ContA-HOI: Towards Physically Plausible Human-Object Interaction Generation via Contact-Aware Modeling

Zhe Li[1] Jiakun Li[1] Mingqi Gao[2] Jinyu Yang[3] Wei Wang[4] Feng Zheng[1,5*]

[1]Southern University of Science and Technology [2]University of Sheffield
[3]tapall.ai [4]ZTE Corporation [5]Spatialtemporal AI

Abstract

In human-object interaction (HOI), physical contact between the body and objects is a primary determinant of realism and plausibility. Prior HOI methods typically encode relations via global joint-to-centroid or joint-to-boundary. Such strategies neglect contact anchors that are essential for defining joint-to-contact relations-where and how HOI occurs, thereby implicitly reducing the problem to nearest-distance optimization. Without explicit contact anchors and joint-to-contact dynamics, previous models drift toward artifacts: human-object penetration or unnatural object floating. We argue that modeling contact relationships by contact anchors is important for generating realistic HOIs, as it directly captures where and how humans physically interact with objects rather than merely minimizing spatial proximity. To address these limitations, we propose Contact-Aware HOI (ContA-HOI), a progressive framework that decomposes HOI generation into three synergistic stages: discovering where contact occurs, modeling how contact evolves, and guiding generation with contact constraints. First, a Contact Affordance Predictor (CAP) addresses the "where" by predicting precise object-surface contact anchors from text, human pose, and object geometry. Second, these anchors seed a Contact Relation Field (CRF) that captures "how" by modeling spatiotemporal dynamics of joint-to-contact relations throughout the interaction. Finally, a Contact Dynamics Model (CDM) learns a prior CRF evolution pattern and guides motion diffusion sampling by aligning the generated motion's CRF with this learned prior. On the FullBodyManipulation dataset, ContA-HOI yields more realistic and physically plausible HOIs, improving foot sliding and contact percentage over recent baselines.

1. Introduction

Human-object interaction generation (HOI) has emerged as a fundamental challenge in computer animation [14, 16, 17] and embodied AI [4, 15, 32, 35]. Text-driven HOI generation seeks to synthesize realistic, semantically coherent motions for both humans and objects directly from language, providing fine-grained controllability over intent, roles, and context [21, 26, 30, 34, 36]. While recent diffusion-based models [6, 12, 29] have made remarkable strides in enabling the generation of HOI [2, 19, 31, 33, 39], producing physically plausible interactions remains a fundamental challenge.

Existing HOI methods typically rely on global distance measures—joint-to-centroid or joint-to-boundary—that fail to capture *where* on the object contact should occur and *how* these contacts evolve temporally. These distance-based proxies optimize for spatial proximity rather than meaningful interaction, lacking contact-anchored spatial constraints and temporal dynamics constraints. Without explicit modeling of where and how human-object interacts, previous methods drift toward unrealistic HOI such as human-object penetration and unnatural object floating.

These limitations highlight the necessity of explicitly modeling the *where* and *how* contact occurs, which are critical to physically plausible HOI. Consider the examples in Figure 1: when lifting and rotating a box, the *where*—precise contact points on the box's sides (red dots, top)—determines hand placement. Without identifying these specific contact locations, methods resort to minimizing distances to object centroids, leading to unrealistic hovering or penetration. The *how*—the temporal evolution of joint-to-contact relations shown in the CRF distance plots—governs the coordination between hands throughout different interaction phases. By identifying and focusing on these critical contact regions, we can capture the true constraints that govern physical interaction: where bodies make contact with objects and how these contacts evolve to accomplish tasks.

To address these limitations, we propose ContA-HOI, a

[*]Corresponding author.

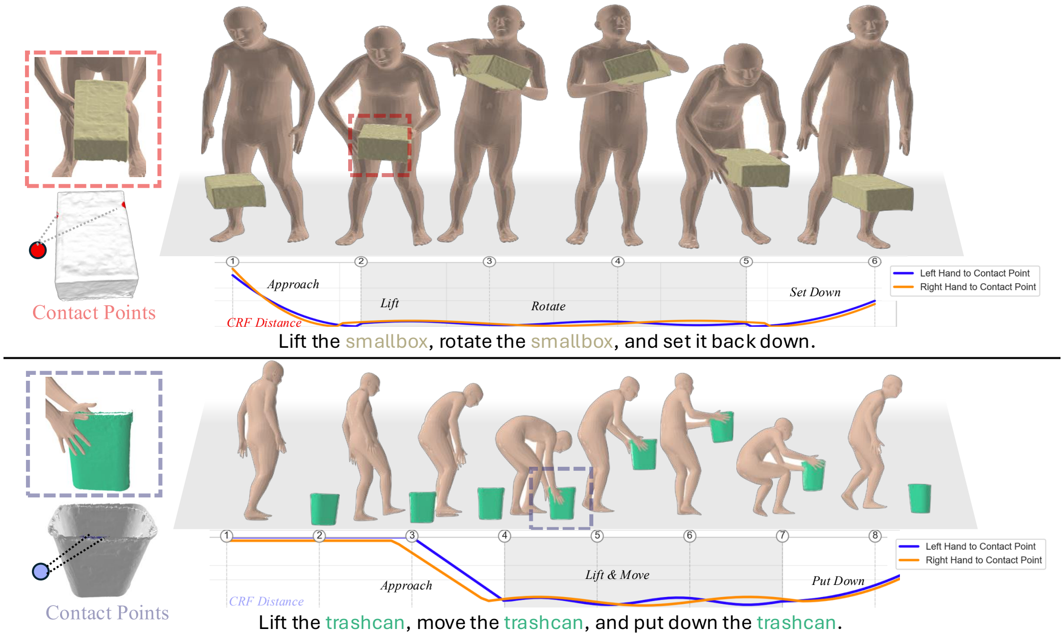
Contact Points
① ② ③ ④ ⑤ ⑥
Approach
Lift
Rotate
Set Down
CRF Distance
Left Hand to Contact Point
Right Hand to Contact Point
Lift the smallbox, rotate the smallbox, and set it back down.
Contact Points
① ② ③ ④ ⑤ ⑥ ⑦ ⑧
Approach
Lift & Move
Put Down
CRF Distance
Left Hand to Contact Point
Right Hand to Contact Point
Lift the trashcan, move the trashcan, and put down the trashcan.

Figure 1. Given initial human–object states and a text description, our framework generates synchronized motions with physically plausible contact. CAP localizes object contact points, while CRF tracks joint-to-contact distances across interaction phases, ensuring realistic manipulation without floating or penetration artifacts.

progressive contact-aware framework that decomposes HOI generation into three synergistic stages: discovering where contact occurs, modeling how contact evolves, and guiding generation with contact constraints. First, a Contact Affordance Predictor (CAP) addresses the ”where” by predicting precise object-surface contact anchors from text, human pose, and object geometry. CAP employs hierarchical attention that contextualizes human features with language before attending to object surfaces, producing contact likelihood maps validated through world-coordinate feasibility constraints. Second, these anchors seed a Contact Relation Field (CRF) that captures ”how” by modeling spatiotemporal dynamics of joint-to-contact relations throughout the interaction. Unlike dense representations, which compute all possible distances, CRF adaptively focuses on task-relevant relationships, creating a compact yet expressive representation of interaction dynamics. Finally, a Contact Dynamics Model (CDM) learns prior CRF evolution patterns from real HOI data and guides motion diffusion sampling by aligning the generated motions’ CRF with these learned priors. This guidance mechanism repeatedly steers the sampling process toward physically plausible interactions.

Our main contributions are summarized as follows:

- We propose **ContA-HOI**, a contact-aware framework that explicitly models **where** and **how** HOI occurs through three synergistic components, achieving physically plausible HOI generation.
- We introduce a progressive HOI pipeline: a **Contact Affordance Predictor (CAP)** that localizes where contact occurs, a **Contact Relation Field (CRF)** that models how joint-to-contact relations evolve, and a **Contact Dynamics Model (CDM)** learning these dynamics to guide diffusion sampling toward physically plausible HOIs.
- **ContA-HOI** achieves state-of-the-art performance on the FullBodyManipulation dataset, with notable improvements in reducing foot sliding and increasing contact percentage compared to recent baselines.

2. Related Work

2.1. Human-Object Interaction Generation

The field of human-object interaction generation has evolved from isolated human motion synthesis to integrated approaches that jointly model humans and objects. Early text-to-motion methods like MDM [31] and MotionDiffuse [38] achieved impressive results for human-only motion generation using diffusion models. These methods established the foundation of using transformer architectures and classifier-free guidance for motion synthesis. However, they lack the capability to model object dynamics and human-object interactions, limiting their applicability to real-world scenarios where humans constantly interact with their environment.

Recent works have shifted toward joint human-object synthesis. HOI-Diff [26] decomposes the problem into dual-branch diffusion for motion generation and affordance prediction for contact estimation. This modular approach allows for specialized modeling of different aspects but may struggle with maintaining consistency between branches. CHOIS [20] introduces controllable generation through sparse object waypoints, demonstrating that geometric constraints can guide realistic interactions. However, these waypoint-based methods require manual specification or rely on predefined trajectories, limiting their flexibility.

More advanced approaches like CG-HOI [7] explicitly model contact as proxy guidance, using contact maps between human body surface and object geometry. InterDreamer [23] pushes the boundary further by achieving zero-shot text-to-3D dynamic interactions through compositional generation. While these methods improve interaction realism, they typically process all possible human-object relations uniformly through dense distance fields or full contact maps. Our approach differs fundamentally by adaptively selecting only the most relevant relations for interaction, enabling more focused and efficient learning.

2.2. Human-Object Relation Modeling

Modeling spatial-temporal relations between humans and objects is crucial for realistic interaction synthesis. Early approaches relied on simple heuristics such as object centroids or nearest-point distances. These oversimplified representations fail to capture the rich geometric relationships in complex interactions. Recent methods have proposed more sophisticated relation modeling techniques.

NIFTY [18] introduces neural object interaction fields that output distances to valid interaction manifolds, providing continuous guidance for motion generation. FORCE [39] models interactions through physics-based force-resistance relationships, capturing how humans adapt their motions based on object properties. These methods demonstrate the importance of relation modeling but often require expensive computation of full distance fields or complex physics simulation.

In the vision domain, the HOT dataset [3] introduces detailed contact heatmaps with body-part-specific labels, advancing contact representation beyond simple binary masks. The recent InteractVLM [9] leverages large VLM for 3D contact estimation, showing that foundation models can provide priors for interaction understanding. However, these vision-based methods [5, 11, 13, 37] focus on static contact detection in single frames, lacking the temporal modeling necessary for dynamic interaction generation. While these approaches have advanced relation modeling, they share a common limitation: treating all spatial relationships with equal importance. In contrast, our approach recognizes that interactions are inherently a small subset of human-object relations that are relevant for any given action. By learning to identify and focus on these critical relations through contact-aware importance sampling, we achieve more efficient and effective relation modeling that directly translates to higher-quality interaction generation.

3. Method

Generating physically plausible human-object interactions requires precise modeling of contact relationships—both *where* contact occurs on object surfaces and *how* these contacts evolve temporally during interaction. Existing methods that rely on global distance measure fail to capture these critical aspects, leading to unrealistic artifacts such as floating objects or penetration issues.

Our approach begins with a Contact Affordance Predictor (CAP) that identifies precise contact regions on object surfaces, moving beyond coarse centroid approximations to establish *where* interactions should occur. These predicted contact anchors then seed a Contact Relation Field (CRF) that captures *how* human joints relate to these contact points throughout the interaction sequence. Finally, a Contact Dynamics Model (CDM) learns the temporal evolution patterns of these contact relationships from real HOI data and guides the diffusion-based motion generation process to maintain physically plausible contacts. Figure 2 illustrates the complete pipeline of our approach.

In the following sections, we first introduce the data representation and preliminary concepts (Section 3.1), then detail each component of our framework: CAP (Section 3.2), CRF construction (Section 3.3), CDM (Section 3.4), and the guided generation process (Section 3.5).

3.1. Preliminary

Data Representation. We represent human motion as $\mathbf{X} \in \mathbb{R}^{N \times D_h}$, where N denotes the number of frames and D_h represents the pose dimension. Each frame $\mathbf{x}_n$ consists of 24 SMPL-X [25] joint positions $\mathbf{Q}_n = \{q_{1,n}, ..., q_{24,n}\}$ with $q_{j,n} \in \mathbb{R}^3$, along with 6D continuous rotations [40] for each joint. The SMPL-X parametric model enables accurate human mesh reconstruction from these pose parameters, providing detailed body surface geometry necessary for contact modeling.

Object motion is represented as $\mathbb{O} \in \mathbb{R}^{N \times D_o}$, where $D_o = 12$ encompasses the object's centroid position $\mathbf{p}_c \in \mathbb{R}^3$ and 9D rotation representation. The object geometry is encoded using a combination of mesh vertices $\mathcal{V} = \{v_i\}_{i=1}^K$ and surface normals, providing rich geometric information for contact prediction.

Contact Representation. Unlike prior works that use binary contact labels or dense distance fields, we introduce a structured contact representation that explicitly models contact anchors $\mathcal{C} = \{c_1, ..., c_k\}$ on object surfaces. Each contact anchor $c_k \in \mathbb{R}^3$ represents a potential interaction point

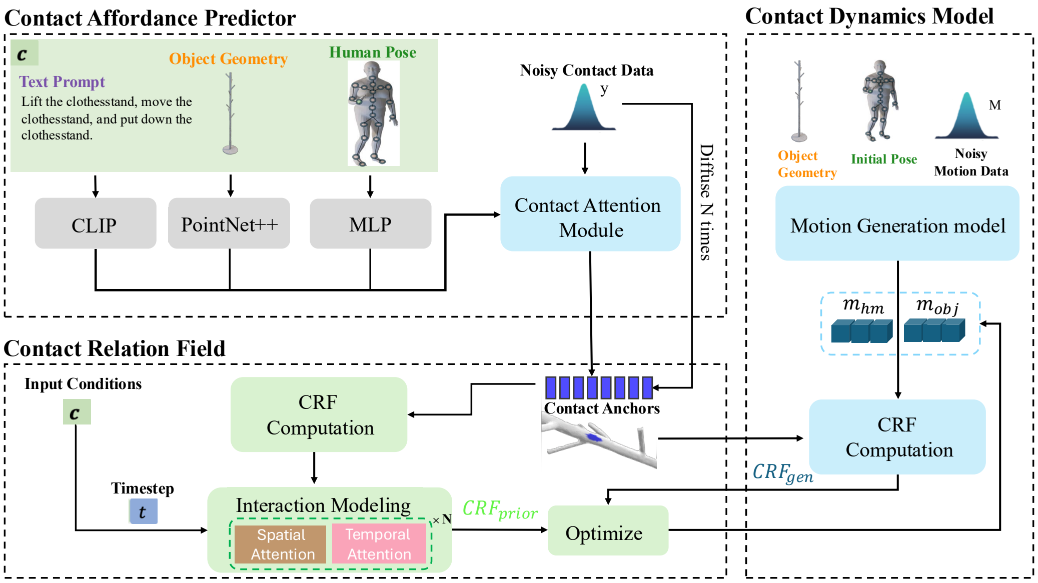

Figure 2. Overview of ContA-HOI. The framework consists of three components: 1) Contact Affordance Predictor (CAP) predicts where contact occurs, 2) Contact Relation Field (CRF) encodes how contact evolves, and 3) Contact Dynamics Model (CDM) guides diffusion sampling for physically plausible HOI generation

on the object mesh, identified through our Contact Affordance Predictor based on text semantics and object geometry. The contact state at frame n is characterized by the spatial relationships between human joints and these anchors, forming the basis of our Contact Relation Field.

Diffusion Framework. The diffusion-based motion generation model, which has shown remarkable success in generating high-quality, diverse motions. The forward diffusion process progressively adds Gaussian noise to clean motion data $\mathbf{M}_0 = \{\mathbf{X}_0, \mathbb{O}_0\}$ over T timesteps:

$$q(\mathbf{M}_t|\mathbf{M}_{t-1}) = \mathcal{N}(\mathbf{M}_t; \sqrt{1-\beta_t}\mathbf{M}_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

where β_t follows a predefined variance schedule.

The reverse process learns to denoise the data conditioned on text and contact information:

$$p_\theta(\mathbf{M}_{t-1}|\mathbf{M}_t, \mathbf{c}) = \mathcal{N}(\mathbf{M}_{t-1}; \mu_\theta(\mathbf{M}_t, t, \mathbf{c}), \sigma_t^2\mathbf{I}), \quad (2)$$

where $\mathbf{c} = \{\mathbf{e}_{\text{text}}, \mathbf{X}_0, \mathbb{O}_0, \mathcal{V}\}$ represents the information including text, initial states, and object geometry.

The key innovation of ContA-HOI lies in how we augment this standard diffusion framework with explicit contact modeling through CAP, CRF, and CDM, ensuring that the generated motions respect physical contact constraints while maintaining semantic alignment with the input text.

3.2. Contact Affordance Predictor (CAP)

The Contact Affordance Predictor (CAP) addresses the fundamental question of where contact should occur on object surfaces during interaction. Unlike previous methods that rely on coarse distance measures to object centroids, CAP predicts precise contact anchors on object surfaces by jointly reasoning about text semantics, human pose configuration, and object geometry.

CAP employs a hierarchical attention mechanism that mirrors human interaction planning: language informs which body parts to engage, which then determines where to contact the object. Unlike prior methods that map all human joints to an object's centroid and overlook critical contact constraints, this design explicitly models the contact regions in HOI. The module processes three input: text embedding $\mathbf{e}_{\text{text}} \in \mathbb{R}^{512}$ from a frozen CLIP encoder capturing interaction semantics, human pose $\mathbf{h} \in \mathbb{R}^{24\times3}$ representing 24 SMPL-X [25] joints positions, and object geometry information $\mathbf{O}$ from PointNet++ [27] encoding.

The hierarchical processing implements a two-stage attention mechanism. First, we compute language-contextualized human features through cross-attention, where the text embedding modulates which body parts are relevant—"kick the trashcan" emphasizes foot joints while "lift the trashcan" highlights hand configurations. Second, these contextualized human features attend to object geometry encoded by PointNet++ [27] to produce contact probabilities $\mathbf{P}_{\text{contact}}$. Each probability indicates the likelihood of an object point serving as a contact anchor. We select the top-k points with highest probabilities as contact anchors for contact relation field construction. To ensure predicted contacts are physically feasible, we introduce a contact validity loss that enforces consistency between predicted object contacts and designated human joints:

$$\mathcal{L}_{\text{validity}} = \sum_{l \in L_{\text{contact}}} \max(0, d^l_{\text{min}} - \tau_{\text{contact}}), \tag{3}$$

where L_{active} denotes active limbs based on contact labels from the OMOMO [19] dataset (e.g., left hand, right hand, left foot, right foot), d^l_{min} is the minimum distance between limb l and its nearest predicted contact in world coordinates, and $\tau_{\text{contact}} = 0.05m$ is the reachability threshold determined empirically from biomechanical constraints. This loss penalizes contact predictions that are spatially infeasible given the current pose, guiding CAP to learn physically plausible contact affordances.

3.3. Contact Relation Field Construction

The Contact Relation Field (CRF) captures *how* human joints dynamically relate to predicted contact regions throughout an interaction. Building upon the contact anchors identified by CAP, CRF models the spatiotemporal evolution of joint-to-contact relationships, providing a compact yet expressive representation for guiding motion generation. Given a human-object interaction sequence of N frames with human joints $\mathbf{Q} = \{q_1, ..., q_J\}$ ($J = 24$ SMPL-X [25] joints) and the predict contact anchors $\mathcal{C} = \{c_1, ..., c_k\}$ from CAP, we construct the CRF as:

$$\text{CRF} = [d_{i,k,n}], d_{i,k,n} = \|q_{i,n} - c_{k,n}\|_2, \tag{4}$$

where $d_{i,k,n}$ is the distance between joint i and contact anchor k at frame n. This representation offers crucial advantages over global distance fields used in prior work. In contrast, CRF focuses only on the $24 \times K$ relationships represents task-relevant contact regions identified by CAP.

3.4. Contact Dynamics Model (CDM)

The Contact Dynamics Model learns to predict realistic CRF evolution over time, providing a learned prior for physically plausible interactions. Unlike the motion generation model that operates in full pose space, CDM focuses specifically on learning the dynamics of contact relationships, enabling effective guidance during inference.

CDM employs a conditional diffusion model [24] with spatiotemporal attention to capture CRF dynamics. The model learns how contact relationships evolve throughout an interaction sequence.

The forward diffusion process progressively adds noise to the clean CRF:

$$q(\text{CRF}_t|\text{CRF}_{t-1}) = \mathcal{N}(\text{CRF}_t; \sqrt{1-\beta_t}\text{CRF}_{t-1}, \beta_t\mathbf{I}), \tag{5}$$

where β_t follows a cosine schedule from 10^{-4} to 0.02 over $T = 1000$ timesteps, providing smooth noise addition. The model learns to reverse this process:

$$p_\theta(\text{CRF}_{t-1}|\text{CRF}_t, \mathbf{c}) = \mathcal{N}(\text{CRF}_{t-1}; \mu_\theta(\text{CRF}_t, t, \mathbf{c}), \sigma_t^2\mathbf{I}), \tag{6}$$

where $\mathbf{c} = \{\mathbf{e}_{\text{text}}, \mathbf{h}_0, \mathbf{O}\}$ are conditioning text embeddings, initial human pose, and object geometry.

The denoising network employs spatiotemporal attention with conditional cross-attention to incorporate semantic and geometric conditions, ensuring the predicted CRF dynamics align with the intended action and are compatible with the object's geometry and initial human configuration.

Training Objective. The CDM is trained using a denoising objective. The network directly predicts the clean CRF:

$$\mathcal{L}_{\text{CDM}} = \mathbb{E}_{t,\epsilon}\left[\|\text{CRF}_0 - f_\theta(\text{CRF}_t, t, \mathbf{c})\|^2\right], \tag{7}$$

where f_θ is the denoising network, $\text{CRF}_t = \sqrt{\bar{\alpha}_t}\text{CRF}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ is the noisy CRF at timestep t, with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\bar{\alpha}_t = \prod_{s=1}^t(1-\beta_s)$.

3.5. Guided Motion Generation

During inference, we use the learned CDM to guide the motion generation process toward physically plausible interactions. At each denoising step t, we compute the CRF from current motion predictions and align it with the CDM prior. We then optimize the motion to minimize CRF discrepancy,

$$\mathcal{L}_{\text{guide}} = \|\text{CRF}_{\text{gen}} - \text{CRF}_{\text{prior}}\|^2. \tag{8}$$

and use L-BFGS optimization for motion update:

$$\tilde{\mathbf{M}_t} = \mathbf{M}_t - \eta\nabla_{\mathbf{M}_t}\mathcal{L}_{\text{guide}}. \tag{9}$$

Through this contact-aware framework, ContA-HOI transforms the challenging problem of HOI generation from operating in high-dimensional pose space to reasoning about sparse contact relationships, enabling more efficient learning and superior generation quality compared to traditional distance-based approaches.

4. Experiments

4.1. Dataset and Settings

Dataset. We evaluate our method on the FullBodyManipulation dataset [19], which provides comprehensive human-object interaction data essential for training and evaluating contact-aware generation models. The dataset contains 10 hours of high-quality motion capture data featuring 17 subjects interacting with 15 diverse objects, ranging from small items like small boxes to large furniture like large tables. Each interaction sequence includes synchronized human motion (captured as SMPL-X parameters) and object motion (12D pose trajectories), along with detailed textual descriptions that specify the action semantics and interaction intent. The rich diversity of objects and interaction

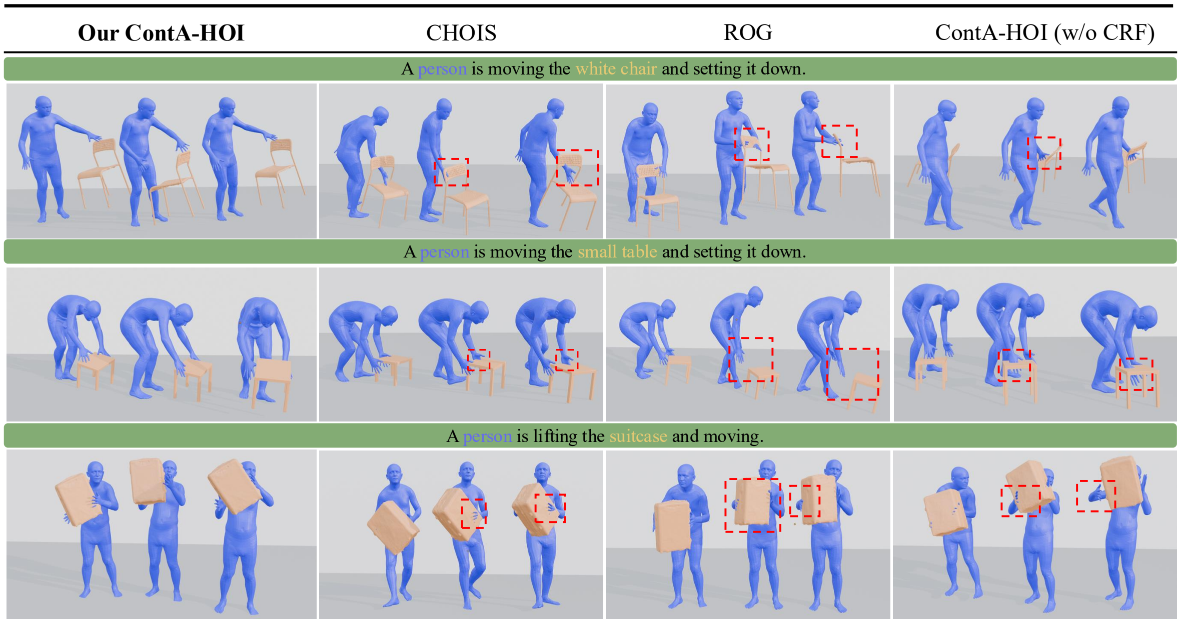
Our ContA-HOI
CHOIS
ROG
ContA-HOI (w/o CRF)
A person is moving the white chair and setting it down.
A person is moving the small table and setting it down.
A person is lifting the suitcase and moving.

Figure 3. Qualitative comparison of our proposed ContA-HOI with state-of-the-art methods CHOIS [20] and ROG [34], as well as an ablated variant ContA-HOI (w/o CRF). Each row shows generated human–object interactions under different text prompts. Our method produces physically plausible interactions with stable contact, while baseline methods often suffer from artifacts such as penetration or human–object separation, highlighted by red dashed boxes. The ablated variant without CRF also exhibits similar floating or penetration issues, demonstrating the importance of CRF for maintaining realistic contact.

Table 1. Interaction synthesis on the FullBodyManipulation [19] dataset.

Method	Human Motion			Interaction		
	Foot Sliding↓	R-precision↑	FID↓	Contact%↑	Collision%↓	MDev ↓
Interdiff [33]	0.42	0.08	20.80	0.22	0.17	23.12
MDM [31]	0.46	0.51	6.16	0.31	**0.19**	12.43
CHOIS [20]	0.35	**0.65**	**5.29**	0.44	0.25	15.32
ROG [34]	0.41	0.62	6.35	0.45	0.22	**8.30**
ContA-HOI (Ours)	**0.34**	0.64	6.21	**0.49**	0.23	10.53

types makes this dataset particularly suitable for evaluating contact-aware generation methods, as it encompasses both simple single-contact interactions (e.g., pushing) and complex multi-contact scenarios (e.g., lifting and carrying).

Evaluation Metrics. Following the standard split protocol established by existing works [20, 34], we adopt a comprehensive set of metrics to evaluate different aspects of generated interactions: **Foot Sliding (FS)**, quantifies unrealistic foot movements when feet should remain stationary, measuring the average per-frame displacement of foot joints during ground contact. **R-Precision**, which valuates semantic alignment between generated motions and input text descriptions using a retrieval-based approach. **Fréchet Inception Distance (FID)**, which measures the distributional similarity between generated and real motions in the feature space, where lower values indicate better quality and more realistic motion distributions. Following [20], we use **Contact Percentage (Contact%)** to calculate the percentage of frames where meaningful contacts occur. This metric directly evaluates our method's ability to maintain appropriate contact. **Collision Percentage (Collision%)**, which measures undesirable penetrations by computing the percentage of frames where human vertices penetrate the object mesh beyond a threshold, assessing physical plausibility. **Motion Deviation (MDev) [10]**, which evaluates motion consistency by measuring the directional difference between hand and object movements during contact periods, reflecting the coordination quality.

Implementation Details. Following MDM [31], we encode text prompts using the ViT-B/32 [8] variant of

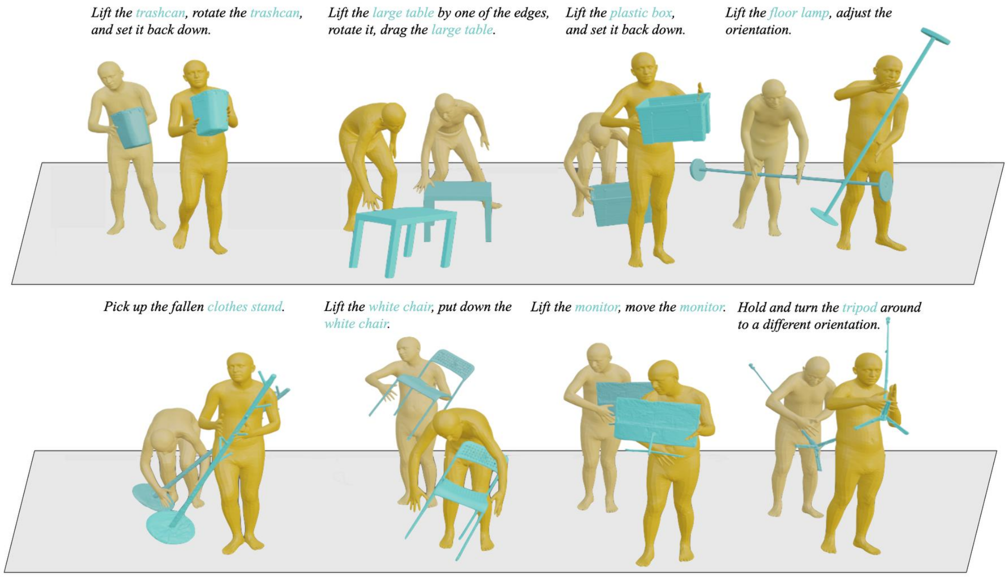
Lift the trashcan, rotate the trashcan, and set it back down.
Lift the large table by one of the edges, rotate it, drag the large table.
Lift the plastic box, and set it back down.
Lift the floor lamp, adjust the orientation.
Pick up the fallen clothes stand.
Lift the white chair, put down the white chair.
Lift the monitor, move the monitor.
Hold and turn the tripod around to a different orientation.

Figure 4. Additional qualitative results of ContA-HOI across diverse human–object interaction tasks.

CLIP [28] to obtain 512-dimensional text embeddings. Our motion generation model employs a transformer-based architecture with 8 layers, hidden dimension of 512, and 8 attention heads per layer. The contact affordance predictor consists of a lightweight 4-layer transformer with cross-attention between text and pose features, utilizing a hidden dimension of 256. For the contact dynamics model that learns CRF evolution, we adopt a conditional U-Net architecture with spatial and temporal attention blocks, operating on the sparse contact relation fields with 4 spatial and 4 temporal attention heads. Both the motion generation model and contact dynamics model are trained using the AdamW [22] optimizer with learning rate 1×10^{-4}, batch size 32, and weight decay of 0.01. The contact predictor is pre-trained for 100 epochs before joint training to ensure stable contact predictions. We use DDPM sampling with 1,000 diffusion steps during training and 50 steps during inference for efficiency. For CRF construction, we select top $k = 24$ contact pairs based on predicted contact probabilities. During inference, we apply contact-aware guidance in the last 10 denoising steps using the L-BFGS optimizer [1] with 5 iterations per step and a learning rate of 0.01. The guidance weight is set to $\lambda = 0.1$ to balance generation quality and contact adherence. All experiments are conducted on NVIDIA RTX A6000 GPUs.

Baselines. We compare our method against several state-of-the-art approaches: 1) InterDiff [33], a method predicts human–object interactions by leveraging the preceding 10 frames. 2) MDM [31], a foundational text-to-motion model that we extend to generate both human and object motions by expanding input/output dimensions. 3) CHOIS [20] that employs waypoint-based guidance for controllable generation. 4) ROG [34] proposed object geometry keypoint sampling to construct a distance-based joint-to-boundaries representation for human-object interactions. For a fair comparison, all baselines are trained on the same dataset split without their original control signals (e.g., by removing waypoints in CHOIS and retraining without the input signal), ensuring they operate under the same input condition as our method. This setup better reflects real-world applications where detailed control signals are typically unavailable. All models use the same train/test split and evaluation protocols to ensure comparable results.

4.2. Quantitative Results

Table 1 presents comprehensive quantitative results on the FullBodyManipulation dataset. Our method demonstrates significant improvements across all evaluation categories:

Human Motion Quality. ContA-HOI achieves the lowest foot sliding score (0.34), outperforming all baselines including CHOIS [20] and ROG [34]. Our R-Precision of 0.64 is competitive with CHOIS while surpassing ROG and MDM. The FID score of 6.21 indicates good motion quality, though CHOIS achieves a slightly better score of 5.29. These metrics demonstrate that our contact-aware approach generates stable and naturally aligned human motions.

Physical Plausibility. Most notably, ContA-HOI excels in human-object interaction metrics. We achieve the highest contact percentage, surpassing ROG and CHOIS. The collision percentage remains competitive, with ROG achieving a slightly lower rate.

Overall Performance. The results validate that explicit contact modeling through our Contact Affordance Predictor, Contact Relation Field representation, and Contact Dynamics Model leads to more physically plausible human-object interactions. While different methods excel in specific metrics, ContA-HOI achieves a strong balance across all evaluation dimensions, particularly excelling in contact establishment and foot stability—critical factors for realistic HOI generation. These improvements demonstrate the effectiveness of our progressive framework that explicitly models where and how contact occurs, guiding the generation process toward physically realistic interactions.

4.3. Qualitative Results

Figure 3 presents visual comparisons of generated interactions across different methods. We analyze three scenarios that highlight the advantages of our approach:

"A person is moving the white chair and setting it down": Our ContA-HOI method demonstrates stable hand-chair contact throughout the interaction, with natural hand placement on the chair back. CHOIS exhibits floating artifacts (marked in red boxes) where hands fail to maintain proper contact with the chair. ROG shows significant spatial misalignment with hands hovering away from the chair surface. The ablated version (ContA-HOI w/o CRF) suffers from object floating, highlighting the importance of our Contact Relation Field in maintaining realistic interactions.

"A person is moving the small table and setting it down": This scenario requires precise coordination for table manipulation. Our method achieves natural bending posture with hands firmly grasping the table edges. In contrast, CHOIS shows penetration issues (red boxes) where hands pass through the table surface. ROG struggles with proper hand positioning, resulting in unrealistic floating near the table. The ablation without CRF demonstrates degraded performance with inconsistent hand-table contact, validating the effectiveness of our contact modeling approach.

"A person is lifting the suitcase and moving": For this carrying task, ContA-HOI generates natural grasping with appropriate body posture adjustment for balance. CHOIS, ROG, and the ablated version exhibit severe contact failures (red boxes), with hands either penetrating or floating near the suitcase handle.

The superior performance across these diverse scenarios demonstrates that our contact-aware approach—through the synergistic combination of Contact Affordance Predictor, Contact Relation Field, and Contact Dynamics Model—enables significantly more realistic interaction synthesis compared to existing methods that rely on distance-based or boundary-based representations. We provide additional qualitative results of ContA-HOI across diverse interaction tasks in Figure 4.

Table 2. Ablation study on key components of ContA-HOI on the FullBodyManipulation dataset. We progressively add each component to evaluate the individual contributions.

Method	R-Prec ↑	FS ↓	C% ↑	Coll% ↓	MDev ↓
+ Contact	0.56	0.52	0.41	**0.20**	16.32
+ Validity loss	0.58	0.42	0.40	0.22	12.12
+ Guidance loss	0.62	0.38	0.47	0.21	12.83
Full Model (ContA-HOI)	**0.64**	**0.34**	**0.49**	0.23	**10.53**

4.4. Ablation Studies

Table 2 presents our ablation study analyzing the contribution of each component: *Contact Anchors as Input:* Using CAP-predicted contact anchors as input conditions ("+Contact") establishes the foundation for contact-aware generation. *Contact Validity Loss:* Adding the validity loss in CAP training ("+Validity Loss") enforces physically feasible contact predictions. *CRF Guidance without Optimization:* Incorporating CRF guidance from CDM during inference ("+Guidance Loss") substantially improves performance. The complete ContA-HOI framework with L-BFGS optimization achieves the best overall performance. These results demonstrate that each component addresses specific challenges: contact anchors provide spatial grounding, validity loss ensures physical feasibility, CRF guidance captures temporal dynamics, and L-BFGS optimization refines the final output. The synergistic combination of all components is essential for generating physically plausible human-object interactions.

5. Conclusion and Limitations

In this work, we introduce ContA-HOI, a progressive contact-aware framework designed to generate physically plausible human-object interactions by explicitly modeling where and how contact occurs. We begin by developing a Contact Affordance Predictor (CAP) that identifies precise contact anchors on object surfaces, moving beyond coarse centroid-based representations. Building on these anchors, we construct a Contact Relation Field (CRF) that captures the spatiotemporal dynamics of joint-to-contact relationships throughout the interaction. Finally, we develop a Contact Dynamics Model (CDM) that learns prior CRF evolution patterns and guides generation through iterative refinement, ensuring realistic contact constraints while maintaining semantic alignment with text descriptions.

Limitations include computational overhead: the L-BFGS optimization increases generation time compared to single-pass methods. Additionally, our approach relies on accurate object geometry representations, which may limit the applicability to scenarios with incomplete object meshes. The current framework also focuses on rigid object interactions, leaving articulated object manipulation as future work.

References

[1] Raghu Bollapragada, Dheevatsa Mudigere, Jorge Nocedal, Hao-Jun Michael Shi, and Ping Tak Peter Tang. A progressive batching L-BFGS method for machine learning. In *ICML*, pages 619–628. PMLR, 2018. 7

[2] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *CVPR*, pages 1577–1585. IEEE, 2024. 1

[3] Yixin Chen, Sai Kumar Dwivedi, Michael J. Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *CVPR*, pages 17100–17110. IEEE, 2023. 3

[4] Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. Expressive whole-body control for humanoid robots. In *Robotics: Science and Systems*, 2024. 1

[5] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *CVPR*, pages 1778–1787. Computer Vision Foundation / IEEE, 2021. 3

[6] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 1

[7] Christian Diller and Angela Dai. CG-HOI: contact-guided 3d human-object interaction generation. In *CVPR*, pages 19888–19901. IEEE, 2024. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021. 6

[9] Sai Kumar Dwivedi, Dimitrije Antic, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J. Black, and Dimitrios Tzionas. Interactvlm: 3d interaction reasoning from 2d foundational models. In *CVPR*, pages 22605–22615. Computer Vision Foundation / IEEE, 2025. 3

[10] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, pages 12943–12954. IEEE, 2023. 6

[11] Pei Geng, Jian Yang, and Shanshan Zhang. HORP: human-object relation priors guided HOI detection. In *CVPR*, pages 25325–25335. Computer Vision Foundation / IEEE, 2025. 3

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1

[13] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction from multi-view RGB-D images. *Int. J. Comput. Vis.*, 132(7):2551–2566, 2024. 3

[14] Inwoo Hwang, Bing Zhou, Young Min Kim, Jian Wang, and Chuan Guo. Scenemi: Motion in-betweening for modeling human-scene interactions. *CoRR*, abs/2503.16289, 2025. 1

[15] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, Xinda Xue, Qinghang Su, Huaihai Lyu, Xiaolong Zheng, Jiaming Liu, Zhongyuan Wang, and Shanghang Zhang. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *CVPR*, pages 1724–1734. Computer Vision Foundation / IEEE, 2025. 1

[16] Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia*, pages 33:1–33:11. ACM, 2024. 1

[17] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *CVPR*, pages 1737–1747. IEEE, 2024. 1

[18] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas J. Guibas. NIFTY: neural object interaction fields for guided human motion synthesis. In *CVPR*, pages 947–957. IEEE, 2024. 3

[19] Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Trans. Graph.*, 42(6): 197:1–197:11, 2023. 1, 5, 6

[20] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *ECCV (41)*, pages 54–72. Springer, 2024. 3, 6, 7, 1

[21] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *Int. J. Comput. Vis.*, 132(9): 3463–3483, 2024. 1

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net, 2019. 7

[23] Yuke Lou, Yiming Wang, Zhen Wu, Rui Zhao, Wenjia Wang, Mingyi Shi, and Taku Komura. Zero-shot human-object interaction synthesis with multimodal priors. *CoRR*, abs/2503.20118, 2025. 3

[24] Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. VDT: general-purpose video diffusion transformers via mask modeling. In *ICLR*. OpenReview.net, 2024. 5

[25] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. 3, 4, 5

[26] 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. *CoRR*, abs/2312.06553, 2023. 1, 3

[27] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, pages 5099–5108, 2017. 4

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 7

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021. 1

[30] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Sören Pirk, and Michael J. Black. GRIP: generating interaction poses using spatial cues and latent consistency. In *3DV*, pages 933–943. IEEE, 2024. 1

[31] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*. OpenReview.net, 2023. 1, 2, 6, 7

[32] Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, Jian Zhang, Lei Zhang, and Ping Tan. Skillmimic: Learning basketball interaction skills from demonstrations. In *CVPR*, pages 17540–17549. Computer Vision Foundation / IEEE, 2025. 1

[33] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, pages 14882–14894. IEEE, 2023. 1, 6, 7

[34] Mengqing Xue, Yifei Liu, Ling Guo, Shaoli Huang, and Changxing Ding. Guiding human-object interactions with rich geometry and relations. In *CVPR*, pages 22714–22723. Computer Vision Foundation / IEEE, 2025. 1, 6, 7

[35] Runyi Yu, Yinhuai Wang, Qihan Zhao, Hok Wai Tsui, Jingbo Wang, Ping Tan, and Qifeng Chen. Skillmimic-v2: Learning robust and generalizable interaction skills from sparse and noisy demonstrations. *CoRR*, abs/2505.02094, 2025. 1

[36] Ling-An Zeng, Guohong Huang, Yi-Lin Wei, Shengbo Gu, Yu-Ming Tang, Jingke Meng, and Wei-Shi Zheng. Chainhoi: Joint-based kinematic chain modeling for human-object interaction generation. In *CVPR*, pages 12358–12369. Computer Vision Foundation / IEEE, 2025. 1

[37] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV (12)*, pages 34–51. Springer, 2020. 3

[38] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(6):4115–4128, 2024. 2

[39] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez-Pellitero, and Gerard Pons-Moll. FORCE: dataset and method for intuitive physics guided human-object interaction. *CoRR*, abs/2403.11237, 2024. 1, 3

[40] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753. Computer Vision Foundation / IEEE, 2019. 3