream policy training, which makes large-scale data expansion and policy learning more direct and stable. In contrast, methods such as MaskedManipulator and InterMimic Tessler et al. [2025], Xu et al. [2025] typically require explicit and implicit retargeting from multi-subject mocap data to one unified humanoid shape, along with complex hand-joint contact reward engineering, before stable large-scale HOI tracking. Our integrated Synth-to-Track pipeline avoids this mocap-shape retargeting and uses a basic human-object co-tracking reward without elaborate joint-level contact terms.

Under the same object set, number of training trajectories, the same ManipTracker trained on ManipSynth shows stronger generalization than when trained on mocap data. ManipTracker achieves over 90% tracking success rate on training data and maintains strong performance on the test split, demonstrating effective generalization from purely synthetic training data.

Table 1: Comparison with representative physics-based HOI methods. We compare scalability of data generation, reliance on demonstrations, retargeting requirement, and data sources.

Method	Year	Scalable Data	Demonstration-Based	Retarget-Free	Data Sources
PhysHOI Wang et al. [2023]	2023	✗	✓	✗	Mocap
PMP Bae et al. [2023]	2023	✗	✗	✗	No
Braun et al. Braun et al. [2024]	2024	✗	✗	✓	No
Li et al. Li et al. [2025]	2025	✗	✓	✗	Mocap
Omnigrasp Luo et al. [2024]	2024	✗	✗	✓	No
InterMimic Xu et al. [2025]	2025	✗	✓	✗	Mocap
SkillMimic-V2 Yu et al. [2025]	2025	✗	✓	✗	Mocap
MaskedManipulator Tessler et al. [2025]	2025	✗	✓	✗	Mocap
Wen et al. Wen et al. [2025b]	2025	✓	✓	✗	Video
DeVI Kim et al. [2026]	2026	✓	✓	✗	Video
ManipTracker (Ours)	2026	✓	✓	✓	Synthetic

In summary, our contributions are threefold.

1. We propose ManipSynth, which is a framework for whole-body loco-manipulation data synthesis under a unified formulation.
2. We establish a continually expandable 4D HOI dataset via ManipSynth. Its lower penetration and higher contact coverage make HOI tracking training more scalable.
3. Leveraging ManipSynth synthetic demonstrations, we develop ManipTracker, a general-purpose HOI tracking controller trained with task-agnostic imitation rewards.

2 Related Work

2.1 Demonstration Data for Loco-Manipulations

Demonstration data for loco-manipulations has evolved from controlled motion-capture datasets to large-scale and in-the-wild resources. Mocap-based HOI datasets, such as GRAB Taheri et al. [2020], OMOMO Li et al. [2023], and HUMOTO Lu et al. [2025], provide high-quality body-object trajectories and contact details, but they are costly to collect and difficult to deploy in portable, large-scale settings. Beyond mocap, reconstruction-based resources such as BEHAVE Bhatnagar et al. [2022], InterCap Huang et al. [2022], Open3DHOI Wen et al. [2025a] and Open4DHOI Wen et al. [2025b] improve scene diversity and real-world coverage, but still suffer from multi-view RGB or RGB-D estimation errors in hand-object contacts and object poses, as well as severe occlusion in human-object interaction regions. Another data-collection paradigm uses high-density dome capture systems, such as NeuralDome pipelines Zhang et al. [2022], to improve reconstruction fidelity; however, these setups require dense camera infrastructure and incur substantial hardware and operational cost. Taken together, these three acquisition paradigms face a common triad of bottlenecks: high data cost, reconstruction noise with occlusion, and heavy hardware dependence. Theoretically, purely synthetic data is substantially more effcient to produce and easier to scale up, yet principled rule-driven synthesis of physically grounded HOI trajectories remains largely underexplored.

2.2 Reinforcement Learning for Loco-Manipulations

As shown in Table 1, recent physics-based RL methods for HOI can be broadly grouped into goal-conditioned whole-body grasping and data-driven interaction imitation. Methods such as PMP Bae et al. [2023], physically plausible full-body HOI synthesis Braun et al. [2024], and Omnigrasp Luo et al. [2024] typically decouple strong whole-body motion priors from interaction optimization, and then solve target behaviors with explicitly designed task rewards. Although this design can generalize well in grasping-centric settings, it does not readily scale to broader and more compositional loco-manipulation skill sets, such as bimanual handover. In contrast, data-driven interaction imitation methods, including PhysHOI Wang et al. [2023], SkillMimic Wang et al. [2025b], SkillMimic-V2 Yu et al. [2025], InterMimic Xu et al. [2025], and DeVI Kim et al. [2026], learn controllers by imitating

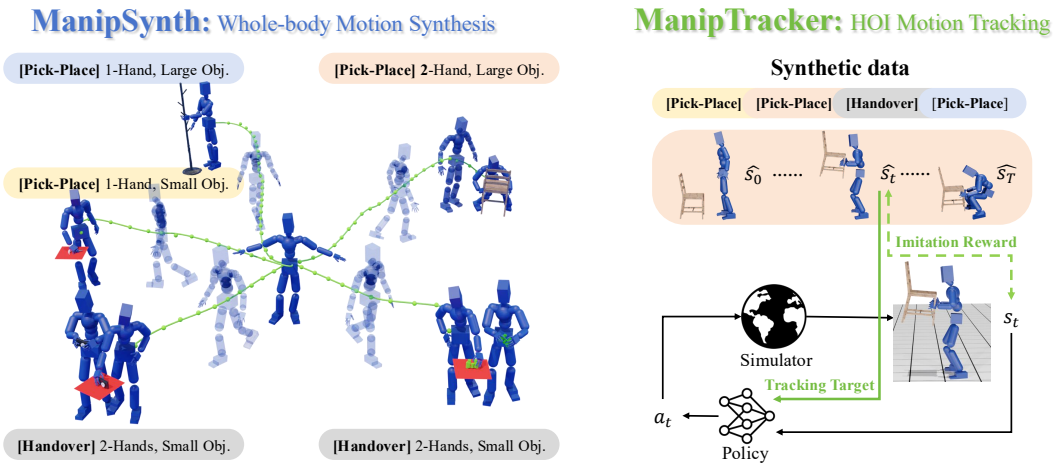


Figure 2: Overview of ManipSynth and ManipTracker for scalable whole-body loco-manipulation. Left: ManipSynth generates structured, contact-aware, and trackable synthetic HOI demonstrations. Right: ManipTracker learns a unified HOI tracking policy from these synthetic demonstrations in simulation.

demonstration trajectories and show strong performance on dynamic interactions, providing a low-level control foundation for higher-level goal planning Tessler et al. [2025], Xu et al. [2026]. However, these methods still largely depend on mocap-derived demonstrations collected from multiple human subjects with different body shapes. When training in a simulator with one unified humanoid shape, this setup typically requires retargeting, and the induced retargeting errors can degrade interaction fidelity and materially hinder stable scale-up toward broader object categories and skill families.

3 Method

ManipSynth and ManipTracker are designed as a data-to-control pipeline. ManipSynth generates structured humanoid-object interaction (HOI) demonstrations with explicit contact states, while ManipTracker learns to track these coupled human-object references under physics. Our method contains five components and follows the visual layout in Fig. 2. The left branch is ManipSynth, including (i) compositional skill representation, (ii) object-centric grasp anchor synthesis (Part II in Fig. 3), (iii) whole-body trajectory synthesis (Part I in Fig. 3), and (iv) contact-aware skill instantiation, which composes the previous outputs into structured demonstrations. The right branch is ManipTracker, which takes these demonstrations as references and learns a physics-based HOI tracking policy.

3.1 Problem Formulation

Let $\mathcal{H}$ denote a humanoid morphology, $\mathcal{O}$ an object geometry, $z \in \mathcal{Z}$ a skill type, and ξ task parameters (e.g., initial and target object pose, grasp side, transport path, and interaction duration). ManipSynth defines a generator $\mathcal{G} : (\mathcal{H}, \mathcal{O}, z, \xi) \mapsto \hat{\tau}$ that outputs a coupled HOI trajectory

$$\hat{\tau} = \{(\hat{\mathbf{x}}_t^h, \hat{\mathbf{x}}_t^o, \hat{\mathbf{c}}_t^h, \hat{\mathbf{c}}_t^o)\}_{t=1}^T. \tag{1}$$

Here $\hat{\mathbf{x}}_t^h$ denotes humanoid state (root pose, body joints, hand articulation), $\hat{\mathbf{x}}_t^o$ denotes object state, and $\hat{\mathbf{c}}_t^h, \hat{\mathbf{c}}_t^o$ denote human and object contact states.

3.2 Compositional Skill Representation

This component defines the compositional control interface in the left branch of Fig. 2, which organizes the synthesis process before Part I/Part II are instantiated. We represent each manipulation skill as a sequence of hand-object contact modes $m_t \in \{0, L, R, LR\}$, where 0 indicates no hand contact, L and R indicate single-hand contact, and LR indicates bimanual contact. This gives a

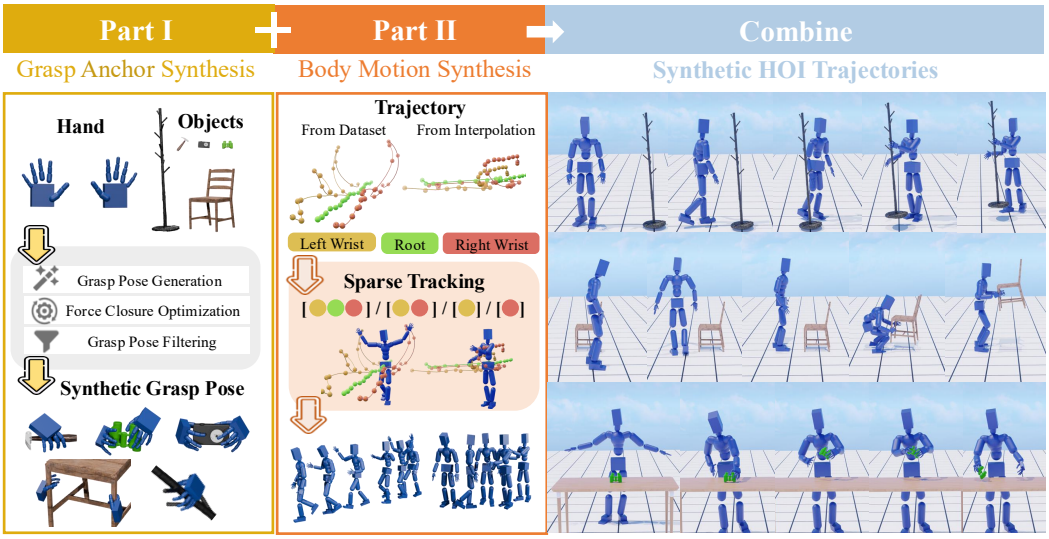


Figure 3: ManipSynth pipeline. Part I synthesizes object-centric grasp anchors and applies quality filtering. These reusable object-frame anchors convert object motion into sparse wrist targets that guide Part II to generate whole-body trajectories from root and wrist controls. The composed outputs are structured synthetic HOI demonstrations with aligned humanoid, object, and contact trajectories.

unified representation for all target skills: single-hand manipulation follows $0 \to h \to h \to 0$ with $h \in \{L, R\}$, bimanual manipulation follows $0 \to LR \to LR \to 0$, and bimanual handover follows $0 \to L \to LR \to R \to 0$ (or its symmetric right-to-left counterpart). This formulation converts whole-body loco-manipulation synthesis into composition over contact modes, grasp anchors, object trajectories, and sparse whole-body tracking, rather than skill-specific full-body trajectory design.

3.3 Object-Centric Grasp Anchor Synthesis

This subsection corresponds to Part II of Fig. 3, where object-centric grasp anchors are generated and filtered. An additional visual illustration of this stage is provided in the supplementary. For each object, we synthesize object-centric grasp anchors that define reusable contact states. For hand $h \in \{L, R\}$, a single-hand grasp anchor is $g_h = (W_h, \phi_h^g, c_h)$, where W_h is wrist pose in object frame, ϕ_h^g is the grasp hand articulation, and c_h is the contact pattern.

For single-hand candidates, we follow force-closure grasp optimization Wang et al. [2022], Chen et al. [2025a,b] and rank anchors by $S_{g_h} = Q_{\text{fc}}(g_h, \mathcal{O}) - \lambda_p E_p(g_h, \mathcal{O}) - \lambda_j E_j(g_h)$. We keep anchors with high S_{g_h} under penetration and joint-feasibility constraints (e.g., $E_p < \epsilon_p$). For small objects, we additionally apply simulation-based gravity-resistance validation. For large-object bimanual manipulation, we do not select left and right anchors independently. Instead, we construct bimanual anchor pairs $g_{LR} = (g_L, g_R)$ and rank them with a pair-level score inspired by recent bimanual grasp synthesis method Yang et al. [2026]:

$$S_{\text{bi}} = S_{g_L} + S_{g_R} + \lambda_o Q_o + \lambda_s Q_s - \lambda_c E_c - \lambda_i E_i - \lambda_r E_r. \tag{2}$$

Here Q_o and Q_s encourage complementary support and sufficient spatial separation, while E_c, E_i, and E_r denote center-of-mass support imbalance penalty, inter-hand collision penalty, and bimanual reachability penalty, respectively. We keep only pair candidates that satisfy penetration, collision, and reachability thresholds. We further filter anchors by geometric and physical validity. Specifically, we discard anchors with excessive penetration, infeasible joint states, or unstable contact.

During trajectory synthesis, we sample from valid single-hand anchors or valid bimanual pairs instead of optimizing a new grasp for every motion. During contact phases, given object trajectory T_t^o, each active wrist target is computed as $T_t^{w_h} = T_t^o W_h$, which preserves hand-object relative geometry across different paths and durations.

3.4 Whole-Body Trajectory Synthesis

This subsection corresponds to Part I of Fig. 3, where sparse root-wrist controls are converted into full-body references. Direct optimization in full humanoid joint space is inefficient for long-horizon HOI synthesis. We therefore synthesize whole-body motion from sparse task-relevant controls $u_t = \{T_t^{root}, T_t^{w_L}, T_t^{w_R}\}$. Depending on contact mode m_t, only a subset of controls is active (e.g., root-only during approach, one wrist for single-hand transport, both wrists for bimanual transport).

To keep sparse wrist trajectories collision-free before contact, we follow a pre-grasp planning principle similar to Chen et al. [2025b]. We do not plan approach motion directly to the contact anchor W_h. For each anchor, we first define a pre-grasp anchor W_h^{pre} by retreating a small offset along the approach direction, then use the pre-grasp target $T_t^{w_h,pre} = T_t^o W_h^{pre}$ during approach. The final short transition from W_h^{pre} to W_h is used only for contact establishment. This means the approach phase keeps positive clearance, and only the final short phase enters intended contact.

Given sparse targets, a partially conditioned tracking policy $\boldsymbol{\pi}^{pc}$ Luo et al. [2023], Wu et al. [2025] produces whole-body references:

$$\hat{\mathbf{x}}_t^h = \boldsymbol{\pi}^{pc}(\hat{\mathbf{x}}_{t-1}^h, u_{t:t+H}). \tag{3}$$

Hand articulation is phase-wise interpolated between open and grasp articulations by $\phi_t^h = (1 - \alpha_t)\phi_{\text{open}} + \alpha_t \phi_h^g$, where ϕ_t^h denotes finger poses (hand joint articulation) at frame t and α_t is phase-dependent. This yields a unified synthesis interface across locomotion, grasping, transport, release, and handover transitions.

3.5 Contact-Aware Skill Instantiation

This subsection composes Part I and Part II outputs in Fig. 3 into final structured HOI demonstrations. Given contact-mode sequence, grasp anchors, and sparse tracking, we instantiate each skill by constructing object trajectory, wrist targets, and hand articulation schedules.

Single-hand manipulation. Using mode sequence $0 \to h \to h \to 0$ with $h \in \{L, R\}$, the active wrist is coupled to object motion in transport by $T_t^{w_h} = T_t^o W_h$. Finger articulation follows $\phi_t^h = (1 - \alpha_t)\phi_{\text{open}} + \alpha_t \phi_h^g$ for grasp/release transitions.

Bimanual manipulation. Using mode sequence $0 \to LR \to LR \to 0$, two wrist targets are coupled to the same object trajectory, i.e., $T_t^{w_L} = T_t^o W_L$ and $T_t^{w_R} = T_t^o W_R$. To reduce small accumulated dual-wrist drift (consistent with the two-wrist tracking gap shown in Table 2), we apply arm-level IK correction in contact-critical phases:

$$\mathbf{q}_t^{arm,*} = \arg\min_{\mathbf{q}^{arm}} \sum_{h \in \mathcal{A}_t} \|\Delta(FK_h(\mathbf{q}^{arm}), T_t^{w_h})\|^2 + \lambda \|\mathbf{q}^{arm} - \tilde{\mathbf{q}}_t^{arm}\|^2, \tag{4}$$

where $FK_h(\mathbf{q}^{arm})$ maps arm joint configuration to the wrist pose of hand h, $\Delta(\cdot, \cdot)$ denotes wrist-pose residual (position and rotation), $\mathcal{A}_t$ is the active-hand set, and $\tilde{\mathbf{q}}_t^{arm}$ is the arm pose from sparse tracking.

Bimanual handover. We use the sequence $0 \to L \to LR \to R \to 0$: giver-hand transport, short bimanual overlap, then receiver-hand transport after giver release. Both giver and receiver targets are generated from object-centric anchors using $T_t^{w_h} = T_t^o W_h$, with mode-dependent activation.

3.6 ManipTracker with Synthetic Demonstrations

ManipSynth explicitly combines sparse whole-body motion generation with hand-object contact motion synthesis to produce full-joint HOI trajectories (Eq. 1). We factorize object-centric anchors, pre-grasp motions, and object trajectories so that a small set of interaction templates can be expanded into many training demonstrations through structured randomization, which makes the synthesis pipeline naturally scalable. Given synthetic demonstrations from ManipSynth, we train ManipTracker as a goal-conditioned HOI tracking policy in simulation. We control a humanoid with 51 actuated joints and model tracking as an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma)$, where policy $\pi_\theta(\mathbf{a}_t \mid \mathbf{o}_t, \mathbf{g}_t)$ predicts residual joint targets.

Given reference trajectory $\hat{\tau} = \{(\hat{\mathbf{x}}_t^h, \hat{\mathbf{x}}_t^o, \hat{\mathbf{c}}_t^h, \hat{\mathbf{c}}_t^o)\}_{t=1}^T$, policy sampling and residual command are

$$\mathbf{a}_t \sim \pi_\theta(\cdot \mid \mathbf{o}_t, \mathbf{g}_t) = \mathcal{N}\big(\boldsymbol{\mu}_\theta(\mathbf{o}_t, \mathbf{g}_t), \text{diag}(\boldsymbol{\sigma}_\theta^2)\big), \tag{5}$$

Table 2: Sparse whole body tracking results on AMASS train and test (10595 and 137 sequences).

	AMASS-Train					AMASS-Test				
Joints	Succ ↑	$E_{g\text{-}mpjpe}$ ↓	E_{mpjpe} ↓	E_{acc} ↓	E_{vel} ↓	Succ ↑	$E_{g\text{-}mpjpe}$ ↓	E_{mpjpe} ↓	E_{acc} ↓	E_{vel} ↓
Root	95.51%	59.88	0	6.95	6.14	63.50%	219.60	0	11.71	14.10
Left wrist	93.29%	66.36	92.43	13.27	18.03	53.28%	201.16	225.49	27.05	27.70
Right wrist	92.43%	66.67	101.50	12.66	16.64	51.82%	214.26	252.19	25.34	25.44
Left wrist + Root	97.74%	49.97	15.78	8.60	10.92	78.83%	179.07	42.96	17.52	17.69
Right wrist + Root	97.67%	49.77	16.24	8.22	10.04	78.83%	142.95	34.75	17.06	16.60
Two wrists	97.50%	51.52	48.18	11.44	15.01	67.15%	204.07	123.76	24.27	24.94
Two wrists + Root	98.09%	47.60	20.70	9.06	11.31	81.75%	163.60	49.61	18.28	18.47

and residual command is $\mathbf{q}_t^{\mathrm{cmd}} = \hat{\mathbf{q}}_t^{\mathrm{ref}} + \Delta\mathbf{q}_t$ with $\Delta\mathbf{q}_t = \mathbf{a}_t$, where $\mathbf{q}_t^{\mathrm{cmd}}$ is executed by a low-level PD controller.

The observation is decomposed as $\mathbf{o}_t = \left[\mathbf{o}_t^h, \mathbf{o}_t^o\right]$, where $\mathbf{o}_t^h$ contains local humanoid kinematics, contact forces, and human-reference residuals, and $\mathbf{o}_t^o$ contains local object pose and velocity, transformed object keypoints, and object-reference residuals. The goal is a short future window from the reference:

$$\mathbf{g}_t = \left[\hat{\mathbf{s}}_t, \hat{\mathbf{s}}_{t+1}, \ldots, \hat{\mathbf{s}}_{t+H}\right], \quad \hat{\mathbf{s}}_\tau = \left[\hat{\mathbf{x}}_\tau^h, \hat{\mathbf{x}}_\tau^o, \hat{\mathbf{c}}_\tau^h, \hat{\mathbf{c}}_\tau^o\right]. \tag{6}$$

All kinematic features are represented in the root-heading local frame for numerical stability.

We use a unified HOI imitation reward:

$$r_t = r_t^h \cdot r_t^o \cdot r_t^i \cdot r_t^b, \tag{7}$$

where r_t^h tracks humanoid motion, r_t^o tracks object motion, r_t^i preserves human-object relative geometry, and r_t^b is a binary object-interaction indicator. This task-agnostic reward reduces per-skill reward redesign and shifts the main scaling bottleneck toward the availability of diverse, contact-consistent, and trackable demonstrations. Compared with prior pipelines that rely on inferred interaction-marker promotion and penalty terms (e.g., InterMimic) or phased hand-guidance terms across approach, engagement, and release (e.g., MaskedManipulator), our formulation uses only a basic human-object co-tracking reward and still achieves a high tracking success rate.

4 Experiments

4.1 Experimental Setup

Implementation Settings. Our pipeline contains two learned components: the sparse whole-body controller $\boldsymbol{\pi}^{pc}$ used by ManipSynth for reference motion generation, and the downstream HOI tracking policy ManipTracker. For $\boldsymbol{\pi}^{pc}$, we build on PULSE Luo et al. [2023] by reusing released checkpoints and training only its third-stage for sparse whole body tracking. Both $\boldsymbol{\pi}^{pc}$ and ManipTracker are trained for 5000 epochs on a single RTX 4090 GPU, where ManipTracker uses synthetic loco-manipulation demonstrations generated by ManipSynth. Both ManipSynth and ManipTracker are implemented in Isaac Gym Makoviychuk et al. [2021] with PD joint control. Detailed simulation hyperparameters, contact parameters, reward weights, and initialization strategies are summarized in the appendix.

Metrics. To evaluate the performance of ManipTracker, we report Success Rate, Human Tracking Error, Object Tracking Error, and Tracking Length. Success Rate measures the fraction of reference sequences for which the policy succeeds at least once, averaged across all references. Human and Object Tracking Errors are reported as position and rotation errors between simulated and reference trajectories (averaged over the best-performing trial). Tracking Length is the duration (in seconds) that the imitation remains stable before triggering the interaction termination conditions defined in the appendix.

4.2 Sparse Whole-Body Tracking Results

We evaluate sparse whole-body tracking by training the third-stage PULSE controller Luo et al. [2023] with task-relevant joint trajectories as conditions. During training, conditioning joints are randomly sampled from subsets of root, left wrist, and right wrist, and each condition includes the

Table 3: Quantitative comparisons on data coverage and quality. Time denotes the average runtime of ManipSynth for generating one trajectory of approximately 400 frames.

	OMOMO				*ManipSynth* (Ours)				
Evaluation metrics	Contact Coverage ↑	Foot Sliding [mm] ↓	Penetration [mm] ↓	Jerk m/s^3 ↓	Contact Coverage ↑	Foot Sliding [mm] ↓	Penetration [mm] ↓	Jerk m/s^3 ↓	Time [s]
Woodchair	4.15%	0.674	**11.81**	**60.33**/30.03	**20.43%**	**0.141**	13.07	120.73/**17.20**	5.4
Clothesstand	3.94%	0.603	14.50	**56.91**/29.30	**16.01%**	**0.126**	**4.74**	123.40/**8.88**	5.2
Largetable	3.78%	0.799	11.31	**59.71/27.47**	**19.05%**	**0.134**	**6.49**	121.40/**27.47**	5.5
Average	3.96%	0.692	12.54	**58.98**/28.93	**18.50%**	**0.134**	**6.54**	121.84/**17.85**	5.37

Table 4: HOI tracking comparison using ManipTracker trained separately on ManipSynth or OMOMO. For cross-domain and all-test evaluation, Succ.* and Track Len.* are measured with randomized initialization, since direct cross-domain testing is substantially more difficult.

Train Set	Train-Set Quality		In-Domain Train (Fit)		In-Domain Test		Cross-Domain Test		All-Test ($\mathcal{D}_{all}$)	
	Pen. ↓	Jerk ↓	Succ. ↑	Track Len. ↑	Succ. ↑	Track Len. ↑	Succ.* ↑	Track Len.* ↑	Succ.* ↑	Track Len.* ↑
ManipSynth	**9.78**	121.06	**87.00%**	**12.43**	80.00%	11.23	**32.50%**	**5.68**	**56.25%**	**8.46**
OMOMO	11.56	**60.02**	85.00%	12.39	**85.00%**	**12.33**	22.50%	4.56	53.75%	8.45

current reference state plus a short look-ahead window. The sparse target trajectories are sourced conditionally: single-joint trajectories are sampled from the randomized generators in Rempe et al. [2023], while multi-joint trajectories are extracted from AMASS full-body motions Mahmood et al. [2019]. Table 2 compares different conditioning configurations. Root-plus-wrist conditioning works best, reaching over 95% success on AMASS-Train and around 80% on AMASS-Test, while single-joint or wrist-only conditioning is generally less reliable, likely because such sparse signals under-constrain many full-body motions. Qualitative sparse-tracking results and analysis are provided in the appendix.

4.3 Synthetic Data Results

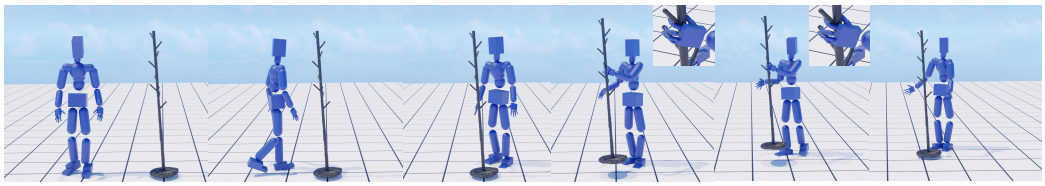
We evaluate ManipSynth on four common interaction modes. Qualitative results are shown in Fig. 4, and quantitative results are reported in Table 3 using contact coverage, penetration, foot sliding, and jerk (human/object). Contact coverage reflects the diversity of human-object contact patterns, penetration reflects contact quality, and foot sliding plus jerk reflect motion smoothness. Higher coverage is better, while lower values of the other metrics are better. Across objects, ManipSynth achieves about $5\times$ higher contact coverage and lower average penetration than OMOMO, with weaker foot sliding and jerk likely due to residual jitter from sparse whole-body reconstruction. Detailed metric computation is provided in the appendix.

4.4 Baseline Comparison

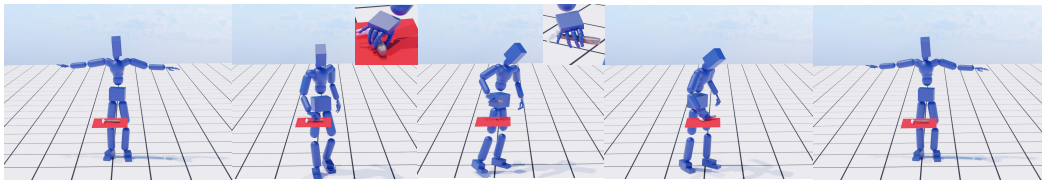
To compare against a real-data baseline, we use the filtered OMOMO subset released by InterMimic and select 100 trajectories from LargeTable and Woodchair, where interactions come from the same humanoid shape. In parallel, we use the same shape to synthesize 100 ManipSynth trajectories on LargeTable and Woodchair. We then train the same ManipTracker policy separately on ManipSynth and OMOMO, and evaluate on both in-domain and cross-domain test splits. Quantitative baseline metrics are reported in the corresponding comparison table, and more training curves are provided in the appendix. Overall, training on ManipSynth yields lower-penetration training data and stronger cross-domain and merged-test generalization, while OMOMO remains slightly better on the in-domain test split.

4.5 HOI Tracking Performance and Ablation Study

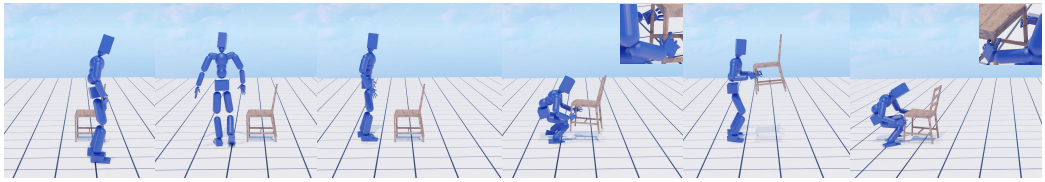
We evaluate ManipTracker on three representative objects using synthetic training splits generated by ManipSynth. Table 5 shows 96.83% training success rate, indicating that the synthesized training trajectories have reliable quality and are mostly trackable. Notably, test trajectories use different grasp poses, namely grasping different object regions, suggesting non-trivial transfer from purely synthetic training data. Table 6 shows that the unified HOI imitation design is important: removing the binary object-interaction term r^b leads to a clear performance drop, while full ManipTracker performs best overall. Adding the extra joint-contact reward r^c remains competitive and can help on



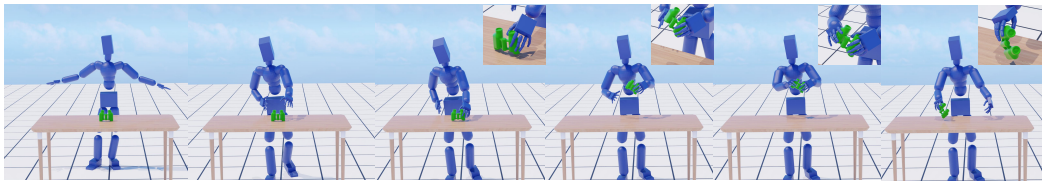
(a) Single-hand manipulation: picking up and relocating the clothesstand



(b) Single-hand manipulation: picking up and placing the hammer



(c) Bimanual manipulation: picking up and relocating the woodchair



(d) Bimanual manipulation: transferring the binocular between hands

Figure 4: Qualitative samples of synthetic loco-manipulation from ManipSynth.

Table 5: HOI Tracking Performance on synthetic data of three different objects.

	Train				Test			
Evaluation metrics	Sequence Success ↑	Object Tracking Error [mm] ↓	MPJPE [mm] ↓	Tracking Length [s] ↑	Sequence Success ↑	Object Tracking Error [mm] ↓	MPJPE [mm] ↓	Tracking Length [s] ↑
Woodchair	98.70%	30.0	54.1	13.2	62.50%	38.2	121.0	10.6
Clothesstand	95.40%	55.3	57.7	10.7	91.67%	83.9	68.1	10.5
Table	96.39%	65.7	54.5	13.1	80.00%	75.2	106.1	11.8
Average	96.83%	50.3	55.4	12.3	78.06%	65.8	98.4	11.0

Table 6: Ablation study on different reward components for HOI tracking on two different objects. w/ r^c denotes the variant that adds the human joint contact reward adopted from InterMimic while keeping the observation space and training mechanism unchanged.

Reward	*Woodchair*				*Clothesstand*			
	Sequence Success ↑	Object Tracking Error [mm] ↓	MPJPE [mm] ↓	Tracking Length [s] ↑	Sequence Success ↑	Object Tracking Error [mm] ↓	MPJPE [mm] ↓	Tracking Length [s] ↑
w/o r^b	64.94%	52.5	84.2	12.1	48.28%	92.5	86.6	9.4
w/o $r^{interact}$	32.47%	54.7	113.4	11.2	78.16%	95.1	116.5	10.2
w/o r^{object}	76.62%	41.2	87.7	12.1	71.26%	84.9	103.9	9.9
w/ r^c	96.43%	**29.6**	54.2	13.1	**95.70%**	56.2	**55.6**	**10.8**
ManipTracker	**98.70%**	30.0	**54.1**	**13.2**	95.40%	**55.3**	57.7	10.7

contact-sensitive cases, but its gains are not consistent across objects, suggesting that explicit joint-level contact supervision is not the main bottleneck once human-object co-tracking and interaction consistency are already enforced. Overall, these results show strong HOI tracking performance with lower-penetration synthetic data, without heavily engineered joint-level contact rewards.

5 Conclusion

We presented **ManipSynth**, a unified and scalable framework for synthesizing whole-body loco-manipulation data. By composing learning-based sparse whole-body tracking with contact-aware hand-object motion synthesis, ManipSynth generates diverse HOI trajectories with improved contact fidelity. Experiments show that ManipSynth achieves about $\mathbf{5}\times$ higher contact coverage and lower average penetration than OMOMO across representative objects and interaction types. Using these synthetic demonstrations, we trained **ManipTracker**, a general-purpose HOI tracking controller with task-agnostic imitation rewards, acheving over 90% tracking success on training data while maintaining effective generalization on held-out interactions.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Project Number 62595774), the MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (KLIP-HuMaCo), and the HPC Platform of ShanghaiTech University.

Competing Interests. The authors declare no competing interests.

References

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 *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 464–473. IEEE, 2024.

Jiayi Chen, Yubin Ke, Lin Peng, and He Wang. Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy. *arXiv preprint arXiv:2504.18829*, 2025a.

Jiayi Chen, Yubin Ke, and He Wang. Bodex: Scalable and efficient robotic dexterous grasp synthesis using bilevel optimization. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 01–08. IEEE, 2025b.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025c.

Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Wenhao Zhang, Heming Cui, Zhizheng Zhang, and He Wang. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. 2025. URL `https://arxiv.org/abs/2505.03233`.

Caelan Garrett, Ajay Mandlekar, Bowen Wen, and Dieter Fox. Skillmimicgen: Automated demonstration generation for efficient skill learning and deployment. In *8th Annual Conference on Robot Learning*, 2024.

Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *German Conference on Pattern Recognition*, 2022.

Hyeonwoo Kim, Jeonghwan Kim, Kyungwon Cho, and Hanbyul Joo. Devi: Physics-based dexterous human-object interaction via synthetic video imitation. *arXiv preprint arXiv:2604.20841*, 2026.

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

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27673–27682, 2025.

Yuhang Lin, Yijia Xie, Jiahong Xie, Yuehao Huang, Ruoyu Wang, Jiajun Lv, Yukai Ma, and Xingxing Zuo. Simgenhoi: Physically realistic whole-body humanoid-object interaction via generative modeling and reinforcement learning. *arXiv preprint arXiv:2508.14120*, 2025.

Xueyi Liu, Kangbo Lyu, Jieqiong Zhang, Tao Du, and Li Yi. Quasisim: Parameterized quasi-physical simulators for dexterous manipulations transfer. *arXiv preprint arXiv:2404.07988*, 2024.

Xueyi Liu, Jianibieke Adalibieke, Qianwei Han, Yuzhe Qin, and Li Yi. Dextrack: Towards generalizable neural tracking control for dexterous manipulation from human references. *arXiv preprint arXiv:2502.09614*, 2025a.

Xueyi Liu, He Wang, and Li Yi. Dexndm: Closing the reality gap for dexterous in-hand rotation via joint-wise neural dynamics model. *arXiv preprint arXiv:2510.08556*, 2025b.

Yuke Lou, Yiming Wang, Zhen Wu, Rui Zhao, Wenjia Wang, Mingyi Shi, and Taku Komura. Zero-shot human-object interaction synthesis with multimodal priors, 2025. URL `https://arxiv.org/abs/2503.20118`.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. Humoto: A 4d dataset of mocap human object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10886–10897, 2025.

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.

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, 2024.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. URL `https://amass.is.tue.mpg.de`.

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.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13756–13766, 2023.

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020.

Chen Tessler, Yifeng Jiang, Erwin Coumans, Zhengyi Luo, Xue Bin Peng, and Gal Chechik. Maskedmanipulator: Versatile whole-body control for loco-manipulation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025.

Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, et al. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. *arXiv preprint arXiv:2511.16651*, 2025.

Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. *arXiv preprint arXiv:2210.02697*, 2022.

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

Yinhuai Wang, Runyi Yu, Hok Wai Tsui, Xiaoyi Lin, Hui Zhang, Qihan Zhao, Ke Fan, Miao Li, Jie Song, Jingbo Wang, et al. Learning generalizable hand-object tracking from synthetic demonstrations. *arXiv preprint arXiv:2512.19583*, 2025a.

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*, pages 17540–17549, 2025b.

Boran Wen, Dingbang Huang, Zichen Zhang, Jiahong Zhou, Jianbin Deng, Jingyu Gong, Yulong Chen, Lizhuang Ma, and Yong-Lu Li. Reconstructing in-the-wild open-vocabulary human-object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17426–17436, 2025a.

Boran Wen, Ye Lu, Keyan Wan, Sirui Wang, Jiahong Zhou, Junxuan Liang, Xinpeng Liu, Bang Xiao, Dingbang Huang, Ruiyang Liu, et al. Efficient and scalable monocular human-object interaction motion reconstruction. *arXiv preprint arXiv:2512.00960*, 2025b.

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*, pages 13214–13224, 2025.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12266–12277, 2025.

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

Sizhe Yang, Yiman Xie, Zhixuan Liang, Yang Tian, Jia Zeng, Dahua Lin, and Jiangmiao Pang. Ultradexgrasp: Learning universal dexterous grasping for bimanual robots with synthetic data. *arXiv preprint arXiv:2603.05312*, 2026.

Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, 2023a.

Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023b.

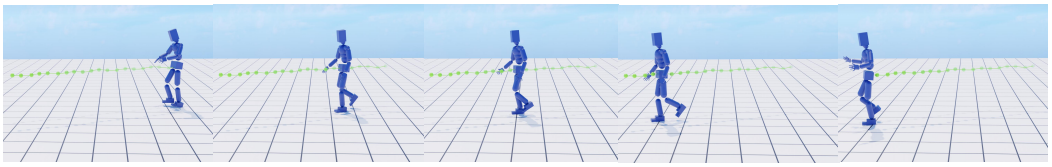
Runyi Yu, Yinhuai Wang, Qihan Zhao, Hok Wai Tsui, Jingbo Wang, Ping Tan, and Qifeng Chen. Skillmimic-v2: Learning robust and generalizable interaction skills from sparse and noisy demonstrations. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025.

Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. *arXiv preprint arXiv:2212.07626*, 2022.

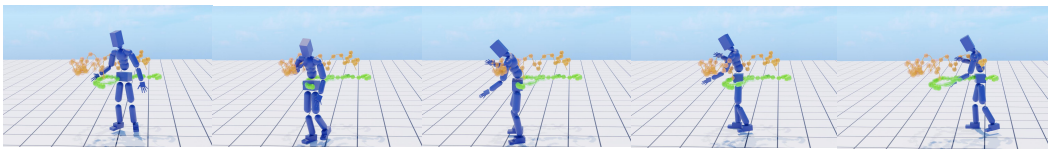
Zhenhao Zhang, Jiaxin Liu, Ye Shi, and Jingya Wang. Unihm: Unified dexterous hand manipulation with vision language model. In *The Fourteenth International Conference on Learning Representations*.

Zhenhao Zhang, Ye Shi, Lingxiao Yang, Suting Ni, Qi Ye, and Jingya Wang. Openhoi: Open-world hand-object interaction synthesis with multimodal large language model. *arXiv preprint arXiv:2505.18947*, 2025.

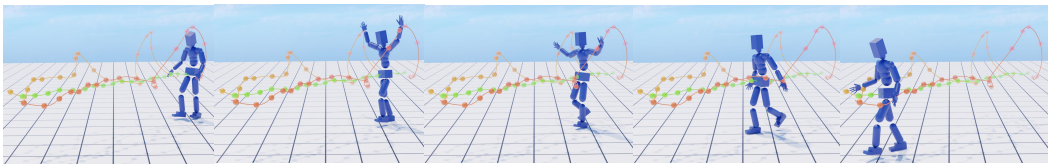
Kefei Zhu, Fengshuo Bai, YuanHao Xiang, Yishuai Cai, Xinglin Chen, Ruochong Li, Xingtao Wang, Hao Dong, Yaodong Yang, Xiaopeng Fan, et al. Dexflywheel: A scalable and self-improving data generation framework for dexterous manipulation. *arXiv preprint arXiv:2509.23829*, 2025.



(a) Root tracking



(b) Root and right wrist tracking



(c) Root and two wrists tracking

Figure 5: Qualitative results of sparse whole-body tracking with different conditioning joints.

Scaling Whole-Body Loco-Manipulation through Compositional Data Synthesis

Supplementary

A Simulation Setup

We use two closely related but not identical simulation configurations in our pipeline, since whole-body motion synthesis and downstream HOI tracking place different demands on physics fidelity and control frequency. For whole-body motion synthesis, the simulator mainly serves to generate scalable reference motions under stable locomotion and contact dynamics. For HOI tracking, the environment additionally needs to handle object interaction, contact-sensitive policy learning, and the training-time augmentation strategies described later in this supplement. Table 7 summarizes the core simulation hyperparameters used for whole-body motion synthesis, while Table 8 summarizes the simulation, contact-material, reward-weight, and initialization-disturbance settings used for HOI tracking training.

B More ManipSynth Details and Examples

ManipSynth is built by composing a whole-body motion synthesis stage with an object-centric grasp anchor synthesis stage. Figure 5 first illustrates the whole-body motion synthesis component: from sparse controls, the controller produces stable gait, coordinated upper-body motion, and smooth phase transitions, providing reusable full-body trajectories for downstream interaction composition.

Figure 6 then provides additional detail on object-centric grasp anchor synthesis, where reusable grasp candidates are constructed and filtered in the object frame. By combining these object-centric grasp anchors with the synthesized whole-body motion priors, ManipSynth forms compositional loco-manipulation demonstrations with aligned human motion, object motion, and contact structure. Figure 7 shows how this composition supports structured data augmentation on the same interaction template through variations in grasp pose, pre-grasp motion, and object transport path, while Fig. 8

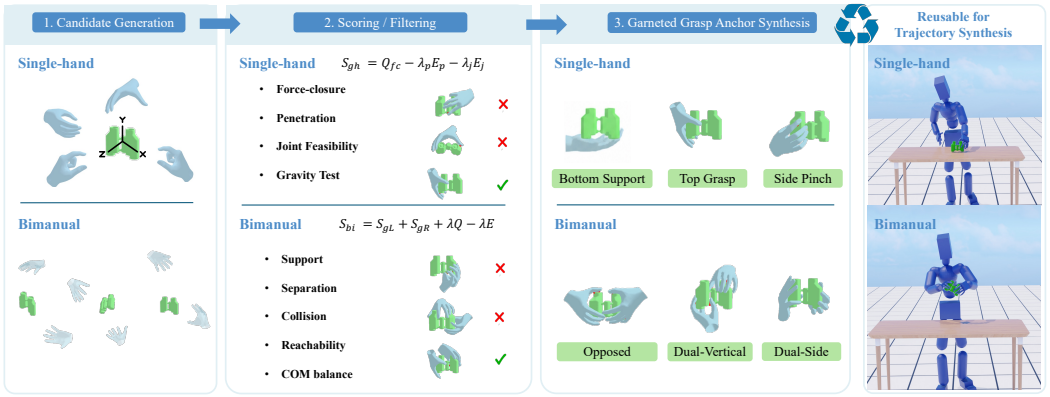


Figure 6: Illustration of object-centric grasp anchor synthesis. For each object, ManipSynth constructs reusable grasp candidates in the object frame and filters them by geometric feasibility and contact quality before downstream trajectory composition.

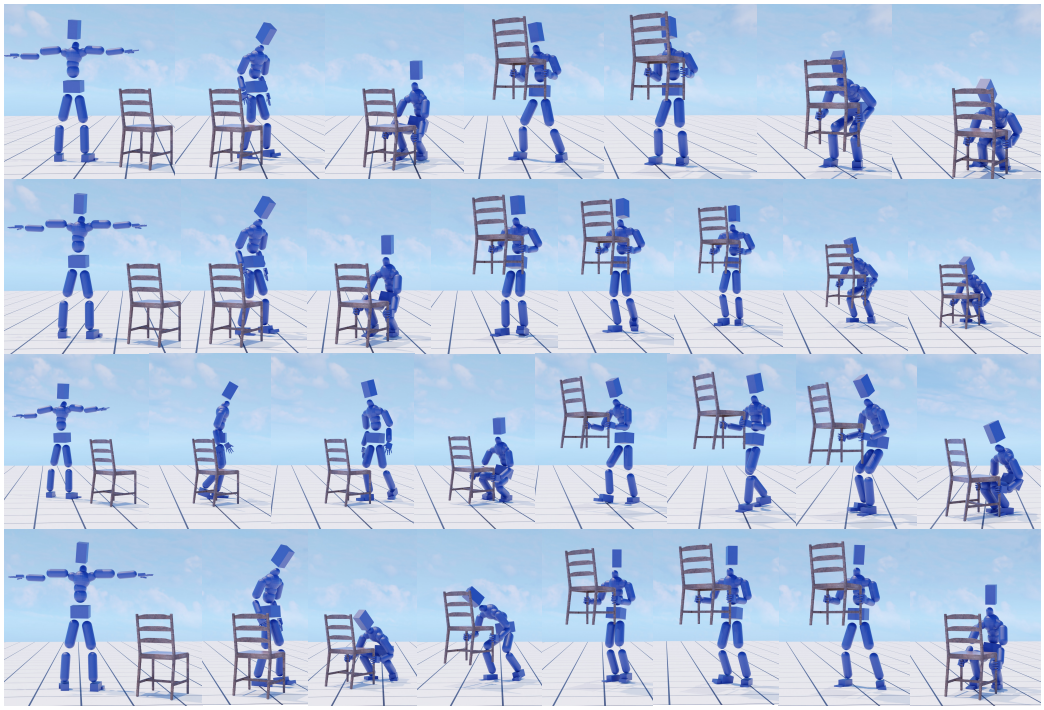


Figure 7: Additional ManipSynth examples on the `woodchair` object. Starting from the same interaction template, we augment the grasp poses, pre-grasp motions, and object transport trajectories to synthesize diverse yet structured HOI demonstrations.

highlights that the same framework can also extend beyond pick-and-place to other manipulation behaviors such as box pushing.

C HOI Tracking Training Details

Our HOI tracking training uses different stabilization strategies for large-object and small-object regimes. For large-object interactions, the main challenge is not only reference mismatch but also physical validity of the reset state: because the object is heavy and often requires sustained whole-body support or bi-manual coordination, small inconsistencies at initialization can quickly lead to penetration, unstable support contacts, or early rollout failure. We therefore adopt Physical State

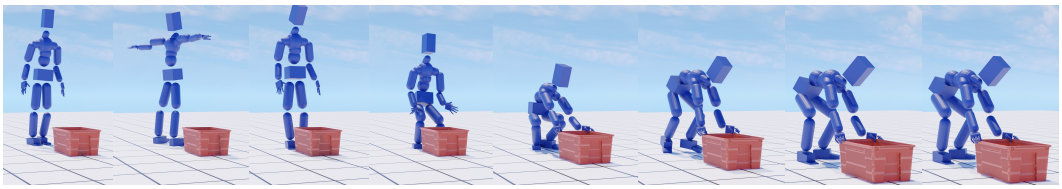


Figure 8: ManipSynth examples beyond pick-and-place. In addition to grasp-and-transport behaviors, the framework can also synthesize box-pushing interactions with coordinated whole-body motion.

Initialization (PSI), which improves training by reusing physically feasible states collected from successful rollouts rather than relying exclusively on direct reference-state resets.

For small-object interactions, the failure mode is different. These motions are typically more sensitive to local hand articulation, object pose offsets, and transient contact perturbations than to long-horizon support instability. Accordingly, instead of emphasizing physically validated reset buffers, we apply online initialization disturbance during training. Concretely, we perturb a subset of initial states with randomized object pose, object velocity, and hand-state noise, so that the policy repeatedly learns to recover from nearby off-reference states and return to the target HOI trajectory. The detailed formulations of PSI and initialization disturbance are provided below, and the simulator settings are collected in the tables at the end of this supplement.

Physical State Initialization for large-object HOI tracking. We adopt Physical State Initialization (PSI) as a training strategy mainly for large-object HOI tracking. Compared with small-object manipulation, large-object interactions are substantially more sensitive to initialization quality: the object has larger inertia, often requires sustained arm–torso coordination or bi-manual support, and is more likely to exhibit contact mismatch if the episode starts from a kinematically plausible but physically inconsistent reference frame. As a result, directly applying reference-state initialization to synthesized trajectories can lead to immediate penetration, unstable support contacts, or rapid divergence between the humanoid and the object before the policy has a chance to recover.

To mitigate this issue, we maintain a buffer of physically validated rollout states collected during training. Each buffer entry stores a simulator state together with its aligned reference index,

$$\mathcal{B}_{\mathrm{PSI}} = \{(\mathbf{s}_t, \kappa_t)\}, \tag{8}$$

where $\mathbf{s}_t$ contains the humanoid state, object state, and other simulator variables needed for reset, and κ_t records the associated demonstration sequence and frame index so that the future reference window remains temporally aligned after reset. During training, the initial state of an episode is sampled from a mixture of standard reference initialization and buffered physical states:

$$\mathbf{s}_0 \sim (1-\rho)\, p_{\mathrm{ref}} + \rho\, p_{\mathrm{psi}}, \tag{9}$$

where p_{ref} denotes sampling from the demonstration frames and p_{psi} denotes sampling from $\mathcal{B}_{\mathrm{PSI}}$. We apply PSI sparsely, namely only for a subset of resets, so that training still covers the original motion distribution while benefiting from physically feasible restart states.

The key update rule is that only high-quality rollout states are allowed to re-enter the initialization buffer. For a rollout fragment starting at frame t, we compute its discounted return

$$G_t = \sum_{k=t}^{T} \gamma^{k-t} r_k. \tag{10}$$

States with sufficiently high G_t and without severe early failure are treated as valid physical states and inserted into $\mathcal{B}_{\mathrm{PSI}}$, while old entries are discarded in a queue-like manner to keep the buffer compact and up to date. Intuitively, this mechanism bootstraps from successful partial interactions: once the policy discovers a physically coherent carrying or handover configuration for a large object, that state can later be reused as a training start point, making it much easier to practice subsequent phases such as transport, turning, and placement.

This strategy is particularly helpful for large objects because the most difficult portions of the motion often occur after contact has already been established. Without PSI, many episodes terminate early

near the grasp-acquisition phase, which greatly reduces supervision for later-stage behaviors. PSI shifts part of the training distribution toward feasible post-contact states, allowing the policy to repeatedly rehearse long-horizon whole-body object transport under realistic contact and balance constraints. In our setting, PSI therefore acts as a practical bridge between imperfect synthetic references and stable large-object HOI policy optimization.

Initialization disturbance for small-object HOI tracking. For small-object HOI tracking, we additionally use online data augmentation by perturbing the initial state at reset time. The goal is not to create arbitrary corruptions, but to expose the policy to nearby off-reference states from which it must recover and rejoin the original demonstration. This is especially useful for small-object skills because the object pose, hand articulation, and contact condition can change rapidly over short time horizons, making the controller sensitive to small initialization errors.

We use a Bernoulli-gated noise operator

$$\epsilon(\sigma, p) = z \odot u, \qquad z \sim \text{Bernoulli}(p), \quad u \sim \mathcal{U}([-\sigma, \sigma]), \tag{11}$$

where the binary variable z is sampled per environment instance and determines whether that reset is perturbed. Thus, only a subset of training episodes is disturbed, while the others remain on the original reference distribution. At the object level, we perturb the initial linear and angular velocities as

$$\tilde{\mathbf{v}}_0^o = \mathbf{v}_0^o + \epsilon(0.02, p), \qquad \tilde{\boldsymbol{\omega}}_0^o = \boldsymbol{\omega}_0^o + \epsilon(0.02, p). \tag{12}$$

We then apply contact-conditioned pose perturbations. When the initial frame is non-contact, we use a larger translational disturbance and allow full 3-axis rotational perturbation:

$$\tilde{\mathbf{p}}_0^o = \mathbf{p}_0^o + \epsilon(0.05, p), \qquad \tilde{\mathbf{q}}_0^o = \mathcal{R}_{xyz}(\epsilon(\pi/5, p)) \otimes \mathbf{q}_0^o, \tag{13}$$

with the additional constraint that the vertical position disturbance is upward-only. When the object is already in contact, we use a smaller positional perturbation and restrict orientation noise to yaw:

$$\tilde{\mathbf{p}}_0^o = \mathbf{p}_0^o + \epsilon(0.02, p), \qquad \tilde{\mathbf{q}}_0^o = \mathcal{R}_z(\epsilon(\pi/8, p)) \otimes \mathbf{q}_0^o. \tag{14}$$

This asymmetric design reflects the fact that disturbing a contacted small object too aggressively can immediately destroy the grasp geometry, whereas non-contact frames permit broader perturbation.

On the human side, we perturb the hand articulation states while leaving the rest of the body initialization unchanged. In practice, we add random offsets to the left- and right-hand DoF positions and velocities:

$$\tilde{\mathbf{q}}_0^{h,\text{hand}} = \mathbf{q}_0^{h,\text{hand}} + \epsilon(\pi/8, p), \qquad \tilde{\dot{\mathbf{q}}}_0^{h,\text{hand}} = \dot{\mathbf{q}}_0^{h,\text{hand}} + \epsilon(0.1, p). \tag{15}$$

These disturbances encourage the policy to correct small finger configuration errors and object-state deviations at the beginning of an episode. Overall, this reset-time augmentation substantially enriches the effective training distribution for small-object manipulation and improves robustness to imperfect grasps, pose offsets, and transient contact inconsistencies.

D HOI tracking reward detail

Following Eq. (7) in the main paper, we use a unified imitation reward for human-object interaction (HOI) tracking:

$$r_t = r_t^h \cdot r_t^o \cdot r_t^i \cdot r_t^b. \tag{16}$$

The reward is evaluated after each simulator transition using the current simulated state and the aligned reference frame. For brevity, we omit the one-step index shift and write all quantities at time t. Let $\mathcal{B}$ denote the set of tracked humanoid rigid bodies, $\mathcal{K}$ the set of canonical object keypoints, and $\mathcal{C}$ the set of human links that are allowed to interact with the object. We use $(\mathbf{p}_t^b, \mathbf{R}_t^b)$ and $(\hat{\mathbf{p}}_t^b, \hat{\mathbf{R}}_t^b)$ for the simulated and reference pose of body $b \in \mathcal{B}$, and $(\mathbf{p}_t^o, \mathbf{R}_t^o)$ and $(\hat{\mathbf{p}}_t^o, \hat{\mathbf{R}}_t^o)$ for the simulated and reference object pose. All imitation sub-rewards share the same exponential form

$$r_t^\alpha = \exp(-\lambda_\alpha e_t^\alpha), \tag{17}$$

where e_t^α is the corresponding tracking error and $\lambda_\alpha > 0$ is a fixed weight.

Human tracking reward. The human term encourages the humanoid to match the reference whole-body motion in both global body positions and orientations:

$$r_t^h = r_t^{hp} \cdot r_t^{hr}, \tag{18}$$

$$e_t^{hp} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \left\| \mathbf{p}_t^b - \hat{\mathbf{p}}_t^b \right\|_2^2, \tag{19}$$

$$e_t^{hr} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \left\| \log\left(\mathbf{R}_t^b \hat{\mathbf{R}}_t^{b\top} \right) \right\|_2^2. \tag{20}$$

Here $\log(\cdot)$ denotes the rotation logarithm map, so e_t^{hr} measures geodesic orientation mismatch on $\mathrm{SO}(3)$. Compared with hand-centric trackers, this term supervises the entire humanoid kinematic chain and is therefore crucial for preserving locomotion rhythm, balance, torso coordination, and arm posture simultaneously.

Object tracking reward. The object term aligns the rigid object motion with the demonstration:

$$r_t^o = r_t^{op} \cdot r_t^{or}, \tag{21}$$

$$e_t^{op} = \| \mathbf{p}_t^o - \hat{\mathbf{p}}_t^o \|_2^2, \tag{22}$$

$$e_t^{or} = \left\| \log\left(\mathbf{R}_t^o \hat{\mathbf{R}}_t^{o\top} \right) \right\|_2^2. \tag{23}$$

This term ensures that the tracked policy follows the demonstrated object transport, placement, and reorientation behavior instead of only matching the human motion.

Interaction reward. Accurate HOI tracking additionally requires preserving the relative geometry between the human and the object. To this end, we define a set of interaction pairs $\mathcal{P} \subseteq \mathcal{B} \times \mathcal{K}$ between interaction-relevant human body parts and object keypoints. For an object keypoint $\bar{\mathbf{k}}^m$ in the object local frame, its world-space position is

$$\mathbf{k}_t^m = \mathbf{R}_t^o \bar{\mathbf{k}}^m + \mathbf{p}_t^o. \tag{24}$$

We then measure the consistency of human-object relative vectors:

$$\mathbf{d}_t^{b,m} = \mathbf{k}_t^m - \mathbf{p}_t^b, \quad \hat{\mathbf{d}}_t^{b,m} = \hat{\mathbf{k}}_t^m - \hat{\mathbf{p}}_t^b, \tag{25}$$

$$e_t^i = \frac{1}{|\mathcal{P}|} \sum_{(b,m) \in \mathcal{P}} \left\| \mathbf{d}_t^{b,m} - \hat{\mathbf{d}}_t^{b,m} \right\|_2^2, \quad r_t^i = \exp(-\lambda_i e_t^i). \tag{26}$$

This is the whole-body HOI analogue of relative hand-object tracking used in dexterous manipulation. It stabilizes the demonstrated coupling between the object and the relevant human links, especially during grasp acquisition, carrying, handover, and release, where small absolute errors can accumulate into large interaction failures.

Binary object-interaction indicator. In our final implementation, this term only supervises whether the object is in contact, rather than matching per-link human contact patterns. Let $c_t^o \in \{0, 1\}$ denote the simulated object-contact indicator extracted from the current observation, and let $\hat{\mathbf{c}}_t^o$ denote the reference object-contact labels. Following the code, we first collapse the reference contact vector into a single binary target

$$\hat{c}_t^o = \mathbb{I}[\max \hat{\mathbf{c}}_t^o > 0.1]. \tag{27}$$

The resulting indicator reward compares the simulated object-contact state against this binary reference target:

$$e_t^b = |c_t^o - \hat{c}_t^o|, \quad r_t^b = \exp(-\lambda_b e_t^b), \tag{28}$$

where λ_b corresponds to the configured object-contact weight (`cg_obj`). This binary object-interaction indicator complements the interaction-geometry reward by explicitly distinguishing contact and non-contact phases at the object level, which is sufficient in our setting to stabilize grasp acquisition, object transport, and release without imposing additional human-link contact supervision.

E Limitations and Future Work

Although, for the same object instances, our synthesized data achieve substantially higher contact coverage and lower penetration than OMOMO, the current framework still has several important limitations. First, the interaction patterns that can be generated reliably remain relatively narrow. Our current compositional pipeline is effective for structured modes such as pick-and-place and handover, but it is less straightforward to synthesize whole-body motion for interactions whose body-object coupling is dominated by sustained external force and scene-dependent contact transitions, such as pushing or dragging. **A promising future direction** is to incorporate single-frame human-object interaction snapshots from mocap data as anchors, and then perform data augmentation under approximately preserved human-object relative configurations, so as to seed more diverse interaction families without requiring full motion capture sequences. Second, the present model is tailored to rigid objects. Consequently, it does not yet support manipulation involving deformable materials or fluids, such as folding garments, where the object state is no longer well captured by a single rigid-body pose and contact evolution becomes substantially more complex. Addressing this limitation will likely depend on advances in simulation backends that can represent deformable and fluid interactions with sufficient fidelity and efficiency for large-scale policy learning.

Table 8: Simulation, contact-material, reward-weight, and initialization-disturbance settings used for HOI tracking training.

Setting	Choice
Physics integration step	$1/120$ s
Policy action interval	$1/30$ s
Physics solver	TGS
Position solver iterations	4
Velocity solver iterations	1
Continuous collision detection	Enabled
Humanoid contact offset	0.02
Humanoid rest offset	0.0
Humanoid maximum depenetration velocity	100.0
Object density	200
Object static / dynamic friction	0.6 / 0.6
Object restitution	0.05
Ground static / dynamic friction	0.9 / 0.9
Ground restitution	0.1
Object collision approximation	Convex decomposition
Maximum convex hulls per object	64
Reward Weights	
Human position reward weight (`p`)	30.0
Human rotation reward weight (`r`)	2.5
Human position-velocity reward weight (`pv`)	0.0
Human angular-velocity reward weight (`rv`)	0.0
Object position reward weight (`op`)	5.0
Object rotation reward weight (`or`)	0.1
Object position-velocity reward weight (`opv`)	0.1
Object angular-velocity reward weight (`orv`)	0.0
Interaction-geometry reward weight (`ig`)	5.0
Object-contact reward weight (`cg_obj`)	5.0
Human motion energy weight (`eg1`)	2.0×10^{-5}
Initialization Disturbance (Small-Object HOI)	
Initialization disturbance probability (`state_noise_prob`)	0.2
Object linear / angular velocity noise scale	0.02 / 0.02
Non-contact object position noise scale	0.05
Non-contact object rotation range	$\pi/5$
Contact object position noise scale	0.02
Contact object yaw-noise range	$\pi/8$
Hand DoF position noise scale	$\pi/8$
Hand DoF velocity noise scale	0.1

Table 7: Core simulator settings used for whole-body motion synthesis.

Setting	Choice
Physics integration step	$1/60$ s
Control update interval	$1/30$ s
Physics substeps per control interval	2
Physics solver	TGS
Position solver iterations	4
Velocity solver iterations	0
Contact offset	0.02
Rest offset	0.0
Bounce threshold velocity	0.2
Maximum depenetration velocity	10.0
Gravity direction and magnitude	$(0, 0, -9.81)$
Ground static / dynamic friction	1.0 / 1.0
Ground restitution	0.0

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize the paper's main claims on scalable synthetic HOI data generation and downstream tracking, and these claims are supported by the quantitative and qualitative results in Secs. 3–4 and the appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes an explicit "Limitations and Future Work" section in the appendix discussing the current interaction scope, object assumptions, and future extensions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper is an empirical systems and synthesis paper and does not present formal theorems or proof-based theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the synthesis pipeline, reward formulation, training setup, evaluation protocol, and simulation settings in Secs. 3–4 and the appendix, which are the key ingredients underlying the reported claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The current submission does not provide an anonymized public release of code or data. Reproducibility is supported through the methodological and experimental details in the paper and appendix.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: The paper provides the main training setup, data splits, reward terms, and simulation hyperparameters, but it does not yet document all implementation details such as the full optimizer configuration and all tuning choices.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports single-run quantitative results and does not include error bars, confidence intervals, or formal statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper states that the learned components were trained on a single RTX 4090 GPU and reports synthesis runtime for one trajectory, but it does not provide a full accounting of memory usage or total project compute.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, the research and reporting in this submission conform to the NeurIPS Code of Ethics and do not involve undisclosed deviations from it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

Justification: The current paper does not include a dedicated discussion of broader societal impacts. The work is primarily presented as a technical contribution on synthetic humanoid manipulation data and tracking.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The submission does not release high-risk generative models or scraped datasets whose deployment would require dedicated misuse safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: Existing datasets and prior methods are cited in the paper, but the submission does not yet explicitly list the licenses and terms of use for all external assets it builds on.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The paper introduces new synthetic data and trained policies, but the current submission does not provide standalone release documentation for these assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or newly conducted research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not report newly conducted experiments with human subjects that would require IRB review or participant risk disclosure.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not an important or non-standard component of the core methodology described in this submission.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.