HOIAnimator: Generating Text-prompt Human-object Animations using Novel Perceptive Diffusion Models

Wenfeng Song[1], Xinyu Zhang[1], Shuai Li[2,3*], Yang Gao[3], Aimin Hao[3,5],
Xia Hou[1], Chenglizhao Chen[4], Ning Li[1], Hong Qin[6†]

[1]Beijing Information Science and Technology University [2]Zhongguancun Laboratory, China

[3]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University

[4]College of Computer Science and Technology, China University of Petroleum (East China)

[5]Research Unit of Virtual Human and Virtual Surgery (2019RU004), Chinese Academy of Medical Sciences

[6]Department of Computer Science, Stony Brook University (SUNY at Stony Brook), Stony Brook, New York 11794-2424, USA

https://zxylinkstart.github.io/HOIAnimator-Web/

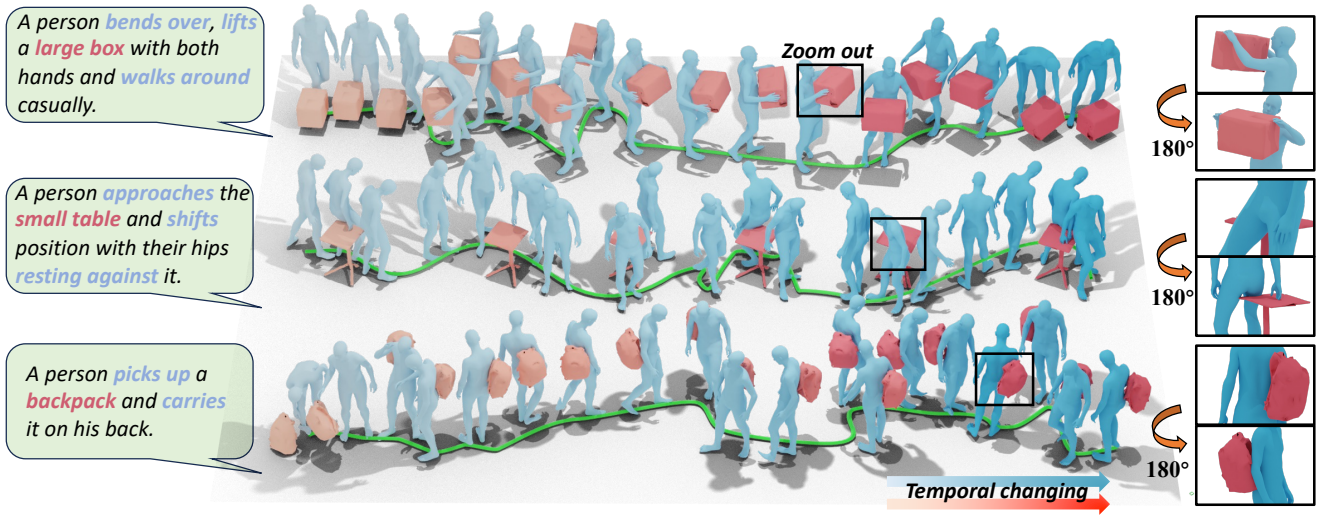

Figure 1. Our HOIAnimator excels in turning text descriptions into realistic animations of human-object interactions. It's adept at depicting a variety of actions, such as bending, lifting boxes, and picking up bags, with believable contact between the human and the objects.

Abstract

To date, the quest to rapidly and effectively produce human-object interaction (HOI) animations directly from textual descriptions stands at the forefront of computer vision research. The underlying challenge demands both a discriminating interpretation of language and a comprehensive physics-centric model supporting real-world dynamics. To ameliorate, this paper advocates HOIAnimator, a novel and interactive diffusion model with perception ability and also ingeniously crafted to revolutionize the animation of complex interactions from linguistic narratives. The effectiveness of our model is anchored in two ground-breaking innovations: (1) Our Perceptive Diffusion Models (PDM) brings together two types of models: one focused on human movements and the other on objects. This combination allows for animations where humans and objects move in concert with each other, making the overall motion more realistic. Additionally, we propose a Perceptive Message Passing (PMP) mechanism to enhance the communication bridging the two models, ensuring that the animations are smooth and unified; (2) We devise an Interaction Contact Field (ICF), a sophisticated model that implicitly captures the essence of HOIs. Beyond mere predictive contact points, the ICF assesses the proximity of human and object to their respective environment, informed by a probabilistic distribution of interactions learned throughout the denoising phase. Our comprehensive evaluation showcases HOIanimator's superior ability to produce dynamic, context-aware animations that surpass existing benchmarks in text-driven animation synthesis.

1. Introduction and Motivation

In the dynamic landscape of AI-guided creation (AIGC), 3D animation has emerged as a crucial and challenging domain. This challenge is epitomized in human-object interactions

*,† Corresponding authors

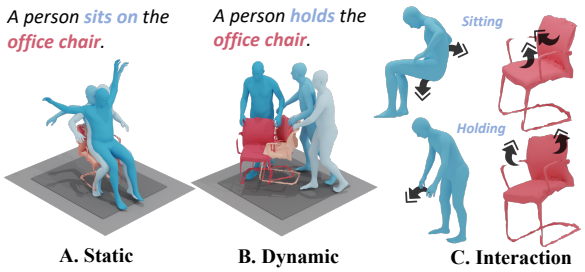

Figure 2. **Navigating the complexity of HOIAnimator.** (A): the 'static' interaction is depicted with a stationary office chair, showcasing a human sitting on the chair. (B): the 'dynamic' interaction portrays both the human and the object in motion, exemplified by the act of holding an object. (C): Arrows denote forces and trajectories involved in HOI.

(HOIs), which aim to generate realistic animations from textual descriptions. The intricacy lies in translating written language into visual narratives that accurately capture the nuanced dynamics between humans and objects. Achieving this requires a sophisticated understanding of linguistic cues and a deep knowledge of physical interaction principles. The intersection of language and physics makes 3D animation creation challenging and innovative.

The field of animation creation, driven by advancements in natural language processing and generative modeling as demonstrated by Li et al.[20] and Zhang et al.[48], faces a formidable challenge: **effectively mapping the low-dimensional latent spaces of textual descriptions to the high-dimensional, complex spaces of human and object motions**. While recent endeavors like those by previous effort [6, 25, 35] showcase the potential of converting text prompts into visual content, the intricate task of accurately rendering realistic dynamics from concise text remains largely unmastered. This issue is particularly evident when the model interprets a simple text-based action like 'holding' a bag but cannot adequately replicate the diverse, context-dependent ways this action might manifest, such as the various methods of carrying a bag.

Building upon the foundation laid by innovative projects like InterGen [21], Scene Diffuser [15], Narrator [42], and InterDiff [41], the field has advanced significantly in generating animations through interactions with scenarios or human figures. Yet, a formidable challenge persists: **accurately modeling and animating the complex forces and reciprocal influences between humans and objects**. As illustrated in Fig. 2, the relationship between a human and an office chair, for example, is not merely a static or one-dimensional interaction. It is a dynamic interplay, influenced by various factors such as the intent behind the interaction, the trajectory of movement, and the contextual use of the object, all of which are dictated by the narrative text.

Our HOIAnimator introduces dual perceptive diffusion models (PDM) to simplify the first challenge, adeptly capturing spatial dynamics between humans and objects, creating animations consistent with the textual prompts. HOIAnimator utilizes two diffusion models: a human-centric model for human movements and an object-centric model for object dynamics. The models communicate with each other through a novel Perceptive Message Passing (PMP) mechanism. The PMP adaptively learns the weight and bias of object clues embedded into human motion flow. This collaborative approach ensures the active engagement of both entities in the animation, leading to more complete and accurate representations of the narrative.

To tackle the challenge of accurately representing complex forces in HOI, we introduce a novel concept of Interaction Contact Field (ICF), a model that learns the patterns of contact as described in text prompts through a diffusion model. This diffusion model is skilled at interpreting textual cues and translating them into ICF, effectively capturing the spatial dynamics of HOIs. The ICF offers a comprehensive perspective on interaction probabilities, advancing beyond traditional collision detection methods. By incorporating a probabilistic approach that considers object affordance, human intent, and ergonomics, the ICF is able to predict potential points of interaction. This leads to animations that are both dynamic and adaptive, more accurately mirroring the complexities of real-world object manipulation.

To summarize, our contributions are listed as follows.

- We propose **a brand new HOIAnimator**, a framework that utilizes dual Perceptive Diffusion Models, human-centric model and object-centric model, to accurately render human-object interactions in animations. The core of PMD is a cutting-edge PMP mechanism designed to enable seamless and effective communication between human and object-centric models, ensuring lifelike and engaging animations.
- We present **a simple yet powerful ICF** to proactively identify and assess potential contact points between entities. The key is learning the distribution of interaction probability between humans and objects, and mapping text-based interaction cues into spatial dynamics. Our method goes beyond basic collision detection, using a probabilistic field informed by object characteristics, human intentions, and ergonomics. This leads to animations that are dynamically responsive and closely mimic real-world object interactions.
- We conduct extensive experiments in both public and wild datasets. The results demonstrate the superior ability of our HOIAnimator to generate human-object interaction animations. Our dataset and project will be public.

2. Related Work

2.1. Text-to-animation Generation

Animation generation commonly employed neural network models such as the Variational AutoEncoder (VAE) [11,

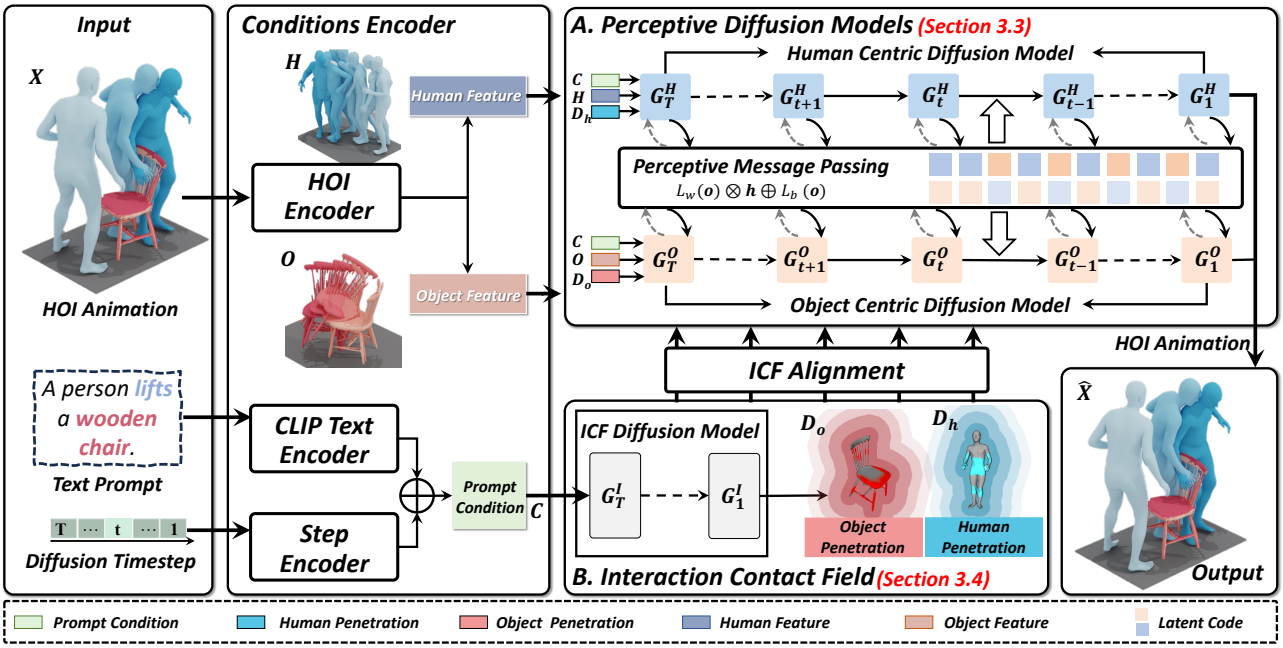

Figure 3. **Method overview**. We propose the HOIAnimator with two key parts: (1) Perceptive Diffusion Models (PDM). This part combines the movements of both people and objects in the animation, making sure they move together in a realistic way. (2) Interaction Contact Field (ICF). The ICF provides the clues that humans and objects interact and contact each other (Training phase of HOIAnimator).

27], Vector Quantized-Variational AutoEncoder (VQ-VAE) [12, 28], and Generative Adversarial Networks (GAN) [23, 43, 44] to acquire representations and patterns necessary for generating animations from text descriptions. These models were trained on textual descriptions and generated representations and patterns for animations. However, the recent introduction of diffusion models [14, 22, 32, 47] greatly improved the ability to reason about text and represent animation [1, 2, 5, 39, 48]. For instance, MDM [36] introduced a transformer-based generative model that was adapted for the many-to-many nature of the domain. Building on MDM, SinMDM [29] learned the internal motifs of a single motion sequence with arbitrary topology and synthesized motions of arbitrary length that were faithful to them. Furthermore, PriorMDM [31] proposed using a pre-trained diffusion-based model as a generative prior to fine-tuning for few-shot and zero-shot settings. Inspired by the two-person generation, we propose a bidirectional diffusion model to generate HOI animations. This model utilizes the diffusion process for both inference and generation, enabling it to handle the intricate relationships and uncertainties associated with humans and objects.

2.2. Human-object Dynamic Interaction

Current research prominently focuses on unraveling the intricacies of human-object interactions. Recent studies [9, 13, 37, 45] explored the detailed modeling of whole-body interactions. However, most works [4, 24, 34, 40, 46] focused on human-object relations within static environments, where objects are treated as passive. In contrast, recent works [15, 19, 41] integrated objects and scenes as dynamic components in motion prediction models. In multi-human interactions, some works [21, 31] introduced diffusion-based approaches for generating text-driven interaction motions for two people. The complex nature of interactions between humans and objects, significantly different from multi-human interactions, is further complicated by the disparity in data features. Addressing this, several works [7, 33, 38, 49] attempted to unravel these complexities. Utilizing the expanding array of 3D datasets that capture human interaction [3, 8, 16, 17, 24, 24, 42], our work introduces a novel text-prompt paradigm for streamlining the generation of HOI animations.

3. New Methodology

3.1. Overview

Our HOIAnimator aims to achieve end-to-end conversion from text description to 3D HOI animations. As illustrated in Fig. 3, the pipeline of HOIAnimator begins with a novel representation for HOI animation, described in Section 3.2, which underpins our text-prompt-based animation generator, aimed at minimizing inconsistencies between humans and objects. Subsequently, to synchronize text-driven prompts with corresponding dynamic visual representations, we introduce the PDM, a novel interactive diffusion mechanism specifically described in Section 3.3. Finally, to optimize the details between the surfaces of humans and objects, we introduce the ICF, engineered to evaluate the probability of contact within the interaction spaces

afforded by objects in Section 3.4.

3.2. HOIAnimator Definition and Preliminaries

Our HOIAnimator is specifically designed to handle the animation with the positions of 3D coordinates for both humans and objects. We define the **HOI Animation Definition** in rigid mathematics. Meanwhile, we introduce **Diffusion Model for HOI Animation Generation** for generating HOI animation. We utilize diffusion models' generative power, employing a stochastic diffusion process for dynamic, precise HOI animations.

HOI Animation Definition. Generating the spatial coordinates of humans and objects and providing pose information for both are essential for creating consistent animations. Inconsistencies often occur when various methods are used to represent these elements. For instance, human bodies are commonly represented using SMPL-H [26], while objects are typically depicted through translation and rotation. To resolve this, we propose a unified approach involving four key parameters: the human shape parameter ($\beta \in \mathbb{R}^{10}$), the human pose parameter ($\theta \in \mathbb{R}^{159}$), the object's translation parameter ($\tau \in \mathbb{R}^3$), and the object's rotation parameter ($\gamma \in \mathbb{R}^3$). By integrating these parameters, we form the HOI animation, denoted as $x_{1:i} = \{\beta, \theta, \tau, \gamma\}$, where $x_i \in \mathbb{R}^{175}$ represents the pose state in frame i. $i \in [0, N]$ and N is the maximum animation length. The HOI animation effectively captures the dynamics of human-object interactions in animation.

Diffusion Model for HOI Generation. Drawing inspiration from previous works [10, 18, 30], we opt for the diffusion model [15] to generate HOI animations. Similarly to the text-driven motion generation task, our training set for text-driven HOI animations consists of pairs $(x_i, text_i)$, where $text_i$ is the textual description of the HOI animation (x_i). During inference, given a textual description of the animation, we can generate an animation that matches the description. We build our text-driven HOI animation pipeline based on diffusion models. This diffusion can be modeled as a Markov noising process ($\{x^t_{1:i}\}^T_{t=0}$), gradually adding Gaussian noise to the ground truth ($x^0_{1:i}$) until it eventually becomes pure Gaussian noise ($x^T_{1:i}$):

$$q(x^t_{1:i}|x^{t-1}_{1:i}) = \mathcal{N}(\sqrt{1-\alpha_t}x^{t-1}_{1:i}, \alpha_t \mathbf{I}), \tag{1}$$

where t denotes step, $t \in [1, T]$, $\alpha_t \in [0, 1]$ are fixed set of values, generated by formula [14]. Thus, at a sufficiently large step T, α_t approaches 1, at which point it can be approximated as a Gaussian distribution $x^T_{1:i} \sim \mathcal{N}(0, \mathbf{I})$.

3.3. Perceptive Diffusion Models

Specific correlations between humans and objects are essential to address the challenge of significant differences in pose parameters between humans and objects. Extended from the single diffusion model [36], our approach employs the PDM for more nuanced processing as depicted in Fig. 3-A. PDM consists of two specialized components: (1) The human centric diffusion model and object centric diffusion model. (2) PMP exchanges clues for the two separate diffusion models as depicted in Fig. 4.

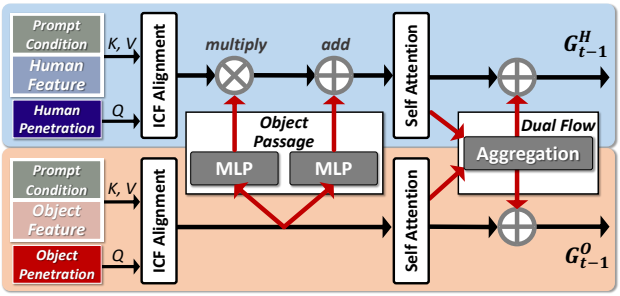

Figure 4. **Perceptive Message Passing.** Between object and human centric diffusion models, we use object passage and dual flow to adjust the features of humans and objects dynamically.

Human and Object Centric Diffusion Models. In PDM, the human centric diffusion model adapts to refine human motion, emphasizing human movement dynamics. In contrast, the object centric diffusion model focuses on optimizing object trajectories, ensuring precision in object motion. We duplicate the original animation sequence ($x_{1:i} = \{\beta, \theta, \tau, \gamma\}$), resulting in two identical copies. Each copy is then specialized: one for the object sequences ($x_{obj} = \{\tau, \gamma\}$) and the other for human sequences ($x_{hum} = \{\beta, \theta\}$). We distinctively handle the feature of the human sequence (**H**) and the feature of the object sequence (**O**). This separation allows for tailored processing of each sequence type. Further, as elaborated in Equation 1, we develop two diffusion models: the object centric diffusion, denoted as G^O, and the human centric diffusion, denoted as G^H. We get the final output ($\hat{x}_{obj}, \hat{x}_{hum}$) as:

$$\begin{aligned} \hat{x}_{obj} &= G^O\big(E_{hoi}(x_{obj}), E_{text}(text) + E_{step}(t)\big), \\ \hat{x}_{hum} &= G^H\big(E_{hoi}(x_{hum}), E_{text}(text) + E_{step}(t)\big), \end{aligned} \tag{2}$$

where E_{step} is diffusion step encoder. E_{text} is text encoder. E_{hoi} is HOIs encoder. The G^O, G^H predict the final clean animation in each sampling step. We further break down this objective into distinct components: rotation and translation losses, applicable to both humans and objects. We then focus on optimizing the diffusion model for humans and objects separately, addressing each aspect as follows:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{human} + \mathcal{L}_{obj} \\ &= \mathbb{E}_{t\sim[1:T]}[\|x_{hum} - \hat{x}_{hum}\|_2 + \|x_{obj} - \hat{x}_{obj}\|_2], \end{aligned} \tag{3}$$

where $\mathbb{E}_{t\sim[1:T]}$ denotes the average loss for all time steps T. Hence, it is possible to systematically eliminate noise from both the human and object sequences collectively, yielding a coherent HOIs aligned with the provided text condition.

Perceptive Message Passing. PMP facilitates efficient information exchange between these two components. This

exchange is enhanced by two mechanisms: object passage, which deals with the movement of objects passage, and dual flow, which intertwines the processing of human and object motions. Together, these elements of PDM ensure a comprehensive and accurate representation of motion dynamics, avoiding oversimplifications in single diffusion models.

In the PMP, the first step is utilizing the object passage method. This is pivotal for improving the incorporation of object-specific details into human motion. Object passage processes object latent code ($\mathbf{o}$) to modify human latent code ($\mathbf{h}$) dynamically ($\mathbf{o}$ and $\mathbf{h}$ have a detailed description in Section 3.4). This dynamic adjustment aligns human-centered data effectively. The adaptive mechanism is instrumental in enhancing the model's proficiency in capturing the complex relationship between human kinetics and object dynamics, a key factor in preserving the realism of the generated HOI animation. The object passage ($F_{obj}(\mathbf{h}|\varphi,\phi)$) can be written as:

$$\begin{aligned}\mathbf{h}' &= F_{obj}(\mathbf{h}|\varphi,\phi)\\ &= \varphi\cdot\mathbf{h}+\phi\\ &= L_w(\mathbf{o})\cdot\mathbf{h}+L_b(\mathbf{o}),\end{aligned} \tag{4}$$

where L_w and L_b are two fully connected networks. φ and ϕ denote dynamically adjusted weights and biases. The input processed by the converter, which is responsible for refining the human position sequence, is influenced by its intrinsic characteristics and dynamic interactions with object movement. This approach ensures that the generated object positional sequences harmonize with contemporaneous human motion, resulting in a faithful portrayal of human-object interactions.

The second step in PMP involves implementing the dual flow approach. This method boosts bidirectional communication between the human and object centric diffusion modules. Before the decoding phase of these modules, we integrate the human latent features ($\mathbf{h}'$) and object latent features ($\mathbf{o}'$) using an aggregation module (F_{dual}). These integrated features are then added to the original features as residuals, enhancing the overall process. This integration can be articulated as:

$$<\hat{\mathbf{h}},\hat{\mathbf{o}}>=F_{dual}(<\mathbf{o}',\mathbf{h}'>,<\mathbf{h}',\mathbf{o}'>)\oplus<\mathbf{o}',\mathbf{h}'>, \tag{5}$$

where $<,>$ represents a pair of input data in a specified order, which is compatible with certain mathematical operations. $\oplus$ is the element-wise add, which is to learn the residual value. $F_{dual}(a,b)$ denotes the concatenation of the two feature vectors a and b. First, we perform self-attention on a,b to obtain features. After transformation by the aggregation module, the features are truncated to match the dimensional of a. Last, $\hat{\mathbf{h}}$ and $\hat{\mathbf{o}}$ serve as feature inputs to the latent decoder, reconstructing them into $\hat{x}^0_{1:i}$.

3.4. Interaction Contact Field

Our primary objective is to create realistic and interactive HOI animations. Previous models [21, 31, 41] often struggle to capture this information about the contact between humans and objects. To address this challenge, we introduce a revolutionary method: ICF. The ICF is meticulously designed to calculate the probability of contact between human bodies and specific object regions crucial for interaction, as depicted in Fig. 3-B. By focusing on these contact probabilities, the ICF facilitates the generation of realistic HOI animations. This notably improves the realism of interactions. Further strengthening this innovation is a sophisticated ICF embedding scheme tailored for both granular (ICF) and comprehensive (HOI animation) latent spaces. This scheme ensures exceptional precision in capturing and visualizing the intricacies of HOIs.

We randomly sample 1,500 points from each mesh, with a focus on maintaining consistent vertex indices for humans or objects of the same category. This ensures uniformity and comparability across different samples.

ICF Prediction. Simply calculating the SDF for humans and objects yields only basic positional data. However, during interactions between humans and objects, more intricate details such as contact and penetration are crucial. To address these complex interactions, we propose calculating the ICF via the contact area. Specifically, we assess the contact and penetration states within the length (S) of HOI animation. For humans, we represent the vertices as v_h. Similarly, for objects, the vertices are denoted as v_o representing the object vertex count. Then, we randomly sample 1,500 points from each mesh, focusing on maintaining consistent vertex indices for humans or objects of the same category. We calculate the nearest distance to the human contact information (C_h), assigning a symbol to represent object penetration ($\mathbf{D}_o$). The calculation of human penetration ($\mathbf{D}_h$), follows a similar approach. Therefore, $\mathbf{D}_{<h,o>}$ can be defined as follows:

$$\begin{aligned}\mathbf{D}_{<h,o>} &= F_{ICF}\Big(\underbrace{C_{<h,o>}}_{\Uparrow},sample(v_{<o,h>},N)\Big),\\ &\overbrace{\begin{aligned}C_h[j] &= \parallel v_h[j]-v_o[i]\parallel_2, j=1,...,V_h\\ C_o[i] &= \parallel v_o[i]-v_h[j]\parallel_2, i=1,...,V_o\end{aligned}}\end{aligned} \tag{6}$$

where $v_h[i]\in\mathbb{R}^3, v_h[j]\in\mathbb{R}^3$ are j-th and k-th vertex on humans and objects, respectively. $F_{ICF}()$ is the function that directly computes the signed distance field. $sample()$ is a sequence of point clouds. $\mathbf{D}_h,\mathbf{D}_o\in\mathbb{R}^N$ provide us with information on the spatial relationship between the object mesh and the human mesh. Then we pre-train the ICF diffusion model $G^I(text)$ similar to Equation 7, and the corresponding interaction contact field can be generated

through text description. We predict the ICF process as:

$$< \hat{\mathbf{D}}_h, \hat{\mathbf{D}}_o > = G^I\big(\mathbf{D}_{<h,o>}, E_{text}(text) + E_{step}(t)\big), \tag{7}$$

where $text$ remains consistent with the text of HOI animation. In this way, we can get the ICF based on textual cues, which helps the PDM learn the probability of contact.

ICF Alignment for HOI. We incorporate the ICF embedding scheme to effectively align the interaction contact field $(\hat{\mathbf{D}}_h, \hat{\mathbf{D}}_o)$ with the HOI animations, as shown in Fig. 4. Using an object latent code $(\mathbf{o})$ as an example, we merge the object feature $(\mathbf{O})$ with the condition $(\mathbf{c})$ to form a combined feature map $\mathbf{L} = \{\mathbf{c}, \mathbf{O}\}$. Following this, we employ cross-attention $(Attn)$ to calculate the desired attention weights, which are crucial for integrating the human and object features in the HOI animation. The ICF Embedding process can be formulated as:

$$\begin{aligned} \mathbf{o} &= Attn(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}, \\ \mathbf{Q} &= \mathbf{W}^Q\hat{\mathbf{D}}_o, \mathbf{Q} = \mathbf{W}^K\mathbf{L}, \mathbf{V} = \mathbf{W}^V\mathbf{L}, \end{aligned} \tag{8}$$

where $\mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d_s} \times \mathbb{R}^{d_k}$ and $\mathbf{W}^Q \in \mathbb{R}^{d_q} \times \mathbb{R}^{d_k}$ are trainable weights. d_s, d_q and d_k are the channel numbers of the corresponding weights. ICF alignment guides the generation of HOI animation, ensuring excellent accuracy in generating the complexity of HOI animations.

4. Experiments

This section presents our HOIAnimator's implementation details and experimental results, comparing it with previous state-of-the-art methods. In addition, it includes an ablation study and a user study. More results are provided in the supplementary material.

4.1. Implementation Details

Datasets. Our dataset is compiled from publicly available Behave [3] and InterCap [16] datasets, which consist of motion captured from 3D human interactions. These datasets offer diverse human interaction actions, featuring common objects such as tables, backpacks, and chairs and typical interactive actions such as carrying, sitting on, and playing with. However, these datasets are limited because they are labeled in a multi-class format and lack detailed textual descriptions. To enhance our dataset for this application, we undertake three normalization steps. First, we standardize the frame rate of all motions to a consistent 30 FPS: (1) We shorten sequences over 10 seconds to a length of 6-10 seconds by trimming the end. (2) Sequences between 6-10 seconds may also be trimmed, but we keep them longer than 6 seconds. (3) Sequences under 6 seconds are extended with extra frames to reach 6 seconds. This standardizes the length of all the sequences, helping our animation process. Finally, these animations are aligned with our HOI animation template, ensuring uniformity and coherence in the dataset. Following this, we describe the actions in complete sentences and annotate them using the SpaCy . The final step in our data preparation process involves manual post-processing, where we meticulously filter out any anomalies in the textual descriptions, ensuring the dataset's quality and relevance to the HOI animation.

Evaluation Metrics. We follow the performance measures [11] for quantitative evaluations, including Frechet Inception Distance (FID), R Precision, Diversity, and Multi-Modal Distance (MM Dist). Additionally, our evaluation is broadened to include an examination of the Vertex distance and Penetration score [19] of the generated objects. (1) FID measures the quality of HOI animation generation by contrasting features of real and synthetically generated HOI animation. (2) R precision quantifies the alignment between generated HOI animations and their textual descriptions, ranking actual text within the top 1, 2, or 3 positions. (3) Diversity evaluates the range and depth of the HOI animation produced. (4) MM Dist calculates the average Euclidean distance between motion features and corresponding textual descriptions. (5) Vertex distance evaluates generation quality by comparing distances between vertices in real and generated objects. (6)Penetration score assesses realism based on the human-object interaction proximity in the animations.

Parameters. For the HOI Encoder, we utilize a 2-layer linear architecture with a latent dimensional of 1024. Regarding the ICF, as well as object and human centric diffusion modes, we use a 4-layer transformer with a latent dimension of 512. For the variance settings in 3 diffusion models, we preset the variance value to increase linearly from 0.0001 to 0.02 within T = 1000 noise steps. The text encoder incorporates a frozen CLIP ViT-B/32 model complemented by two additional transformer encoder layers. The Adam optimization algorithm is employed to train the model, with a learning rate set at 0.0002. Training is executed on 4 NVIDIA 3090Ti GPUs, with a batch size of 64 per GPU. The model is trained over 250,000 steps.

4.2. Quantitative Evaluation

Baselines. In this study, we propose a novel approach to the generation of HOI animations with prompt text and compare it with several state-of-the-art methods, including MoitonCLIP [35], T2M [11], MDM [15], PriorMDM [31], MLD [6], InterGen [21].

To the best of our knowledge, rare existing work has explored text-driven 3D HOI animation generation. To thoroughly evaluate the effectiveness of our HOIAnimator, we conduct a comprehensive comparison with the above-mentioned state-of-the-art. Our method takes textual descriptions as input and produces HOI animations. For PriorMDM and InterGen, we retained the core structure of

https://spacy.io/models

Methods	R Precision↑			Vertex Distance↓	FID↓	MM Dist↓	Diversity→	Penetration↑
	Top 1	Top 2	Top 3					
Real motions	$0.508^{\pm0.004}$	$0.725^{\pm0.005}$	$0.821^{\pm0.006}$	–	$0.012^{\pm0.002}$	$6.754^{\pm0.005}$	$9.534^{\pm0.065}$	–
MoitonCLIP [35]	$0.322^{\pm0.006}$	$0.493^{\pm0.005}$	$0.614^{\pm0.005}$	$0.979^{\pm0.110}$	$1.389^{\pm0.049}$	$10.424^{\pm0.009}$	$8.192^{\pm0.075}$	$0.529^{\pm0.003}$
T2M [11]	$0.384^{\pm0.005}$	$0.582^{\pm0.006}$	$0.673^{\pm0.005}$	$0.813^{\pm0.003}$	$0.944^{\pm0.042}$	$8.492^{\pm0.011}$	$8.724^{\pm0.132}$	$0.561^{\pm0.006}$
MDM [15]	$0.363^{\pm0.007}$	$0.573^{\pm0.006}$	$0.692^{\pm0.006}$	$0.783^{\pm0.021}$	$0.859^{\pm0.080}$	$9.382^{\pm0.017}$	$\mathbf{9.537}^{\pm0.043}$	$0.568^{\pm0.003}$
MLD [12]	$0.448^{\pm0.007}$	$0.628^{\pm0.006}$	$0.701^{\pm0.006}$	$0.711^{\pm0.005}$	$0.859^{\pm0.080}$	$8.382^{\pm0.017}$	$8.543^{\pm0.132}$	$0.578^{\pm0.003}$
PriorMDM [31]	$0.461^{\pm0.006}$	$0.636^{\pm0.005}$	$0.727^{\pm0.035}$	$0.683^{\pm0.073}$	$0.853^{\pm0.028}$	$8.776^{\pm0.012}$	$9.213^{\pm0.042}$	$0.601^{\pm0.002}$
InterGen [6]	$0.491^{\pm0.005}$	$0.652^{\pm0.005}$	$0.734^{\pm0.005}$	$0.523^{\pm0.005}$	$0.717^{\pm0.055}$	$7.932^{\pm0.021}$	$9.344^{\pm0.023}$	$0.613^{\pm0.001}$
Ours	$\mathbf{0.526}^{\pm0.006}$	$\mathbf{0.719}^{\pm0.006}$	$\mathbf{0.781}^{\pm0.005}$	$\mathbf{0.118}^{\pm0.063}$	$\mathbf{0.623}^{\pm0.063}$	$\mathbf{7.521}^{\pm0.014}$	$9.526^{\pm0.029}$	$\mathbf{0.643}^{\pm0.001}$

Table 1. **Quantitative evaluation on BEHAVE [3].** To ensure a fair comparison, we conducted 20 experiments. $x^{\pm y}$ denotes that x represents the average value of the metric, while y corresponds to the confidence interval 95% around this mean. '↑' ('↓','→') indicates that the values are better if the metric is larger (smaller, closer); The **bold fonts** denote best performers. The results show that the HOI animations synthesized by our model outperform other baselines in terms of semantic matching.

communication diffusion but tailored the parts involving interactions between two humans to match the format of our dataset. For MLD, we employed a VAE to encode our dataset into a latent code representation. Subsequently, we utilized a diffusion model to generate the latent code, which was then decoded to produce the final HOI animations.

Quantitative Results and Analysis. Tab. 1 shows our quantitative comparison results with 6 baselines on BEHAVE. Our HOIAnimator marks a notable advancement over InterGen, as evidenced by measurable improvements across several key metrics. Firstly, it demonstrates enhanced precision (Top-3), boosting the score from 0.734 to 0.781. Furthermore, there is a significant enhancement in the vertices distance metric, with a reduction in the score from 0.523 to 0.118, reflecting a more accurate representation. In addition, the HOIAnimator has achieved greater fidelity in generated animations, evidenced by a decrease in the FID score from 0.717 to 0.623 and a 3% improvement in the penetration score. These advancements collectively signify a substantial improvement in the performance and quality of our HOIAnimator.

Methods	Precision↑	FID ↓	Penetration↑
Real motions	$0.821^{\pm0.005}$	$0.012^{\pm0.002}$	–
w/o ICF	$0.756^{\pm0.004}$	$0.714^{\pm0.027}$	$0.615^{\pm0.003}$
w/o PMP	$0.723^{\pm0.006}$	$0.789^{\pm0.042}$	$0.593^{\pm0.002}$
w/o PDM	$0.696^{\pm0.008}$	$0.823^{\pm0.034}$	$0.564^{\pm0.003}$
Ours	$\mathbf{0.781}^{\pm0.005}$	$\mathbf{0.623}^{\pm0.063}$	$\mathbf{0.643}^{\pm0.001}$

Table 2. **Ablation study.** We show precision (Top-3), FID, and penetration. Our configuration can achieve the best results.

4.3. Ablation Study

In this section, we examine the roles of three crucial components in our method: PDM, PMP, and ICF. We present the comparative results in Tab. 2 and Fig. 5.

PDM. We evaluate the effect of PDM through an ablation study. Specifically, we benchmark our HOIAnimator ('Ours') against a variant devoid of the PDM ('w/o PDM'), which employs a single diffusion model. Our results are visually represented in Fig. 5, demonstrating that our proposed method is more effective in accurately capturing the

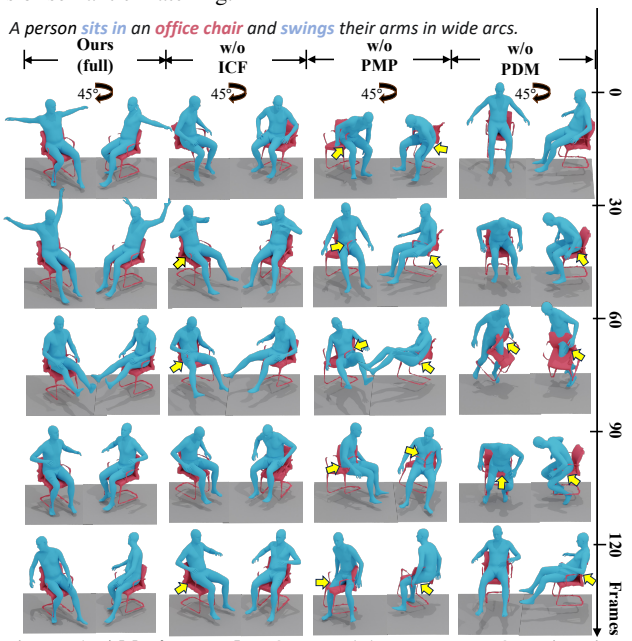

Figure 5. **Ablation study**. Our model generates HOI animations from text descriptions. Simultaneously, we apply a 45° rotation to the right side of each result and use yellow arrows to point out interaction errors to facilitate the comparison of their quality.

spatial relationships between humans and objects. Furthermore, when assessed in terms of precision (Top-3), Frechet Inception Distance (FID), and penetration score, our configuration outperforms the 'w/o PDM' model, indicating its superior performance.

PMP. To evaluate the effectiveness of our proposed PMP. This section evaluates our model in the absence ('w/o PMP') and our approach. Fig. 5 clearly illustrates that while the spatial arrangement between individuals and objects appears normal, there is an absence of interactive dynamics.

ICF. We remove the ICF ('w/o ICF'). As depicted in Fig. 5, the exclusion of the components results in generally acceptable HOI animations; nevertheless, these animations are marred by specific inaccuracies, such as unrealistic penetrations. This observation highlights the enhanced positional accuracy afforded by our proposed configuration.

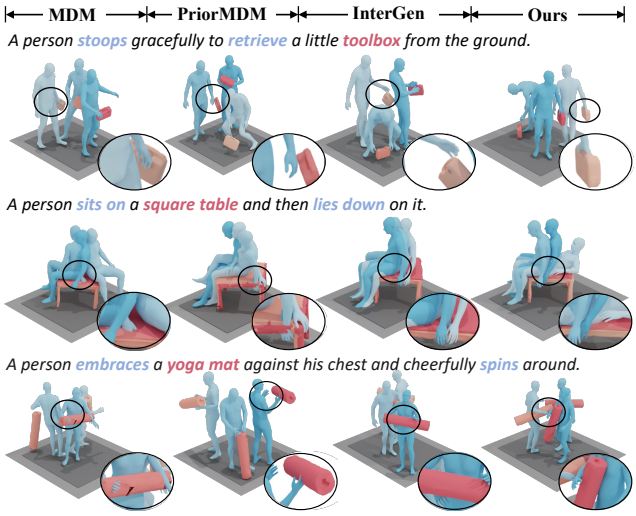

Figure 6. **Qualitative evaluation**. We present zoomed-in details highlighted within black boxes. For any specified text description, only our HOIAnimator is capable of accurately depicting the spatial relationships and the dynamic interactions involved.

4.4. Qualitative Evaluation

To illustrate the effectiveness of HOIAnimator, we provide a qualitative comparison between previous works [12, 15, 31] and HOIAnimator. As shown in Fig. 6, HOIAnimator stands out as the only method capable of effectively translating textual descriptions that encompass the positional relationships and interactive dynamics between humans and objects. In comparison, MDM struggles with precise spatial positioning of individuals and objects. Although Prior MDM is adept at classifying human-object interaction (HOI) actions, it lacks in detailing the nuances of interaction dynamics. On the other hand, InterGen effectively understands these dynamics, yet it does not consistently execute interactions accurately. Through examples, HOIAnimator effectively structures and generates complex interactions.

4.5. User Study

To further evaluate the quality of our generated HOI animations, we conduct a user study to evaluate the quality of HOI animations. In the study, we randomly select 9 motion labels and 20 object labels and combine them to create 10 meaningful descriptions of HOI animations. Based on these coherent HOI descriptions, we generate synthetic animations and shuffle them using scoring methods for presentation. After that, we ask users to rate the synthetic animations on three aspects: (1) Semantic Matching: The generated animations match the semantics of the given text descriptions. (2) Interaction Score: The quality of the poses and interactions between humans and objects in the animation. (3) Realism: The level of realism in the motion of the characters. As shown in Fig. 7, the results demonstrate that our method surpasses other baselines in terms of semantic matching, interaction score, and realism.

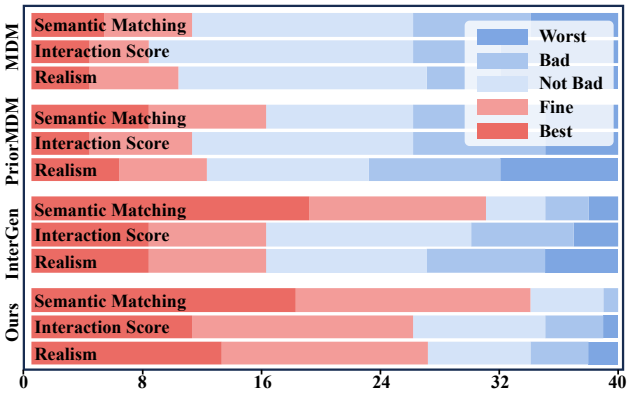

Figure 7. **User study.** The color bars in the figure indicate the percentage of the scores. The X-axis represents the number of participants.

4.6. Limitation and Discussion

Although HOIAnimator generates realistic HOI animation, it still has some limitations (see Sup. Mat.). First, the HOIAnimator is less adept at depicting complex sequences of interactions. Second, it is limited to scenes involving multiple objects interacting simultaneously. Furthermore, our method does not support nonrigid objects animations (e.g., water). Including the deformation prior into a current framework for high-quality deformable generation is promising, but it requires much more diverse training data.

5. Conclusion and Future Work

In this paper, we propose the HOIAnimator to convert text instructions into detailed animations of human and object interactions. Our key innovation is the PDM, which could closely align human and object movements with their corresponding text descriptions. Additionally, we have developed an ICF. This field actively influences animation, ensuring it mirrors the precise and diverse nature of interactions observed in the real world. The results demonstrate that HOIAnimator excels at creating dynamic and context-aware animations. In future work, we will improve HOIAnimator to better handle complex, sequential actions and interactions involving multiple objects.

Acknowledgments

This paper is supported by National Natural Science Foundation of China (62102036, 62272021), Beijing Natural Science Foundation (L232102, 4222024), R&D Program of Beijing Municipal Education Commission (KM202211232003), Beijing Science and Technology Plan Project Z231100005923039, National Key R&D Program of China (No. 2023YFF1203803), USA NSF IIS-1715985 and USA NSF IIS-1812606 (awarded to Hong QIN).

References

[1] Hyemin Ahn, Esteve Valls Mascaro, and Dongheui Lee. Can we use diffusion probabilistic models for 3d motion prediction? In *2023 IEEE International Conference on Robotics and Automation*, pages 9837–9843. IEEE, 2023. 3

[2] German Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2317–2327, 2023. 3

[3] 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. 3, 6, 7

[4] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision*, pages 387–404, 2020. 3

[5] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9544–9555, 2023. 3

[6] 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. 2, 6, 7

[7] 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

[8] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 3

[9] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, pages 1–12, 2023. 3

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 4

[11] 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*, pages 5152–5161, 2022. 2, 6, 7

[12] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597, 2022. 3, 7, 8

[13] Sanjay Haresh, Xiaohao Sun, Hanxiao Jiang, Angel X Chang, and Manolis Savva. Articulated 3d human-object interactions from rgb videos: An empirical analysis of approaches and challenges. In *International Conference on 3D Vision*, pages 312–321, 2022. 3

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3, 4

[15] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 2, 3, 4, 6, 7, 8

[16] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pages 281–299, 2022. 3, 6

[17] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9365–9376, 2023. 3

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[19] 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. 3, 6

[20] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429, 2022. 2

[21] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 2, 3, 5, 6

[22] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439, 2022. 3

[23] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. *arXiv preprint arXiv:2305.10973*, 2023. 3

[24] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17152–17162, 2023. 3

[25] Minho Park, Jooyeol Yun, Seunghwan Choi, and Jaegul Choo. Learning to generate semantic layouts for higher

text-image correspondence in text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7591–7600, 2023. 2

[26] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 4

[27] 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. 3

[28] 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, 2022. 3

[29] 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

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 4

[31] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 3, 5, 6, 7, 8

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

[33] Wenfeng Song, Xinyu Zhang, Yuting Guo, Shuai Li, Aimin Hao, and Hong Qin. Automatic Generation of 3D Scene Animation Based on Dynamic Knowledge Graphs and Contextual Encoding. *International Journal of Computer Vision*, 131(11):2816–2844, 2023. 3

[34] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21179–21189, 2023. 3

[35] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374, 2022. 2, 6, 7

[36] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023. 3, 4

[37] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20460–20469, 2022. 3

[38] 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 *International Conference on 3D Vision*, pages 353–362, 2022. 3

[39] 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

[40] 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

[41] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023. 2, 3, 5

[42] Haibiao Xuan, Xiongzheng Li, Jinsong Zhang, Yebin Liu Hongwen Zhang, and Kun Li. Narrator: Towards natural control of human-scene interaction generation via relationship reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22268–22278, 2023. 2, 3

[43] Honghui Yang, Tong He, Jiaheng Liu, Hua Chen, Boxi Wu, Binbin Lin, Xiaofei He, and Wanli Ouyang. Gd-mae: generative decoder for mae pre-training on lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9403–9414, 2023. 3

[44] Xiao Yang, Chang Liu, Longlong Xu, Yikai Wang, Yinpeng Dong, Ning Chen, Hang Su, and Jun Zhu. Towards effective adversarial textured 3d meshes on physical face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4119–4128, 2023. 3

[45] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8834–8845, 2023. 3

[46] 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 *European Conference on Computer Vision*, pages 34–51, 2020. 3

[47] 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

[48] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 364–373, 2023. 2, 3

[49] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19568–19577, 2022. 3