Stability-Driven Motion Generation for Object-Guided Human-Human Co-Manipulation

Jiahao Xu[1] Xiaohan Yuan[2] Xingchen Wu[1] Chongyang Xu[3] Kun Li[1] Buzhen Huang[1*]

[1]Tianjin University [2]National University of Singapore [3]Sichuan University

Figure 1. Given an object mesh and its trajectory (green), our method generates coordinated motions that are consistent with the trajectory while remaining natural and physically plausible for co-manipulation.

Abstract

Co-manipulation requires multiple humans to synchronize their motions with a shared object while ensuring reasonable interactions, maintaining natural poses, and preserving stable states. However, most existing motion generation approaches are designed for single-character scenarios or fail to account for payload-induced dynamics. In this work, we propose a flow-matching framework that ensures the generated co-manipulation motions align with the intended goals while maintaining naturalness and effectiveness. Specifically, we first introduce a generative model that derives explicit manipulation strategies from the object's affordance and spatial configuration, which guide the motion flow toward successful manipulation. To improve motion quality, we then design an adversarial interaction prior that promotes natural individual poses and realistic interperson interactions during co-manipulation. In addition, we also incorporate a stability-driven simulation into the flow matching process, which refines unstable interaction states through sampling-based optimization and directly adjusts the vector field regression to promote more effective manipulation. The experimental results demonstrate that our method achieves higher contact accuracy, lower penetration, and better distributional fidelity compared to state-of-the-art human-object interaction baselines. The code is available at https://github.com/boycehbz/StaCOM.

1. Introduction

Modeling coordinated human motion in interactive environments is a key problem in computer graphics, virtual reality, and robotics. The challenge intensifies in scenarios where multiple agents interact with shared objects (*e.g.*, two-person lifting or collaborative manipulation), since each agent must adapt to both the object's motion and the partner's behavior. Such human-human co-manipulation involves tightly coupled triadic interactions,

[1]Corresponding author: Buzhen Huang, hbz@tju.edu.cn.

which require synchronized motion between humans and objects and physical feasibility during joint manipulation.

However, most existing motion generation methods are tailored to single-person scenarios, where an individual moves in response to static environments or predefined object trajectories [20, 42]. These models lack the mechanisms for inter-agent communication and mutual adaptation, making them unsuitable for collaborative tasks in co-manipulation. On the other hand, recent multi-person motion generation methods [22, 24] typically focus on social or dance-like interactions without involving object manipulation. Directly adapting these frameworks to co-manipulation cannot ensure reasonable physical dynamics and coherent human–object coordination. Consequently, developing motion generation models that can capture both inter-agent coordination and realistic human–object dynamics for multi-person co-manipulation remains an open research question.

We observe that real-world co-manipulation motions need to satisfy three key aspects: **1) Intention**: the manipulation strategy should be determined based on the object's shape, affordance, and goal state to achieve the desired manipulation. **2) Naturalness**: human motion should be natural and responsive to the partner's actions. **3) Effectiveness**: the transport process should be stable and comply with physical laws. Based on these observations, we propose a flow-matching framework that integrates object affordances, motion priors, and physics-based feedback into the motion generation process. Guided by these principles, our model generates co-manipulation motions that better satisfy intention, naturalness, and physical plausibility.

Specifically, we adopt flow-matching [25] as our basic framework, which learns a continuous vector field to map noise into clean motion data under object trajectory guidance. We further employ BPS representation [38] to enhance the model's perception of dynamic object pose and shape. Building upon this, we introduce an affordance-informed manipulation strategy to guide motion generation during inference. Since humans can produce the same object motion through diverse movements, this strategy generates graspability fields conditioned on object affordances, and the resulting contact anchors serve as explicit gradient cues that attract hands toward high-probability regions on the object surface, yielding geometrically consistent yet diverse manipulation plans. Nonetheless, relying solely on the manipulation strategy often results in stiff or poorly synchronized poses. We therefore develop an adversarial interaction prior that scores dual-human motion naturalness by analyzing joint rotations, inter-agent timing, and role symmetry. In contrast to previous discriminators [6, 15], this prior focuses on collaborative cues and penalizes misaligned reactions, so the flow receives gradients that preserve coordinated behaviors. In addition, we incorporate a stability-driven simulation to inject physics-based feedback, where a sampling-based formulation refines unstable poses and steers the generator toward motions that maintain grasp stability and limit payload drift. The main contributions of this work are summarized as follows:

- We propose a flow-matching framework to consider intention, naturalness, and effectiveness principles to generate physically plausible and socially coordinated human co-manipulations.
- We introduce an affordance-informed manipulation strategy and an interaction prior that ensure natural interactions while producing plausible manipulations.
- We consider co-manipulation motion with a focus on stability and propose a sampling-based simulation to jointly refine the interaction and motion.

2. Related Work

Human-Object Interaction Generation. Human-object interaction (HOI) generation has evolved from contact-aware diffusion to affordance-driven reasoning, yet most methods still emphasize single-actor scenes [21, 28, 37, 45, 47, 52]. InterDiff [51] and CG-HOI [3] inject contact objectives into diffusion pipelines, whereas OMOMO [20], NIFTY [16], and NAP [17] reconstruct interactions from object trajectories or articulated priors to provide semantic control. With the availability of advanced human–object–human datasets [29, 50], recent works have begun to explore collaborative manipulation. OnlineHOI [14] introduces a Mamba-based network for generating human motions conditioned on object geometry. SyncDiff [9] promotes phase-aligned multi-body motion, while COLLAGE [2] samples cooperative contacts for dual agents. However, these methods rely solely on stochastic denoising without incorporating explicit physical feedback. CooHOI [4] achieves physically plausible cooperative human–object interactions through reinforcement learning, but its policy fails to generalize across different objects and tasks. In contrast to prior works, our approach advances human co-manipulation motion generation by integrating deterministic flow matching with affordance-guided gradients and adversarial coordination priors. A generalizable sampling-based simulation module is further introduced to improve manipulation stability.

Multi-person Interaction Generation. Motion generation for multi-person interaction focuses on modeling dynamic interactions between individuals [1, 13, 32, 53, 59]. However, existing methods have long inference times that are not applicable to real-time generation of scenarios. To improve the model inference speed, Li *et al*. [18] introduced a factorization strategy, which decomposes the joint probability of interaction motion into three independent distri-

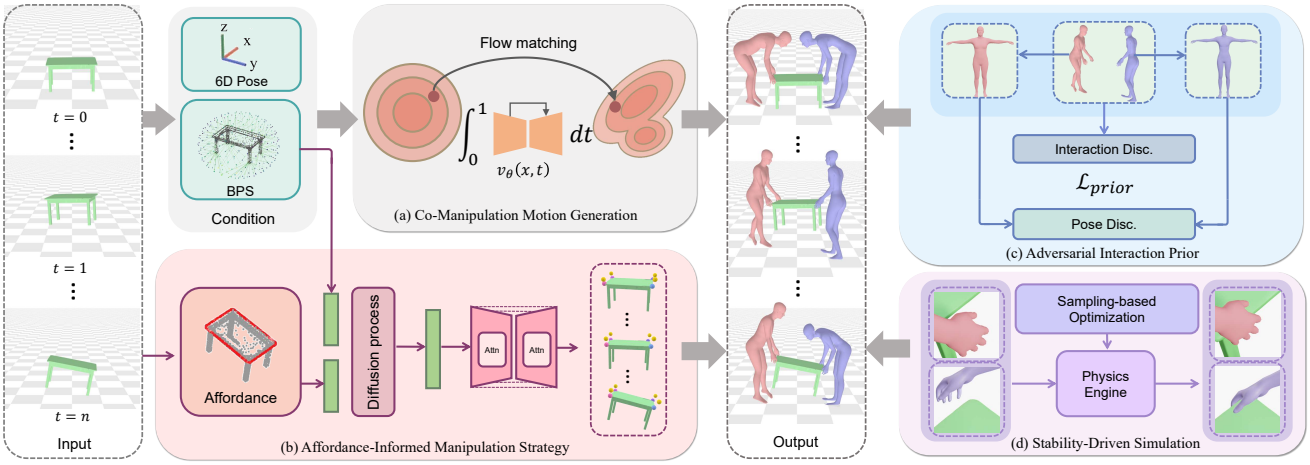

Figure 2. **Overview.** Given an input object trajectory, our method generates co-manipulation motions conditioned on object 6D poses and their BPS features (a). To ensure that the motions are consistent with the object trajectory, an affordance-informed manipulation strategy (b) is introduced to produce explicit contact signals as flow guidance. Building on this design, we further propose an adversarial interaction prior (c) and a stability-driven simulation (d) to enhance motion quality. Note that the contact strategy (b), flow matching (a), and interaction prior (c) are trained separately in advance, while all components are executed jointly at inference time.

butions and effectively captures the geometric and topological relationships in the interaction between two people. Some works generate actions conditional on textual descriptions [46, 49], combining boosted VAE models and a diffusion-based framework to capture spatio-temporal dependencies to enhance the modeling of two-player interactions [12, 19, 24, 27, 48]. Other approaches integrate physical constraints such as contact optimization [22] and collision detection [11, 44] to alleviate human penetration and improve realism. While these methods can generate higher-quality two-person interaction actions, they do not scale to more complex multi-person or even interaction actions with objects. Directly applying generic multi-person motion generation frameworks to co-manipulation scenarios may lead to severe artifacts.

Physics-based Motion Generation. Physics-grounded motion synthesis blends data-driven priors with simulators or analytic constraints to ensure physical plausibility [43]. Optimization-based controllers such as [10, 26] integrate tracking objectives with physics solvers to reconstruct human motions under contact and torque limits. PhysDiff [56] augments diffusion with physics-constrained gradient guidance. Reinforcement-learned character controllers such as DeepMimic [35] and AMP [36] further couple motion capture priors with physics-based policies to maintain balance, enforce contact stability, and suppress penetration. The technology is improved in recent years with more advanced designs like universal representation [30, 31], diverse reward functions [8, 55], and motion priors [33, 40, 41, 57]. Despite these advances, existing methods primarily address single-agent locomotion or isolated human-object interactions. In contrast, we operate on dual-human co-manipulation, combining gradient-based contact objectives with simulator-in-the-loop selection to handle tightly coupled human-object dynamics without additional policy training.

3. Method

We aim to generate coordinated human–human co-manipulation motions conditioned on an object's geometry and trajectory. An overview is illustrated in Fig. 2. Our framework first employs a flow-matching model to generate coordinated human motions conditioned on the object's configuration (Sec. 3.2). Built upon this model, we introduce a diffusion-based module that generates contact strategies guided by object affordances, providing explicit cues to refine the motion flow (Sec. 3.3). We further incorporate an adversarial regularizer (Sec. 3.4) and a stability-driven simulation module (Sec. 3.5) to enhance the realism and physical stability of the generated motions. The details of each component are described in the following sections.

3.1. Interaction representation

We represent each individual with SMPL-X [34] tuple $\mathbf{x}_t^{(a)} = (\boldsymbol{\theta}_t^{(a)}, \boldsymbol{\beta}^{(a)}, \gamma_t^{(a)})$ at frame t, where $\boldsymbol{\theta}^{(a)} \in \mathbb{R}^{J\times 6}$ denotes joint rotations in 6D representation [58]. $\gamma^{(a)} \in \mathbb{R}^3$ and $\boldsymbol{\beta}^{(a)} \in \mathbb{R}^{10}$ are global translation and shape coefficients for agent $a \in \{1, 2\}$. Concatenating the two agents yields the interaction sequence $\mathbf{x} = \{(\mathbf{x}_t^{(1)}, \mathbf{x}_t^{(2)})\}_{t=0}^{T}$. The object is described by its mesh $\mathcal{O}$ and rigid motion trajectory $\{(R_t^o, \mathbf{d}_t^o)\}_{t=0}^{T}$. We also employ a Basis Point Set (BPS) descriptor $\mathbf{b}_t \in \mathbb{R}^{1024}$ [38] computed from a fixed sampling of $\mathcal{O}$ to provide per-frame shape context. Contact information is encoded as $\mathcal{C}_t = \{(\mathbf{p}_{a,h}^t, \mathbf{n}_{a,h}^t, \boldsymbol{\delta}_{a,h}^t, s_{a,h}^t)\}$,

where $\mathbf{p}_{a,h}^t$ and $\mathbf{n}_{a,h}^t$ are the world-space position and normal of the contact point for hand $h \in \{\text{left}, \text{right}\}$ of agent a. $\boldsymbol{\delta}_{a,h}^t$ is the local offset relative to the mesh vertex, and $s_{a,h}^t \in \{0, 1\}$ indicates contact validity. To model graspability priors for contact estimation, we also predict object affordance $\mathcal{A} = \{(\mathbf{q}_k, \alpha_k)\}$ on sampled surface points $\mathbf{q}_k \in \mathcal{O}$, where α_k is the affordance probability for the kth point.

3.2. Co-manipulation motion generation

Human-human co-manipulation requires satisfying multiple criteria, including goal alignment, motion naturalness, and task effectiveness. To model such realistic and coordinated interactions, we adopt flow matching [25] as our foundational framework, which formulates motion generation as a velocity field regression that transports a noise sample $\mathbf{x}_0$ toward the data distribution $\mathbf{x}_1$. Flow matching enables stable likelihood-based training and avoids stochastic sampling via a deterministic vector field, which can efficiently incorporate various conditions as guidance to steer motion generation.

We therefore propose a transformer-based flow f_θ to estimate the instantaneous interaction velocity from states $\mathbf{x}_\tau$ and conditioning $\mathbf{c}$, where τ is a continuous parameter ranging from 0 to 1. The condition $\mathbf{c}$ concatenates the object pose descriptors $\{(R_t^o, \mathbf{d}_t^o)\}$, BPS embeddings $\{\mathbf{b}\}$, and cached contact anchors before projection. The flow network predicts a velocity that transports the current state toward the data manifold. The update is written as

$$\mathbf{x}_{\tau+\Delta\tau} = \mathbf{x}_\tau + \Delta\tau\, f_\theta(\mathbf{x}_\tau, \tau, \mathbf{c}). \tag{1}$$

The network performs K Euler integration steps to evolve the initial noise $\mathbf{x}_0$ into the reconstructed motion. The training procedure minimizes the mean-squared flow objective:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{\tau,\mathbf{x}_\tau}\Big[\big\| f_\theta(\mathbf{x}_\tau, \tau, \mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0)\big\|_2^2\Big], \tag{2}$$

which follows the continuous flow-matching derivation. To reduce the sensitivity to outliers in articulated joints, we additionally supervise an element-wise L_1 loss on the decoded SMPL-X parameters,

$$\mathcal{L}_{\text{SMPL}} = \mathbb{E}_\tau\Big[\big\|\hat{\mathbf{x}}_1 - \mathbf{x}_1^{\text{gt}}\big\|_1\Big], \tag{3}$$

where $\hat{\mathbf{x}}_1$ denotes the reconstructed SMPL-X parameters predicted at flow step τ and $\mathbf{x}_1^{\text{gt}}$ is the target state. This clear separation keeps the flow objective focused on estimating the continuous velocity field while the additional L_1 term stabilizes the articulated-body decoding. To further suppress foot-sliding artifacts, we incorporate a foot-contact loss [42]:

$$\mathcal{L}_{\text{foot}} = \Big\|(\mathbf{J}_f^{t+1} - \mathbf{J}_f^t)\cdot f^t\Big\|_2^2, \tag{4}$$

where $\mathbf{J}_f$ denotes the 3D position of the foot joints, and $f^t \in \{0, 1\}$ is the binary foot-contact mask at frame t. This loss penalizes foot displacement during contact frames, effectively suppressing foot-sliding artifacts. We optimize the weighted sum of the flow, SMPL, foot-contact, and prior objectives

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{flow}} + \mathcal{L}_{\text{SMPL}} + \mathcal{L}_{\text{foot}} + \mathcal{L}_{\text{prior}}. \tag{5}$$

$\mathcal{L}_{\text{prior}}$ serves as an adversarial loss that encourages realistic and coherent motion generation, and its formulation is detailed in Sec. 3.4. The resulting motions are decoded to SMPL-X meshes, and the gradient-based contact refinement from Sec. 3.3 keeps the synthesized wrists aligned with the stored hand-object anchors for coherent bimanual manipulation.

3.3. Manipulation strategy generation

Although the flow matching network can generate plausible interactive motions conditioned on specific object information and contact points, humans can produce the same object motion with varying movements. To this end, we propose an affordance-informed contact prediction network to generate diverse strategies for a specific manipulation task. For a given object geometry, we first train a regression network [23] with dense contact annotations to predict the affordance probability α_k for sampled surface points. As shown in Fig. 2 (b), the predicted affordance and BPS features are used as conditions to a diffusion model to predict the contact strategy $\mathcal{C} = \{(\mathbf{p}, \mathbf{n}, \boldsymbol{\delta}, s)\}$ from pure noises. The diffusion model is supervised with the following constraints:

$$\mathcal{L}_{\text{str}} = \mathcal{L}_{\text{anchor}} + \mathcal{L}_{\text{normal}} + \mathcal{L}_{\text{aff}}. \tag{6}$$

The contact-anchor loss constrains the generated anchors $\hat{\mathbf{p}}$ to stay close to the ground-truth contact points $\mathbf{p}$:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{Z_{\text{pos}}}\sum_{t,a,h} s_{a,h}^t \big\|\hat{\mathbf{p}}_{a,h}^t - \mathbf{p}_{a,h}^t\big\|_2^2, \tag{7}$$

where s is a binary flag indicating whether contact occurs, and $Z_{\text{pos}} = \sum_{t,a,h} s_{a,h}^t$ serves as a normalization factor over valid anchors. In parallel, a normal-alignment objective is also applied:

$$\mathcal{L}_{\text{normal}} = \frac{1}{Z_{\text{pos}}}\sum_{t,a,h} s_{a,h}^t\big(1 - \hat{\mathbf{n}}_{a,h}^t \cdot \mathbf{n}_{a,h}^t\big). \tag{8}$$

To highlight the affordance guidance, we introduce a simple regularizer that encourages each sampled contact to lie in graspable regions suggested by the affordance predictor.

$$\mathcal{L}_{\text{aff}} = -\frac{1}{Z_{\text{pos}}}\sum_{t,a,h} s_{a,h}^t \log \alpha(\hat{\mathbf{p}}_{a,h}^t), \tag{9}$$

Evaluating the learned affordance field $\alpha(\cdot)$ at the predicted anchor $\hat{\mathbf{p}}_{a,h}^{t}$ encourages the diffusion model to sample from surface zones with high graspability scores. These constraints ensure that the generated manipulation strategies remain consistent with positional, directional, and affordance cues while still allowing diverse motion variations.

The predicted contacts are then used to guide the motion flow. Specifically, we minimize a differentiable distance loss that keeps the human wrists close to the contact points during flow matching. Let $\mathcal{V}_{a,h}$ be the validity indicator for hand h of agent a and $\hat{\mathbf{p}}_{a,h}$ the corresponding contact anchor. Given the wrist positions $\mathbf{w}_{a,h}$ decoded from the current state, we evaluate

$$\mathcal{L}_{\text{contact}} = \frac{1}{Z}\sum_{a,h}\mathcal{V}_{a,h}\left\|\mathbf{w}_{a,h} - \hat{\mathbf{p}}_{a,h}\right\|_2^2, \tag{10}$$

where $Z = \sum_{a,h}\mathcal{V}_{a,h}$ normalizes over active constraints. During each Euler step, we adjust the flow prediction by the loss gradient,

$$\tilde{f}_\theta(\mathbf{x}_\tau) = f_\theta(\mathbf{x}_\tau) - \gamma\,\nabla_{\mathbf{x}_\tau}\mathcal{L}_{\text{contact}}, \tag{11}$$

With the manipulation strategy and contact guidance, the flow matching network produces trajectories that follow the desired object motion yet remain consistent with the affordance-weighted contact plans.

3.4. Adversarial interaction prior

A realistic human-human co-manipulation should reflect natural interactions. However, the motion may exhibit artifacts when relying solely on contact guidance, thereby affecting its naturalness. We therefore adopt two adversarial discriminators operating at both the individual pose and paired interaction levels to improve motion quality. As shown in Fig. 2 (c), the pose prior $\mathcal{D}_\phi^{\text{body}}$ focuses on individual poses and takes as input per-joint rotation matrices and SMPL shape coefficients. It applies 1×1 convolutions across the 21 joint rotation blocks to extract joint-wise realism cues, which are then fused with shape-aware MLP branches to produce realness score. In parallel, the interaction prior $\mathcal{D}_\phi^{\text{int}}$ processes the concatenation of dual-agent rotations, relative root transformations, and fused shape descriptors to capture inter-person coordination cues that cannot be inferred from individual body observations.

Specifically, ground-truth poses from datasets and generated poses are used as real and fake samples, respectively. However, since individual poses in Core4D may contain artifacts, Core4D samples are excluded from the real sample set in the individual pose prior $\mathcal{D}_\phi^{\text{body}}$. Likewise, the interaction discriminator is trained with paired trajectories from datasets as positive samples and synthesized dual-agent sequences as negative ones, encouraging $\mathcal{D}_\phi^{\text{int}}$ to focus on cooperative behaviors observed in real co-manipulation data.

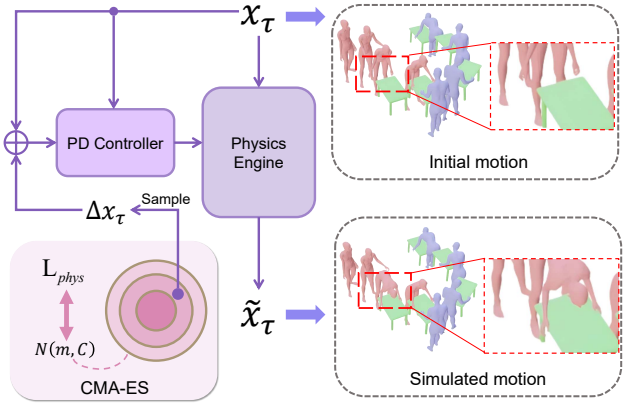

Figure 3. **Stability-driven simulation pipeline.** The CMA-ES algorithm samples corrective offsets $\Delta\mathbf{x}_\tau$ for the flow-matching outputs $\mathbf{x}_\tau$. The corrected motions are then fed into the physics engine equipped with a PD controller, and the simulated results are used in the next Euler integration step.

Each discriminator is optimized with a non-saturating binary cross-entropy objective:

$$\mathcal{L}_{\text{prior}}^{(k)} = -\mathbb{E}_{(\mathbf{R},\boldsymbol{\beta})\sim\mathcal{D}_{\text{real}}^{(k)}}\left[\log\mathcal{D}_\phi^k(\mathbf{R},\boldsymbol{\beta})\right] \tag{12}$$

$$-\,\mathbb{E}_{(\tilde{\mathbf{R}},\tilde{\boldsymbol{\beta}})\sim\mathcal{D}_{\text{gen}}^{(k)}}\left[\log(1-\mathcal{D}_\phi^k(\tilde{\mathbf{R}},\tilde{\boldsymbol{\beta}}))\right], \tag{13}$$

where $k\in\{\text{body},\text{int}\}$ indexes the single-body and interaction priors, and $\mathbf{R}$ denotes the input of the prior. The losses sum to $\mathcal{L}_{\text{prior}} = \mathcal{L}_{\text{prior}}^{(\text{body})} + \mathcal{L}_{\text{prior}}^{(\text{int})}$ and the gradients propagate through the flow decoder to encourage realistic articulation and coordination during training.

Beyond training, the learned priors can also enhance motion naturalness during flow matching. The trained discriminators are reused as evaluators that guide the sampling process, where the predicted state is refined using the aggregated gradients $\nabla_{\mathbf{x}_\tau}\log\mathcal{D}_\phi^k$, *i.e.*, gradients that encourage the sampled poses to align with realistic human motion patterns.

$$\tilde{f}_\theta(\mathbf{x}_\tau) = f_\theta(\mathbf{x}_\tau) + \eta\sum_{k\in\{\text{body},\text{int}\}}\nabla_{\mathbf{x}_\tau}\log\mathcal{D}_\phi^k, \tag{14}$$

where η is a guidance weight. This correction guides the integration toward regions favored by the learned priors while preserving the base velocity field.

3.5. Stability-driven simulation

Although the generated co-manipulation motions align with the input object trajectory and exhibit plausible interactions, they often suffer from severe floating and penetration artifacts between hands and manipulated objects, leading to object instability and physically implausible motion. To address this limitation, we further introduce a stability-driven simulation. Since the current RL-based policies [4, 33] cannot be generalized to diverse objects and tasks, we adopt

sampling-based optimization [5, 10, 26] to achieve the refinement. During the flow matching process, we execute the simulation in the penultimate integration step, and then the simulated results are directly used for the last step.

Specifically, we convert $\mathbf{x}_\tau$ into SMPL-X parameters and instantiate humanoid models based on the SMPL-X meshes. The decoded pose and translation are used to initialize the humanoids in the physics engine. Without further correction, artifacts such as floating or penetration may cause the object to fall during simulation. Therefore, we use a proportional-derivative (PD) controller with CMA-ES algorithm [7] to adjust the body poses. As shown in Fig. 3, based on the initial parameters $\mathbf{x}_\tau$, we sample a correction $\Delta\mathbf{x}_\tau$ from a multivariate normal distribution $\mathcal{N}(\boldsymbol{m}, \boldsymbol{C})$ to construct the desired targets $\bar{\mathbf{x}}_\tau = \mathbf{x}_\tau + \Delta\mathbf{x}_\tau$ for the PD controller. The distribution $\mathcal{N}(\boldsymbol{m}, \boldsymbol{C})$ is initialized as a standard normal and updated during CMA-ES optimization. During simulation, the PD controller produces joint torques to drive the humanoids toward the target poses. We then evaluate each sample with several cost functions.

$$\mathcal{L}_{\text{phys}} = \mathcal{L}_{\text{sim}} + \mathcal{L}_{\text{sta}}. \tag{15}$$

The similarity loss evaluate the similarity between simulated body and object poses:

$$\mathcal{L}_{\text{sim}} = \|\tilde{\mathbf{x}}_\tau - \mathbf{x}_\tau\|_2^2 + \|\tilde{R}^o - R^o\|_2^2 + \|\tilde{\mathbf{d}}^o - \mathbf{d}^o\|_2^2, \tag{16}$$

where $\tilde{\mathbf{x}}_\tau$ and $\{\tilde{R}^o, \tilde{\mathbf{d}}^o\}$ are the simulated results for humanoids and object, respectively. The stability loss is formulated as:

$$\mathcal{L}_{\text{sta}} = \frac{\left\|\vec{f}(t) - M_o\vec{a}\right\|_2^2}{\|M_o\vec{g}\|_2^2} + \frac{\|\vec{\mu}(t) - I_o\vec{\alpha}\|_2^2}{\|I_o\vec{\alpha}\|_2^2} + e^{-m(t)}, \tag{17}$$

where $\vec{f}(t)$ and $\vec{\mu}(t)$ are resultant external force and torque. M_o and I_o are mass and inertia matrix of the object. $\vec{a}$ and $\vec{\alpha}$ are linear and angular acceleration calculated from input trajectories. $e^{-m(t)}$ is an energy regularization [35], and $\vec{g}$ is the gravity.

We refine the sampling distribution through CMA-ES optimization and take the simulation result with the lowest cost $\tilde{\mathbf{x}}_\tau$ as the final output. The simulated trajectories are fed into the next integration step. With the simulation, this refinement leads to more physically plausible and stable co-manipulation behaviors.

With the manipulation strategy, interaction prior, and simulation, our method can finally generate intention-driven manipulation with natural and effective motions.

4. Experiments

In this section, we first introduce the datasets and evaluation metrics used in our experiments, followed by implementation details for reproducibility. We then compare our approach with state-of-the-art methods to demonstrate its effectiveness. Finally, we perform ablation studies to analyze the contribution of each key component.

4.1. Datasets

Core4D [29] is a large-scale dataset focusing on collaborative human–object–human interactions. We extract the ground-truth object trajectories from its sequences and use them as input to our 3D motion generation framework. We adopt the official training and testing splits defined in Core4D. **Inter-X** [48] is a large-scale dataset designed for versatile human–human interaction analysis. We utilize this dataset to enhance the quality and diversity of our motion generation model.

4.2. Metrics

We adopt several metrics to evaluate motion and interaction quality. For interaction, we use interactive Distance Field (IDF) [54] to measure the fidelity of human-object spatial relationships via the mean-squared error between predicted and ground-truth implicit distance fields sampled around the object. In addition, **Contact Accuracy (Contact Acc.)** [29] evaluates how well binary hand-object contacts are reproduced using frame-level precision. **Penetration (Pene.)** [11] quantifies physical plausibility by averaging signed penetration depths between reconstructed human meshes and the object. To evaluate motion quality, we report the **Fréchet Inception Distance (FID)** for distributional realism and **Diversity (Div.)** defined as the average pairwise distance between generated sequences.

4.3. Implementation Details

We train and evaluate our framework using a single NVIDIA RTX 4090 GPU with 24 GB memory, an Intel Xeon Platinum CPU, and 90 GB of RAM. Training follows the default configuration with a batch size of 10, a learning rate of 1×10^{-4}, AdamW optimization, and a cyclic cosine scheduler. Flow-matching inference uses $K = 10$ Euler integration steps with a stability refinement. Physical simulation leverages PyBullet, where the physics simulator runs at 240 Hz while the proportional-derivative controller issues targets at 60 Hz to match the contact-conditioned motion sampling rate. The flow matching model and physics simulation take 1.19 s and 3 min, respectively, to generate a 128-frame motion sequence.

4.4. Comparison to State-of-the-Art Methods

Since no existing open-source human–human–object interaction methods are directly comparable to ours, we adapt several state-of-the-art human–human and human–object interaction methods as baselines. To align with our object-guided task, we modify ComMDM [39] and InterGen [24]

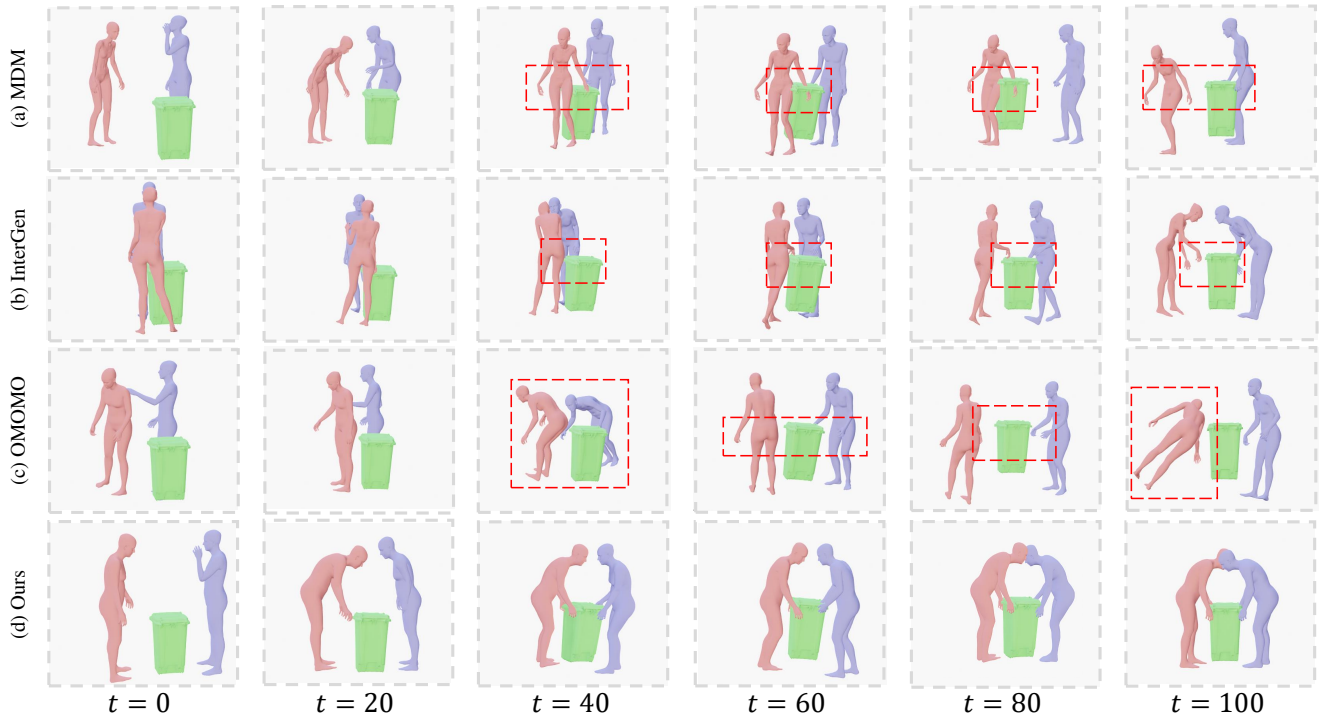

Figure 4. Qualitative comparison on Core4D-S1, showing manipulations generated by ComMDM, InterGen, and OMOMO (a–c), as well as our approach (d), at key timestamps $t \in 0, 20, 40, 60, 80, 100$. Our results (d) maintain coordinated grasps and stable payload alignment, whereas previous methods exhibit slipping contacts or delayed responses when the green object changes its pose.

Table 1. Performance comparison on the Core4D dataset. Metrics marked with ↑ indicate higher is better, while ↓ indicates lower is better.

Method	Core4D-S1					Core4D-S2				
	IDF ↓	Contact Acc. ↑	FID ↓	Div. ↑	Pene. ↓	IDF ↓	Contact Acc. ↑	FID ↓	Div. ↑	Pene. ↓
ComMDM [39]	0.41	0.11	52.5	1.13	0.19	0.43	0.12	49.4	1.13	0.21
OMOMO [20]	0.38	0.21	45.8	1.08	0.15	0.37	0.23	44.4	1.10	0.15
InterGen [24]	0.47	0.13	35.4	**1.21**	0.11	0.47	0.10	30.2	**1.18**	0.12
Ours	**0.22**	**0.44**	**25.5**	1.15	**0.05**	**0.20**	**0.46**	**21.6**	**1.18**	**0.06**

by replacing their text-conditioning inputs with our object 6D poses and BPS features. Furthermore, we extend OMOMO [20], an object-guided human–object interaction generation method, to handle dual-human scenarios.

Tab. 1 show a quantitative comparison on Core4D test sets. Since ComMDM and InterGen do not explicitly incorporate object trajectories, their performance on IDF and Contact Acc. is inferior to ours. Although OMOMO imposes contact constraints, it still exhibits unstable dual-motion coordination and frequent penetrations due to the heuristic extension from single-character to dual-human modeling. Furthermore, the affordance-informed manipulation strategy offers diverse and semantically consistent guidance, leading to improved FID and comparable diversity, which demonstrates the effectiveness of our approach in generating realistic and varied motions. Moreover, by incorporating physical simulation, our method also achieves lower penetration rates and improved physical plausibility compared to other kinematics-based baselines.

The qualitative comparison in Fig. 4 shows that when two people cooperate to move an object, both ComMDM and OMOMO suffer from orientation misalignment and fail to generate natural and coordinated interactions with the same object. Although InterGen employs a Transformer with cross-attention to improve the naturalness of dual-human actions, the lack of contact constraints between the humans and the object leads to significant hand–object misalignment. In contrast, our method, guided by affordance priors and physical simulation, maintains stable contact among both humans and the object during t=40–100. Moreover, benefiting from the adversarial interaction prior, our method consistently produces natural and physically plausible interactive postures throughout the manipulation process.

Figure 5. Cooperative motions produced by our framework. The two characters remain synchronized while steering and lifting the green object along the given trajectory, exhibiting fine-grained grasp readjustments throughout the interaction.

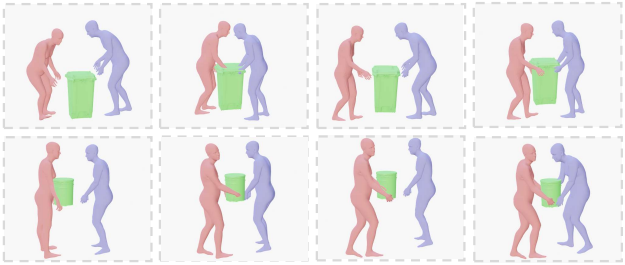

Figure 6. Ablation of key components on Core4D-S1. The vanilla flow matching model fails to generate realistic interactions. Built upon this baseline, the affordance-informed strategy and interaction prior further improve motion realism and naturalness. Moreover, the physics-based simulation enhances physical plausibility.

Table 2. Ablation studies on Core4D-S1 dataset. "Flow Matching" uses 6D poses as the trajectory without additional modules, and "+" indicates the inclusion of the corresponding module on top of the Flow Matching baseline.

Method	IDF ↓	Contact Acc. ↑	FID ↓	Div. ↑	Pene. ↓
Flow Matching	0.25	0.35	26.3	1.16	0.15
+ BPS feature	0.24	0.37	26.1	1.15	0.14
+ Contact	0.24	0.40	26.0	1.16	0.20
+ GT Contact	0.25	0.42	26.1	1.15	0.20
+ Individual Prior	0.22	0.34	25.5	1.15	0.16
+ Interaction Prior	0.23	0.35	25.4	1.16	0.16
+ Simulation	0.23	0.42	28.6	1.15	0.02
Ours	0.22	0.44	25.5	1.15	0.05

4.5. Ablation Study

Physics-based simulation. The stability-driven simulation module rolls out the predicted motions using a PD controller and optimizes torques via CMA-ES to correct unstable postures before the final denoising step. Although the simulated motions are physically plausible, they may still exhibit unnatural poses due to the lack of prior knowledge (*i.e.*, FID). Therefore, the motions are further fine-tuned with the last integration step of the Flow Matching model. With this pipeline, Tab. 2 and Fig. 6 show that our method generates realistic motions while maintaining physical plausibility. Removing the simulation leads to a notable drop in contact accuracy from 0.44 to 0.37 and a significant increase in penetration depth from 0.05 to 0.16, confirming that the physics feedback is essential for stable and accurate co-manipulation. The FID remains largely unchanged, demonstrating that the final flow-matching step successfully recovers motion naturalness after physics-based correction.

Adversarial interaction prior. The adversarial interaction prior consists of two components: an individual pose discriminator that refines local articulation, and an interaction prior that encourages coordinated full-body interactions. Compared with the vanilla Flow Matching model in Tab. 2, adding the individual pose prior decreases IDF from 0.25 to 0.22 and FID from 26.3 to 25.5, improving single-pose generation quality despite a temporary drop in contact accuracy. Adding the dual-agent prior subsequently increases contact accuracy to 0.35 while maintaining FID at 25.4, indicating that adversarial guidance enhances interactions without sacrificing realism.

Manipulation Strategy. The affordance module predicts dense scores on the object surface and guides the diffusion sampler to generate contact anchors that satisfy positional, normal, and availability constraints. These anchors are then used to refine hand–object alignment during motion generation. Using the predicted anchors to optimize hand–object contact increases contact accuracy from 0.35 to 0.40, while FID remains at 26.0, as shown in Tab. 2 under "+ Contact". This demonstrates that availability-aware anchors can significantly improve hand–object alignment.

5. Conclusion

We propose a collaborative manipulation generation framework that conditions dual SMPLX humans on object geometry and trajectories. To ensure that the generated motion aligns with the input trajectories, we introduce an affordance-informed manipulation strategy that provides explicit contact guidance for the flow matching model. Furthermore, an adversarial interaction prior and a physics-based simulation are incorporated to further enhance motion realism and physical plausibility. With these modules, our method can generate realistic and natural human–human co-manipulation motions from given object information.

Acknowledgements This work was supported by Science Fund for Distinguished Young Scholars of Tianjin under Grant 22JCJQJC00040.

References

[1] Ziyi Chang, He Wang, George Koulieris, and Hubert PH Shum. Large-scale multi-character interaction synthesis. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2025. 2

[2] Divyanshu Daiya, Damon Conover, and Aniket Bera. Collage: Collaborative human-agent interaction generation using hierarchical latent diffusion and language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8203–8210. IEEE, 2025. 2

[3] 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*, pages 19888–19901, 2024. 2

[4] 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. *Advances in Neural Information Processing Systems*, 37:79741–79763, 2024. 2, 5

[5] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13106–13115, 2022. 6

[6] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 2

[7] Nikolaus Hansen. The cma evolution strategy: A tutorial. *arXiv preprint arXiv:1604.00772*, 2016. 6

[8] 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*, pages 1–9, 2023. 3

[9] Wenkun He, Yun Liu, Ruitao Liu, and Li Yi. Syncdiff: Synchronized motion diffusion for multi-body human-object interaction synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11731–11743, 2025. 2

[10] Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, and Yangang Wang. Neural mocon: Neural motion control for physically plausible human motion capture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6417–6426, 2022. 3, 6

[11] Buzhen Huang, Chen Li, Chongyang Xu, Liang Pan, Yangang Wang, and Gim Hee Lee. Closely interactive human reconstruction with proxemics and physics-guided adaption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1011–1021, 2024. 3, 6

[12] Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. Intermask: 3d human interaction generation via collaborative masked modeling. *arXiv preprint arXiv:2410.10010*, 2024. 3

[13] Kaiyang Ji, Ye Shi, Zichen Jin, Kangyi Chen, Lan Xu, Yuexin Ma, Jingyi Yu, and Jingya Wang. Towards immersive human-x interaction: A real-time framework for physically plausible motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10173–10183, 2025. 2

[14] Yihong Ji, Yunze Liu, Yiyao Zhuo, Weijiang Yu, Fei Ma, Joshua Zhexue Huang, and Fei Yu. Onlinehoi: Towards online human-object interaction generation and perception. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9395–9403, 2025. 2

[15] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2

[16] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 947–957, 2024. 2

[17] Jiahui Lei, Congyue Deng, William B Shen, Leonidas J Guibas, and Kostas Daniilidis. Nap: Neural 3d articulated object prior. *Advances in Neural Information Processing Systems*, 36:31878–31894, 2023. 2

[18] Baiyi Li, Edmond SL Ho, Hubert PH Shum, and He Wang. Two-person interaction augmentation with skeleton priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 2

[19] Boyuan Li, Xihua Wang, Ruihua Song, and Wenbing Huang. Two-in-one: Unified multi-person interactive motion generation by latent diffusion transformer. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 3

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

[21] 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*, pages 54–72. Springer, 2024. 2

[22] Ronghui Li, Youliang Zhang, Yachao Zhang, Yuxiang Zhang, Mingyang Su, Jie Guo, Ziwei Liu, Yebin Liu, and Xiu Li. Interdance: Reactive 3d dance generation with realistic duet interactions. *arXiv preprint arXiv:2412.16982*, 2024. 2, 3

[23] Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-seng Chua. Laso: Language-guided affordance segmentation on 3d object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14251–14260, 2024. 4

[24] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion genera-

tion under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024. 2, 3, 6, 7

[25] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2, 4

[26] Libin Liu, KangKang Yin, and Baining Guo. Improving sampling-based motion control. In *Computer Graphics Forum*, pages 415–423. Wiley Online Library, 2015. 3, 6

[27] Shaowei Liu, Chuan Guo, Bing Zhou, and Jian Wang. Ponimator: Unfolding interactive pose for versatile human-human interaction animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12068–12077, 2025. 3

[28] Yun Liu, Bowen Yang, Licheng Zhong, He Wang, and Li Yi. Mimicking-bench: A benchmark for generalizable humanoid-scene interaction learning via human mimicking. *arXiv preprint arXiv:2412.17730*, 2024. 2

[29] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1769–1782, 2025. 2, 6

[30] 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*, pages 10895–10904, 2023. 3

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

[32] Sakuya Ota, Qing Yu, Kent Fujiwara, Satoshi Ikehata, and Ikuro Sato. Pino: Person-interaction noise optimization for long-duration and customizable motion generation of arbitrary-sized groups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10676–10685, 2025. 2

[33] 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*, pages 5379–5391, 2025. 3, 5

[34] 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*, pages 10975–10985, 2019. 3

[35] 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. 3, 6

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

[37] Huaijin Pi, Sida Peng, Minghui Yang, Xiaowei Zhou, and Hujun Bao. Hierarchical generation of human-object interactions with diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15061–15073, 2023. 2

[38] 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*, pages 4332–4341, 2019. 2, 3

[39] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 6, 7

[40] Yi Shi, Jingbo Wang, Xuekun Jiang, Bingkun Lin, Bo Dai, and Xue Bin Peng. Interactive character control with auto-regressive motion diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–14, 2024. 3

[41] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 43(6):1–21, 2024. 3

[42] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 4

[43] 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. 3

[44] Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. Intercontrol: Zero-shot human interaction generation by controlling every joint. *Advances in Neural Information Processing Systems*, 37:105397–105424, 2024. 3

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

[46] Qingxuan Wu, Zhiyang Dou, Chuan Guo, Yiming Huang, Qiao Feng, Bing Zhou, Jian Wang, and Lingjie Liu. Text2interact: High-fidelity and diverse text-to-two-person interaction generation. *arXiv preprint arXiv:2510.06504*, 2025. 3

[47] Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. Human-object interaction from human-level instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11176–11186, 2025. 2

[48] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22260–22271, 2024. 3, 6

[49] Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. Regennet: Towards human action-reaction synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1759–1769, 2024. 3

[50] Liang Xu, Chengqun Yang, Zili Lin, Fei Xu, Yifan Liu, Congsheng Xu, Yiyi Zhang, Jie Qin, Xingdong Sheng, Yunhui

Liu, et al. Perceiving and acting in first-person: A dataset and benchmark for egocentric human-object-human interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12535–12548, 2025. 2

[51] 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*, pages 14928–14940, 2023. 2

[52] 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*, pages 7048–7060, 2025. 2

[53] Wenning Xu, Shiyu Fan, Paul Henderson, and Edmond SL Ho. Multi-person interaction generation from two-person motion priors. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025. 2

[54] Mengqing Xue, Yifei Liu, Ling Guo, Shaoli Huang, and Changxing Ding. Guiding human-object interactions with rich geometry and relations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22714–22723, 2025. 6

[55] Ye Yuan, Viktor Makoviychuk, Y Guo, S Fidler, X Peng, and K Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM Trans. Graph*, 42(4):66, 2023. 3

[56] 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*, pages 16010–16021, 2023. 3

[57] Ziyu Zhang, Sergey Bashkirov, Dun Yang, Yi Shi, Michael Taylor, and Xue Bin Peng. Physics-based motion imitation with adversarial differential discriminators. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–12, 2025. 3

[58] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 3

[59] Chen Zhu, Buzhen Huang, Zijing Wu, Binghui Zuo, and Yangang Wang. E-react: Towards emotionally controlled synthesis of human reactions. *arXiv preprint arXiv:2508.06093*, 2025. 2