InterDiff: Generating 3D Human-Object Interactions with Physics-Informed Diffusion

Sirui Xu Zhengyuan Li Yu-Xiong Wang* Liang-Yan Gui*
University of Illinois at Urbana-Champaign
{siruixu2, zli138, yxw, lgui}@illinois.edu
https://sirui-xu.github.io/InterDiff/

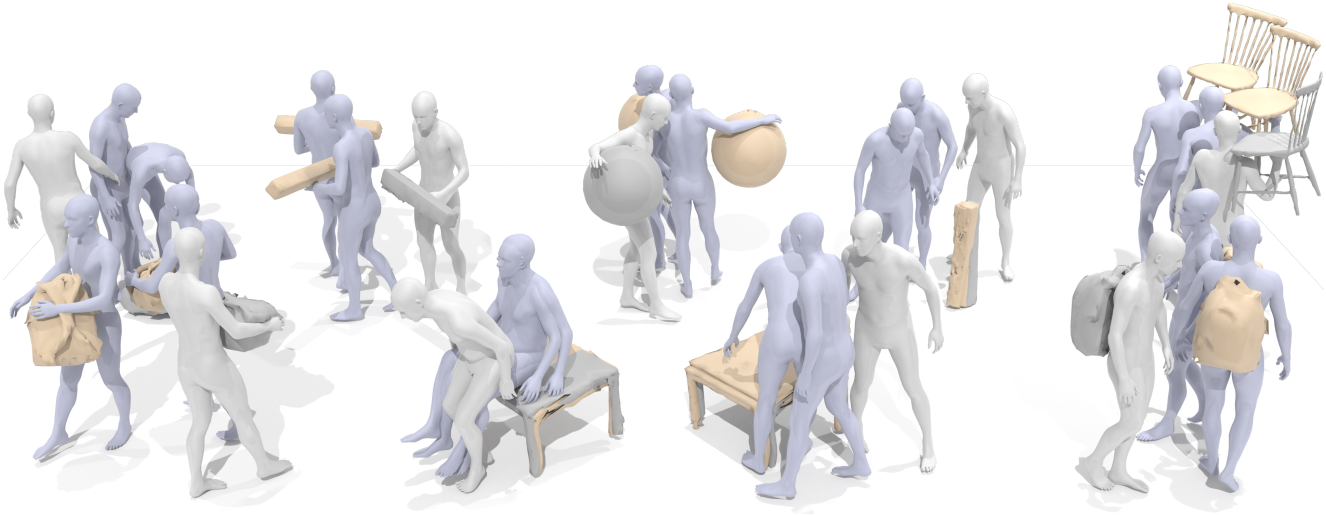


Figure 1. **A novel task of predicting 3D human-object interactions.** We provide 9 HOI sequences sampled every 40 frames at 30 FPS. Conditioned on past HOIs in gray meshes, our model generates long-term, diverse, and vivid HOIs, represented by the colored meshes.

Abstract

This paper addresses a novel task of anticipating 3D human-object interactions (HOIs). Most existing research on HOI synthesis lacks comprehensive whole-body interactions with dynamic objects, e.g., often limited to manipulating small or static objects. Our task is significantly more challenging, as it requires modeling dynamic objects with various shapes, capturing whole-body motion, and ensuring physically valid interactions. To this end, we propose InterDiff, a framework comprising two key steps: (i) interaction diffusion, where we leverage a diffusion model to encode the distribution of future human-object interactions; (ii) interaction correction, where we introduce a physics-informed predictor to correct denoised HOIs in a diffusion step. Our key insight is to inject prior knowledge that the interactions under reference with respect to contact points follow a simple pattern and are easily predictable. Experiments on multiple human-object interaction datasets demonstrate the effectiveness of our method for this task, capable of producing realistic, vivid, and remarkably long-term 3D HOI predictions.

1. Introduction

Being able to "look into the future" is a remarkable cognitive hallmark of humans. Not only can we anticipate how people will move or behave in the near future, but we can also forecast how our actions will interact with the ever-changing environment based on past information. An automated system that accurately forecasts 3D human-object interactions (HOIs) would have significant implications for various fields, such as robotics, animation, and computer vision. However, existing work on HOI synthesis does not adequately reflect the real-world complexity, *e.g.*, examining hand-object interactions from an ego-centric view [51, 53], synthesizing interactions of grasping small objects [20], representing HOIs in simplified skeletons [13, 71, 88], or overlooking object dynamics [46, 81, 97].

To overcome such limitations, in this work, we reformulate the task of *human-object interaction prediction*, where

*Equal contribution.

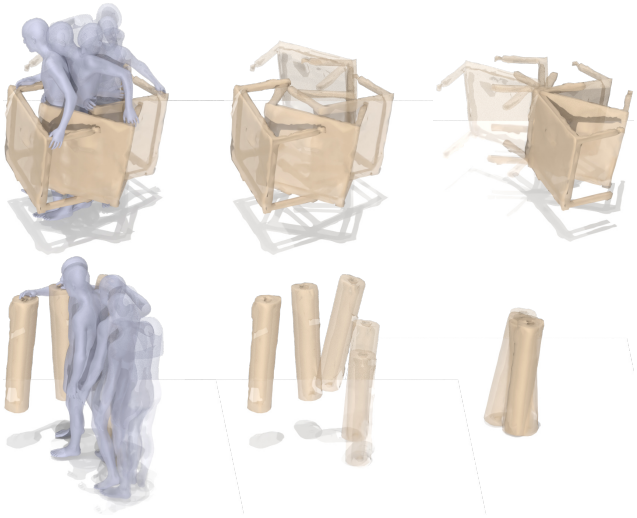


Figure 2. We present ground truth HOI sequences (**left**), object motions (**middle**), and objects relative to the contacts after coordinate transformations (**right**). Our key insight is to inject coordinate transformations into a diffusion model, as the relative motion shows simpler patterns that are easier to predict, *e.g.*, rotating around a fixed axis (**top**) or being almost stationary (**bottom**).

we aim to model and forecast 3D mesh-based whole-body movements and object dynamics simultaneously, as shown in Figure 1. This task presents several unique real-world challenges. **(i) High Complexity**: it requires the modeling of both full-body and object dynamics, which is further complicated by the considerable variability in object shapes. **(ii) Physical Validity**: the predicted interaction should be *physically* plausible. Specifically, the human body should naturally conform to the surface of the object when in contact, while avoiding any penetration.

A naïve approach would be to directly extend existing deep generative models that have been developed for human motion prediction, as exemplified by motion diffusion models [85], to capture the distribution of future human-object interactions. However, these models fail to incorporate the underlying physical laws that would ensure perceptually realistic predictions, thus introducing artifacts such as floating contact and penetration. This problem is amplified when autoregressive inference is utilized to synthesize long-term interactions, as errors accumulate over time. To this end, most existing research on 3D HOI synthesis [20,81,97] relies on post-hoc optimization to inject physical constraints. HOI synthesis has also been explored with simulators [3, 27, 50, 60] to ensure physical properties. Although plausible interactions can be generated, effort is required to build a physics simulation environment, *e.g.*, registering objects with diverse shapes, as well as frictions, stiffness, and masses, which are hardly present in motion capture datasets. Moreover, considerable time is needed to train control policies to track realistic interactions.

Rather than relying on post-optimization or physics simulation, we introduce a pure learning-based method that utilizes a diffusion model with intuitive physics directly injected, which we call "InterDiff." Our approach is based on the key observation that *the short-term relative motion of an object with respect to the contact point follows a simple and nearly deterministic pattern*, despite the complexity of the overall interaction. For example, when juggling balls, their path reflects a complex pattern under the global coordinate system, due to the movement of the juggler. Yet, each ball simply moves up and down with respect to the juggler's hand. We provide further illustrations of relative motion extracted from the BEHAVE dataset [5] in Figure 2.

Inspired by this, our InterDiff incorporates two components as follows. **(i) Interaction Diffusion**: a Denoising Diffusion Probabilistic Model (DDPM)-based generator [29] that models the distribution of future human-object interactions. **(ii) Interaction Correction**: a *novel physics-informed interaction predictor* that synthesizes the object's *relative motion* with respect to regions in contact on the human body. We enhance this predictor by promoting simple motion patterns for objects and encouraging contact fitting of surfaces, which largely mitigates the artifacts of interactions produced by the diffusion model. By injecting the plausible dynamics back into the diffusion model iteratively, InterDiff generates vivid human motion sequences with realistic interactions for various 3D dynamic objects. *Another attractive property* of InterDiff is that its two components can be trained separately, while naturally conforming during inference without fine-tuning.

Our **contributions** are three-fold. **(i)** To the best of our knowledge, we are the *first* to tackle the task of mesh-based 3D HOI prediction. **(ii)** We propose the *first* diffusion framework that leverages past motion and shape information to generate future human-object interactions. **(iii)** We introduce a simple yet effective HOI corrector that incorporates physics priors and thus produces plausible interactions to infill the denoising generation. Extensive experiments validate the effectiveness of our framework, particularly for *out-of-distribution objects, and long-term autoregressive inference* where input past HOIs may be unseen in the training data. We attribute our improved generalizability to our important design strategies, such as the promotion of simple motion patterns and the anticipated interaction within a local reference system.

2. Related Work

Denoising Diffusion Models. Denoising diffusion models [29, 52, 76, 77] are equipped with a stochastic diffusion process that gradually introduces noise into a sample from the data distribution, following thermodynamic principles, and then generates denoised samples through a reverse iterative procedure. Recent work has extended them to the task

of human motion generation [2, 4, 10, 11, 15, 34, 39, 69, 75, 80, 85, 86, 95, 113, 115, 116, 120]. For instance, MDM [85] utilizes a transformer architecture to predict clean motion in the reverse process. We extend their framework to our HOI prediction task. To generate conditional samples, a common strategy involves repeatedly injecting available information into the diffusion process. A similar idea applies to motion diffusion models [69,75,85] for motion infilling. Compared with PhysDiff [110], which injects a motion imitation policy based on physics simulation into the diffusion process, we leverage a much simpler interaction correction step that is informed of the appropriate coordinate system to yield plausible interactions at a lower cost.

Human-Object Interaction. Despite recent advancements in human-object interaction learning, existing research has primarily focused on HOI detection [12, 21, 36, 96, 99, 124, 127], reconstruction [31, 42, 64, 92, 98, 112, 114], and generating humans that interact with static scenes [8, 26, 34, 83, 89–91, 93, 121, 122]. Most attempts have been made to synthesize only hand-object interactions in computer graphics [49, 67, 111], computer vision [14, 23, 37, 40, 45, 48, 82, 106, 123, 125], and robotics [7, 16, 32, 49]. Generating whole-body interactions, such as approaching and manipulating static [46, 81, 97, 117], articulated [47, 104], and dynamic objects [20] has also been a growing topic. The task of synthesizing humans interacting with dynamic objects has been explored based on first-person vision [51, 53] and skeletal representations [13, 71, 88] on skeleton-based datasets [44, 55, 56]. In humanoid control, progress on full-body HOI synthesis has been made with kinematic-based approaches [78, 79] and in the application of physics simulation environments [3, 9, 27, 50, 60, 62, 100, 101, 105]. However, most approaches have significant limitations regarding action and object variation, such as focusing on approaching or manipulating objects on *e.g.*, the GRAB [82] dataset. Recent datasets [5, 19, 35, 38, 112] are established to address the above limitations and provide 3D interactions with richer objects and actions, setting the stage for achieving our task.

Human Motion and Object Dynamics. Generative modeling, including variational autoencoders [43], generative adversarial networks [22], normalizing flows [73], and diffusion models [76, 77], has witnessed significant progress recently, leading to attempts for skeleton-based human motion prediction [4, 6, 17, 24, 57, 103, 108, 109]. Moreover, research has expanded beyond skeleton generation and utilized statistical models such as SMPL [54] to generate 3D body animations [25, 58, 65, 66, 84, 102, 118, 119]. Our study employs SMPL parameters to drive the 3D mesh of the human body on the BEHAVE dataset [5], while also extending the method to skeleton-based datasets [88], demonstrating its broad applicability. Predicting object dynamics has also received increasing attention [18, 61, 72, 107, 128]. Different from solely predicting the human motion or the object dynamics, our method jointly models their interactions.

3. Methodology

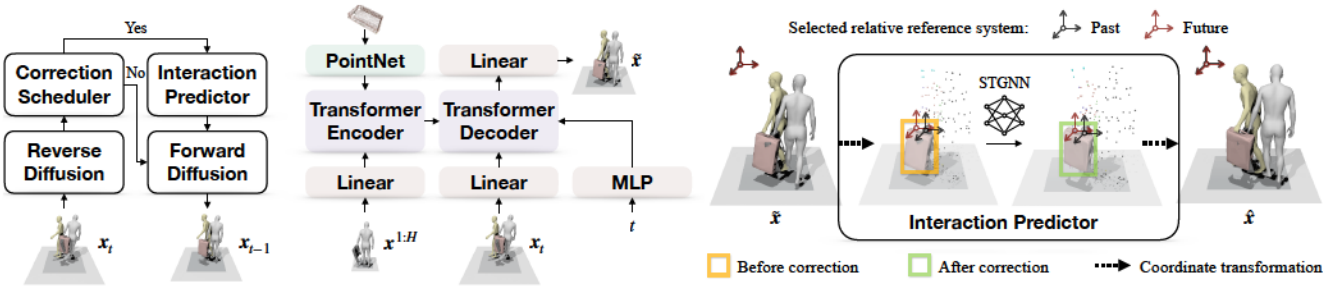
Problem Formulation: Human-Object Interaction Prediction. We denote a 3D HOI sequence with H historical frames and F future frames as $\boldsymbol{x} = [\boldsymbol{x}^1, \boldsymbol{x}^2, \dots, \boldsymbol{x}^{H+F}]$, where $\boldsymbol{x}^i$ consists of human pose state $\boldsymbol{h}^i$ and object pose state $\boldsymbol{o}^i$. Human pose state $\boldsymbol{h}^i \in \mathbb{R}^{J \times D_h}$ is defined by J joints with a D_h-dimensional representation at each joint, which can be joint position, rotation, velocity, or their combination. Object pose state $\boldsymbol{o}^i$ has D_o features, including *e.g.*, the position of the center, and the rotation of the object w.r.t. the template. Note that the specific meanings of these states are dataset-dependent and will be explained in detail in Sec. 4. Given object shape information $\boldsymbol{c}$, our goal is to predict a 3D HOI sequence $\boldsymbol{x}_0$ that is (i) close to the ground truth $\boldsymbol{x}$ in future F frames, and (ii) physically valid.

Overview. As shown in Figure 3, InterDiff consists of interaction diffusion and correction. In Sec. 3.1, we introduce interaction diffusion, which includes the forward and reverse diffusion processes. We explain how we extract shape information for the diffusion model. We then detail interaction correction in Sec. 3.2, including correction schedule and interaction prediction steps. Our *key insight* is applying interaction correction to implausible denoised HOIs. Given a denoised HOI after each reverse diffusion process, the correction scheduler determines if this denoised HOI needs correction, and infers a *reference system* based on contact information extracted from this intermediate result (Sec. 3.2.1). If the correction is needed, we pass the denoised HOI and the inferred reference system to an interaction predictor, which forecasts plausible object motion under the identified reference. Afterward, we inject this plausible motion back into the denoised HOI for further denoising iterations (Sec. 3.2.2). Notably, interaction diffusion and correction do not need to be coupled *during training*. Instead, they can be composed during inference *without fine-tuning*.

3.1. Interaction Diffusion

Basic Diffusion Model. Our approach incorporates a diffusion model, generating samples from isotropic Gaussian noise by iteratively removing the noise at each step. More specifically, to model a distribution $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$, the forward diffusion process follows a Markov chain of T steps, giving rise to a series of time-dependent distributions $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$. These distributions are generated by gradually injecting noise into the samples until the distribution of $\boldsymbol{x}_T$ is close to $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Formally, this process is denoted as

$$\begin{aligned} q(\boldsymbol{x}_1, \dots, \boldsymbol{x}_T|\boldsymbol{x}_0) &= \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \\ q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) &= \mathcal{N}(\sqrt{\beta_t}\boldsymbol{x}_{t-1} + (1-\beta_t)\mathbf{I}), \end{aligned} \tag{1}$$



(i) Overall Framework **(ii)** Reverse Diffusion **(iii)** Interaction Correction

Figure 3. **Overview of InterDiff.** **(i)** We combine a Correction Scheduler and an Interaction Predictor with the diffusion framework to correct a denoised HOI. The Correction Scheduler determines whether the current denoised HOI needs correction. If so, we fuse the additional prediction generated by the Interaction Predictor into the denoised HOI. **(ii)** Our reverse diffusion employs a transformer architecture conditioned on the encoded object shape and the past HOI. **(iii)** We transform object motion under the reference system selected by the Correction Scheduler, predict future motion via STGNN, and transform it back to the ground system. Markers are in point clouds.

where $\beta_t \in (0, 1)$ is the variance of the Gaussian noise injected at time t, and we define $\beta_0 = 0$.

Here, we adopt the Denoising Diffusion Probabilistic Model (DDPM) [29] for motion prediction, given that it can sample $\boldsymbol{x}_t$ directly from $\boldsymbol{x}_0$ without intermediate steps:

$$\begin{aligned} q(\boldsymbol{x}_t|\boldsymbol{x}_0) &= \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + (1-\bar{\alpha}_t)\mathbf{I}) \\ \boldsymbol{x}_t &= \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \end{aligned} \tag{2}$$

where $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{t'=0}^{t} \alpha_{t'}$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The reverse process of diffusion gradually cleans $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ back to $\boldsymbol{x}_0$. Following [69, 70, 85], we directly recover the clean signal $\tilde{\boldsymbol{x}}$ at each step, instead of predicting the noise $\boldsymbol{\epsilon}$ [29] that was added to $\boldsymbol{x}_0$ in Eq. 2. This iterative process at step t is formulated as

$$\begin{aligned} \tilde{\boldsymbol{x}} &= \mathcal{G}(\boldsymbol{x}_t, t, \boldsymbol{c}) \\ \boldsymbol{x}_{t-1} &= \sqrt{\bar{\alpha}_{t-1}}\tilde{\boldsymbol{x}} + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}, \end{aligned} \tag{3}$$

where $\mathcal{G}$ is a network estimating $\tilde{\boldsymbol{x}}$ given the noised signal $\boldsymbol{x}_t$ and the condition $\boldsymbol{c}$ at step t, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Interaction Diffusion Model. While most existing *human motion diffusion models* [69, 85, 110, 115] can predict future frames by infilling ground truth past motion into the denoised motion at each diffusion step, we observe that encoding the historical motion $\boldsymbol{x}^{1:H}$ as a condition leads to better performance in our task. A similar design is used in [4, 94] but for conventional human motion prediction. Now our model $\mathcal{G}$ includes a transformer encoder that encodes $\boldsymbol{x}^{1:H}$ along with the object shape embedding $\boldsymbol{c}$ from a PointNet [68], shown in Figure 3(b). The input denoised HOI $\boldsymbol{x}_t$ at time step t is linearly projected into the transformer combined with a standard positional embedding. A feed-forward network also maps the time step t to the same dimension. The decoder then maps these representations to the estimated clean HOI $\tilde{\boldsymbol{x}}$. $\mathcal{G}$ is optimized by the objective:

$$\mathcal{L}_r = \mathbb{E}_{t\sim[1,T]} \|\mathcal{G}(\boldsymbol{x}_t, t, \boldsymbol{c}) - \boldsymbol{x}\|_2^2. \tag{4}$$

We further disentangle this objective into rotation and translation losses for both human state $\boldsymbol{h}$ and object state $\boldsymbol{o}$, and re-weight these losses. We also introduce velocity regularizations, as detailed in the Supplementary.

3.2. Interaction Correction

Given that deep networks do not inherently model fundamental physical laws, generating plausible interactions even for state-of-the-art generative models trained on large-scale datasets can be challenging. Instead of relying on post-hoc optimization or physics simulation to promote physical validity, we *embed an interaction correction step within the diffusion framework*. This is motivated by the fact that the diffusion model produces intermediate HOI $\tilde{\boldsymbol{x}}$ at each diffusion step, allowing us to blend plausible dynamics into implausible regions and still generate seamless results. Remarkably, we achieve such plausible interactions with a simple *physics-informed* correction step. This is greatly attributed to the essential inductive bias induced – *e.g.*, even though human and object motion can be complicated, the relative object motion in an appropriate reference system follows a simple pattern that is easier to predict.

3.2.1 Correction Schedule

Similar to PhysDiff [110], we only consider performing corrections every few diffusion steps in late iterations, as early denoising iterations primarily produce noise with limited information.

For 3D HOIs represented by meshes, we set additional constraints based on geometric clues to determine the steps to apply corrections. As demonstrated in Algorithm 1, given *the current denoised HOI* $\tilde{\boldsymbol{x}}$, we first obtain the contact and penetration states in the future F frames. Let $\boldsymbol{v}_h \in \mathbb{R}^{F\times V_h\times 3}$ be the human vertices where V_h is the number of vertices, and $\mathbf{sdf}$ be a series of human body's signed distance fields [63] in the future F frames. Specifically, for

Algorithm 1 InterDiff: given a diffusion model $\mathcal{G}$, a correction scheduler $\mathcal{S}$, an interaction predictor $\mathcal{P}$, hyperparameters $\epsilon_1, \epsilon_2, \{\bar{\alpha}_t\}_{t=1}^T$

Input: condition $\boldsymbol{c}$
Output: the clean HOI $\boldsymbol{x}_0$ with correction
$\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for t from T to 0 **do**
    # Reverse Diffusion
    $\tilde{\boldsymbol{x}} \leftarrow \mathcal{G}(\boldsymbol{x}_t, t, \boldsymbol{c})$
    # Correction Schedule
    Obtain the contact and penetration states $\boldsymbol{C}, \boldsymbol{P}$
    if $\mathcal{S}(\boldsymbol{P}, \epsilon_1, \boldsymbol{C}, \epsilon_2, t)$ **then**
        Obtain the reference system s
        # Interaction Prediction
        $\hat{\boldsymbol{x}} \leftarrow \mathcal{P}(\tilde{\boldsymbol{x}}, s)$
        # Interaction Blending
        $\tilde{\boldsymbol{x}} \leftarrow \tilde{\boldsymbol{x}} \times \frac{t}{T} + \hat{\boldsymbol{x}} \times (1 - \frac{t}{T})$
    end if
    # Forward Diffusion
    $\boldsymbol{x}_{t-1} \sim \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}}\tilde{\boldsymbol{x}}, (1 - \bar{\alpha}_{t-1})\mathbf{I})$
end for
return $\boldsymbol{x}_0$

the SMPL representations [54], the vertices and sdf can be derived from the skinning function, using the body shape and pose parameters in $\tilde{\boldsymbol{x}}$ as input. We can also obtain the future sequence of object point clouds $\boldsymbol{v}_o \in \mathbb{R}^{F \times V_o \times 3}$ from the object state in the denoised HOI $\tilde{\boldsymbol{x}}$, where V_o is the number of object vertices. Based on the distance measurement, the contact state $\boldsymbol{C} \in \mathbb{R}^{F \times V_h}$ and the penetration state $\boldsymbol{P} \in \mathbb{R}^F$ are defined as,

$$\begin{aligned} \boldsymbol{C}^i[j] &= \min_{k=1,\dots,V_o} \|\boldsymbol{v}_h^i[j] - \boldsymbol{v}_o^i[k]\|_2, \ j = 1, \dots, V_h \\ \boldsymbol{P}^i &= \sum_{k=1,\dots,V_o} -\min\{\text{sdf}(\boldsymbol{v}_o^i[k]), 0\}, \end{aligned} \tag{5}$$

where $\boldsymbol{v}_h^i[j] \in \mathbb{R}^3$, $\boldsymbol{v}_o^i[k] \in \mathbb{R}^3$ are j-th and k-th vertex on human and object at frame $i \in \{H+1, \dots, H+F\}$, respectively.

The correction scheduler $\mathcal{S}$ serves two main functions. One is to determine whether the current denoised HOI $\tilde{\boldsymbol{x}}$ requires correction. Guided by the contact information from $\tilde{\boldsymbol{x}}$, we only perform correction when the diffusion model is likely to make a mistake – (i) *penetration already exists*, defined as $\|\boldsymbol{P}\| > \epsilon_1$; or (ii) *no contact happens*, defined as $\min_j \|\boldsymbol{C}[j]\| > \epsilon_2$. ϵ_1 and ϵ_2 are two hyperparameters. We only apply these constraints to mesh-represented HOIs, as the contact in skeletal HOIs is ill-defined.

The second function is to decide which reference system to use in Sec 3.2.2. We define a set of markers $\mathcal{M}$ [118] to index 67 human vertices as potential reference points for efficiency, instead of using all the V_h vertices. We operate on the contact state $\boldsymbol{C}$ and get the index of the reference system s, as follows:

$$s = \begin{cases} -1, & \text{if } \min\limits_{j \in \mathcal{M}} \|\boldsymbol{C}[j]\| \geq \epsilon_2 \\ \arg\min\limits_{j \in \mathcal{M}} \|\boldsymbol{C}[j]\|, & \text{o.w.} \end{cases} \tag{6}$$

This selection process means that we retain the default ground reference system if there is no contact; otherwise, we determine the reference point as the marker on the human body surface that is in contact with the object.

For skeletal HOIs, we follow a similar way to define the contact state $\boldsymbol{C} \in \mathbb{R}^{F \times J_h}$ for the purpose of capturing the reference system, despite its ill-posedness, as follows:

$$\boldsymbol{C}^i[j] = \min_{k=1,\dots,J_o} \|\boldsymbol{j}_h^i[j] - \boldsymbol{j}_o^i[k]\|_2, \ j = 1, \dots, J_h, \tag{7}$$

where $\boldsymbol{j}_h \in \mathbb{R}^{F \times J_h \times 3}$ represents J_h human joints and $\boldsymbol{j}_o \in \mathbb{R}^{F \times J_o \times 3}$ represents J_o object keypoints. We define the reference system s based on joints rather than markers:

$$s = \begin{cases} -1, & \text{if } \min\limits_{j=1,\dots,J_h} \|\boldsymbol{C}[j]\| \geq \epsilon_2 \\ \arg\min\limits_{j=1,\dots,J_h} \|\boldsymbol{C}[j]\|, & \text{o.w.} \end{cases} \tag{8}$$

3.2.2 Interaction Prediction

Given the past object motion and the trajectories of human markers/joints in both *past and future*, we now predict future object motions under different references. We first apply coordinate transformations to the past object motion (which is by default under the ground reference system) and obtain relative motions with respect to *all* markers/joints. Then we formulate the object motions, either under the ground reference system or relative to each marker/joint, *collectively* as a spatial-temporal graph $\boldsymbol{G}^{1:H}$. For example, given $|\mathcal{M}|$ markers, we define $\boldsymbol{G}^{1:H} \in \mathbb{R}^{H \times (1+|\mathcal{M}|) \times D_o}$, where D_o is the number of features for object poses as defined previously. Here, $1 + |\mathcal{M}|$ correspond to 1 ground reference system and $|\mathcal{M}|$ marker-based reference systems.

Given the spatial-temporal graph $\boldsymbol{G}^{1:H}$, we use a spatial-temporal graph neural network (STGNN) [103] to process the past motion graph $\boldsymbol{G}^{1:H}$ and obtain $\boldsymbol{G}^{H:H+F}$ that represents future object motions in these systems. Then, we acquire a specific object relative motion $\boldsymbol{G}^{H:H+F}[s+1]$, under the reference system s specified in Sec. 3.2.1. We transform this predicted reference motion back to the ground system. The resulting object motion is defined as $\hat{\boldsymbol{x}} = \mathcal{P}(\tilde{\boldsymbol{x}}, s)$, where the interaction predictor $\mathcal{P}$ performs the above operations. We blend the original denoised HOI $\tilde{\boldsymbol{x}}$ with this newly obtained HOI $\hat{\boldsymbol{x}}$, denoted as $\tilde{\boldsymbol{x}} \times \frac{t}{T} + \hat{\boldsymbol{x}} \times (1 - \frac{t}{T})$.

Informed by the reference system, we argue that the motion after coordinate transformation follows a simpler pattern and becomes easier for the network to predict and

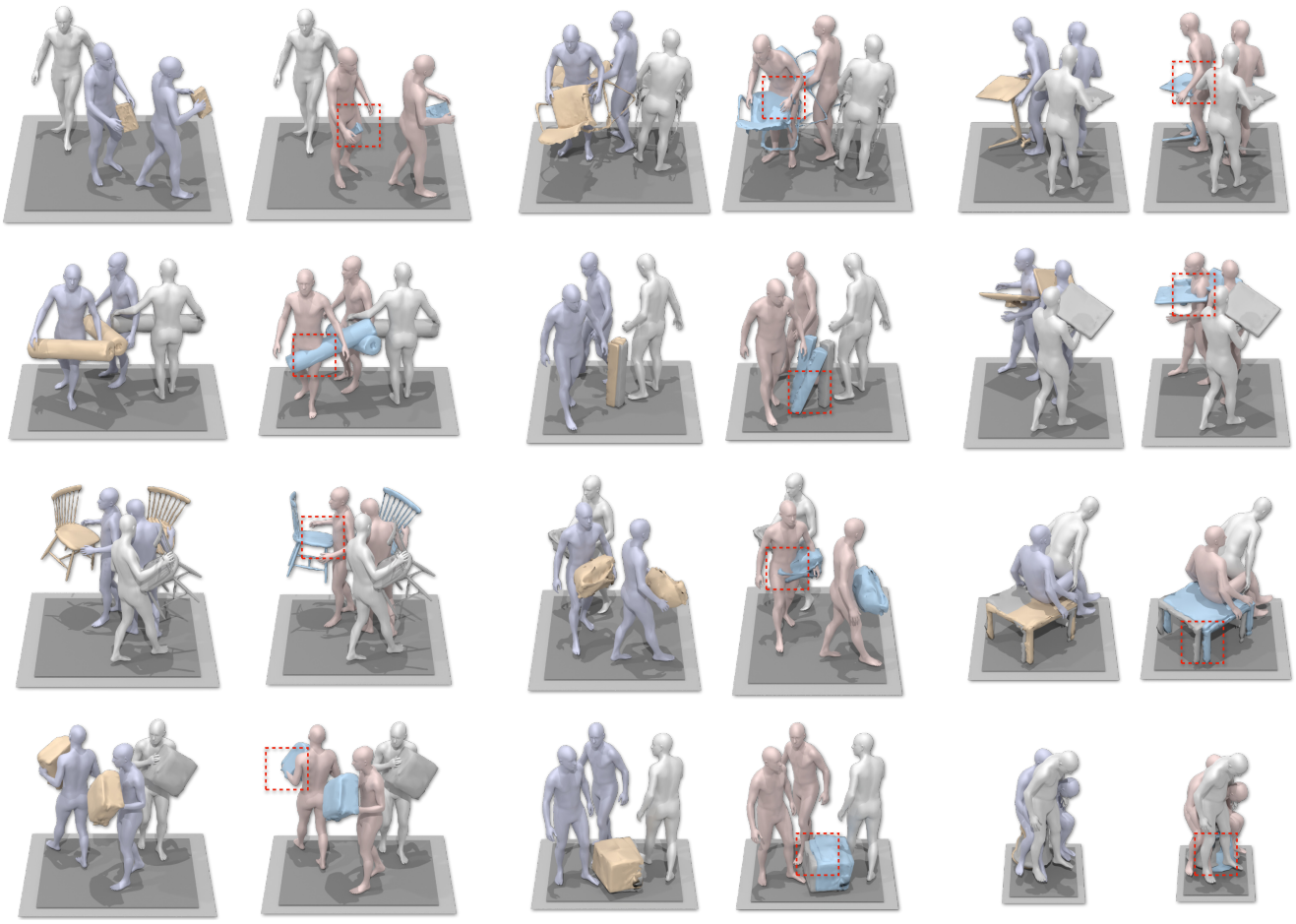


Figure 4. **Qualitative comparisons** on the BEHAVE dataset [5]. We show starting HOIs in gray and predicted HOIs sampled every 40 frames (30 FPS). The blue and red human meshes denote the results from InterDiff with and without interaction correction, respectively. The injected correction step helps mitigate contact floating and penetration artifacts, and maintain static objects when there is no contact.

maintain physical validity. To further promote the simple motion pattern, in this STGNN, we use DCT/IDCT [1] as a preprocessing step in accordance with [59]. We also find that a small number of frequency bases work well for predicting relative object motion. To decouple STGNN training from diffusion, we directly use *clean HOI* data for training and perform inference on *denoised HOIs*. We introduce learning objectives to promote contact and penalize penetration, which are detailed in the Supplementary.

4. Experiments

4.1. Experimental Setup

Datasets. We conduct the evaluation on three datasets pertaining to 3D human-object interaction. BEHAVE [5] encompasses recordings of 8 individuals engaging with 20 ordinary objects at a frame rate of 30 Hz. SMPL-H [54, 74] is used to represent the human, and we represent the object pose in 6D rotation [126] and translation. We adhere to the official train and test split originally proposed in HOI detection [5]. During training, our model forecasts 25 future frames after being provided with 10 past frames, and we can generate longer motion sequences autoregressively during inference. GRAB [82] is a dataset that records whole-body humans grasping small objects, including 1,334 videos, which we downsample to 30 FPS. We investigate the *cross-dataset generalizability* – we train our method on the BEHAVE dataset and test it on the GRAB dataset. The Human-Object Interaction dataset [88] comprises 6 individuals and 384,000 frames recorded at a framerate of 240 Hz. We follow the official data preprocess-

Table 1. **Quantitative results** on the BEHAVE dataset [5], demonstrating the effectiveness of our diffusion model and the correction.

Method	BEHAVE [5]			
	MPJPE-H ↓	Trans. Err. ↓	Rot. Err. ↓	Pene. ↓
InterRNN	165	139	267	314
InterVAE	145	125	268	222
InterDiff w/o correction (**Ours**)	**140**	**123**	256	228
InterDiff (full) (**Ours**)	**140**	**123**	**226**	**164**

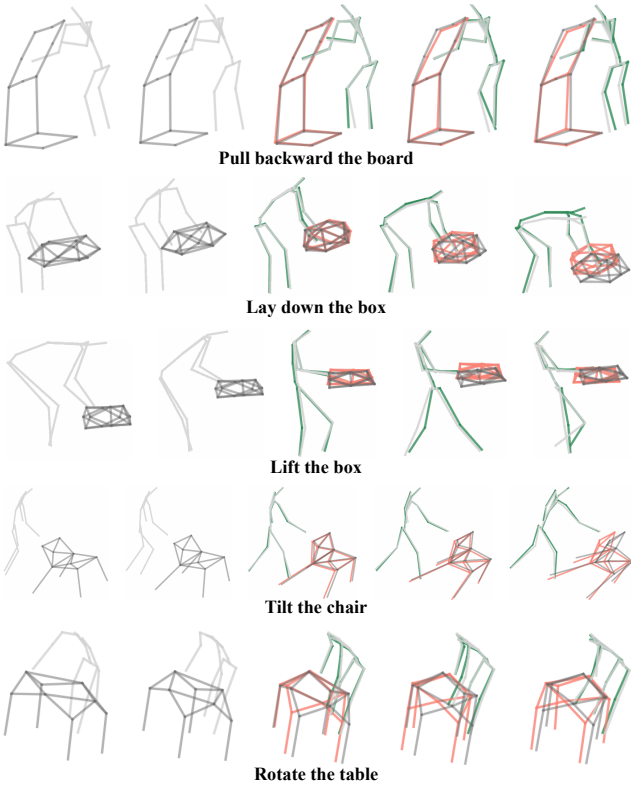


Figure 5. **Qualitative results** on interactions with *unseen objects* on the Human-Object Interaction dataset [88]. The predicted skeletons and objects are green and red respectively while GT is gray. We show five frames at 0.4, 0.8, 1.2, 1.6, and 2.0s.

Table 2. **Quantitative results** on the Human-Object Interaction dataset [88]. We evaluate our model in challenging scenarios with *unseen instances* in the training data. The results show the effectiveness and generalizability of InterDiff and the correction. * marks results directly reported from [88].

Method	MPJPE-H ↓	MPJPE-O ↓	Trans. Err. ↓	Rot. Err. ↓
HO-GCN* [88]	111	153	123	303
CAHMP* [13]	107	167	N/A	N/A
InterRNN	124	127	109	151
CAHMP [13]	111	132	111	164
InterVAE	108	125	100	178
InterDiff w/o Correction (**Ours**)	**105**	117	92	158
InterDiff w/ Correction (**Ours**)	**105**	**84**	**60**	**120**

ing guidelines and extract the sequences at a rate of 10 Hz. In total, 18,352 interactive sequences with a length of 20 frames are obtained, and 582 of these sequences include objects that are not seen during training and are directly used for evaluation. Following [88], our model is trained to forecast 10 future frames given 10 past frames. Unlike the above two datasets, we employ a 21-joint skeleton to represent the human pose, and 12 key points for objects.

Metrics. Based on the established evaluation metrics in the literature [28, 33, 41, 88], we introduce a set of metrics to evaluate this new task as follows: (a) **MPJPE-H**: the average l_2 distance between the predicted and ground truth joint positions in the global space. For SMPL-represented HOIs, joint positions can be obtained through forward kinematics. (b) **Trans. Err.**: the average l_2 distance between the predicted and ground truth object translations. (c) **Rot. Err.**: the average l_1 distance between the predicted and ground truth quaternions of the object. (d) **MPJPE-O**: the average l_2 distance between the predicted and ground truth object key points in the global space. This is only reported for the Human-Object Interaction dataset, where the object is abstracted into key points. (e) **Pene.**: the average percentage of object vertices with non-negative human signed distance function [63] values. Note that (a)(b)(d) are in mm, (c) is in 10^{-3} radius, and (e) is in $10^{-2}\%$. In Table 3, to evaluate diverse predictions, we sample multiple candidate predictions for each historical motion, and report the best results (Best-of-Many [6]) over the candidates for each metric.

Table 3. **Quantitative results** on the BEHAVE dataset [5]. We generate *multiple predictions* and report the lowest score across different samples. Here we focus on long-term forecasting, where we *autoregressively generate 100 frames* of future interactions. Our method with interaction correction outperforms pure diffusion, and the improvement is more significant with more samples.

# of samples	InterDiff (ours)	Best-of-Many			
		MPJPE-H ↓	Trans. Err. ↓	Rot. Err. ↓	Pene. ↓
1	w/o correction	400	384	644	236
	full	**392**	**374**	**632**	**88**
2	w/o correction	382	365	636	211
	full	**374**	**350**	**601**	**83**
5	w/o correction	371	349	610	191
	full	**361**	**331**	**545**	**65**
10	w/o correction	361	341	601	187
	full	**348**	**318**	**523**	**59**

Baselines. As our work introduces a new task, a baseline directly from prior research is not readily available. To facilitate comparisons with existing work, we adapt the following baselines from tasks of *human motion generation* and *object dynamic prediction*. (a) **InterVAE**: transformer-based VAEs [65, 66, 87] have been widely adopted for human motion prediction and synthesis. We employ this framework and extend it to our human-object interaction setting. (b) **InterRNN**: we adopt an LSTM-based [30, 72] predictor and enable the prediction of HOIs. For skeletal representations, we further include (c) **CAHMP** [13]: as there is no publicly available codebase, we implemented their methods according to the paper; (d) **HO-GCN** [88]: since the code for the methods is not yet available, we report the results in their paper, as shown in Table 2, marked with *.

Implementation Details. The interaction diffusion model comprises 8 transformer [87] layers for the encoder and decoder, with training involving a batch size of 32, a latent dimension of 256, and 500 epochs. The interaction predictor includes 10 frequency bases for DCT/IDCT [1], with training conducted using a batch size of 32, and 500 epochs. For the Human-Object Interaction dataset, we do not apply contact and penetration losses since they are not applicable for

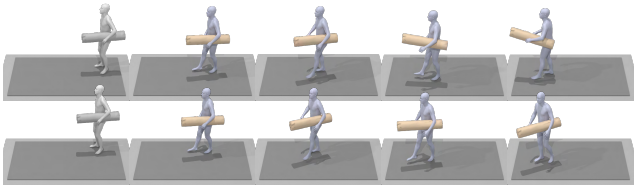


Figure 6. **Qualitative results** on the BEHAVE dataset [5]. We place two different samples of the predicted interactions. Our approach can generate diverse and legitimate predictions.

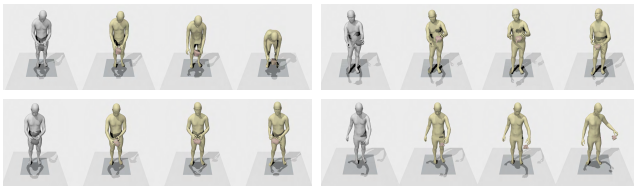


Figure 7. **Generalization** of InterDiff on the GRAB dataset [82]. The predicted human bodies and objects are in color while the past interactions are in gray. We visualize four frames of each sequence at 0, 0.33, 0.66, and 1.0s. Our approach can directly generalize to this different dataset containing novel small-size objects.

Table 4. **User study** on the BEHAVE dataset [5]. We obtain pairwise human voting results comparing our method with baselines and alternatives introduced in Sec. 4.4. Under human evaluation, the full model outperforms baselines regarding physical fidelity.

Model pair	Physical fidality			
	ground truth	InterDiff (full)	w/o correction	w/o relative
ground truth	N/A	73.0%	69.6%	88.8%
InterDiff (full)	27.0%	N/A	**67.8%**	**67.8%**
w/o correction	30.4%	32.2%	N/A	67.2%
w/o relative	11.2%	32.2%	32.8%	N/A

skeleton representation. For autoregressive inference, we use predicted last few frames as the past motion and generate the next prediction, *etc*. Additional implementation details are provided in the Supplementary.

4.2. Quantitative Results

We compare our proposed method InterDiff with two baselines on the BEHAVE (Table 1) and Human-Object Interaction (Table 2) datasets. We demonstrate the superiority of InterDiff over the two baselines in all metrics for both datasets. Moreover, we observe and validate that the performance of InterDiff is improved by incorporating interaction correction. Specifically, the correction step results in more plausible interactions with reduced penetration artifacts, as demonstrated in Table 1. Additionally, InterDiff with interaction correction provides more precise object motions. In Table 3, following standard Best-of-Many evaluation [6], we demonstrate that as more predictions are sampled, the best predictions are closer to the ground truth. Note that our full method shows a significant improvement over pure diffusion with more samples and longer horizons. Furthermore, the precise object motion generated by our InterDiff in turn improves the accuracy of predicted human motion even with the same interaction diffusion model, as evidenced in Table 3. The reason behind this is that by injecting more accurate object motion into the diffusion model, human motion generated by the diffusion is also positively affected and can be corrected.

4.3. Qualitative Results

Consistent with the quantitative results, we observe that our approach InterDiff with the interaction correction yields more plausible HOI predictions than the one without interaction correction across various cases, as illustrated in Figure 4. Furthermore, the efficacy of our method extends to the Human-Object Interaction dataset (Figure 5), showing that our approach accommodates the skeletal representation and effectively predicts future HOIs across *a diverse range of actions and unseen objects* with convincing outcomes. In Figure 7, we generalize our method trained on the BEHAVE dataset to the GRAB dataset. Our approach effectively adapts to the new dataset that focuses on grasping small objects *without any fine-tuning*, further validating the generalizability of our approach. Figure 6 illustrates that our approach can generate diverse and legitimate HOIs. We provide additional demo videos on the project website.

To assess the motion plausibility, we conduct a double-blind user study, as shown in Table 4. We design pairwise evaluations between ground truth, InterDiff (full), InterDiff without interaction correction ('w/o correction'), and InterDiff with the correction step yet not having coordinate transformation involved ('w/o relative'). We generate 100 frames of future interactions for comparisons. With a total of 30 pairwise comparisons, 23 human judges are asked to determine which interaction is more realistic. Our method has a success rate of 67.8% against baselines.

4.4. Ablation Study

We conduct an ablation study (Figures 8 and 9) to evaluate the efficacy of different components in our proposed interaction correction (Sec. 3.2). Figure 8 displays smooth long-term sequences generated by each ablated variant. We show that our full model, InterDiff, produces plausible long-term HOIs, while InterDiff without the correction step results in contact floating and penetration artifacts. Then we ablate on each component inside the correction step. In the absence of reference system transformation ('w/o relative'), the object motion integrated into the diffusion fails to align with the contact's motion, resulting in significant contact penetration, especially in the long term. This highlights the critical role of the reference system in the interaction correction step. Furthermore, removing contact and penetration losses ('w/o contact') also leads to unrealistic outcomes. Finally, blindly applying correction without considering the quality of intermediate denoised results ('w/o

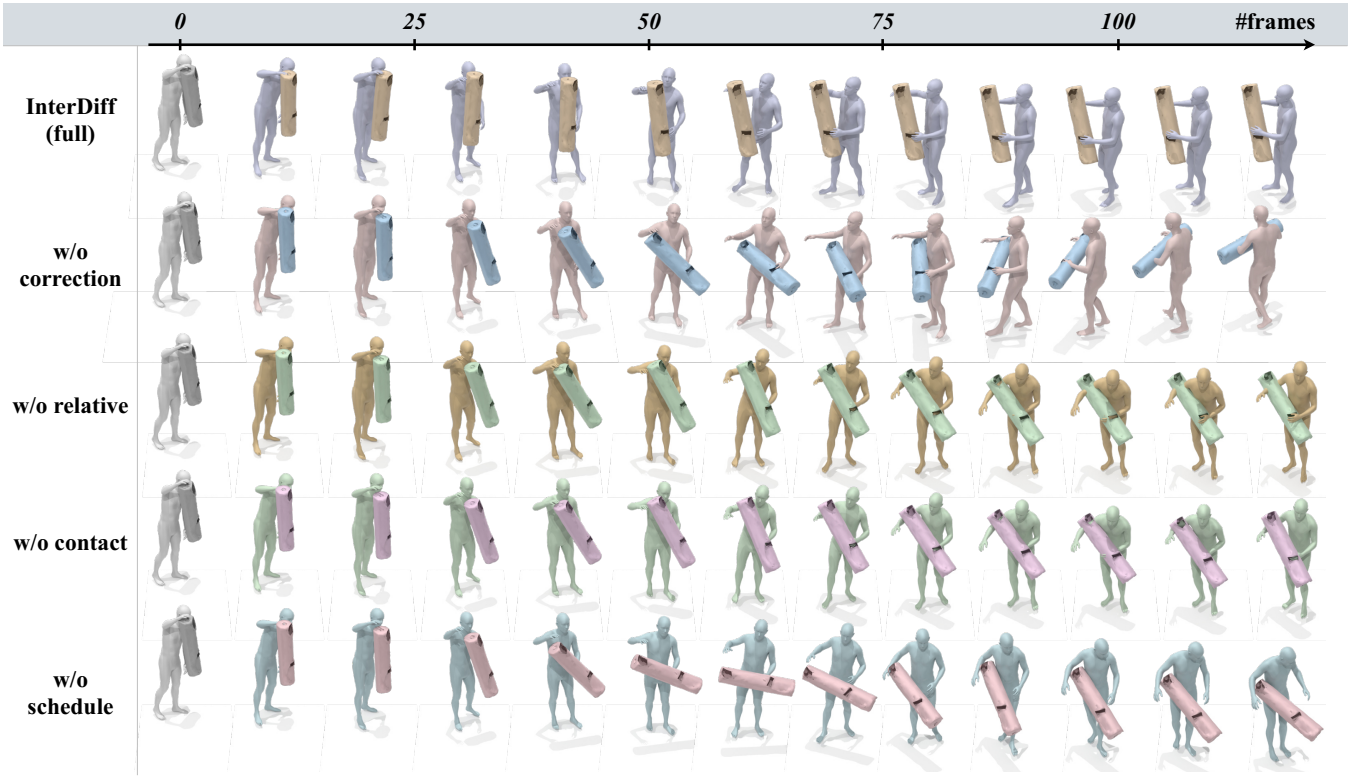


Figure 8. **Ablation study** on the BEHAVE dataset [5]. We show starting HOIs in gray and predicted HOIs sampled every 10 frames (30 FPS), up to 4 seconds. The ablated variants of our InterDiff produce HOIs containing contact floating and penetration artifacts.

schedule') may lead to the contact floating, as applying correction may destabilize the good quality of the original motion. In Figure 9, we further provide quantitative evidence of the effectiveness of our method, where the full model significantly outperforms the variants, especially in correcting the accumulation of errors in long-term autoregressive generation. Additional ablations are available in the Supplementary, including an evaluation of the effectiveness of DCT/IDCT in promoting simple motion patterns.

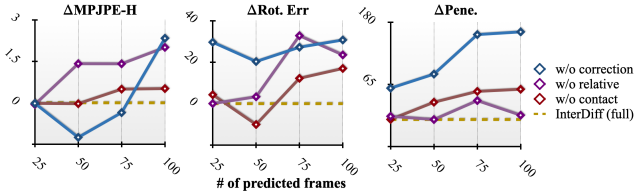


Figure 9. **Ablation study** on the BEHAVE dataset [5]. We compare our pipeline with various alternatives introduced in Sec. 4.4. We normalize the scores of 'full model' to 0. The results shows the superiority of 'full model' over others in the long horizon.

5. Discussions

We propose a novel task, coined as 3D human-object interaction prediction, considering the intricate real-world challenges associated with this domain. To ensure the validity of physical interactions, we introduce an interaction diffusion framework, InterDiff, which effectively generates vivid interactions while simultaneously reducing common artifacts such as contact floating and penetration, with minimal additional computational cost. Our approach shows effectiveness in this novel task and thus holds significant potential for a wide range of real-world applications. A future direction would be generalizing our work to human interaction with more complex environments, such as with more than one dynamic object, more complicated objects, *e.g.*, articulated and deformable objects, and with other humans.

Limitations. We've demonstrated that our InterDiff framework is able to produce high-quality and diverse HOI predictions, without the use of post-optimization and physics simulators. Artifacts such as contact inconsistency are still observed in some generated results, though the artifacts are largely alleviated by interaction correction. Nonetheless, InterDiff with correction provides effective results that can be directly applied post-optimization to improve quality. More illustrations are available on the project website.

Acknowledgement. This work was supported in part by NSF Grant 2106825, NIFA Award 2020-67021-32799, the Jump ARCHES endowment, the NCSA Fellows program, the Illinois-Insper Partnership, and the Amazon Research Award. This work used NVIDIA GPUs at NCSA Delta through allocations CIS220014 and CIS230012 from the ACCESS program.

References

[1] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 1974. 6, 7

[2] Hyemin Ahn and Dongheui Mascaro, Valls Esteve an Lee. Can we use diffusion probabilistic models for 3d motion prediction? In *ICRA*, 2023. 3

[3] Jinseok Bae, Jungdam Won, Donggeun Lim, Cheol-Hui Min, and Young Min Kim. Pmp: Learning to physically interact with environments using part-wise motion priors. In *SIGGRAPH*, 2023. 2, 3

[4] German Barquero, Sergio Escalera, and Cristina Palmero. BeLFusion: Latent diffusion for behavior-driven human motion prediction. In *ICCV*, 2023. 3, 4

[5] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *CVPR*, 2022. 2, 3, 6, 7, 8, 9

[6] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Accurate and diverse sampling of sequences based on a "best of many" sample objective. In *CVPR*, 2018. 3, 7, 8

[7] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. ContactGrasp: Functional multi-finger grasp synthesis from contact. In *IROS*, 2019. 3

[8] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020. 3

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

[10] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. HumanMAC: Masked motion completion for human motion prediction. In *ICCV*, 2023. 3

[11] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, 2023. 3

[12] Yixin Chen, Sai Kumar Dwivedi, Michael J Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *CVPR*, 2023. 3

[13] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *CVPR*, 2020. 1, 3, 7

[14] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, 2020. 3

[15] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. MoFusion: A framework for denoising-diffusion-based motion synthesis. In *CVPR*, pages 9760–9770, 2023. 3

[16] Renaud Detry, Dirk Kraft, Anders Glent Buch, Norbert Krüger, and Justus Piater. Refining grasp affordance models by experience. In *ICRA*, 2010. 3

[17] Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, M. J. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016. 3

[18] Danny Driess, Zhiao Huang, Yunzhu Li, Russ Tedrake, and Marc Toussaint. Learning multi-object dynamics with compositional neural radiance fields. *arXiv preprint arXiv:2202.11855*, 2022. 3

[19] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023. 3

[20] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. IMoS: Intent-driven full-body motion synthesis for human-object interactions. *arXiv preprint arXiv:2212.07555*, 2022. 1, 2, 3

[21] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018. 3

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3

[23] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. ContactOpt: Optimizing contact to improve grasps. In *CVPR*, 2021. 3

[24] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R. Venkatesh Babu. DeLiGAN : Generative adversarial networks for diverse and limited data. In *CVPR*, 2017. 3

[25] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021. 3

[26] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *CVPR*, 2021. 3

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

[28] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *ACCV*, 2013. 7

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 4

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997. 7

[31] Zhi Hou, Baosheng Yu, and Dacheng Tao. Compositional 3d human-object neural animation. *arXiv preprint arXiv:2304.14070*, 2023. 3

[32] Kaijen Hsiao and Tomas Lozano-Perez. Imitation learning of whole-body grasps. In *international conference on intelligent robots and systems*, 2006. 3

[33] Heng-Wei Hsu, Tung-Yu Wu, Sheng Wan, Wing Hung Wong, and Chen-Yi Lee. QuatNet: Quaternion-based head pose estimation with multiregression loss. *IEEE Transactions on Multimedia*, 2019. 7

[34] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *CVPR*, 2023. 3

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

[36] Jingwei Ji, Rishi Desai, and Juan Carlos Niebles. Detecting human-object relationships in videos. In *ICCV*, 2021. 3

[37] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, 2021. 3

[38] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. CHAIRS: Towards full-body articulated human-object interaction. *arXiv preprint arXiv:2212.10621*, 2022. 3

[39] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. GMD: Controllable human motion synthesis via guided diffusion models. *arXiv preprint arXiv:2305.12577*, 2023. 3

[40] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *3DV*, 2020. 3

[41] Manuel Kaufmann, Yi Zhao, Chengcheng Tang, Lingling Tao, Christopher Twigg, Jie Song, Robert Wang, and Otmar Hilliges. EM-POSE: 3d human pose estimation from sparse electromagnetic trackers. In *ICCV*, 2021. 7

[42] Taeksoo Kim, Shunsuke Saito, and Hanbyul Joo. NCHO: Unsupervised learning for neural 3d composition of humans and objects. In *ICCV*, 2023. 3

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

[44] Franziska Krebs, Andre Meixner, Isabel Patzer, and Tamim Asfour. The kit bimanual manipulation dataset. In *Humanoids*, 2021. 3

[45] Paul G Kry and Dinesh K Pai. Interaction capture and synthesis. *ACM Transactions on Graphics*, 25(3):872–880, 2006. 3

[46] 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. *arXiv preprint arXiv:2307.07511*, 2023. 1, 3

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

[48] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. *arXiv preprint arXiv:2303.13129*, 2023. 3

[49] Ying Li, Jiaxin L Fu, and Nancy S Pollard. Data-driven grasp synthesis using shape matching and task-based pruning. *IEEE Transactions on visualization and computer graphics*, 13(4):732–747, 2007. 3

[50] Libin Liu and Jessica Hodgins. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2, 3

[51] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *ECCV*, 2020. 1, 3

[52] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 2

[53] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *CVPR*, 2022. 1, 3

[54] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics*, 2015. 3, 5, 6

[55] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *ICAR*, 2015. 3

[56] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. Unifying representations and large-scale whole-body motion databases for studying human motion. *IEEE Transactions on Robotics*, 32(4):796–809, 2016. 3

[57] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *CVPR*, 2021. 3

[58] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Weakly-supervised action transition learning for stochastic human motion prediction. In *CVPR*, 2022. 3

[59] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *ICCV*, 2019. 6

[60] Josh Merel, Saran Tunyasuvunakool, Arun Ahuja, Yuval Tassa, Leonard Hasenclever, Vu Pham, Tom Erez, Greg Wayne, and Nicolas Heess. Catch & carry: reusable neural controllers for vision-guided whole-body tasks. *ACM Transactions on Graphics (TOG)*, 39(4):39–1, 2020. 2, 3

[61] Damian Mrowca, Chengxu Zhuang, Elias Wang, Nick Haber, Li F Fei-Fei, Josh Tenenbaum, and Daniel L Yamins. Flexible neural representation for physics prediction. In *NeurIPS*, 2018. 3

[62] Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes. *arXiv preprint arXiv:2308.09036*, 2023. 3

[63] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 4, 7

[64] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *CVPR*, 2023. 3

[65] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, 2021. 3, 7

[66] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, 2022. 3, 7

[67] Nancy S Pollard and Victor Brian Zordan. Physically based grasping control from example. In *SIGGRAPH*, 2005. 3

[68] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 4

[69] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or. Single motion diffusion. *arXiv preprint arXiv:2302.05905*, 2023. 3, 4

[70] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. 4

[71] Haziq Razali and Yiannis Demiris. Action-conditioned generation of bimanual object manipulation sequences. In *AAAI*, 2023. 1, 3

[72] Davis Rempe, Srinath Sridhar, He Wang, and Leonidas J. Guibas. Predicting the physical dynamics of unseen 3d objects. In *WACV*, 2020. 3, 7

[73] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICLR*, 2015. 3

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

[75] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 3

[76] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2, 3

[77] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

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

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

[80] Jiarui Sun and Girish Chowdhary. Towards globally consistent stochastic human motion prediction via motion diffusion. *arXiv preprint arXiv:2305.12554*, 2023. 3

[81] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. GOAL: Generating 4d whole-body motion for hand-object grasping. In *CVPR*, 2022. 1, 2, 3

[82] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 3, 6, 8

[83] Purva Tendulkar, Dídac Surís, and Carl Vondrick. FLEX: Full-body grasping without full-body grasps. In *CVPR*, 2023. 3

[84] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. MotionCLIP: Exposing human motion generation to CLIP space. *arXiv preprint arXiv:2203.08063*, 2022. 3

[85] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 2, 3, 4

[86] Sibo Tian, Minghui Zheng, and Xiao Liang. TransFusion: A practical and effective transformer-based diffusion model for 3d human motion prediction. *arXiv preprint arXiv:2307.16106*, 2023. 3

[87] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 7

[88] Weilin Wan, Lei Yang, Lingjie Liu, Zhuoying Zhang, Ruixing Jia, Yi-King Choi, Jia Pan, Christian Theobalt, Taku Komura, and Wenping Wang. Learn to predict how humans manipulate large-sized objects from interactive motions. *IEEE Robotics and Automation Letters*, 2022. 1, 3, 6, 7

[89] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *CVPR*, 2022. 3

[90] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *CVPR*, 2021. 3

[91] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *CVPR*, 2021. 3

[92] Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *3DV*, 2022. 3

[93] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3d scenes. In *NeurIPS*, 2022. 3

[94] Dong Wei, Huaijiang Sun, Bin Li, Jianfeng Lu, Weiqing Li, Xiaoning Sun, and Shengxiang Hu. Human joint kinematics diffusion-refinement for stochastic motion prediction. In *AAAI*, 2023. 4

[95] Dong Wei, Xiaoning Sun, Huaijiang Sun, Bin Li, Shengxiang Hu, Weiqing Li, and Jianfeng Lu. Understanding text-driven motion synthesis with keyframe collaboration via diffusion models. *arXiv preprint arXiv:2305.13773*, 2023. 3

[96] Xiaoqian Wu, Yong-Lu Li, Xinpeng Liu, Junyi Zhang, Yuzhe Wu, and Cewu Lu. Mining cross-person cues for body-part interactiveness learning in hoi detection. In *ECCV*, 2022. 3

[97] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. SAGA: Stochastic whole-body grasping with contact. In *ECCV*, 2022. 1, 2, 3

[98] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *ECCV*, 2022. 3

[99] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *CVPR*, 2023. 3

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

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

[102] Sirui Xu, Yu-Xiong Wang, and Liangyan Gui. Stochastic multi-person 3d motion forecasting. In *ICLR*, 2023. 3

[103] Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *ECCV*, 2022. 3, 5

[104] Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. D3D-HOI: Dynamic 3d human-object interactions from videos. *arXiv preprint arXiv:2108.08420*, 2021. 3

[105] Zeshi Yang, Kangkang Yin, and Libin Liu. Learning to use chopsticks in diverse gripping styles. *ACM Transactions on Graphics (TOG)*, 41(4):1–17, 2022. 3

[106] 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*, 2023. 3

[107] Yufei Ye, Maneesh Singh, Abhinav Gupta, and Shubham Tulsiani. Compositional video prediction. In *ICCV*, 2019. 3

[108] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967*, 2019. 3

[109] Ye Yuan and Kris Kitani. DLow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, 2020. 3

[110] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-guided human motion diffusion model. In *ICCV*, 2023. 3, 4

[111] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. ManipNet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics*, 40(4):1–14, 2021. 3

[112] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. NeuralDome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023. 3

[113] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023. 3

[114] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020. 3

[115] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 3, 4

[116] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. ReMoDiffuse: Retrieval-augmented motion diffusion model. In *ICCV*, 2023. 3

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

[118] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *CVPR*, 2021. 3, 5

[119] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *CVPR*, 2022. 3

[120] Zihan Zhang, Richard Liu, Kfir Aberman, and Rana Hanocka. TEDi: Temporally-entangled diffusion for long-term motion synthesis. *arXiv preprint arXiv:2307.15042*, 2023. 3

[121] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *ECCV*, 2022. 3

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

[123] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. CAMS: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *CVPR*, 2023. 3

[124] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *CVPR*, 2022. 3

[125] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *ECCV*, 2022. 3

[126] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 6

[127] Fangrui Zhu, Yiming Xie, Weidi Xie, and Huaizu Jiang. Diagnosing human-object interaction detectors. *arXiv preprint arXiv:2308.08529*, 2023. 3

[128] Guangxiang Zhu, Zhiao Huang, and Chongjie Zhang. Object-oriented dynamics predictor. In *NeurIPS*, 2018. 3