

# PAMotion: Physics-Aware Motion Generation for Full-Body Interaction with Multiple Objects

Yan Di<sup>1\*</sup>, Yuheng Li<sup>3\*</sup>, Yaoxing Wang<sup>3</sup>, Mengge Liu<sup>2</sup>, Shan Gao<sup>3</sup>, Xiangyang Ji<sup>2</sup>  
<sup>1</sup>Harbin Institute of Technology, <sup>2</sup>Tsinghua University, <sup>3</sup>Northwestern Polytechnical University  
diyan@hit.edu.cn, gaoshan@nwpu.edu.cn, xyji@tsinghua.edu.cn

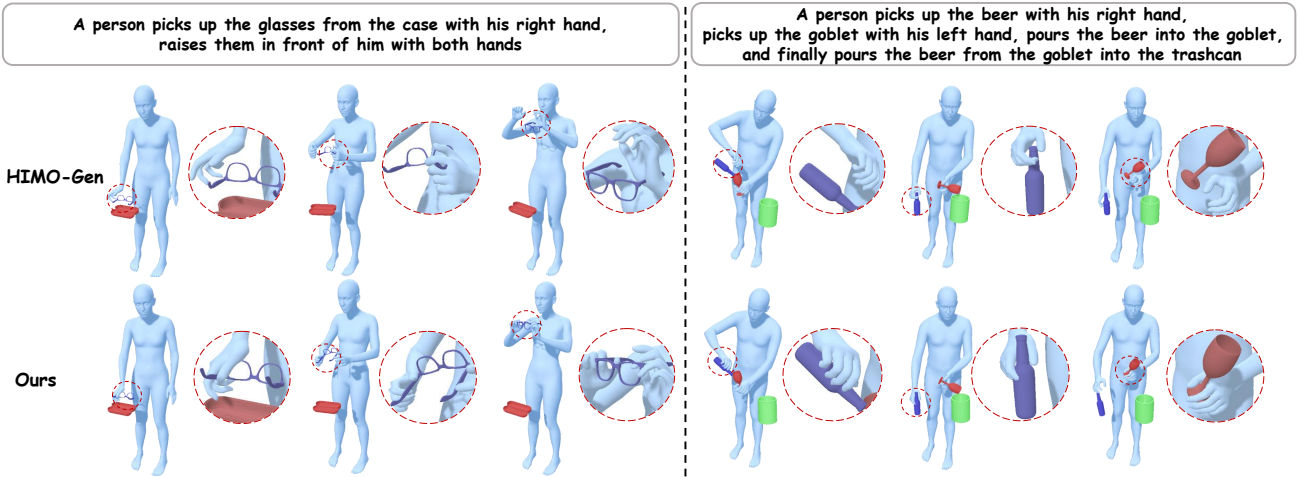


Figure 1. Qualitative Comparison on HIMO dataset [53]. Our method PAMotion produces physically plausible human–object interactions, while the baseline method HIMO-Gen [53] often exhibits floating or penetration artifacts.

## Abstract

We present **PAMotion**, a physics-aware diffusion framework for generating realistic full-body human interactions with multiple objects. Existing diffusion-based methods that jointly synthesize human and object motions often struggle to capture the intricate physical interactions—especially those involving complex hand–object contacts. To address this issue, in this paper, we begin with our key observation: in everyday, slow-motion scenarios, object accelerations inherently reveal the underlying physical interactions. If an object’s acceleration aligns with gravity, it is likely in free motion with no physical contact from human or other objects; otherwise, it must be in contact—directly or indirectly—with the human body. Building on this intuition, PAMotion jointly models full-body human motion, object motion, and their corresponding accelerations, enforcing physical plausibility through a physics-aware interaction loss. In this loss, we softly penalize violations of consistency between object acceleration and human-object

contact states. PAMotion follows a coarse-to-fine pipeline: we first synthesize global torso and object translations, then conditionally refine hand motions and object rotations, achieving both high-level motion-text consistency and low-level physical fidelity. Experiments on two challenging datasets HIMO and ParaHome demonstrate that PAMotion achieves state-of-the-art performance in generating realistic, physically consistent full-body manipulation sequences involving multiple objects. Our code is released at <https://github.com/liyuheng520/PAMotion>.

## 1. Introduction

Generating realistic human–object interactions (HOIs) is a fundamental problem in computer vision and graphics, with broad applications in AR/VR [67, 77], robotics [39, 69], and human behavior understanding [7, 124, 132]. Given the recent progress of diffusion-based generative models, it has become possible to synthesize high-quality motion sequences of a single human manipulating one target ob-

ject [34, 64, 78, 87, 98]. However, real-world human-object interaction is typically far more complex. To achieve a goal specified by a language command, humans usually need to manipulate multiple target objects simultaneously. A straightforward idea is to directly extend previous single-human single-object methods, such as HIMO [53], which adopts a dual-branch diffusion network to jointly predict kinematic human and object motions. However, such methods often fail to produce physically plausible results, hands may penetrate objects, objects may float in mid-air, or contact timing may not align with physical forces. The key reason behind this limitation lies in the intricate nature of multi-object interactions: treating the interactions as simple kinematic signals without capturing the underlying physics leads to physically inconsistent generation.

In this paper, we aim to explicitly understand why objects move the way they do, that is, the physical causes behind motion. We start from a simple yet insightful observation: in everyday, slow-motion human-object interactions, object accelerations inherently reveal the underlying physical contact states. We introduce it with three cases. First, if an object’s acceleration aligns with gravity, it is most likely in free fall with no human or object contact. Second, if an object remains stationary in mid-air, the external force from a human or another object balances gravity, implying physical contact. Third, if an object’s acceleration deviates from gravity, it implies that human exerts force on it directly or indirectly by using other objects. These observations allow us to bridge the gap between kinematic generation and physical reasoning within a generative framework.

Building upon this insight, we introduce **PAMotion**, a **Physics-Aware Motion** diffusion framework that jointly synthesizes full-body human motions and multiple object trajectories while ensuring physical plausibility, conditioned on text commands. Overall PAMotion follows a coarse-to-fine generation pipeline. Since humans usually describe tasks from a high-level perspective, the purpose of coarse stage is to better fulfill the goals implied by the language command. For example, a common instruction like “move the desk lamp from the table to the bed” does not specify the fine details of hand motion or object orientation. Directly generating the complete motion of the human body, including both torso and hands, along with fine-grained object motions can cause the network to become unbalanced between achieving the goals specified by the language command and focusing on physical details. Therefore, at the coarse stage, we first synthesize global torso and object translations to capture high-level motion patterns, conditioned on initial states of human and objects, language commands, and objects geometry. At the fine stage, we generate fine-grained hand pose and object orientations conditioned on the results of coarse stage, mainly complying to physical plausibility. We employ the structure of HIMO [53] for the

coarse stage, while for the fine stage, we adopt a new dual-branch network architecture, one for handling hand pose and the other for generating object-related signals. Then we turn the three observations into a soft physics-aware interaction loss, which constrains the contact of human (typically hand) and object according to the object acceleration states.

We evaluate PAMotion on two challenging benchmarks, HIMO and ParaHome, which involve complex multi-object manipulation and rich contact dynamics. Quantitative and qualitative results demonstrate that PAMotion achieves state-of-the-art performance in generating realistic, physically consistent human-object interaction sequences. Our approach significantly reduces physically implausible artifacts while maintaining high motion diversity.

In summary, our main contributions are as follows,

1. **Physics-Aware Motion Reasoning:** we introduce the physics-aware interaction loss that uses object accelerations to enforce consistency between physical interactions and human-object contact states, bridging kinematic synthesis and physical reasoning.
2. **Coarse-to-Fine Generation Strategy:** We design a hierarchical generation pipeline that progressively refines global and local motion components for high-quality language-motion consistency and fine-grained physical plausibility.
3. **Superior Performance:** PAMotion achieves superior performance on multiple benchmarks, setting a new standard for physically realistic full-body human-object interaction generation, enabling potential real-world applications.

## 2. Related Work

### 2.1. Text-guided Human Motion Generation

Along with the rapid development of large-scale motion-language datasets [21, 45, 68, 128], text-conditioned human motion generation has experienced remarkable progress. Recent methods [2, 8, 12, 14, 21, 22, 30, 32, 35, 41, 46, 51, 55, 65, 66, 71, 82, 89, 92, 106, 112, 114, 118, 120, 123, 125, 127, 129, 130, 133, 135, 137] leverage diffusion, transformer, and generative language-motion alignment techniques to produce semantically coherent, and physically plausible motion sequences conditioned on natural language. These approaches can synthesize a wide range of human activities [106], from everyday actions to expressive or stylized movements, and even extend to multi-person interaction generation [20, 27, 40, 49, 90, 91], showcasing strong generalization to complex social scenarios. Despite their impressive achievements, most existing models are primarily designed for human-only motion synthesis and do not explicitly account for interactions with surrounding objects or physical constraints arising from the environment. As a result, they often struggle in scenarios that involve rich

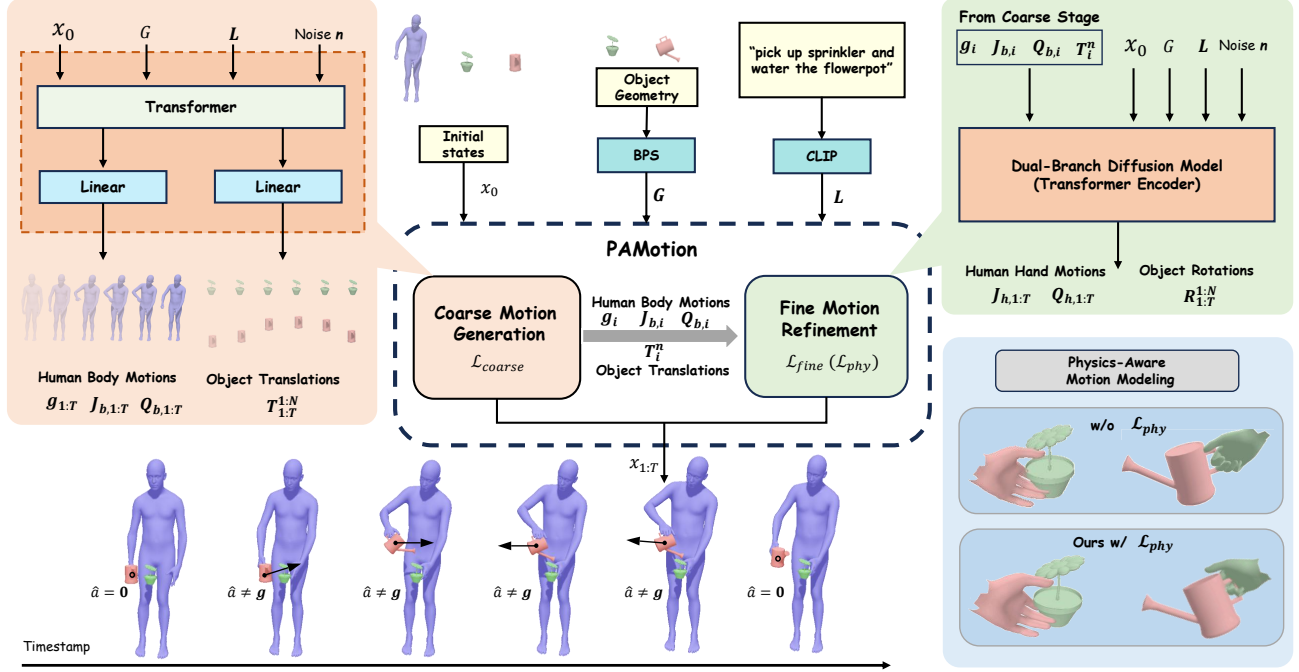


Figure 2. **Illustration of PAMotion.** PAMotion is a two-stage coarse-to-fine conditional diffusion framework. In the Coarse Motion Generation stage, the model predicts coarse, text-aligned global motion by generating global human body translation  $g_i$ , human body motions ( $J_{b,i}, Q_{b,i}$ ) and the object’s translational state  $T_i^n$ , conditioned on initial state  $x_0$ , object geometry  $\mathbf{G}$  and text embedding  $\mathbf{L}$ . We supervise the coarse stage with  $\mathcal{L}_{coarse}$ . In Fine Motion Refinement stage, we generate fine-grained hand articulation ( $J_{h,i}, Q_{h,i}$ ) and the object’s rotational state  $R_i^n$ , conditioned on results from the coarse stage and  $x_0, \mathbf{G}, \mathbf{L}$ . This refinement is guided by  $\mathcal{L}_{fine}$ , which incorporates the Physics-Aware Interaction Loss  $\mathcal{L}_{phy}$  to enforce consistency between object acceleration  $\hat{a}$  and human–object contact states. We demonstrate that  $\mathcal{L}_{phy}$  ensures physically plausible interactions and effectively mitigates hand penetration and object floating.

human–object interactions or require simultaneous manipulation of multiple objects [33, 53], where understanding contact dynamics and environmental affordances becomes crucial for generating realistic and coherent behavior.

## 2.2. Human-Object Interaction Generation

Researchers first focus on hand–object interactions [5, 9, 13, 26, 31, 34, 43, 44, 48, 54, 61, 64, 84, 87, 93, 98, 105, 107, 109, 110, 113, 116, 121, 122, 134, 136], as the hands are the most frequently used body parts for direct physical contact with surrounding objects in daily activities. Early studies [78] in this direction concentrate on grasp generation, hand and object pose estimation [16, 115], and local interaction reasoning from visual or textual cues, enabling fine-grained manipulation dynamics at the hand level.

More recently, research has expanded toward full-body dynamic human–object interaction (HOI) generation [78], aiming to synthesize temporally coherent and physically plausible full-body interactions involving both human motion and object dynamics. Such works can be broadly categorized into kinematics-based approaches [10, 15, 19, 23, 25, 36–38, 42, 57, 63, 73, 75, 76, 80, 85, 94–96, 102, 104, 126], which focus on geometric or motion prior constraints,

and physics-based approaches [1, 4, 6, 11, 24, 47, 52, 58, 60, 81, 83, 86, 88, 97, 100, 101, 103, 111], which explicitly model forces, contacts, and environmental constraints to achieve physically grounded interaction realism. These advancements are largely supported by the emergence of high-quality 3D HOI datasets [3, 17, 28, 29, 33, 42, 50, 53, 56, 79, 99, 108, 117, 119, 131], which capture complex scenes, multi-contact scenarios, and rich object categories, thereby enabling data-driven learning of realistic interaction dynamics at scale. In this paper, we focus on developing a physics-aware approach for modeling human interactions with multiple objects, building upon the HIMO-Gen [53].

## 3. Method

The goal of PAMotion is to generate realistic and physically plausible human–multi-object interaction (HOI) sequences  $X = [x_i], i \in \{1, \dots, T\}$ , conditioned on textual guidance  $\mathbf{L}$ , object geometry  $\mathbf{G} = [g^n], n \in \{1, \dots, N\}$ , and initial state  $x_0$ , where  $T$  denotes total frame number and  $N$  denotes the object number. Each frame  $x_i = (h_i, o_i^n)$  contains the human state  $h_i$  and object states  $[o_i^n], n \in \{1, \dots, N\}$ . We use SMPL-X [62] to represent the human

and  $R_{6D}$  [138] to parameterize human joint rotation. We further decompose  $h_i$  into global translation  $g_i \in \mathbb{R}^3$  and joint poses  $(J_i, Q_i)$ . The joint position  $J_i$  and rotation  $Q_i$  are split into body part  $J_{b,i}, Q_{b,i}$  and hand part  $J_{h,i}, Q_{h,i}$  respectively, with  $J_{b,i} \in \mathbb{R}^{22 \times 3}$ ,  $J_{h,i} \in \mathbb{R}^{30 \times 3}$ ,  $Q_{b,i} \in \mathbb{R}^{22 \times 6}$  and  $Q_{h,i} \in \mathbb{R}^{30 \times 6}$ . The object state  $o_i^n$  consists of translation  $T_i^n \in \mathbb{R}^3$  and rotation  $R_i^n \in \mathbb{R}^6$ .

**Overview:** To generate the full motion sequence, we propose PAMotion, a two-stage coarse-to-fine conditional diffusion framework, as shown in Fig. 2. The first stage predicts coarse global motion aligned with the text, generating  $(g_i, J_{b,i}, Q_{b,i})$  and object translation state  $T_i^n$ . The second stage then refines fine-grained physical details, including hand articulation  $(J_{h,i}, Q_{h,i})$  and object rotation state  $R_i^n$ . This refinement is guided by a Physics-Aware Interaction Loss  $\mathcal{L}_{\text{phy}}$ , which enforces consistency between object acceleration and human-object contact states, ensuring physically plausible interactions and effectively preventing hand penetration and object floating.

### 3.1. Physics-Aware Motion Modeling

We first introduce the basic theory of our physics-aware approach and then describe how we formalize this theory into a concise yet powerful Physics-Aware Interaction Loss ( $\mathcal{L}_{\text{phy}}$ ) to ensure physical plausibility between human and object motions in the generated sequences.

Our key insight is that an object’s acceleration serves as a critical indicator of its contact state with humans or other objects in everyday, slow-motion interactions. We leverage this observation to bridge the gap between purely kinematic motion generation and physically grounded interaction reasoning. This observation can be categorized into three representative cases. First, as illustrated in Fig. 3 (a), when the object (apple) is solely influenced by gravity, such as during free fall or along a parabolic trajectory—its acceleration  $\hat{\mathbf{a}}$  should approximately equal the gravitational acceleration  $\mathbf{g}$  (i.e.,  $\hat{\mathbf{a}} \approx \mathbf{g}$ ). In this condition, the object has no functional contact with the human body or any other object. We refer to this as the **Free-Motion State**. Second, as shown in Fig. 3 (b), when the object (apple) is held stationary in mid-air ( $\hat{\mathbf{a}} = \mathbf{0}$ ), the gravitational force acting on it is counterbalanced by the upward force exerted by the hand, resulting in a state of static equilibrium. Third, as illustrated in Fig. 3 (c), when the object (knife) is used to cut the apple, it experiences external forces from both the hand and the apple, causing its acceleration to deviate from  $\mathbf{g}$  (i.e.,  $\hat{\mathbf{a}} \neq \mathbf{g}$ ). In both the second and third cases, the object’s acceleration reveals the presence of contact forces, indicating direct or indirect interaction with the human. We therefore define these cases collectively as the **Contact-Motion State**.

Directly enforcing  $\hat{\mathbf{a}} = \mathbf{g}$  or constructing contact losses can be unstable and impractical in real-world scenarios. This instability arises from that (i) the human hand can un-

dergo subtle deformations, (ii) the measured acceleration is often noisy, and (iii) our physical assumptions may not always hold. Nevertheless, the loss function should remain effective under these imperfections. To address these challenges, we propose a more concise and robust loss formulation that enforces the underlying physical intuition in a soft manner. Specifically, we establish a differentiable constraint between the object’s acceleration  $\hat{\mathbf{a}}$  and the minimum distance  $d_t$  from the object to other objects or the hands. For simplicity, we only consider the hands and omit contact with the torso. Intuitively, this formulation ensures that when an object exhibits non-trivial acceleration ( $\hat{\mathbf{a}} \neq \mathbf{g}$ ), that is, when it is in the *Contact-Motion State*, it must maintain a close, non-penetrating contact with the hands or other objects. Based on this principle, we introduce the *Physics-Aware Interaction Loss*  $\mathcal{L}_{\text{phy}}$ , applied during the fine-grained generation stage, formalized as:

$$\mathcal{L}_{\text{phy}} = \mathbb{E}_t \left[ \left| \log \left( \frac{d_t}{\beta} \right) \right| \cdot |(a_t - g) \cdot t| \right] \quad (1)$$

where  $t$  indexes the frame time, and  $\beta$  is a hyperparameter allowing for subtle deformation of hands and objects.

This loss  $\mathcal{L}_{\text{phy}}$  works as follows,

- **When the object is Free-Motion State** ( $a_t \approx g$ ): Regardless of the value of  $|\log(d_t/0.01)|$ , it is scaled by a near-zero value  $|(a_t - g) \cdot t|$ , driving  $\mathcal{L}_{\text{phy}}$  toward zero. That means the hand and other objects are free to move independently, without need to consider the contact.
- **When the object is in Contact-Motion State:** Here,  $|(a_t - g) \cdot t|$  acts as a positive weight. To minimize  $\mathcal{L}_{\text{phy}}$ , the model needs to optimize and reduce  $|\log(d_t/\beta)|$ , effectively driving the condition  $d_t \rightarrow \beta$ .

The advantages of this formulation include:

- **Jointly penalizes floating and penetration:** As formulated in Eq. 1, we introduce a heuristic loss that enforces physically plausible contact. When the object’s acceleration  $\hat{\mathbf{a}} \neq \mathbf{g}$ , it is expected to be in contact with the hand or other surrounding objects. Due to the  $\log(\cdot)$  function, the loss increases sharply when penetration occurs, effectively discouraging interpenetration. Conversely, when the object is slightly detached (floating) from other surfaces, the loss increases gradually, providing soft supervision that avoids imposing excessive constraints in situations where physical contact is not intuitively required.
- **Dynamic activation:** The term  $|(a_t - g) \cdot t|$  functions as a dynamic weight that activates the distance constraint only when physically meaningful—specifically, when the object is in the Contact-Motion State.

With this loss function, we effectively couple object motion and physical interactions, guiding the model to generate physically consistent sequences that are free of penetration and maintain proper contact.



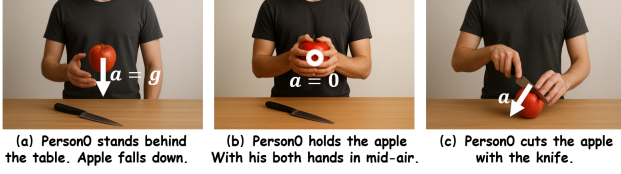


Figure 3. Illustration of Physics-Aware Motion Modeling. Object acceleration  $\hat{a}$  as an indicator of human-object contact state. Free-Motion State: (a) the apple is influenced only by gravity ( $\hat{a} = g$ ). Contact-Motion State: (b) the apple is held in equilibrium ( $\hat{a} = 0$ ); (c) the knife interacts with the apple and the hand, causing  $\hat{a} \neq g$ . We transform these three cases into  $\mathcal{L}_{\text{phy}}$ .

### 3.2. PAMotion: Coarse-to-Fine Generation

Our motion generation pipeline adopts a two-stage coarse-to-fine generation strategy built upon conditional diffusion models. The coarse stage predicts global human-object motion, while the fine stage refines hand articulation, object rotation and enforces physically grounded interactions.

**Motivation.** Since humans typically describe tasks from a high-level perspective, the purpose of the coarse stage is to capture and fulfill the overarching goals implied by the textual command. For instance, an instruction such as “move the desk lamp from the table to the bed” provides an abstract objective without specifying fine-grained details such as hand trajectories or object orientations. Generating the complete motion of the human body, including both torso and hand movements together with detailed object dynamics may lead to an imbalance in the network’s learning process, where it struggles to reconcile high-level goal achievement with low-level physical precision. A natural idea, therefore, is to first estimate the goal-related motions defined by the input text, and then use these motions as conditions to optimize the local physical interactions.

**Stage I: Coarse Global Interaction Synthesis.** This stage generates human body motion ( $g_i^T, J_{b,i}, Q_{b,i}$ ) and object translation state  $T_i^n$ , where  $i \in \{1, \dots, T\}$  denotes the frame index inside the motion sequences and  $n \in \{1, \dots, N\}$  denotes the object index inside total  $N$  objects. We follow HIMO [53] to adopt a dual-branch diffusion network, where one branch is designed to human motion generation and the other to object state generation. Information exchange between the two branches is facilitated through the mutual interaction module [53]. The conditions include textual guidance  $\mathbf{L}$ , object geometry  $\mathbf{G} = [g^n], n \in \{1, \dots, N\}$ , and initial state  $x_0$ .  $\mathbf{L}$  is encoded via a pre-trained CLIP [72] encoder. We encode  $\mathbf{G}$  using the Basis Point Set (BPS) representation [70]. We supervise the generated motion using kinematic and geometric consistency terms. For human body joints  $J_{b,i}$ , we employ an  $\mathcal{L}_2$  loss to constrain both their positions and first-order derivatives, thereby promoting spatial alignment and temporal smooth-

ness.

$$\mathcal{L}_{\text{pv}} = \sum_{i=1}^T \|J_{b,i} - \hat{J}_{b,i}\|_2^2 + \sum_{i=1}^T \|\dot{J}_{b,i} - \hat{\dot{J}}_{b,i}\|_2^2. \quad (2)$$

For  $Q_{b,i}, T_i^n$  and  $g_i^T$ , we use similar formulations as  $\mathcal{L}_{\text{pv}}$ , yielding  $\mathcal{L}_{\text{qv}}, \mathcal{L}_{\text{tv}}$  and  $\mathcal{L}_{\text{gv}}$  respectively. To maintain stable spatial relationships among objects, we additionally supervise relative object distances,

$$\mathcal{L}_{\text{dist}} = \sum_{i=1}^T \|d_{mn,i} - \hat{d}_{mn,i}\|_2^2. \quad (3)$$

where  $d_{mn,i}$  calculate the center distance of objects  $o_i^m$  and  $o_i^n$ . Thereby the total loss for the coarse stage is,

$$\mathcal{L}_{\text{coarse}} = \mathcal{L}_{\text{pv}} + \mathcal{L}_{\text{qv}} + \mathcal{L}_{\text{tv}} + \mathcal{L}_{\text{gv}} + \lambda_0 \mathcal{L}_{\text{dist}} \quad (4)$$

where  $\lambda_0$  is a weighting constant.

**Stage II: Fine-Grained Physical Interaction Generation.** This stage generates human hand motion ( $J_{h,i}, Q_{h,i}$ ) and object rotation state  $R_i^n$ . Besides the conditions employed in Stage I, we also include the coarse stage results ( $g_i^T, J_{b,i}, Q_{b,i}$ ) and  $T_i^n$  as conditions. All the conditions are pre-encoded into latents and integrated into the diffusion network. We also adopt the dual-branch architecture as in Stage I but modifies the last output layer to output corresponding hand motion and object rotation. For acceleration  $\hat{a}_i^n$  of object  $o_i^n$ , we randomly sample 1024 points on the object surface and for any selected point  $p$  on the surface of  $o_i^n$ ,  $\hat{a}_i^n$  calculates as  $\hat{a}_i^n = R_i^n (\dot{\omega} \times p + \omega \times (\omega \times p)) + \dot{T}_i^n$ , where  $\omega$  denotes the angular velocity vector. Since  $\mathcal{L}_{\text{phy}}$  in Eq. 1 computes the magnitude of  $|(a_t - g) \cdot t|$ , for object  $o_i^n$ , we simply select the maximum value of  $|(a_t - g) \cdot t|$  by evaluating all accelerations of all points on  $o_i^n$ . Besides  $\mathcal{L}_{\text{phy}}$ , we supervise ( $J_{h,i}, Q_{h,i}$ ) and  $R_i^n$  in a similar formulation to Eq. 2, yielding  $\mathcal{L}_{\text{pv}}^r, \mathcal{L}_{\text{qv}}^r$  and  $\mathcal{L}_{\text{rv}}^r$ . Therefore, the total loss for the fine stage is,

$$\mathcal{L}_{\text{fine}} = \mathcal{L}_{\text{pv}}^r + \mathcal{L}_{\text{qv}}^r + \mathcal{L}_{\text{rv}}^r + \lambda_1 \mathcal{L}_{\text{phy}} \quad (5)$$

where  $\lambda_1$  is a weighting constant.

## 4. Experiments

### 4.1. Experiment Setup

**Datasets.** We conduct experiments on HIMO [53] and ParaHome [33] datasets. HIMO is a large-scale 4D MoCap dataset focusing on full-body human interactions with multiple objects. It contains 3.3K 4D HOI sequences captured from 34 subjects interacting with 53 daily objects. Each sequence includes fine-grained textual descriptions and temporal segments. We follow the official train/test split for training and evaluation. ParaHome comprises 207 motion

Table 1. Quantitative comparison under two-object and three-object settings on the HIMO dataset. PAMotion outperforms all state-of-the-art methods on nearly all metrics—except MM-Dist in the two-object case—and achieves a clear performance margin in the three-object case. These results demonstrate PAMotion’s strong ability to model complex human–object interactions. ‘↑’: higher is better, ‘↓’: lower is better, ‘→’: closer to ground truth is better.

Method	Two-object				Three-object			
	R-Prec. ↑	FID ↓	MM-Dist ↓	Diversity →	R-Prec. ↑	FID ↓	MM-Dist ↓	Diversity →
Real	0.7988	0.0176	3.5659	11.3973	0.6988	0.1811	3.7696	9.7674
IMoS [19]	0.5013	7.5890	8.7402	7.0033	0.4662	4.9902	7.7702	9.2310
priorMDM [74]	0.5891	7.8517	7.2509	12.5799	0.5137	4.8210	5.8900	9.3402
HIMO-Gen [53]	0.6369	1.4811	<b>3.6491</b>	11.6603	0.5350	4.7712	5.0866	8.9460
<b>Ours</b>	<b>0.6914</b>	<b>0.8285</b>	3.9841	<b>11.4431</b>	<b>0.6750</b>	<b>1.3763</b>	<b>3.7707</b>	<b>9.4573</b>

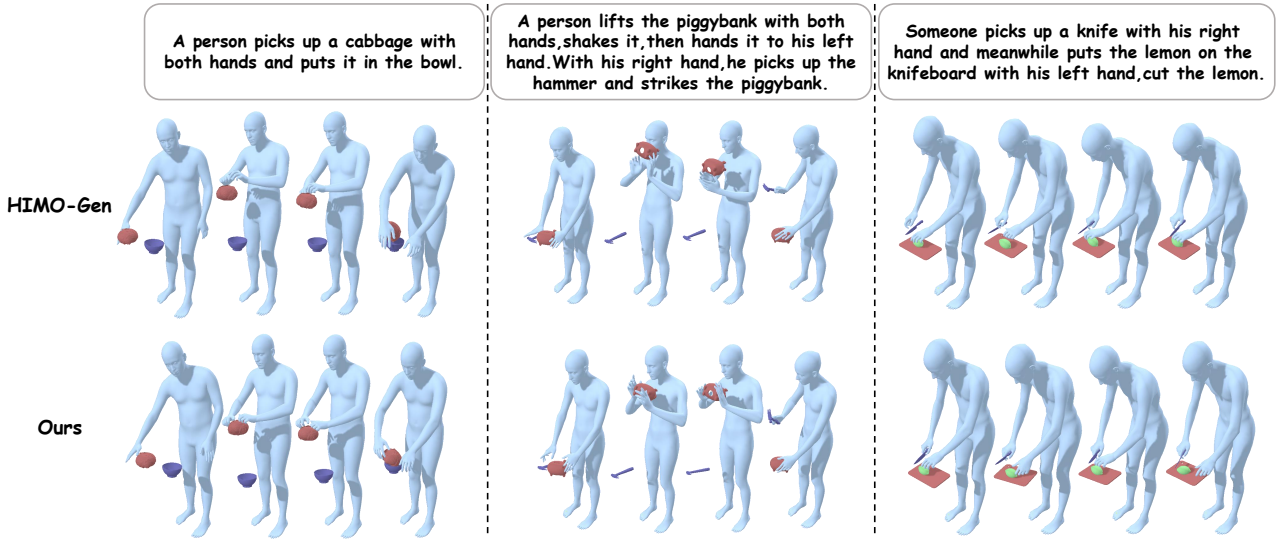


Figure 4. Qualitative Comparison on HIMO dataset [53]. Our method surpasses HIMO-Gen in generating physically plausible human-object interactions.

captures from 38 participants interacting with 22 objects. From these, we randomly select 500 human–object interaction (HOI) sequences, each accompanied by detailed textual descriptions. The dataset is split into 80% for training, 10% for validation, and 10% for testing.

**Implementation Details.** Our model was trained on a single NVIDIA GeForce RTX 3090 GPU with a batch size of 32 for 1000 epochs. The training process took approximately 25 hours in total, with 10 hours for coarse training and 15 hours for fine stage training. Input text is encoded using a frozen CLIP-ViT-B/32 model. The attention module uses 4 heads with a latent dimension of 512. The denoising network consists of  $\ell_{\text{enc}} = 8$  layers of mutual interaction modules. We set  $\{\beta, \lambda_0, \lambda_1\} = \{0.01, 1.0, 0.1\}$  in all experiments unless specified. During training, the fine-stage diffusion network is conditioned on the ground-truth human torso and object translations. During inference, these con-

ditions are replaced with the predictions generated by the coarse stage. PAMotion takes 2.1 seconds to perform 1,000 diffusion steps and generate the motions for all objects and humans.

**Evaluation Metrics.** We adopt the evaluation protocol established by the HIMO benchmark [53] and based on [21] to quantitatively assess the quality of generated HOI sequences. Following the standard setup, all metrics are computed using a pre-trained motion-text feature extractor that captures both human and object motion dynamics. The metrics used in our evaluation include: *R-Precision (Top-3)*: evaluates whether the correct text-motion pair appears among the top three closest matches in the latent feature space, reflecting text-motion alignment. *MM-Dist*: measures the distance between generated motions and their corresponding text embeddings, quantifying semantic consistency. *Fréchet Inception Distance (FID)*: assesses the sim-

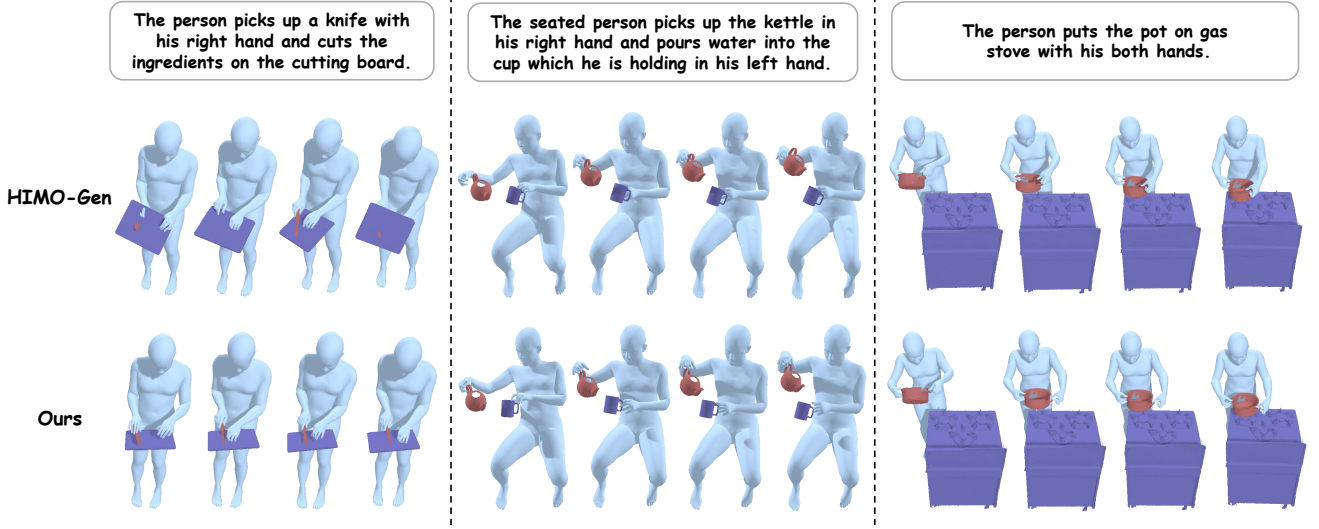


Figure 5. Qualitative Comparison on ParaHome dataset [33]. Our method outperforms HIMO-Gen in generating physically plausible human-object interactions.

ilarity between the distributions of generated motions and real motions in the encoded feature space. *Diversity*: captures the variability of generated motions across different text prompts, indicating the model’s ability to produce diverse outputs.

## 4.2. Comparison with State-of-the-Art Methods

**Baseline.** To comprehensively evaluate the performance of our proposed method, we conduct a direct comparison against the state-of-the-art method HIMO-Gen from the HIMO benchmark [53]. Additionally, we include several key baselines: priorMDM [74], and IMOS [19], which we adapted according to the HIMO setup.

**Results.** In Tab. 1, we compare PAMotion with state-of-the-art methods on the HIMO dataset. As can be clearly observed, our method surpasses all competing approaches, including the recent HIMO-Gen [53], by a significant margin across all metrics except MM-Dist. Under the two-object setting, our method outperforms HIMO-Gen by 5.45% and 0.653 on R-Precision and FID, respectively, demonstrating the effectiveness of our physical interaction loss and coarse-to-fine design. When comparing MM-Dist, our method shows a slight degradation of 0.335 (3.9841 vs. 3.6491), indicating that when the number of objects is relatively small, both HIMO-Gen and our approach can maintain strong motion–text consistency. When moving to the three-object setting, the increased number of objects leads to more complex and challenging human–object interactions. In this scenario, our method clearly outperforms all competitors by a substantial margin, showing an even greater advantage compared to the two-object setting. Specifically, our

approach exceeds the second-best method, HIMO-Gen, by 14.0%, 3.395, and 1.316 on R-Precision, FID, and MM-Dist, respectively. In terms of Diversity, our method also better approximates the real motion distribution, achieving 9.4573 (ours) vs. 9.7674 (Real), compared with 8.9460 (HIMO-Gen) vs. 9.7674 (Real).

Qualitative results shown in Fig. 1 and Fig. 4 further demonstrate the superiority of our approach in generating fine-grained physical interactions. For different objects, our method produces physically plausible human–object interactions, while the baseline method HIMO-Gen often exhibits noticeable floating or penetration artifacts.

In Table 4, we present the results on the ParaHome dataset. Compared to the baseline method HIMO-Gen, our approach consistently outperforms it across all evaluation metrics, with a particularly large improvement in FID (0.7962 vs. 3.2398). Fig. 5 shows qualitative results on ParaHome. Our method consistently surpasses HIMO-Gen in generation quality.

## 4.3. Ablation Study

We conduct ablation experiments to evaluate the effectiveness of the physical interaction loss  $\mathcal{L}_{\text{phy}}$  in both two-object and three-object settings. The results are reported in Tab. 2 and 3. Removing  $\mathcal{L}_{\text{phy}}$  leads to a consistent performance drop across all metrics. In the two-object case, the full model achieves a lower FID (0.8285 vs. 0.9046) and higher R-Precision (0.6914 vs. 0.6758), indicating better visual fidelity and text-motion alignment. Similarly, in the three-object scenario, the full model surpasses the variant without  $\mathcal{L}_{\text{phy}}$ , reducing FID from 1.5736 to 1.3763 and improving

Table 2. Ablation study of  $\mathcal{L}_{\text{phy}}$  on two-object setting on HIMO dataset. The results verify the effectiveness of  $\mathcal{L}_{\text{phy}}$ .

Method	R-Prec. $\uparrow$	FID $\downarrow$	MM-Dist $\downarrow$	Diversity $\rightarrow$
Ours (full)	0.6914	0.8285	3.9841	11.4431
Ours w/o $\mathcal{L}_{\text{phy}}$	0.6758	0.9046	4.0274	11.5996

Table 3. Ablation study of  $\mathcal{L}_{\text{phy}}$  on three-object setting on HIMO dataset. The results verify the effectiveness of  $\mathcal{L}_{\text{phy}}$ .

Method	R-Prec. $\uparrow$	FID $\downarrow$	MM-Dist $\downarrow$	Diversity $\rightarrow$
Ours (full)	0.6750	1.3763	3.7707	9.4573
Ours w/o $\mathcal{L}_{\text{phy}}$	0.6312	1.5736	3.8260	9.4524

Table 4. Quantitative comparison on the ParaHome dataset. PAMotion consistently outperforms the baseline method HIMO-Gen across all metrics. In particular, PAMotion achieves a significantly lower FID score (0.7962 vs. 3.2398), demonstrating its superior generation quality.

Method	R-Prec. $\uparrow$	FID $\downarrow$	MM-Dist $\downarrow$	Diversity $\rightarrow$
Real	0.6818	0.0017	5.3107	6.4100
HIMO-Gen	0.5909	3.2398	5.4455	6.1703
Ours	<b>0.6364</b>	<b>0.7962</b>	<b>5.3356</b>	<b>6.3145</b>

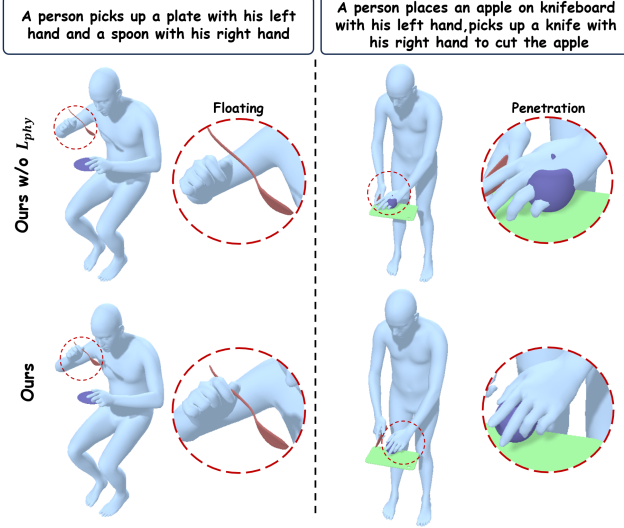


Figure 6. Ablation Study of  $\mathcal{L}_{\text{phy}}$  on HIMO dataset. The results verify that  $\mathcal{L}_{\text{phy}}$  effectively mitigates float and penetration problems in human-object interaction generation.

R-Precision from 0.6312 to 0.6750. Visualization results in Fig. 6 demonstrate that  $\mathcal{L}_{\text{phy}}$  effectively enhances physical plausibility by mitigating floating and penetration problems in human-object interactions.

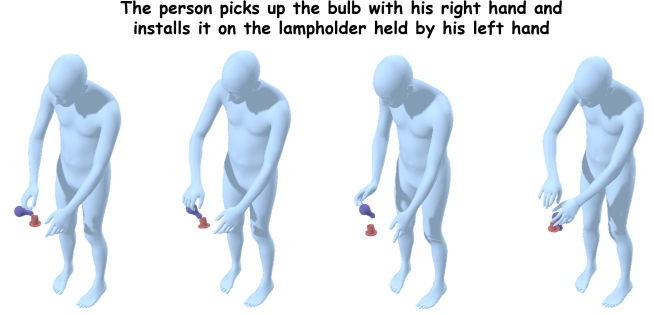


Figure 7. Failure Case on HIMO dataset. Although the hand interacts with the bulb, the grasp pose is physically implausible

#### 4.4. Failure Cases and Limitations.

We illustrate a failure case in Fig. 7. Although  $\mathcal{L}_{\text{phy}}$  enforces that the hand and object maintain direct or indirect contact when the object is in the Contact-Motion State, it does not explicitly constrain the grasp configuration between them. As shown in Fig. 7, the hand interacts with the bulb, but the resulting grasp pose is misaligned and physically implausible. One potential solution is to incorporate a pretrained large-scale grasp model, such as GraspNet [18, 59], to further regularize the grasp pose. However, extending such an approach to human-multi-object interactions introduces additional challenges, including the need to ensure bimanual coordination and inter-hand consistency, which would require substantial data and training time.

## 5. Conclusion

In this work, we presented PAMotion, a Physics-Aware Motion diffusion framework for generating realistic, physically consistent human-object interactions conditioned on textual commands. Unlike previous single-human single-object approaches that treat motion synthesis purely as a kinematic prediction problem, PAMotion explicitly incorporates physical reasoning into the generative process. By leveraging the insight that object accelerations reveal underlying contact states, we introduced a physics-aware interaction loss that enforces consistency between human-object contact and physical dynamics. Combined with a coarse-to-fine generation strategy, our model first captures high-level task semantics and global motion trajectories, and then refines fine-grained hand-object interactions to ensure physical plausibility. Extensive experiments on the HIMO and ParaHome datasets demonstrate that PAMotion achieves state-of-the-art performance across multiple metrics, significantly reducing implausible artifacts such as object floating and penetration while maintaining high diversity and motion-text alignment. In future work, we plan to extend PAMotion to multi-human multi-object motion generation, further advancing its applicability to real-world scenarios.



## References

- [1] Jinseok Bae, Jungdam Won, Donggeun Lim, Cheol-Hui Min, and Young Min Kim. Pmp: Learning to physically interact with environments using part-wise motion priors. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023. [3](#)
- [2] German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 457–469, 2024. [2](#)
- [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](#)
- [4] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 464–473. IEEE, 2024. [3](#)
- [5] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1577–1585, 2024. [3](#)
- [6] Yu-Wei Chao, Jimei Yang, Weifeng Chen, and Jia Deng. Learning to sit: Synthesizing human-chair interactions via hierarchical control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5887–5895, 2021. [3](#)
- [7] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motion-llm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024. [1](#)
- [8] 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](#)
- [9] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. [3](#)
- [10] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6992–7001, 2020. [3](#)
- [11] Jieming Cui, Tengyu Liu, Nian Liu, Yaodong Yang, Yixin Zhu, and Siyuan Huang. Anyskill: Learning open-vocabulary physical skill for interactive agents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 852–862, 2024. [3](#)
- [12] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9760–9770, 2023. [2](#)
- [13] Sisi Dai, Wenhao Li, Haowen Sun, Haibin Huang, Chongyang Ma, Hui Huang, Kai Xu, and Ruizhen Hu. Interfusion: Text-driven generation of 3d human-object interaction. In *European Conference on Computer Vision*, pages 18–35. Springer, 2024. [3](#)
- [14] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*, pages 390–408. Springer, 2024. [2](#)
- [15] Divyanshu Daiya, Damon Conover, and Aniket Bera. Collage: Collaborative human-agent interaction generation using hierarchical latent diffusion and language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8203–8210. IEEE, 2025. [3](#)
- [16] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. So-pose: Exploiting self-occlusion for direct 6d pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12396–12405, 2021. [3](#)
- [17] 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](#)
- [18] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020. [8](#)
- [19] 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. Wiley Online Library, 2023. [3](#), [6](#), [7](#)
- [20] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European Conference on Computer Vision*, pages 418–437. Springer, 2024. [2](#)
- [21] 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](#)
- [22] 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. Springer, 2022. [2](#)
- [23] 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. [3](#)
- [24] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023. [3](#)
- [25] Wenkun He, Yun Liu, Ruitao Liu, and Li Yi. Syncdiff: Synchronized motion diffusion for multi-body human-object interaction synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11731–11743, 2025. [3](#)
- [26] Zhi Hou, Baosheng Yu, and Dacheng Tao. Compositional 3d human-object neural animation. *arXiv preprint arXiv:2304.14070*, 2023. [3](#)
- [27] Buzhen Huang, Chen Li, Chongyang Xu, Liang Pan, Yangang Wang, and Gim Hee Lee. Closely interactive human reconstruction with proxemics and physics-guided adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1011–1021, 2024. [2](#)
- [28] 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. Springer, 2022. [3](#)
- [29] 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](#)
- [30] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Runyi Yu, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Local action-guided motion diffusion model for text-to-motion generation. In *European Conference on Computer Vision*, pages 392–409. Springer, 2024. [2](#)
- [31] Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models. In *European Conference on Computer Vision*, pages 400–419. Springer, 2024. [3](#)
- [32] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8255–8263, 2023. [2](#)
- [33] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1816–1828, 2025. [3](#), [5](#), [7](#)
- [34] Taeksoo Kim, Shunsuke Saito, and Hanbyul Joo. Ncho: Unsupervised learning for neural 3d composition of humans and objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14817–14828, 2023. [2](#), [3](#)
- [35] Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. Priority-centric human motion generation in discrete latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14806–14816, 2023. [2](#)
- [36] Franziska Krebs, Andre Meixner, Isabel Patzer, and Tamim Asfour. The kit bimanual manipulation dataset. In *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, pages 499–506. IEEE, 2021. [3](#)
- [37] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 947–957, 2024.
- [38] Jiye Lee and Hanbyul Joo. Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9663–9674, 2023. [3](#)
- [39] Meng-Lun Lee, Wansong Liu, Sara Behdad, Xiao Liang, and Minghui Zheng. Robot-assisted disassembly sequence planning with real-time human motion prediction. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(1):438–450, 2022. [1](#)
- [40] Baiyi Li, Edmond SL Ho, Hubert PH Shum, and He Wang. Two-person interaction augmentation with skeleton priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. [2](#)
- [41] Chuqiao Li, Julian Chibane, Yannan He, Naama Pearl, Andreas Geiger, and Gerard Pons-Moll. Unimotion: Unifying 3d human motion synthesis and understanding. In *2025 International Conference on 3D Vision (3DV)*, pages 240–249. IEEE, 2025. [2](#)
- [42] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. [3](#)
- [43] Lei Li and Angela Dai. Genzi: Zero-shot 3d human-scene interaction generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20465–20474, 2024. [3](#)
- [44] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3035–3044, 2024. [3](#)
- [45] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36: 25268–25280, 2023. [2](#)
- [46] Hanchao Liu, Xiaohang Zhan, Shaoli Huang, Tai-Jiang Mu, and Ying Shan. Programmable motion generation for open-set motion control tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2024. [2](#)
- [47] Libin Liu and Jessica Hodgins. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. *Acm transactions on graphics (tog)*, 37(4):1–14, 2018. [3](#)

- [48] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20609–20620, 2023. 3
- [49] Yunze Liu, Changxi Chen, and Li Yi. Interactive humanoid: Online full-body motion reaction synthesis with social affordance canonicalization and forecasting. *arXiv preprint arXiv:2312.08983*, 2023. 2
- [50] Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1769–1782, 2025. 3
- [51] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Human-tomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*, 2023. 2
- [52] Zhengyi Luo, Jiashun Wang, Kangni Liu, Haotian Zhang, Chen Tessler, Jingbo Wang, Ye Yuan, Jinkun Cao, Zihui Lin, Fengyi Wang, et al. Smplolympics: Sports environments for physically simulated humanoids. *arXiv preprint arXiv:2407.00187*, 2024. 3
- [53] 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, 2, 3, 5, 6, 7
- [54] Junyi Ma, Jingyi Xu, Xieyuanli Chen, and Hesheng Wang. Diff-ip2d: Diffusion-based hand-object interaction prediction on egocentric videos. *arXiv preprint arXiv:2405.04370*, 2024. 3
- [55] Sihan Ma, Qiong Cao, Jing Zhang, and Dacheng Tao. Contact-aware human motion generation from textual descriptions. *arXiv preprint arXiv:2403.15709*, 2024. 2
- [56] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 3
- [57] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 329–336. IEEE, 2015. 3
- [58] 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. 3
- [59] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2901–2910, 2019. 8
- [60] Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes. In *2024 International Conference on 3D Vision (3DV)*, pages 1498–1507. IEEE, 2024. 3
- [61] Youxin Pang, Ruizhi Shao, Jiajun Zhang, Hanzhang Tu, Yun Liu, Boyao Zhou, Hongwen Zhang, and Yebin Liu. Manivideo: Generating hand-object manipulation video with dexterous and generalizable grasping. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12209–12219, 2025. 3
- [62] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. 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*, 2019. 3
- [63] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2878–2888, 2025. 3
- [64] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4726–4736, 2023. 2, 3
- [65] Mathis Petrovich, Michael J. Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European conference on computer vision*, 2022. 2
- [66] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9488–9497, 2023. 2
- [67] Ken Pfeuffer, Matthias J Geiger, Sarah Prange, Lukas Mecke, Daniel Buschek, and Florian Alt. Behavioural biometrics in vr: Identifying people from body motion and relations in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019. 1
- [68] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2
- [69] Nancy S Pollard, Jessica K Hodgins, Marcia J Riley, and Christopher G Atkeson. Adapting human motion for the control of a humanoid robot. In *Proceedings 2002 IEEE international conference on robotics and automation (Cat. No. 02CH37292)*, pages 1390–1397. IEEE, 2002. 1
- [70] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4332–4341, 2019. 5
- [71] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or. Single motion diffusion. *arXiv preprint arXiv:2302.05905*, 2023. 2
- [72] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learn-



- ing transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5
- [73] Haziq Razali and Yiannis Demiris. Action-conditioned generation of bimanual object manipulation sequences. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2146–2154, 2023. 3
- [74] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 6, 7
- [75] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Transactions on Graphics*, 38(6):178, 2019. 3
- [76] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics*, 39(4), 2020. 3
- [77] Stephan Streuber and Astros Chatziastros. Human interaction in multi-user virtual reality. In *10th International Conference on Humans and Computers (HC 2007)*, pages 1–6. University of Aizu, 2007. 1
- [78] Kewei Sui, Anindita Ghosh, Inwoo Hwang, Bing Zhou, Jian Wang, and Chuan Guo. A survey on human interaction motion generation. *arXiv preprint arXiv:2503.12763*, 2025. 2, 3
- [79] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020. 3
- [80] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13263–13273, 2022. 3
- [81] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 43(6):1–21, 2024. 3
- [82] 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. Springer, 2022. 2
- [83] Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit H Bermano, and Michiel van de Panne. Cload: Closing the loop between simulation and diffusion for multi-task character control. *arXiv preprint arXiv:2410.03441*, 2024. 3
- [84] Jie Tian, Ran Ji, Lingxiao Yang, Suting Ni, Yuexin Ma, Lan Xu, Jingyi Yu, Ye Shi, and Jingya Wang. Gaze-guided hand-object interaction synthesis: Dataset and method. *arXiv preprint arXiv:2403.16169*, 2024. 3
- [85] Weilin Wan, Lei Yang, Lingjie Liu, Zhuoying Zhang, Ruixing Jia, Yi-King Choi, Jia Pan, Christian Theobalt, Taku Komura, and Wenping Wang. Learn to predict how humans manipulate large-sized objects from interactive motions. *IEEE Robotics and Automation Letters*, 7(2):4702–4709, 2022. 3
- [86] Jiashun Wang, Jessica Hodgins, and Jungdam Won. Strategy and skill learning for physics-based table tennis animation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3
- [87] 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 *2022 International Conference on 3D Vision (3DV)*, pages 353–362. IEEE, 2022. 2, 3
- [88] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physshoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023. 3
- [89] Yin Wang, Mu Li, Jiapeng Liu, Zhiying Leng, Frederick WB Li, Ziyao Zhang, and Xiaohui Liang. Fg-t2m++: Lfms-augmented fine-grained text driven human motion generation. *International Journal of Computer Vision*, pages 1–17, 2025. 2
- [90] Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. Timotion: Temporal and interactive framework for efficient human-human motion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7169–7178, 2025. 2
- [91] Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. Intercontrol: Zero-shot human interaction generation by controlling every joint. *Advances in Neural Information Processing Systems*, 37:105397–105424, 2024. 2
- [92] Dong Wei, Xiaoning Sun, Huaijiang Sun, Shengxiang Hu, Bin Li, Weiqing Li, and Jianfeng Lu. Enhanced fine-grained motion diffusion for text-driven human motion synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5876–5884, 2024. 2
- [93] Qingxuan Wu, Zhiyang Dou, Sirui Xu, Soshi Shimada, Chen Wang, Zhengming Yu, Yuan Liu, Cheng Lin, Zeyu Cao, Taku Komura, et al. Dice: End-to-end deformation capture of hand-face interactions from a single image. *arXiv preprint arXiv:2406.17988*, 2024. 3
- [94] Qianyang Wu, Ye Shi, Xiaoshui Huang, Jingyi Yu, Lan Xu, and Jingya Wang. Thor: Text to human-object interaction diffusion via relation intervention. *arXiv preprint arXiv:2403.11208*, 2024. 3
- [95] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision*, pages 257–274. Springer, 2022.
- [96] 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. 3
- [97] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. *arXiv preprint arXiv:2309.07918*, 2023. 3
- [98] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision*, pages 125–145. Springer, 2022. 2, 3



- [99] 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
- [100] Zhaoming Xie, Sebastian Starke, Hung Yu Ling, and Michiel van de Panne. Learning soccer juggling skills with layer-wise mixture-of-experts. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022. 3
- [101] Zhaoming Xie, Jonathan Tseng, Sebastian Starke, Michiel van de Panne, and C Karen Liu. Hierarchical planning and control for box loco-manipulation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–18, 2023. 3
- [102] 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. 3
- [103] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12266–12277, 2025. 3
- [104] Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. D3d-hoi: Dynamic 3d human-object interactions from videos. *arXiv preprint arXiv:2108.08420*, 2021. 3
- [105] Zhu Xu, Qingchao Chen, Yuxin Peng, and Yang Liu. Semantic-aware human object interaction image generation. In *Forty-first International Conference on Machine Learning*, 2024. 3
- [106] Haiwei Xue, Xiangyang Luo, Zhanghao Hu, Xin Zhang, Xunzhi Xiang, Yuqin Dai, Jianzhuang Liu, Zhensong Zhang, Minglei Li, Jian Yang, et al. Human motion video generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2
- [107] ChangHee Yang, ChanHee Kang, Kyeongbo Kong, Hanni Oh, and Suk-Ju Kang. Person in place: Generating associative skeleton-guidance maps for human-object interaction image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8164–8175, 2024. 3
- [108] Jie Yang, Xuesong Niu, Nan Jiang, Ruimao Zhang, and Siyuan Huang. F-hoi: Toward fine-grained semantic-aligned 3d human-object interactions. In *European Conference on Computer Vision*, pages 91–110. Springer, 2024. 3
- [109] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16284–16295, 2024. 3
- [110] Yuhang Yang, Wei Zhai, Chengfeng Wang, Chengjun Yu, Yang Cao, and Zheng-Jun Zha. Egochoir: Capturing 3d human-object interaction regions from egocentric views. *Advances in Neural Information Processing Systems*, 37: 54529–54557, 2024. 3
- [111] Zeshi Yang, Kangkang Yin, and Libin Liu. Learning to use chopsticks in diverse gripping styles. *ACM Transactions on Graphics (TOG)*, 41(4):1–17, 2022. 3
- [112] Payam Jome Yazdian, Eric Liu, Rachel Lagasse, Hamid Mohammadi, Li Cheng, and Angelica Lim. Motionscript: Natural language descriptions for expressive 3d human motions. *arXiv preprint arXiv:2312.12634*, 2023. 2
- [113] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22479–22489, 2023. 3
- [114] Weihao Yuan, Yisheng He, Weichao Shen, Yuan Dong, Xiaodong Gu, Zilong Dong, Liefeng Bo, and Qixing Huang. Mogents: Motion generation based on spatial-temporal joint modeling. *Advances in Neural Information Processing Systems*, 37:130739–130763, 2024. 2
- [115] Chenyangguang Zhang, Guanlong Jiao, Yan Di, Gu Wang, Ziqin Huang, Ruida Zhang, Fabian Manhardt, Bowen Fu, Federico Tombari, and Xiangyang Ji. Moho: Learning single-view hand-held object reconstruction with multi-view occlusion-aware supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9992–10002, 2024. 3
- [116] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *2024 International Conference on 3D Vision (3DV)*, pages 235–246. IEEE, 2024. 3
- [117] 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
- [118] 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. 2
- [119] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-m<sup>3</sup>: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 516–526, 2024. 3
- [120] Jianrong Zhang, Hehe Fan, and Yi Yang. Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17592–17602, 2025. 2
- [121] Jiajun Zhang, Yuxiang Zhang, Liang An, Mengcheng Li, Hongwen Zhang, Zonghai Hu, and Yebin Liu. Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion.

*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

- [122] 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. Springer, 2020. 3
- [123] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems*, 36:13981–13992, 2023. 2
- [124] Pengfei Zhang, Pinxin Liu, Pablo Garrido, Hyeonwoo Kim, and Bindita Chaudhuri. Kinmo: Kinematic-aware human motion understanding and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11187–11197, 2025. 1
- [125] Wenhao Zhang, Meng Chen, and Yebin Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2
- [126] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision*, pages 518–535. Springer, 2022. 3
- [127] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7368–7376, 2024. 2
- [128] Yuhong Zhang, Jing Lin, Ailing Zeng, Guanlin Wu, Shunlin Lu, Yurong Fu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x++: A large-scale multimodal 3d whole-body human motion dataset. *arXiv preprint arXiv:2501.05098*, 2025. 2
- [129] Zihan Zhang, Richard Liu, Rana Hanocka, and Kfir Aberman. Tedi: Temporally-entangled diffusion for long-term motion synthesis. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2
- [130] Zeyu Zhang, Yiran Wang, Wei Mao, Danning Li, Rui Zhao, Biao Wu, Zirui Song, Bohan Zhuang, Ian Reid, and Richard Hartley. Motion anything: Any to motion generation. *arXiv preprint arXiv:2503.06955*, 2025. 2
- [131] Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I’m hoi: Inertia-aware monocular capture of 3d human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 729–741, 2024. 3
- [132] Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Xihan Wei, Liefeng Bo, et al. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025. 1
- [133] Kaifeng Zhao, Gen Li, and Siyu Tang. Dartcontrol: A diffusion-based autoregressive motion model for real-time text-driven motion control. *arXiv preprint arXiv:2410.05260*, 2024. 2
- [134] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2023. 3
- [135] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. In *European Conference on Computer Vision*, pages 405–421. Springer, 2024. 2
- [136] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 3
- [137] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Em dm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pages 18–38. Springer, 2024. 2
- [138] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 4