This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

InterPhys: Physics-aware Human Motion Synthesis in a Dynamic Scene

Chaoyue Xing Wei Mao Miaomiao Liu
Australian National University, Canberra, Australia
{chaoyue.xing, miaomiao.liu}@anu.edu.au; wei.mao.research@gmail.com

Abstract

This paper tackles the problem of physics-aware human motion synthesis in a dynamic scene. Unlike existing works which mainly tend to generate physically unrealistic motions due to limited contact modeling, typically restricted to hands, in this paper, we introduce a physics-aware human motion generation framework that explicitly models the full spectrum of human-related forces, including human-object, human-scene, and internal body dynamics. Our method imposes soft physical constraints to maintain force and torque balance, ensuring physically grounded motion synthesis. We further propose a novel continuous distance-based force model that generalizes contact modeling to arbitrary surfaces, capturing interactions not only with static environments but also with dynamic, moving objects. Extensive experiments show that our approach significantly improves physical plausibility and generalizes well to complex scenes, setting a new benchmark for physically consistent human motion generation.

1. Introduction

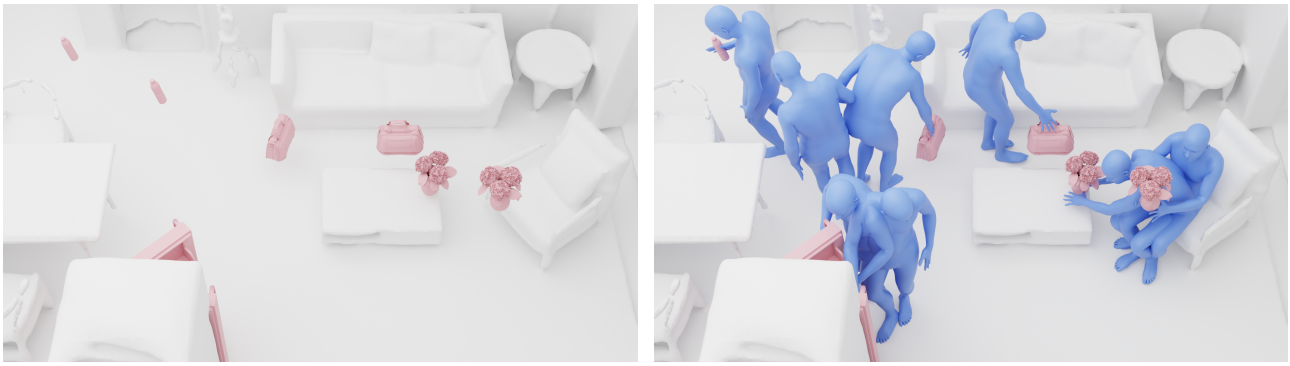
Human motion synthesis is essential to the success of many applications like VR/AR, animation, and embodied AI, and has witnessed significant progress in recent years [1, 4, 9–11, 19, 24, 26, 34, 35, 37, 38, 42, 55–57]. However, these methods overlook a key aspect of human activity, namely, interaction with the surroundings. In this paper, we address the problem of synthesizing human motion interacting with dynamic scenes i.e., scenes with a moving object, a more realistic setting for real applications, as illustrated in Fig. 1.

Although recent advances in generative models have inspired diffusion-based methods [7, 18, 21, 22, 27, 51, 52] to explore synthesizing human motion that interacts with dynamic objects, these methods often rely on simple interaction priors, such as contact or penetration constraints [7, 22, 51, 53], to encourage interaction. While effective in improving the quality of motion synthesis, such works do not accurately model the true physical forces, resulting in artifacts like floating, foot sliding, or unrealistic contact.

To enable physically valid human motion generation, physics simulators and reinforcement learning [14, 29, 44, 46, 49, 54] offer another direction by incorporating physical constraints to produce more realistic results. However, they rely on non-differentiable simulators which prevent seamless integration with end-to-end generative pipelines. To achieve physically plausible human motion in dynamic scenes within an end-to-end framework, PhysPT [58] uses a continuous contact model for motion estimation. However, PhysPT [58] is limited to motion estimation on static planar surfaces, assuming fixed surface normals and decoupled spring systems for normal and tangential forces, which cannot generalize to arbitrary geometries or capture the coupling between normal force and friction.

To overcome these limitations, we propose a novel continuous contact force model that can generalize to arbitrary, dynamic 3D surfaces and better reflects real-world contact physics. Specifically, our model decomposes the contact force into normal and tangential components. For the normal force, inspired by [3, 58], we use a damped spring system. Instead of assuming a fixed upward-facing surface normal (opposite to the direction of gravity), we align it with the local surface geometry to handle arbitrary contact surfaces. We model the tangential force with separate static and kinetic components, with static friction proportional to lateral acceleration and kinetic friction proportional to the normal force. This formulation ensures that normal and frictional forces are coupled and compatible with gradient-based optimization, enabling physically consistent human–object interactions in dynamic scenes.

Given our defined continuous contact force models, we formulate human's dynamics with Euler–Lagrange equations. Specifically, we model the human's interaction not only with static scene but also dynamic objects. Crucially, unlike prior works [21, 22, 51, 58], we explicitly formulate the motion dynamics of the moving object with Euler–Lagrange equations. Following Newton's third law, the force applied by the human to the object is equal and opposite to the reaction force applied by the object to the human. This allows us to integrate the object's dynamics directly



(a) (b)

Figure 1. **Our Task.** Our method takes 3D object motion and a 3D scene as input (a), to synthesize physically consistent 3D human motion interacting with both the moving object and the static background scene (b).

into that of the human, yielding a unified formulation where the human motion is constrained not only by the scene but also by the dynamics of the moving object.

To achieve physics-aware human motion synthesis, we thus propose a two-stage pipeline to integrate our proposed continuous contact model in an end-to-end learning framework for human motion synthesis. In the first stage, we propose to utilize a diffusion model to predict parameters for our contact force model and the human hand trajectory given the static scene and a moving object. They are further used as conditions to generate human motion with another diffusion model in the second stage. We introduce a loss defined by the Euler-Lagrange equations with our continuous contact model to encourage the physical consistency between the generated motion and the dynamic scene. Our contributions can be summarised as follows:

- We introduce a novel continuous contact force model that accurately captures real contact forces and generalizes to arbitrary surfaces and geometries.
- We explicitly model dynamics of moving objects and integrate them into the human dynamics formulation based on Newton's third law, ensuring that reciprocal forces between human and object are consistently preserved.
- We propose a two-stage diffusion-based pipeline for scene-aware human motion synthesis that first predicts parameters for our new continuous contact model and then generates physics-aware human motion.

We evaluate our method on OMOMO dataset [21] and TRUMANS dataset [18] to demonstrate that our approach can achieve state-of-the-art performance and physically plausible results.

2. Related Work

Human-Scene and Human-Object Interaction. The emergence of paired scene-aware [12, 18, 21] and object-aware [16, 22, 53] datasets has fueled interest in synthesizing human motion that interacts realistically with scenes and objects. Existing methods [13, 28, 43, 45, 50] generate scene-aware human motion by leveraging geometric constraints such as collision avoidance, contact priors, and distance field. However, they do not explicitly model interaction physics, often resulting in artifacts such as floating or foot skating. RL-based approaches [14, 29, 32, 49] improved realism but are task-specific, hard to generalize, and non-differentiable. Human-object interaction synthesis initially targeted small, handheld objects, but later expanded to full-body datasets [18, 21, 23, 52]. Recent works [2, 17, 18, 21, 22, 51] incorporated contact or penetration priors to enable dynamic interactions, yet they cannot predict forces, often leading to unrealistic artifacts. RL-based policies [14, 54] have been applied to larger objects but still lack generalization. While some works [8, 36, 40, 41, 58, 59] incorporate differentiable physics into human modeling via body shape conditioning, biomechanical stability, or Lagrangian formulations, they primarily focus on body-level dynamics and interactions with a fixed ground plane, and do not explicitly model interaction forces with external objects or complex scenes.

In contrast, we propose a physics-aware human motion synthesis framework that jointly models full-body motion and physically plausible contact forces in cluttered 3D environments with dynamic objects. Our method integrates a physically grounded, differentiable continuous contact force model, generalizable to arbitrary surfaces, into a two-stage diffusion pipeline: predicting physics parameters and then generating motion. This formulation avoids hard contact assumptions and brittle RL policies, achieving realistic, physically consistent human–scene–object interactions in a differentiable, end-to-end learnable manner.

Contact Modelling. Previous Human-object interaction synthesis methods often rely on binary contact labels, distance thresholds, or vertex correspondences to model con-

tact [5, 6, 39, 47, 48], focusing mainly on the hands [21, 22, 51]. While these approaches encourage plausible contact patterns, they fail to capture continuous interaction forces, leading to artifacts like floating or sliding. Early works [30, 31] directly optimize contact variables or forces, which are not well-suited for learning-based generation. PhysPT [58] introduced a continuous contact model using damped springs, but it assumes planar surfaces and decouples normal and tangential forces, limiting its generalizability.

In contrast, we propose a differentiable continuous contact model that aligns with local surface geometry and explicitly couples normal and frictional forces. This enables physically consistent HOI across arbitrary 3D surfaces and dynamic environments, and integrates seamlessly into generative pipelines.

3. Approach

Let us now introduce our approach to human motion synthesis in a dynamic environment consisting a moving object and a static scene as background. Following the setup in [18, 21], we assume a given object motion $\mathbf{O} \in \mathbb{R}^{T\times B}$ and a static scene represented by 3D voxels $\mathbf{S} \in \{0,1\}^{N_x\times N_y\times N_z}$, where T is the number of frames, B is the object translation and BPS representation following [21] and (N_x, N_y, N_z) defines the volume resolution. Our goal is then to generate a human motion $\mathbf{Q} \in \mathbb{R}^{T\times D}$, where D is the human motion dimension that interacts with the moving object and the scene. In this section, we first formalize the dynamics of human and object motion, then introduce our continuous contact force model, and finally describe our two-stage pipeline that leverages physical principles to generate realistic human motion.

3.1. Preliminary

We introduce preliminaries on human and object dynamics below. It details the mathematical formulation of human motion and human's interactions with the scene in the physical world via Euler-Lagrange equations and prepares the ground for our contact modeling.

Human Motion Dynamics. In the context of modeling human dynamics, a human body is often considered as an object with multiple rigid parts and modeled with rigid body dynamics. Following previous works [21, 58], we use the popular SMPL [25] human model. In SMPL model, a human is represented by a pose parameter $\boldsymbol{\theta} \in \mathbb{R}^{23\times 3}$, a shape parameter $\boldsymbol{\beta} \in \mathbb{R}^{10}$, and a global orientation $\mathbf{R} \in \mathbb{R}^3$, $\mathbf{T} \in \mathbb{R}^3$. Let's denote the human pose as

$$\mathbf{q} = \{\boldsymbol{\theta}, \mathbf{R}, \mathbf{T}\} \in \mathbb{R}^{75} \ . \tag{1}$$

The Euler-Lagrange Equations for the motion of such a human body are then defined as

$$\mathbf{M}_h(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}_h(\mathbf{q}, \dot{\mathbf{q}}) + \mathbf{G}_h(\mathbf{q}) = \boldsymbol{\tau} + \mathbf{J}_{hs}^\top \boldsymbol{\lambda}_s + \mathbf{J}_{ho}^\top \boldsymbol{\lambda}_o \ , \tag{2}$$

where $\dot{\mathbf{q}} \in \mathbb{R}^{75}$ and $\ddot{\mathbf{q}} \in \mathbb{R}^{75}$ are the velocities and accelerations of human joints, respectively. $\mathbf{M}_h(\mathbf{q}) \in \mathbb{R}^{75\times 75}$ represents the human mass matrix that depends on the body mass and segmental inertias. $\mathbf{C}_h \in \mathbb{R}^{75}$ and $\mathbf{G}_h \in \mathbb{R}^{75}$ capture Coriolis/centrifugal and gravitational effects, respectively. On the right-hand side of the equation, $\boldsymbol{\tau} \in \mathbb{R}^{75}$ denotes internal joint torques from e.g., muscles. $\boldsymbol{\lambda}_s \in \mathbb{R}^{3C_s}$ denote the external contact forces from the static scene to human body, and $\boldsymbol{\lambda}_o \in \mathbb{R}^{3C_o}$ represent the contact force from the moving object to human hand. C_s and C_o are the number of contact points on human body and hand respectively. $\mathbf{J}_{hs} \in \mathbb{R}^{3C_s\times 75}$ and $\mathbf{J}_{ho} \in \mathbb{R}^{3C_o\times 75}$ are the contact Jacobian matrix that maps the contact forces to the forces on human joints.

Object Motion Dynamics. Similarly, the motion dynamics of the moving object can be formulated as

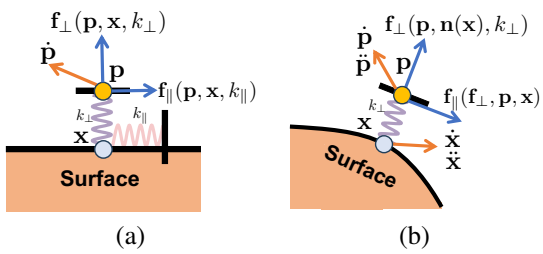
$$\mathbf{M}_o(\mathbf{q}_o)\ddot{\mathbf{q}}_o + \mathbf{C}_o(\mathbf{q}_o, \dot{\mathbf{q}}_o) + \mathbf{G}_o(\mathbf{q}_o) = -\mathbf{J}_o^\top \boldsymbol{\lambda}_o \ , \tag{3}$$

where $\mathbf{q}_o \in \mathbb{R}^6$ denotes the object orientation i.e., rotation and translation, and its velocity and acceleration are represented as $\dot{\mathbf{q}}_o \in \mathbb{R}^6$ and $\ddot{\mathbf{q}}_o \in \mathbb{R}^6$, respectively. $\mathbf{M}_o(\mathbf{q}_o) \in \mathbb{R}^{6\times 6}$ is the object's mass matrix. $\mathbf{C}_o \in \mathbb{R}^6$ is the Coriolis and centrifugal forces, and $\mathbf{G}_o \in \mathbb{R}^6$ represents the gravitational forces. $\mathbf{J}_o \in \mathbb{R}^{3C_o\times 6}$ is the contact Jacobian matrix. Importantly, the contact force $\boldsymbol{\lambda}_o$ here appears with the opposite sign compared to Eq. 2, reflecting Newton's third law: the force that the human hand exerts on the object is equal in magnitude but opposite in direction to the force the object exerts on the hand. This reciprocal relationship tightly couples the object dynamics to the human dynamics. By leveraging this property, we can not only capture how the object responds to human manipulation, but also infer the corresponding reaction forces acting on the human hand. This coupling provides essential constraints for modeling physically consistent human–object interactions.

Note that, the left hand side in Eq. 2 and Eq. 3, is only related to the subject motion, and its intrinsics such as mass, inertia. The contact Jacobian matrices on the right hand side of those equations can be precomputed from a set of possible contact points on the human body and the object. The problem becomes how to model the contact forces. In [3, 58], a continuous contact force model is proposed to capture the contact force by two orthogonal spring systems as shown in Fig. 2(a). Those two independent spring systems do not account for the coupling between the normal and tangential components of contact force and can only model the contact force between the human feet and the ground. To address this, we propose to explicitly model the tangential component as the static friction force and the kinetic friction force, while for the normal component, we still use the spring system. Furthermore, our formulation allows to compute the contact force on arbitrary surfaces. In the next section, we will introduce such contact force

model.

3.2. Continuous Contact Force Model



(a) (b)

Figure 2. **Continous contact force model.** a) The PhysPT model assumes a static ground plane and represents contact force with two independent orthogonal springs, b) Our model generalizes to arbitrary 3D surfaces by incorporating local surface normals for the normal force and explicitly modeling tangential static and kinetic friction that are dependent to the normal force, enabling physically consistent interactions in dynamic scenes.

To capture human dynamics, it is essential to model both joint actuations and contact forces. Modeling contact forces is particularly challenging because the contact status is often unknown and difficult to estimate accurately. Moreover, discrete contact representations introduce non-differentiable processes in force estimation. Although PhysPT [58] mitigates this issue by adopting a continuous contact model inspired by a spring-mass system, it relies on unrealistic assumptions, such as infinite planar surfaces and globally upward-facing normals, and non-physical force directions in the horizontal plane, which limits its applicability. To overcome these limitations, we propose a physically grounded continuous contact force model.

As shown in Fig. 2(b), given a possible contact point on human body $\mathbf{p}$, and its nearest surface point $\mathbf{x}$ on the moving object or the static scene, let's denote their relative position as $\tilde{\mathbf{p}} = \mathbf{p} - \mathbf{x}$. We further decompose such relative position into normal and tangential components as,

$$\tilde{\mathbf{p}}_\perp = (\tilde{\mathbf{p}}^\top \mathbf{n}(\mathbf{x}))\mathbf{n}(\mathbf{x}) \ , \quad \tilde{\mathbf{p}}_\parallel = \tilde{\mathbf{p}} - \tilde{\mathbf{p}}_\perp \ ,$$

where $\mathbf{n}(\mathbf{x})$ is the surface normal at $\mathbf{x}$. The contact force between human and the object/scene is then defined as[1]

$$\boldsymbol{\lambda}(\mathbf{p}) = h(-\alpha(\|\tilde{\mathbf{p}}\| - d_0))h(\beta(\tilde{\mathbf{p}}^\top \mathbf{n}(\mathbf{x}) + d_1))\mathbf{f}(\mathbf{p}) \ , \quad (4)$$

where $h(x) = \frac{1}{1+e^{-x}}$ is a soft gating function. $\alpha > 0$, $\beta > 0$ are hyperparameters that regulate the transition sharpness. The other two hyperparameters d_0 and d_1 control the contact buffer. $\mathbf{f}(\mathbf{p})$ denotes the force from the object/scene to the human. Under such definition, our contact forces will only be activated when human body is close enough to the surface while does not severely penetrates the surface i.e., $\|\tilde{\mathbf{p}}\|$ is small, and $\tilde{\mathbf{p}}^\top \mathbf{n}(\mathbf{x})$ is positive or small negative.

[1]Here, the contact forces are exerted on the human body by the static scene or moving object.

We further decompose the force to normal and tangential components as $\mathbf{f}_\perp$ and $\mathbf{f}_\parallel$, respectively,

$$\mathbf{f}(\mathbf{p}) = \mathbf{f}_\perp(\mathbf{p}) + \mathbf{f}_\parallel(\mathbf{p}) \ . \quad (5)$$

For the normal component of the contact force, we follow [3, 58] to model it with a damped spring system. Unlike their formulation, which can only model forces from the ground where the normal vector always points upward, our formulation incorporates surface normals, enabling us to model forces from arbitrary surfaces. Formally, it is defined as

$$\mathbf{f}_\perp(\mathbf{p}) = k(\mathbf{p})\mathbf{n}(\mathbf{x}) \ , \quad (6)$$

where $k(\mathbf{p})$ is defined as

$$k(\mathbf{p}) = -\kappa(\|\tilde{\mathbf{p}}_\perp\| - d_0) - \delta(\dot{\tilde{\mathbf{p}}}^\top \mathbf{n}(\mathbf{x})), \quad (7)$$

, $\kappa > 0$ and $\delta > 0$ are stiffness and damping coefficients of the spring, $\dot{\tilde{\mathbf{p}}} = \dot{\mathbf{p}} - \dot{\mathbf{x}}$ defines relative velocity between human contact point and its nearest object/scene point.

To capture the tangential component, previous methods [3, 58] also propose to use the damped spring system. However, such system ignores the fact that the tangential component i.e., the friction can depends on the normal force. To address this issue, we propose a drastically different strategy to directly model the static ($\mathbf{f}_s$) and kinetic ($\mathbf{f}_k$) friction.

$$\mathbf{f}_\parallel(\mathbf{p}) = \mathbf{f}_s(\mathbf{p}) + \mathbf{f}_k(\mathbf{p}) \ , \quad (8)$$

where the static friction term $\mathbf{f}_s(\mathbf{p})$ is defined as:

$$\mathbf{f}_s(\mathbf{p}) = h(-\gamma(\|\dot{\tilde{\mathbf{p}}}\| - v_0))\rho|\|\tilde{\mathbf{p}}_\parallel\| - d_0|\mathbf{d}_\parallel, \quad (9)$$

$\mathbf{d}_\parallel \in \mathbb{R}^3$ is the tangential force direction.

Due to the different force analysis of static scenes and moving objects, such tangential force direction is also computed differently. In particular, for a static scene, such direction is defined as the acceleration of contact point projecting onto the tangential plane,

$$\mathbf{d}_\parallel = \frac{\ddot{\mathbf{p}}_\parallel}{\|\ddot{\mathbf{p}}_\parallel\|} \ , \quad (10)$$

where $\ddot{\mathbf{p}}_\parallel$ is the tangential acceleration of the human contact point. Such formulation reflects the fact that the static friction from a static scene supports the motion of the human. For example, during walking, the overall external force that drives the human to move forward is from the ground to the supporting foot. However, for a moving object, it is the opposite i.e., human moves the object. So that, in this scenario, the tangential direction is opposite to the acceleration caused by external forces acting on the moving object, excluding the gravity.

$$\mathbf{d}_\parallel = -\frac{(\mathbf{a} - \mathbf{g})_\parallel}{\|(\mathbf{a} - \mathbf{g})_\parallel\|} \quad (11)$$

where $\mathbf{a}$ is the acceleration of the object. $\mathbf{g}$ is gravitational acceleration. $(\mathbf{a} - \mathbf{g})_\parallel = (\mathbf{a} - \mathbf{g}) - ((\mathbf{a} - \mathbf{g})^\top \mathbf{n}(\mathbf{x}))\mathbf{n}(\mathbf{x})$ is

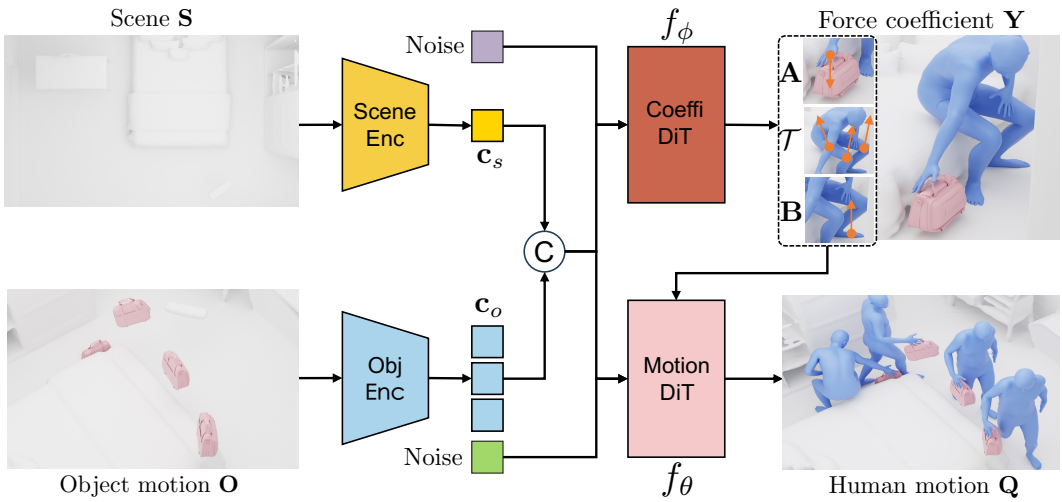


Figure 3. Overview of our pipeline. The input static scene $\mathbf{S}$ and object motion $\mathbf{O}$ is encode to a scene token $\mathbf{c}_s$ and several motion tokens $\mathbf{c}_o$. In stage 1 (Top) a transformer-based diffusion model predicts force coefficients, namely, joint torques $\mathcal{T}$, contact parameters $\mathbf{A}, \mathbf{B}$ and hand trajectories $\mathbf{H}$. In the next stage (Bottom), conditioned on such coefficients, the scene and object motion tokens, another transformer-based diffusion model is employed to generate the human motion that interacts with the scene and object. Thanks to our continuous contact modeling, we can employ a dynamic loss to further encourage physically valid interaction based on Eq. 2.

the projection of overall acceleration of the moving object onto the tangent plane.

Following the physics law, we define the kinetic friction to be linearly related to the normal force and its direction is opposite to the relative velocity, i.e.,

$$\mathbf{f}_k(\mathbf{p}) = -\mu \|\mathbf{f}_\perp(\mathbf{p})\| \frac{\dot{\tilde{\mathbf{p}}}_\parallel}{\|\dot{\tilde{\mathbf{p}}}_\parallel\|}, \tag{12}$$

where $\mu > 0$ is the kinetic friction coefficient and $\dot{\tilde{\mathbf{p}}}_\parallel$ is the projection of relative velocity onto the tangent plane.

In our continuous contact force model, the contact force is formulated as a function of motion, scene geometry, involving model coefficients $(\kappa, \delta, \rho, \mu)$. Note that, since the scene contact varies across time steps and human contact points, those coefficients differ for each contact point at different frame. Let us denote coefficients for C_s human body contact points and C_o hand contact points in a T frame sequence as $\mathbf{A} \in \mathbb{R}^{T \times C_s \times 4}$ and $\mathbf{B} \in \mathbb{R}^{T \times C_o \times 4}$, respectively. We then design a two-stage pipeline whose first stage generates those model coefficients and the human joint torques for our continuous contact force model. The second stage takes model coefficients and joint torques as input to further synthesize human motion. A physics-aware loss is adopted to encourage the consistency between the coefficients and the human motion.

3.3. Our Pipeline

Having established the physical modeling of forces between humans, objects, and scenes, we now describe how these components are integrated into a unified motion synthesis framework. As shown in Fig. 3, our pipeline consists of two stages namely, force coefficients generation and physics-aware human motion synthesis. Specifically, given the static scene $\mathbf{S}$ in voxel representation and object motion $\mathbf{O}$, we follow [18, 21], to first encode the scene and object motion to a C dimensional scene token $\mathbf{c}_s \in \mathbb{R}^C$ and object motion tokens $\mathbf{c}_o \in \mathbb{R}^{T \times C}$, respectively.

Force Coefficients Generation. The first stage of our pipeline employs a diffusion-based network to generate the force coefficients from the scene and object motion tokens. The force coefficients include the internal joint torques $\mathcal{T} \in \mathbb{R}^{T \times 75}$, the force model coefficients for human contact points $\mathbf{A}$ and $\mathbf{B}$. Additionally, we follow previous work [21], to also generate the hand trajectories $\mathbf{H} \in \mathbb{R}^{T \times 6}$, which is proved to be effective in producing accurate hand-object interaction. More formally, we follow [21] to design a transformer-based diffusion model as,

$$\hat{\mathbf{Y}}_0 = f_\phi(\mathbf{Y}_n, \mathbf{c}_s, \mathbf{c}_o, n) \ , \tag{13}$$

where $\hat{\mathbf{Y}}_0 = \{\hat{\mathbf{H}}, \hat{\mathcal{T}}, \hat{\mathbf{A}}, \hat{\mathbf{B}}\}$ is the predicted force coefficients and the hand position. $\mathbf{Y}_n$ is the noisy force coefficients, n is the diffusion step. f_ϕ is the transformer diffusion model. The model is trained with ℓ_1 loss,

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{y}_0, n}\left[\|\hat{\mathbf{Y}}_0 - \mathbf{Y}_0\|_1\right] \ , \tag{14}$$

where $\mathbf{Y}_0 = \{\mathbf{H}, \mathcal{T}, \mathbf{A}, \mathbf{B}\}$ is the ground truth coefficients.

Physics-aware Human Motion Synthesis. Given the force coefficients and hand position, the second stage of our pipeline aims to generate the full-body human motion $\mathbf{Q}$. Similarly, we use another transformer-based diffusion model in this stage,

$$\hat{\mathbf{Q}}_0 = f_\theta(\mathbf{Q}_n, \hat{\mathbf{Y}}_0, \mathbf{c}_s, \mathbf{c}_o, n) \ , \tag{15}$$

Table 1. Comparison of methods on the OMOMO dataset. Lower is better for error metrics; higher is better for precision, recall, and F1.

Methods	HandJPE	MPJPE	MPVPE	T_{root}	O_{root}	Coll.(%)	FS	C_{prec}	C_{rec}	F1
OMOMO [21]	24.01	12.42	16.67	18.44	0.50	0.22	0.38	0.82	0.70	0.72
Interdiff [51]	31.76	16.03	20.19	21.50	0.61	0.20	**0.35**	0.69	0.58	0.63
CHIOS [22]	28.50	14.96	18.73	19.48	0.56	0.24	0.39	0.72	0.63	0.67
InterAct [53]	24.62	12.59	16.71	18.8	0.47	0.21	0.39	0.82	0.71	0.73
Ours	**20.09**	**10.02**	**13.60**	**16.07**	**0.40**	**0.19**	0.41	**0.87**	**0.76**	**0.80**

Table 2. Comparison of methods on the TRUMANS dataset.

Methods	HandJPE	MPJPE	MPVPE	T_{root}	O_{root}	Collision	FS	C_{prec}	C_{rec}	F1	Sc. Pen.
Trumans [18]	47.85	36.20	38.02	33.90	0.45	0.25	**0.39**	0.65	0.54	0.59	33.48
Ours	**38.00**	**31.28**	**34.20**	**29.30**	**0.41**	**0.23**	0.42	**0.75**	**0.66**	**0.69**	**21.03**

where $\mathbf{Q}_n$ is the noisy human motion at n-th diffusion step. We train the model with the ℓ_1 loss,

$$\mathcal{L}_{\text{reco}} = \mathbb{E}_{\mathbf{Q}_n, n}[\|\hat{\mathbf{Q}}_0 - \mathbf{Q}_0\|_1] \ . \tag{16}$$

where $\mathbf{Q}_0 \in \mathbb{R}^{T\times 75}$ is ground truth human motion. To further encourage the generated human motion to be physically valid, we define a dynamic consistency loss using Eq. 2 as

$$\mathcal{L}_{\text{dyn}} = \sum_{t=1}^{T} \|\mathbf{M}_h(\hat{\mathbf{q}}_t)\ddot{\hat{\mathbf{q}}}_t + \mathbf{C}_h(\hat{\mathbf{q}}_t, \dot{\hat{\mathbf{q}}}_t) + \mathbf{G}_h(t) - \mathbf{J}_{hs}(t)^\top \lambda_s(\mathbf{a}_t) - \mathbf{J}_{ho}(t)^\top \lambda_o(\mathbf{b}_t) - \boldsymbol{\tau}_t\|_1,$$

where $\hat{\mathbf{q}}_t, \dot{\hat{\mathbf{q}}}_t, \ddot{\hat{\mathbf{q}}}_t$ denote the predicted human pose, velocity and acceleration, respectively. Velocities and accelerations are computed from the predicted pose sequence. $\mathbf{a}_t \in \mathbb{R}^{C_s\times 4}$ is the parameters for the contact between human body and scene. $\mathbf{b}_t \in \mathbb{R}^{C_o\times 4}$ represents the parameters for the contact between human hand and object.

Finally, the total loss for the second stage combines both objectives:

$$\mathcal{L} = \mathcal{L}_{\text{reco}} + \lambda_{\text{dyn}}\mathcal{L}_{\text{dyn}} \ , \tag{17}$$

where $\lambda_{\text{dyn}} > 0$ is a balancing weight.

4. Experiments

4.1. Datasets

We evaluate our method on two recent human-scene interaction datasets: OMOMO dataset [21] and TRUMANS dataset [18].

OMOMO. The OMOMO dataset [21] is a high-quality motion capture dataset featuring approximately 10 hours of paired human–object interactions on a fixed horizontal plane, involving 15 everyday objects. SMPL-X parameters were extracted via MoSh++, and object poses were estimated using markers attached to the objects. Our setting are following the protocol in [21]. In this dataset, our method accounts for both the interaction between the hands and the moving object, and the interaction between the feet and the ground plane.

TRUMANS. The TRUMANS datase [18] is a large-scale corpus for human–scene interaction modeling, containing 15 hours of motion data (about 1.6M frames at 30 Hz) in richly populated indoor environments, with 20 common object categories. Since we focus on human motion synthesis in a dynamic environment, we use a subset of the dataset to make sure each sequence contains one dynamic object.

4.2. Data Preparation

To train and evaluate our framework, we require ground-truth parameters of the continuous contact force model, such as joint torques and contact coefficients. Given the ground-truth human and object motion sequences, we obtain the internal joint torques $\mathcal{T} \in \mathbb{R}^{T\times 75}$ and the contact model coefficients $\mathbf{A}, \mathbf{B}$ via dynamic optimization following [58]. Specifically, we leverage the human and object dynamics formulations in Eq. 2 and Eq. 3 to minimize the residuals of the Euler–Lagrange equations. This optimization assumes known constants such as gravity $\mathbf{g}$, SMPL-derived segment masses and inertias (regressed from $\boldsymbol{\beta}$), object mass and inertia, and a predefined set of candidate contact points on both the human body and objects. Under these assumptions, we solve for $\mathbf{A}, \mathbf{B}, \mathcal{T}$ to ensure physically consistent supervision signals. Details of the optimization procedure, including implementation and solver settings, are provided in the supplementary material.

4.3. Evaluation Metrics

Metrics for human motion. We adopt the evaluation protocol of [21] to measure the quality of motion using HandJPE, MPJPE, MPVPE, root translation error (T_{root}), root orientation error (O_{root}), and foot sliding (FS). HandJPE, MPJPE, and MPVPE correspond to the mean hand joint position error, mean per-joint position error, and mean per-vertex position error (in cm), respectively. The root translation error is

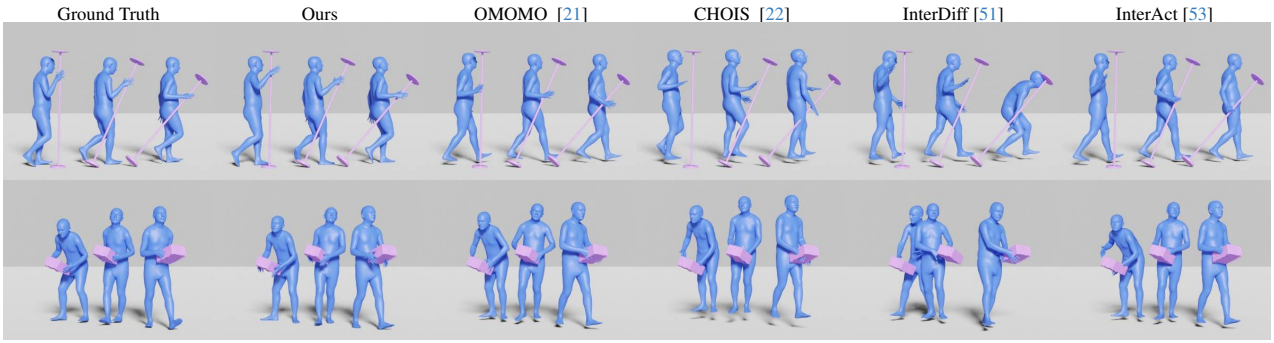


Figure 4. Qualitative comparison on OMOMO. From left to right: object-only context, ground truth, our prediction, and predictions from OMOMO, CHOIS, InterDiff, and InterAct.

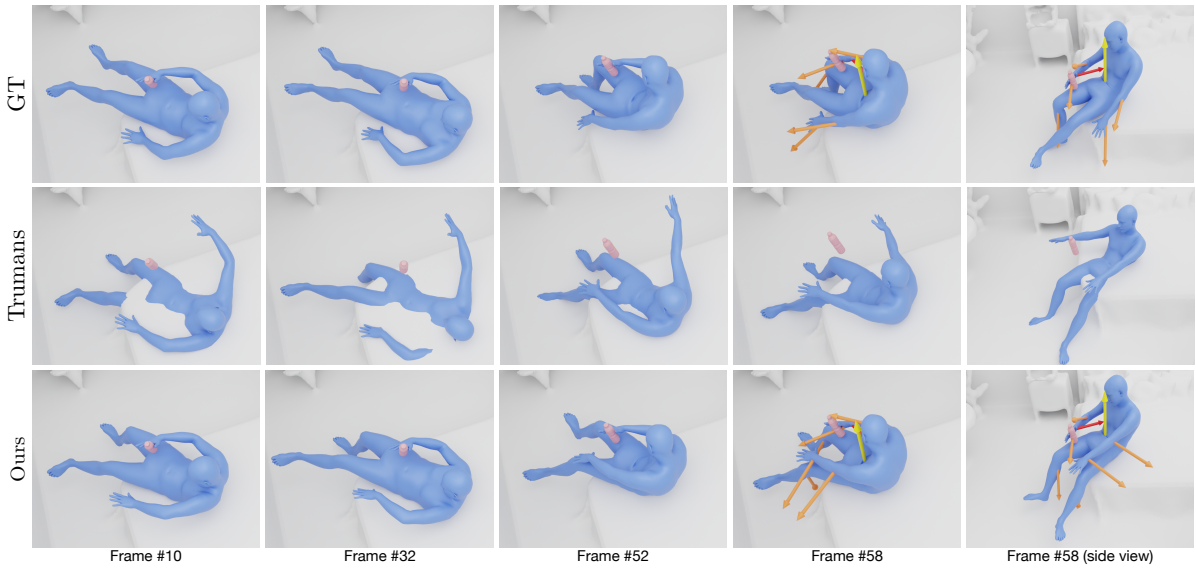


Figure 5. Qualitative comparison on Trumans. Each row shows ground truth, the Turmans baseline, and our method. Arrows illustrate estimated forces: red for forces from the dynamic object to the human, yellow for forces from the static scene to the human, and orange for internal joint forces.

computed as the Euclidean distance between predicted and ground-truth root positions, while the root orientation error is defined as the Frobenius norm between the predicted and ground-truth rotation matrices, i.e., $\|\mathbf{R}_{\text{pred}}\mathbf{R}_{\text{gt}}^{-1} - I\|_2$. We compute foot sliding (FS) following [15].

Metrics for human object interaction. We evaluate human object interactions via Collision Percentage and contact metrics (C_{prec}, C_{rec}, F1) same as [21]. Collision Percentage is the percentage of frames where the synthesized human mesh penetrates the object. The threshold for penetration is 4cm. Contact metrics follow the object detection protocol: we use a 5 cm threshold between hands and the object mesh to obtain the binary contact labels for both the predicted human meshes and the ground truth ones. We then report the precision (C_{prec}), recall (C_{rec}), and F1 score for the contact labels.

Metric for human scene interaction. Following [18], we also report the scene penetration metric to measure the interaction between the human and the static scene.

4.4. Baselines

For the OMOMO dataset, we compare our method with four recent baselines on human motion synthesis. **OMOMO** [21] synthesizes human motion conditioned on object trajectories. **InterDiff** [51] generates human–object interactions given motion histories. **CHOIS** [22] generates human and object motions from language descriptions, initial states, and waypoints. **InterAct** [53] generates human motion based on object sequences. For fair comparison, we adapt InterDiff [51] and CHOIS [22] to generate human motion conditioned solely on object motion, and InterAct [53] using SMPL-X as representation. For OMOMO [21], we directly include the results reported in their paper for comparison.

For the TRUMANS dataset, we use [18] as our baseline. While [18] generates human and object motion from scene, text, action labels, and goals, we adapt it to take the 3D scene and object motion as input for human motion synthesis, using the same object motion encoding as our method.

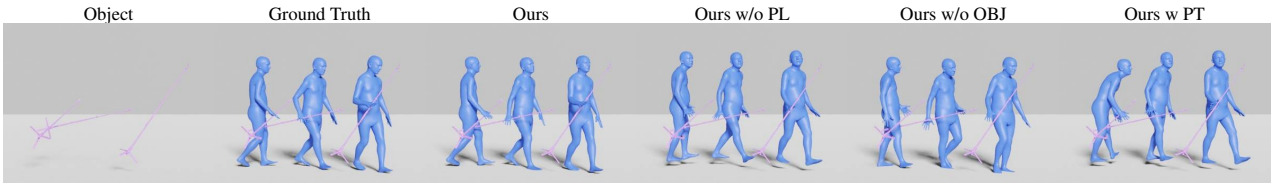


Figure 6. Ablation study comparison on OMOMO.

Table 3. Ablation study of our method. Lower is better for error metrics; higher is better for precision, recall, and F1.

Methods	HandJPE	MPJPE	MPVPE	T_{root}	O_{root}	Collision	FS	C_{prec}	C_{rec}	F1
Ours w/o PL	21.63	11.18	14.05	18.33	0.45	0.21	0.43	0.81	0.67	0.73
Ours w/o OBJ	21.02	10.73	13.89	16.52	0.43	0.22	0.43	0.80	0.69	0.74
Ours w PT	21.47	10.74	13.96	16.83	0.43	0.21	0.44	0.85	0.70	0.76
Ours	**20.09**	**10.02**	**13.60**	**16.07**	**0.40**	**0.19**	**0.41**	**0.87**	**0.76**	**0.80**

4.5. Implementation Details

All models are implemented in PyTorch [33] and trained using the Adam optimizer [20] with an initial learning rate of 0.002. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

4.6. Results

OMOMO. In Tab. 1, we compare our results to those of baselines. Our method consistently outperforms baseline methods for all metrics except for the foot sliding (FS). Although our method has higher foot sliding score, as also discussed in [21], due to the definition of foot sliding, a lower foot sliding score does not necessarily mean a better results. One can achieve a very low foot sliding score to generate a human motion that floating above the ground. The qualitative comparison shown in Fig. 4 also evidences this. For human motion metrics, our method achieves at least 12% better performance across all evaluation metrics comparing to the second best model, i.e., OMOMO [21]. For human object interaction, our results penetrate less with the object while have more precise and complete contact with it. Due to the lack of explicit physics modeling, the baseline methods always produces human motions with feet floating above the ground while the human motion from our method shows a better contact relationship between the feet and the ground. More qualitative results are shown in the supplementary video.

TRUMANS. We report the results on the TRUMANS dataset [18] in Tab. 2. The conclusion remains the same. Our method consistently outperforms the baseline method on all metrics. In this dataset, the static scenes have various geometries. The human scene interaction is not only between human feet and the ground but also between other human body parts and the scene surfaces, e.g., between the bottom of the human and the chair. We also show the qualitative comparison in Fig. 5. The human motion generated by our method have more realistic interaction with the scene and the object. More qualitative results are shown in the supplementary video.

4.7. Ablation Study

We conduct ablation studies on the OMOMO dataset [21] to analyze the contributions of our components. Specifically, we evaluate: (i) our method without the dynamic consistency loss ("Ours w/o PL"), our method without the object dynamic ("Ours w/o OBJ") and (iii) our method using the contact model from PhysPT [58] ("Ours w/ PT"). The results are summarized in Tab. 3. Without dynamic consistency loss, the model performs consistently worse especially for the object contact, having a decease of 7% on F1 score. Although with the contact model proposed in PhysPT [58], the overall performance becomes better than that of "Ours w/o PL", it still consistently underperforms our method.

5. Conclusion

In this paper, we introduced a continuous contact model for dynamic scene aware human motion synthesis. Our formulation overcomes key limitations of existing contact models—most of which operate only on the ground plane and fail to capture realistic friction dynamics—by incorporating surface-normal-conditioned force modeling in the normal direction and explicit static and kinetic friction modeling on the tangent plane. We integrated this contact model into a two-stage, diffusion-based human motion synthesis pipeline that first predicts physics parameters and subsequently generates human motion conditioned on those physics parameters. Our approach achieves state-of-the-art performance on human motion generation in a dynamic scene, demonstrating the importance of explicit physics reasoning for high-fidelity interaction synthesis. Looking forward, we aim to extend our framework to more general and complex scenarios, including interactions involving multiple dynamic objects, articulated tools, and even multi-human collaborative behaviors under diverse environmental contexts. Such extensions would further expand the applicability of physics-aware motion synthesis in embodied AI.

Acknowledgements. This research was supported partially by an ARC Discovery Grant (DP200102274).

References

[1] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, 2023. 1

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 2

[3] Marcus A Brubaker, Leonid Sigal, and David J Fleet. Estimating contact dynamics. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2389–2396. IEEE, 2009. 1, 3, 4

[4] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18000–18010, 2023. 1

[5] Yixin Chen, Sai Kumar Dwivedi, Michael J Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17100–17110, 2023. 3

[6] Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun S Lakshmipathy, Agniv Chatterjee, Michael J Black, and Dimitrios Tzionas. Pico: Reconstructing 3d people in contact with objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1783–1794, 2025. 3

[7] 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. 1

[8] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13190–13200, 2022. 2

[9] Kehong Gong, Dongze Lian, Heng Chang, Chuan Guo, Zihang Jiang, Xinxin Zuo, Michael Bi Mi, and Xinchao Wang. Tm2d: Bimodality driven 3d dance generation via music-text integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9942–9952, 2023. 1

[10] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022.

[11] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 1

[12] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[13] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021. 2

[14] 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. 1, 2

[15] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. *Advances in Neural Information Processing Systems*, 35:4244–4256, 2022. 7

[16] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: joint markerless 3d tracking of humans and objects in interaction from multi-view rgb-d images. *International Journal of Computer Vision*, 132(7): 2551–2566, 2024. 2

[17] Kai Jia, Tengyu Liu, Mingtao Pei, Yixin Zhu, and Siyuan Huang. Primhoi: Compositional human-object interaction via reusable primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11491–11501, 2025. 2

[18] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024. 1, 2, 3, 5, 6, 7, 8

[19] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023. 1

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 8

[21] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 1, 2, 3, 5, 6, 7, 8

[22] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *ECCV*, 2024. 1, 2, 3, 6, 7

[23] Lei Li and Angela Dai. Genzi: Zero-shot 3d human-scene interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20465–20474, 2024. 2

[24] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13401–13412, 2021. 1

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3

[26] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: Quantization-based 3d human motion generation and forecasting. In *European Conference on Computer Vision*, pages 417–435. Springer, 2022. 1

[27] Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, et al. Himo: A new benchmark for full-body human interacting with multiple objects. In *European Conference on Computer Vision*, pages 300–318. Springer, 2024. 1

[28] Wei Mao, Richard I Hartley, Mathieu Salzmann, and Miaomiao Liu. Contact-aware human motion forecasting. *Advances in Neural Information Processing Systems*, 35:7356–7367, 2022. 2

[29] 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. 1, 2

[30] Igor Mordatch, Zoran Popović, and Emanuel Todorov. Contact-invariant optimization for hand manipulation. In *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 137–144, 2012. 3

[31] Igor Mordatch, Jack M Wang, Emanuel Todorov, and Vladlen Koltun. Animating human lower limbs using contact-invariant optimization. *ACM Transactions on Graphics (TOG)*, 32(6):1–8, 2013. 3

[32] 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. 2

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 8

[34] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 1

[35] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 1

[36] Dian Shao, Mingfei Shi, Shengda Xu, Haodong Chen, Yongle Huang, and Binglu Wang. Finephys: Fine-grained human action generation by explicitly incorporating physical laws for effective skeletal guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1905–1916, 2025. 2

[37] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022. 1

[38] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[39] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8001–8013, 2023. 3

[40] Shashank Tripathi, Lea Müller, Chun-Hao P Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 3d human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4713–4725, 2023. 2

[41] Shashank Tripathi, Omid Taheri, Christoph Lassner, Michael Black, Daniel Holden, and Carsten Stoll. Humos: Human motion model conditioned on body shape. In *European Conference on Computer Vision*, pages 133–152. Springer, 2024. 2

[42] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 1

[43] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9401–9411, 2021. 2

[44] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023. 1

[45] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022. 2

[46] 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. 1

[47] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4757–4768, 2023. 3

[48] Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. Intertrack: Tracking human object interaction without object templates. In *2025 International Conference on 3D Vision (3DV)*, pages 1427–1439. IEEE, 2025. 3

[49] Zhaoming Xie, Jonathan Tseng, Sebastian Starke, Michiel van de Panne, and C Karen Liu. Hierarchical planning and control for box loco-manipulation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–18, 2023. 1, 2

[50] Chaoyue Xing, Wei Mao, and Miaomiao Liu. Scene-aware human motion forecasting via mutual distance prediction. In *European Conference on Computer Vision*, pages 128–144. Springer, 2024. 2

[51] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 1, 2, 3, 6, 7

[52] Sirui Xu, Yu-Xiong Wang, Liangyan Gui, et al. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *Advances in Neural Information Processing Systems*, 37:52858–52890, 2024. 1, 2

[53] 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. 1, 2, 6, 7

[54] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. In *CVPR*, 2025. 1, 2

[55] 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. 1

[56] Ling-An Zeng, Guohong Huang, Yi-Lin Wei, Shengbo Gu, Yu-Ming Tang, Jingke Meng, and Wei-Shi Zheng. Chainhoi: Joint-based kinematic chain modeling for human-object interaction generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12358–12369, 2025.

[57] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 1

[58] Yufei Zhang, Jeffrey O Kephart, Zijun Cui, and Qiang Ji. Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2305–2317, 2024. 1, 2, 3, 4, 6, 8

[59] Yufei Zhang, Jeffrey O Kephart, and Qiang Ji. Incorporating physics principles for precise human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6164–6174, 2024. 2