InteractMove: Text-Controlled Human-Object Interaction Generation in 3D Scenes with Movable Objects

Xinhao Cai
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
xinhao.cai@stu.pku.edu.cn

Minghang Zheng
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
minghang@pku.edu.cn

Xin Jin
Beijing Electronic Science and Technology Institute
Beijing , China
jinxinbesti@foxmail.com

Yang Liu*
Wangxuan Institute of Computer Technology
State Key Laboratory of General Artificial Intelligence, Peking University
Beijing, China
yangliu@pku.edu.cn

Abstract

In this paper, we propose a novel task of text-controlled human-object interaction generation in 3D scenes with movable objects. Existing human-scene interaction datasets suffer from insufficient interaction categories and typically only consider interactions with static objects (do not change object positions), and the collection of such datasets with movable objects is difficult and costly. To address this problem, we construct the InteractMove dataset for Movable Human-Object Interaction in 3D Scenes by aligning existing human-object interaction data with scene contexts, featuring three key characteristics: 1) scenes containing multiple movable objects with text-controlled interaction specifications (including same-category distractors requiring spatial and 3D scene context understanding), 2) diverse object types and sizes with varied interaction patterns (one-hand, two-hand, etc.), and 3) physically plausible object manipulation trajectories. With the introduction of various movable objects, this task becomes more challenging, as the model needs to identify objects to be interacted with accurately, learn to interact with objects of different sizes and categories, and avoid collisions between movable objects and the scene. To tackle such challenges, we propose a novel pipeline solution. We first use 3D visual grounding models to identify the interaction object. Then, we propose a hand-object joint affordance learning to predict contact regions for different hand joints and object parts, enabling accurate grasping and manipulation of diverse objects. Finally, we optimize interactions with local-scene modeling and collision avoidance constraints, ensuring physically plausible motions and avoiding collisions between objects and the scene. Comprehensive experiments demonstrate our method's superiority in generating physically plausible, text-compliant interactions compared to existing approaches. The code is available at https://github.com/Cxhcmhhh/InteractMove.

CCS Concepts

• **Computing methodologies** → **Scene understanding**.

Keywords

Diffusion Models, Human-object Interactions, Human Motions, Human-Scene Interactions

ACM Reference Format:
Xinhao Cai, Minghang Zheng, Xin Jin, and Yang Liu. 2025. InteractMove: Text-Controlled Human-Object Interaction Generation in 3D Scenes with Movable Objects. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746027.3754910

1 Introduction

The generation of human motions within scenes is a growing research area with significant applications in VR, AR, video games, and beyond. Recently, there has been increasing interest in generating human motions conditioned on natural language descriptions. However, most prior works either focus on language-driven interactions between humans and isolated objects [5, 7, 18], neglecting the influence of the surrounding scene, or study human-scene interactions [13, 25] without explicitly considering movable objects. This results in limited expressiveness and practicality when deployed in real-world scenarios, where objects are often embedded in complex environments and exhibit various affordances. To bridge this gap, we propose a novel task: text-controlled human-object interaction generation in 3D scenes with movable objects.

In existing Human-Scene Interaction datasets [4, 9, 13, 19, 25], interactions are quite limited, and interactable objects are often fixed and immovable, such as beds and sofas. Furthermore, manually collecting a new large-scale, high-quality 3D dataset is both difficult and costly. Therefore, we introduce the InteractMove dataset

*Corresponding author

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754910

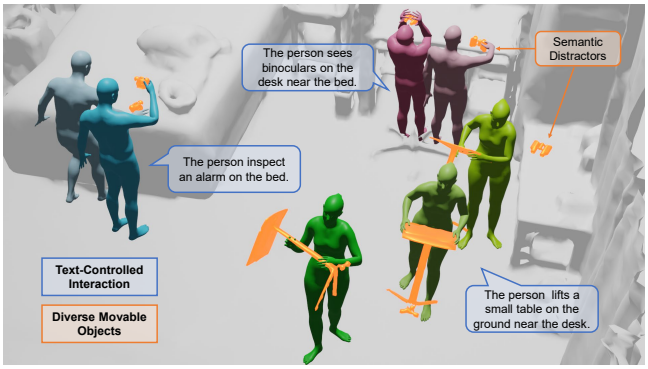

Figure 1: Samples in our dataset. We synthesize a large-scale human-object-interaction-in-scene dataset by aligning captured human-object interaction sequences with various 3D scan scenes. In the dataset, we provide free-form text annotations and interaction with movable objects in high-quality scenes.

constructed by aligning existing human-object interaction data with richly annotated 3D scene contexts as shown in Fig. 1. Our dataset exhibits three key properties: (1) scenes contain multiple movable objects, enabling text-controlled interaction with specified targets, often in the presence of same-category distractors that necessitate spatial understanding; (2) the dataset covers diverse object types and sizes, with interactions that vary in complexity, including one-handed and two-handed actions; and (3) object manipulation trajectories are physically plausible, avoiding collisions.

This new task also introduces the following challenges to be addressed. First, it requires models to comprehend natural language instructions and identify the correct object among multiple, often similar, distractors in the scene. For example, in the scene shown in Fig. 1, there are two binoculars, and the model needs to identify the one described in the text that is on the desk near the bed. Second, the target objects may vary significantly in type and scale, exhibiting various affordances and requiring different interaction strategies. For example, lifting small objects like a cup may only require one hand to interact with, while larger objects like a table need both hands. Even for the same type of object, interaction strategies may differ depending on its specific shape, e.g., a cup with a handle is usually grasped by the handle, whereas a handleless cup is more naturally grasped by its body. Third, the task involves dynamically manipulating objects within 3D scenes while ensuring physical plausibility, which includes avoiding penetrations or collisions with other scene elements, especially moving large objects over long distances with other crowded objects nearby.

To address the above challenges, we propose a novel Affordance-Guided Collision-Aware Interaction Generation (AGCA) framework with carefully designed components that model 3D object grounding, fine-grained hand-object joint affordance learning, and collision-aware motion generation. Specifically, we first employ state-of-the-art 3D visual grounding models to locate the intended object specified by the input text. To capture the diversity of object affordances and interaction strategies, we introduced a hand-object joint affordance learning module, which takes the object mesh as input and predicts the likelihood of interactions occurring between hand joints and object surfaces over time, referred to as hand-object affordance. This fine-grained affordance is used to guide the interaction motion generation, enabling more accurate interactions aligned with object size and interaction semantics. Finally, we incorporate a collision-aware motion generation strategy that voxelizes the region around the interactive object to evaluate spatial accessibility, combined with a collision-aware loss that enforces physically plausible motion and prevents interpenetration, while ensuring the object's trajectory remains synchronized with human control and scene constraints.

In summary, our contributions are threefold: (1) We introduce a new task that focuses on text-conditioned human-object interaction generation in movable-object 3D scenes. (2) We construct a comprehensive dataset for this task with text-controlled interaction and diverse movable objects; we also propose a novel framework for this task with carefully designed components that model 3D object grounding, fine-grained hand-object joint affordance learning, and collision-aware motion generation. (3) Extensive experiments demonstrate the effectiveness of our approach in producing realistic, text-aligned, and physically plausible interaction motions in 3D scene with movable objects.

2 Related Works

2.1 Human-Object Interaction

Datasets capturing human-object interaction(HOI) are crucial for training generative models, yet remain difficult and costly to collect. Datasets like GRAB [22], BEHAVE [3], and CHAIRS [12] employed optical MoCap/IMU systems to capture detailed human-object interactions, including object trajectories. However, these interactions are typically performed in isolation without the presence of a full scene, thus lacking contextual constraints from the environment.

Early works in this region begin with HOI detection[14, 15] or HOI image generation[27]. For human-object interaction generation, early methods like OMOMO [16] rely on object trajectories as input, limiting their applicability in free-form generation tasks. InterDiff [26] introduces object dynamics but focuses on motion prediction conditioned on past human motions. More recent works [7, 18, 21] attempt to generate interactions with isolated movable objects without the presence of a full scene. Several recent methods also incorporate affordance prediction or contact map to inform interaction generation [5, 24]. However, they generally model affordance as a coarse spatial heatmap over object surfaces, neglecting the affordance of hands. They neglect to explicitly model how different hand joints engage with object surfaces, which is important when differentiating between single-handed and two-handed interactions or fine-grained grasp strategies. In contrast, we propose hand-object joint affordance learning, which models fine-grained contact likelihoods between hand joints and object surfaces over time. This enables more accurate and diverse interaction generation aligned with object shape, size, and semantics.

2.2 Human-Scene Interaction

Earlier works begin with scene understanding[28]. Some recent efforts explore large-scale human-scene interaction (HSI). For instance, HUMANISE [25] synthesizes high-quality HSI data within

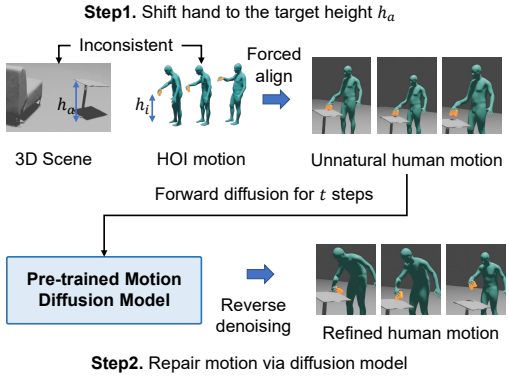

Figure 2: Method of our motion alignment.

virtual environments. Nonetheless, it supports only a limited range of immovable objects and predefined interaction types. TRUMANS[13] is the dataset that is closest to us, as it also includes both scenes and dynamic objects. However, TRUMANS is limited in several key aspects: (1) the set of movable object categories is narrow (20 types), restricting interaction diversity; (2) actions are predefined into 10 coarse categories, lacking diverse interaction types; and (3) it only provide discrete action labels, lacking natural language annotations, thus unable to support language-guided HOI generation. In contrast, our proposed dataset, InteractMove, extends beyond existing efforts by synthesizing realistic HOI data into richly annotated 3D scenes from ScanNet [6], while enabling object movement and fine-grained textual control. It features 71 categories of movable objects embedded in 3D scenes, with interactions spanning 21 types and paired with natural language descriptions. This allows scene-aware and language-controllable interaction generation with dynamic objects, enabling more realistic and physically plausible interactions in complex environments.

Human-scene interaction generation uses conditional Variational Auto-Encoder (cVAE)[20] or diffusion models[10] to generate human-scene interaction based on action labels or text conditions. Earlier works [17, 30, 31] mainly focus on predicting static human poses conditioned on the scene geometry, often for single frames. Later approaches [2, 8, 11] extended this to temporal sequences, enabling more realistic interactions. However, these methods typically rely on action labels or scene cues, without leveraging natural language instructions. HUMANISE [25] proposes text-guided motion generation within static scenes. Yet, it does not handle dynamic object manipulation, as all interactable items are fixed. This severely limits interaction diversity and realism. With the introduction of various movable objects, this task becomes more challenging, as the model needs to accurately identify interactive objects, learn to interact with objects of different sizes and categories, and avoid collisions between objects and the scene. Our work addresses these challenges by introducing a text-controlled generation framework that incorporates 3D object grounding, fine-grained hand-object joint affordance learning, and collision-aware motion generation, enabling precise, diverse, and realistic interactions in complex scenes.

3 InteractMove Dataset

To enable text-controlled human-object interaction generation in 3D scenes with movable objects, we construct InteractMove, a novel dataset that enriches existing human-object interaction (HOI) data with realistic, richly annotated 3D scene contexts. Instead of collecting new data from scratch, which is costly and time-consuming, we automatically align existing motion sequences from BEHAVE [3] and GRAB [22] datasets with 3D scenes to achieve a scalable yet high-quality solution. Our construction process emphasizes the following key principles: (1) **Movable target objects**: Diverse objects are placed in semantically appropriate areas of the scene, including multiple distractors of the same category, to facilitate spatial understanding. (2) **Physically Coherent Motion Alignment**: Human motion sequences are adjusted to achieve realistic interactions with objects at different positions. (3) **Scene-aware Filtering for Physical Plausibility**: The aligned motion-scene pairs are filtered to remove cases violating physical constraints, such as foot-ground detachment, boundary overflow, or human-object collisions.

3.1 Object Placement in 3D Scenes

We first collect object and human-object interaction data from existing HOI datasets BEHAVE [3] and GRAB [22], totaling 71 object categories and 21 interaction types. Taking the interaction of *take picture with camera* as an example, we will discuss the placement process here. To integrate them into realistic 3D environments, we utilize the ScanNet [6] dataset to obtain 3D scene and utilize the Sr3D dataset [1] to obtain object-region annotations and relative spatial relations annotations in ScanNet. For each target interaction, we identify appropriate placement surfaces in the 3D scene. For instance, a camera might be located on the surface of a table. Sr3D[1] provides annotations for such regions and their relative spatial positions like *a table next to a door*. We sample these surfaces within the scene where objects can be placed and ensure that their relative positions are annotated by Sr3D[1]. Then, based on the interaction label provided by the HOI dataset type and location annotations provided by Sr3D[1], we can automatically generate the full interaction textual annotations from templates: *A person takes pictures with the camera on a table next to a door*. For each scene, we also put multiple instances of the same category as the target object on every reasonable surface. For example, if there are k placeable surfaces in the scene, we will randomly select a subset and put cameras on these surfaces when aligning an interaction of *A person takes pictures with the camera*. This requires the model to learn language-conditioned object disambiguation, such as identifying *The camera on a table next to a door*.

3.2 Motion Alignment

One of the core challenges in aligning HOI data with new scene placements is the mismatch between the original object height and its new scene-constrained height (e.g., an object being on a shelf or table). As shown in Fig. 2, to ensure the interaction remains realistic, we adjust the corresponding human motion, particularly hand and arm movements, to match the new object location. We apply a motion inpainting strategy based on a pre-trained motion diffusion model [23], focusing on editing the relevant hand joint trajectories during the phase when the hand comes into contact

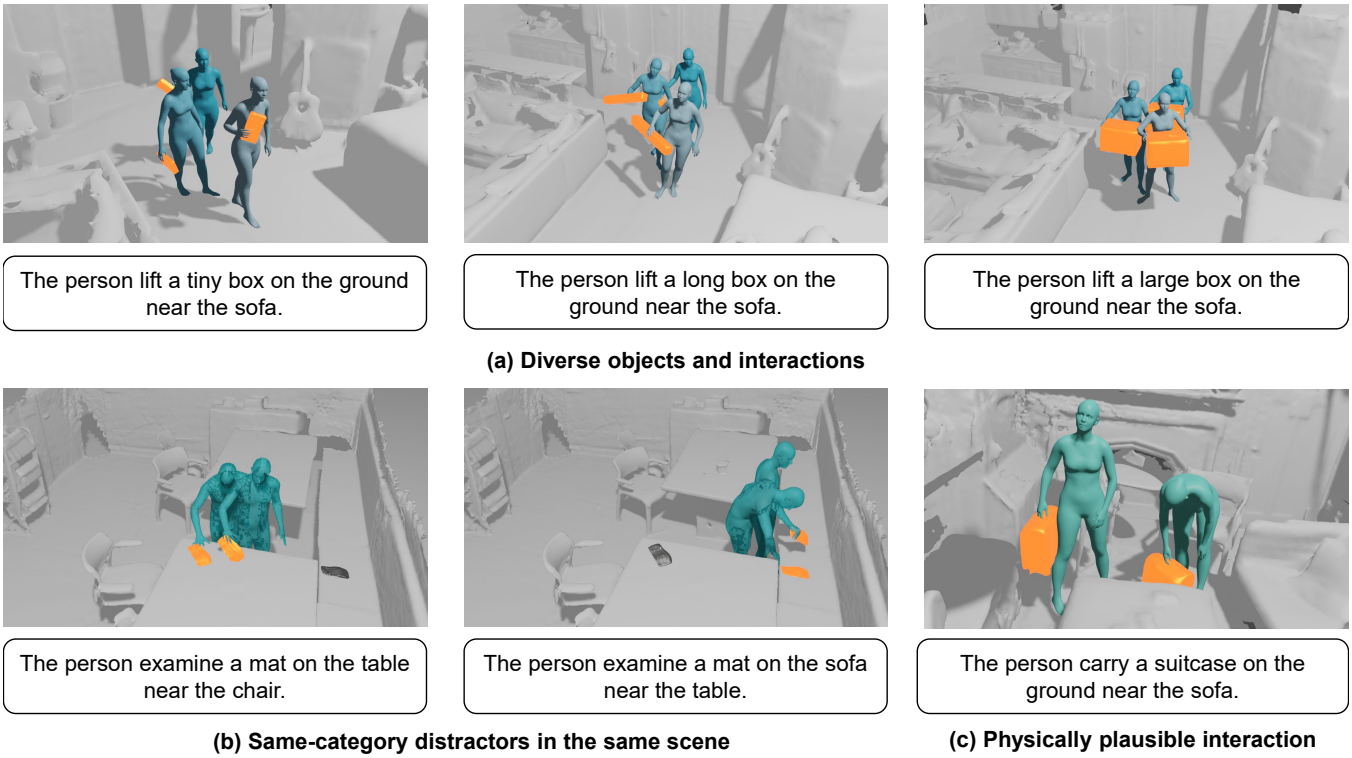

Figure 3: Visualizations of our dataset.

Dataset	Scenes	Movable Objects	Movable Object Categories	Annotated Interaction Types	Frames	Nature Language Annotations
PiGraph[19]	30	×	-	-	0.1M	×
PROX-Q[9]	12	×	-	-	0.1M	×
GTA-IM[4]	49	×	-	-	1.0M	×
CIRCLE[2]	9	×	-	-	4.3M	×
HUMANISE[25]	643	×	-	4	1.2M	✓
TRUMANS[13]	100	✓	20	10	1.6M	×
Ours	618	✓	71	21	2.2M	✓

Table 1: Comparisons of our dataset with other Human-Scene Interaction datasets.

with the object. Specifically, suppose the initial height of the object is h_i, and the adjusted height is h_a (the height of the surface we plan to put the object on). We first identify the moment when the human hand initiates interaction with the object, characterized by two features: the object begins to exhibit motion, and the absolute position of the human hand is proximal to the object. Then, to align the motion with the scene, we first *shift the human hand joints to the target height* h_a as shown in Fig. 2: for the hand joint positions within the T frames before and after the interaction start frame, we forcibly adjust the height based on the formula: $M_t = (h_a - h_i) \times \frac{t}{T}$. This ensures a smooth transition, with the height aligning with the new height when contacting the object. However, this forced adjustment can make the motion sequence unnatural. Then, we *repair the motion using a pretrained diffusion model* as shown in Fig. 2: we apply forward diffusion for t steps to introduce noise, and then leverage a pretrained motion diffusion model [23] to perform reverse denoising, progressively restoring realistic and coherent human motions. This design maximizes the preservation of the majority of the interaction process, with only the approaching phase being modified, thus retaining the valuable original interaction data.

3.3 Physics-based Filtering and Validation

After alignment, we apply strict filtering to ensure all motion-scene combinations are physically plausible and scene-aware. Our filtering criteria include: (1) *Foot-ground contact*: Ensures the human maintains realistic contact with the ground; sequences with abnormal foot elevation or penetration are discarded. (2) *Scene boundary*

constraints: Motion sequences that move the human outside the visible scene space are rejected by monitoring the root joint trajectory. (3) *Collision detection*: We compute distances between the human mesh and nearby objects or walls, removing samples with significant interpenetration. This filtering pipeline guarantees that each retained sequence is compatible with the geometry and physics of the scene, making the dataset suitable for training models on physically realistic 3D human-object interactions.

3.4 Quantitative Statistics and Visualizations

Our InteractMove dataset contains 30.5k interaction sequences across 618 richly annotated indoor 3D scenes, encompassing 71 different types of movable objects, such as cameras, apples, and mugs. Compared with existing Human-Scene Interaction (HSI) datasets, InteractMove exhibits three unique advantages, as visually illustrated in Fig. 3: (1) *Diverse object types and interaction complexity*: As shown in Fig. 3 (a), InteractMove supports a wide range of object categories with diverse object size and interaction strategies (e.g. using one hand or two hands). This significantly enriches the interaction patterns and poses new challenges for motion generation models to adapt to object size, shape, and affordances. (2) *Multiple movable objects per scene*: As shown in Fig. 3 (b), our scenes contain multiple interactable objects of the same category, placed in semantically reasonable locations. This setup introduces same-category distractors, requiring models to perform fine-grained spatial reasoning and accurate object grounding based on text. (3) *Physically plausible interaction*: As shown in Fig. 3 (c), our dataset includes data of humans manipulating objects with large-range movement. Thanks to our scene-aware motion adjustment and filtering pipeline, all motions in our dataset are collision-free and physically reasonable.

As shown in Tab. 1, our dataset outperforms existing human-scene interaction datasets in terms of the number of scenes, scale of interaction frames, and variety of movable objects. Unlike prior datasets that often involve static furniture (e.g., beds or sofas), our InteractMove enables text-guided human-object interactions in 3D scenes with movable objects, along with semantic natural language descriptions. Also, the statistics in Tab. 7 also prove the quality of our synthesized data.

4 Method

4.1 Overview

Our method addresses the task of text-conditioned human-object interaction generation in 3D scenes with moveable objects, producing human motion sequences X and object trajectories Y based on text T and 3D scene information S including a set of object point clouds $O \in \mathbb{R}^{N\times 3}$ for N points of M objects in the scene.

Compared to conventional Human-Scene or Human-Object Interaction generation tasks, our setting introduces unique challenges: the model must (1) identify the target object from free-form language in a 3D scene, (2) adapt the interaction to diverse object geometries and task descriptions, and (3) ensure the generated object trajectory is physically plausible and avoids collisions with the surrounding scene. To tackle these challenges, we propose a novel Affordance-Guided Collision-Aware Interaction Generation (AGCA) framework as shown in Fig. 4 (a). We begin with *3D object grounding* using a pretrained grounding module [29] with the text condition T to identify its point cloud O' for the next stage. Next, we perform *hand-object affordance learning* uses an affordance diffusion module (more details are provided in Sec. 4.4), which takes the object point cloud and text instruction as inputs and generates a fine-grained hand-object joint affordance $A \in \mathbb{R}^{N\times J\times L}$, for N points of the object, J points of the hand, and L frames, to guide plausible hand-object contact by considering the object shape and size. This affordance represents the likelihood of interactions occurring between hand joints and object surfaces over time and is used to guide the interaction motion generation, enabling more accurate interactions aligned with object size and interaction semantics. Finally, we incorporate a *collision-aware motion generation* that voxelizes the region around the interactive object to evaluate spatial accessibility, as the local scene information around the object to be interacted with is more critical for preventing the object from collision through the scene. We also combined with a collision-aware loss that enforces physically plausible motion and prevents interpenetration. Conditioned on the text, local scene, and learned affordance, our model generates physically plausible motion sequences that align with both interaction semantics and environmental constraints.

4.2 Preliminary: Diffusion Models

We utilize the Denoising Diffusion Probabilistic Model (DDPM) [10] to generate both hand-object affordance and motion sequences under conditioning. Given a ground truth signal A_0, the forward diffusion process adds Gaussian noise step-by-step:

$$q(A_t|A_{t-1}) = \mathcal{N}(A_t; \sqrt{1-\beta_t}A_{t-1}, \beta_t I), \tag{1}$$

where β_t is a noise schedule. The closed-form expression for A_t is:

$$A_t = \sqrt{\bar{\alpha}_t}A_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{2}$$

with $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. The reverse process is parameterized by a neural network G_θ which estimates $\hat{A}_0$ conditioned on input A_t and context c:

$$\hat{A}_0 = G_\theta(A_t, c). \tag{3}$$

The model is trained to minimize the mean squared error:

$$\mathcal{L}_{diff} = \mathbb{E}_{A_0,t}\left\|A_0 - \hat{A}_0\right\|_2^2. \tag{4}$$

4.3 3D Object Grounding

To determine which object the human should interact with, we first use a 3D visual grounding model (e.g., ZSVG3D [29]) to locate the object referenced in the input text T. The output is a selected target object point cloud $O' \in \mathbb{R}^{N\times 3}$. Although we use ZSVG3D [29], our pipeline is agnostic to the specific grounding module and can flexibly incorporate future grounding advancements.

4.4 Hand-Object Affordance Learning

Objects in 3D environments vary greatly in their shapes, sizes, and potential interaction strategies. For example, interacting with a small cup may only require one hand, while lifting a heavy box might necessitate both hands, and the hand joints and object parts involved in object interaction also differ. To handle this diversity, our second stage, as shown in Fig. 4(b), takes the object point cloud as inputs and models the fine-grained spatial relationship

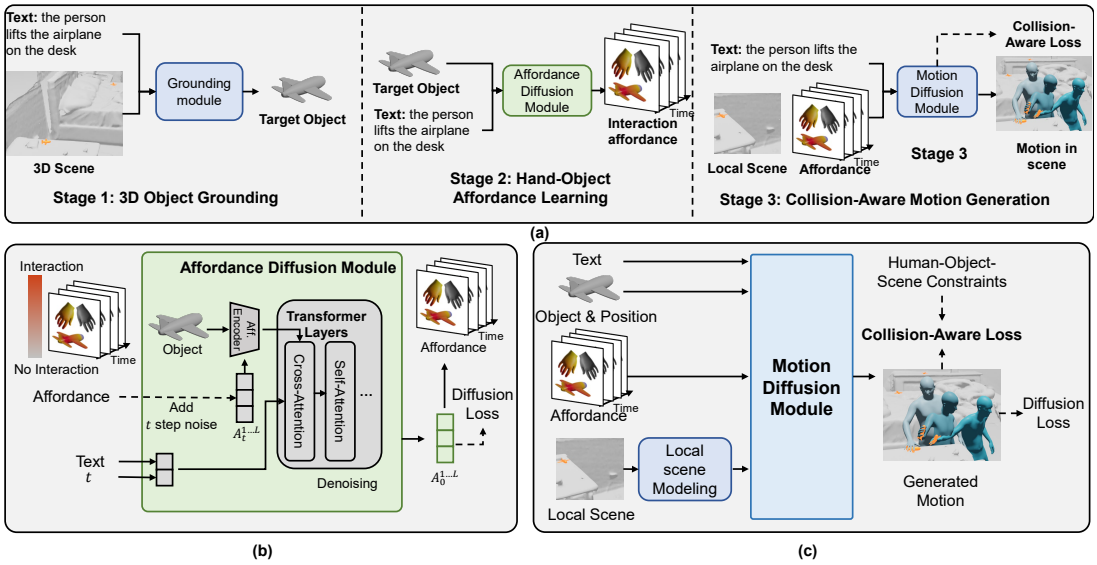

Figure 4: Overview of the proposed framework. (a) Given a text instruction, we first locate the interactive object via a pre-trained grounding model. Then, conditioned on the object point cloud and textual instruction, we generate hand-object affordances. Finally, a collision-aware motion generation module synthesizes human motion and object trajectory, incorporating local scene geometry and learned affordances. (b) Hand-object affordance diffusion module. (c) Collision-Aware motion diffusion module.

Contact loss $\mathcal{L}_{cont}$: encourages meaningful contact (green)

Penetration loss $\mathcal{L}_{pene}$: penalizes penetration between human, object, and scene (blue)

Figure 5: Our Collision-Aware Loss.

between human hands and object surfaces, providing interaction-aware guidance based on the hand-object joint affordances.

Given the object point cloud and text conditions, the model generates the hand-object joint interaction affordance. We calculate the distance between each point of the object and each joint of the human at each frame to get the distance map $d \in \mathbb{R}^{N\times J\times F}$, where N is the number of points in the object, J is the number of human joints, and F is the number of frames. Then we normalize it as $C_{ijn} = \exp\left(-\frac{1}{2}\frac{d_{ijn}}{\sigma^2}\right)$ to assign a higher value to closer point-joint pairs, indicating their high relations. To differentiate interaction types involving single-hand control or bi-hand control, we compute individual affordance scores for each hand and establish a threshold τ to determine hand engagement status. Subsequent normalization based on τ ensures temporal continuity in the resultant affordance signals: $A_{ijn} = \mathbf{1}_{C_{ijn}>\tau} \cdot \frac{C_{ijn}-\tau}{1-\tau}$. A_{ijn} then indicates whether the $j-$th hand joint is involved in the interaction with the $i-$th point of the object in the $n-$th frame or not (if $A_{ijn} = 0$).

To denoise the affordance using a diffusion model, we extract object features from the point cloud using PointNet[32] and fuse them with the noisy affordance via cross-attention. The fused features, along with timestep and text embeddings, form the input to a Transformer decoder, which predicts the final hand-object joint interaction affordance for the subsequent interaction generation stage. Same as Eq(4), we use the diffusion loss to supervise the model training.

4.5 Collision-Aware Motion Generation

Generating physically plausible interactions requires respecting the constraints imposed by the surrounding 3D scene. Thus, we design a collision-aware motion synthesis module guided by local scene modeling and a collision-aware loss. As the local scene information around the object to be interacted with is more critical for preventing the object from collision through the scene, we propose a local scene understanding model that voxelizes the region around the interactive object to evaluate spatial accessibility, providing local scene information for the model. We also combined with a collision-aware loss that enforces physically plausible motion and prevents interpenetration.

Local Scene Modeling. We voxelize the 3D scene into occupancy grids $\mathcal{S}' \in \mathbb{N}^{N_x\times N_y\times N_z}$, indicating whether each voxel is occupied. Around the target object, we extract a region and divide it into patches on the x-y plane. The feature for each patch is derived by pooling occupancy values along the z-axis. These 2D patch features are then encoded using a Vision Transformer (ViT) to obtain local scene feature tokens f_{local} that inform motion synthesis.

Collision-aware Loss. The unique challenge of our proposed task is the complexity of the entities involved, especially the dynamics of objects should be consistent with the scene and human

Method	Goal Distance↓	Multi-modality↑	Physical Realism↑	Non-collision Score↑
MDM[23]	0.904	1.33	0.474	84.97
HUMANISE[25]	0.847	1.17	0.659	95.21
GOAL[21]	0.820	1.25	0.708	96.63
Ours	**0.791**	**1.58**	**0.813**	**98.36**

Table 2: Quantitative evaluations on our dataset.

Method	Multi-modality↑	Physical Realism↑	Non-collision Score↑
TRUMANS[13]	1.29	0.707	98.73
Ours	1.33	0.754	99.03

Table 3: Quantitative evaluations on the TRUMANS dataset. For fairness, we conduct the comparison only on samples involving interactions with movable objects.

motions. Therefore, we introduce a collision-aware loss function composed of two components as shown in Fig. 5: contact loss $\mathcal{L}_{\text{cont}}$ and penetration loss $\mathcal{L}_{\text{pene}}$. The contact loss encourages meaningful contact between hand joints and object surfaces, while the penetration loss penalizes any interpenetration between human body parts, the object, and the scene geometry. Together, these terms enforce physical plausibility and consistency in the generated interactions. The contact loss $\mathcal{L}_{cont}$ is formed as:

$$\mathcal{L}_{cont} = ||d(\tilde{j}, \hat{p}_{\text{obj}})||^2, \tag{5}$$

where $\tilde{j}$ indicates human joints within a distance threshold from the target object and $\hat{p}_{\text{obj}}$ indicates the object points closest to these human joints. The penetration loss $\mathcal{L}_{pene}$ is formed as:

$$\mathcal{L}_{pene} = ||d(\tilde{v}, \hat{p}^{'}_{\text{obj}})||^2, \tag{6}$$

where $\tilde{v}$ indicates human vertices that penetrate the object surface, and $\hat{p}^{'}_{\text{obj}}$ indicates the object points closest to these human vertices.

Considering the necessity of keeping object movement aligned with the scene, we introduce a test-time-penetration constraint. During the denoise process, we recover the human and object point cloud based on the denoise results on step t, and move it along the negative-gradient direction of the test-time penetration loss $\mathcal{L}_{ttp}$. We first filter the vertex set where penetration occurs:

$$\mathcal{P} = \{(i, j) | -\mathbf{n}_j^{\mathrm{T}} \cdot (\mathbf{V}^i_{gen} - \mathbf{V}^j_{scene}) > 0\}, \tag{7}$$

where $\mathbf{V}^i_{gen}$ is the vertex set of the recovered mesh, $\mathbf{V}^i_{gen}$ is the nearest scene vertex set, and $\mathbf{n}_j^{\mathrm{T}}$ is the local normal vector. Then we calculate the test-time penetration loss:

$$\mathcal{L}_{ttp} = \sum_{(i,j)\in\mathcal{P}} ||\mathbf{V}^i - \mathbf{V}^j||_2. \tag{8}$$

Interaction Denoising. We use a diffusion module to generate the interactions. We apply positional encoding to the noisy interaction to obtain interaction features F_{int}. The interaction affordance and the object features are concatenated and fed into the fully connected layers to obtain the F_{obj}. All condition tokens, including F_{int}, F_{obj}, and the local scene feature F_{local}, are concatenated and passed through a transformer encoder to produce the denoised output $\hat{Y}$. The model is trained with a total loss:

$$\mathcal{L}_{total} = \mathcal{L}_{diff} + \lambda_1 \mathcal{L}_{cont} + \lambda_2 \mathcal{L}_{pene} \tag{9}$$

where $\mathcal{L}_{diff}$ is the diffusion reconstruction loss same as Eq(4), $\mathcal{L}_{cont}$ and $\mathcal{L}_{pene}$ are collision-aware losses, and λ_1, λ_2 are hyperparameters. $\mathcal{L}_{ttp}$ is only applied for inference, inhibiting the penetration after every step of the denoising.

Grounding Module	Hand-Object Affordance	Local-Scene Modeling	Goal Dist.↓	Multi-modality↑	Physical Realism↑	Non-collision↑
×	×	✓	1.545	1.73	0.464	90.18
✓	×	×	0.895	1.69	0.524	78.36
✓	×	✓	0.793	1.47	0.570	95.21
✓	✓	×	0.803	**1.87**	0.752	80.24
✓	✓	✓	**0.791**	1.58	**0.813**	**98.36**

Table 4: Ablations of each component in our method.

$\mathcal{L}_{cont}$	$\mathcal{L}_{pene}$	$\mathcal{L}_{ttp}$	Goal Dist.↓	Multi-modality↑	Physical Realism↑	Non-collision↑
×	×	×	0.803	**1.87**	0.752	80.24
×	×	✓	0.793	1.55	0.754	96.15
×	✓	×	0.796	1.72	0.775	90.02
✓	×	×	0.798	1.76	0.788	87.15
✓	✓	×	0.796	1.66	0.808	95.44
✓	✓	✓	**0.791**	1.58	**0.813**	**98.36**

Table 5: Ablations of the collision-aware loss.

Number of Instances	Goal Dist.↓	Multi-modality↑	Physical Realism↑	Non-collision Score↑
Unique	0.405	1.47	0.819	98.77
Multiple	0.793	1.60	0.805	98.21

Table 6: Experiments on the number of instances of the same category as the target object.

5 Experiments

5.1 Evaluation Metrics

We adopt the following metrics to evaluate interaction quality: *1) Goal Distance (Goal Dist).* Measures how well the human interacts with the target object, computed as the average minimum distance between the human body and object surfaces over time. Lower is better. *2) Multimodality.* Assesses the diversity of actions generated from the same prompt and scene. Defined as the average $\mathcal{L}_2$ distance between multiple generated motions. Higher is better. *3) Physical Realism.* Evaluates whether the motion appears physically plausible, using a pre-trained model to score each frame as realistic (1) or not (0). The final score is the average over all frames. Higher is better. *4) Non-collision Score.* Measures the proportion of frames without collisions or penetrations. Higher is better.

Motions	FID↓	Diversity↑	Physical Realism↑
Original Motion	0	0.8629	0.8327
Forced Alignment	0.670	0.8531	0.8186
Refined Motion	0.133	0.8459	0.8213

Table 7: Ablation of our dataset construction. We evaluate the quality of the original HOI motion and our aligned motion (aligned with the 3D scene).

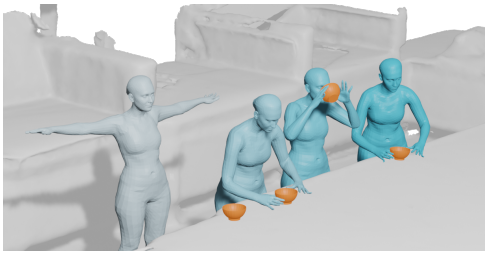

Figure 6: Visualization. The prompt is *The person drinks the bowl on the desk near the sofa.*

5.2 Quantitative Results

We present the quantitative results in our InteractMove dataset in Tab.2. We can observe that: *1) Goal Distance:* Our method achieves the best Goal Distance performance, demonstrating its ability to generate accurate interactions with the correct target object. *2) Physical Realism:* The results show our advantages in the physical plausibility of interactions, which we attribute to the joint modeling of hand-object affordance. *3) Non-collision Score:* Our method yields fewer collisions with the scene, confirming the effectiveness of our collision-aware motion generation design. *4) Multimodality:* Our approach achieves significantly higher diversity compared to previous methods, while still satisfying other constraints, indicating strong capability in generating diverse yet plausible interactions.

We further evaluate our method on the TRUMANS dataset [13] in Tab. 3. Unlike our dataset, TRUMANS includes only 20 interactive objects and 10 predefined interaction types, with discrete action labels instead of free-form language descriptions. While our method is designed for text-controlled interaction motion generation, it is also compatible with label-based inputs. To fit our task, we conduct the comparison only on samples involving interactions with movable objects. Since the TRUMANS dataset provides the target object location, we omit the Goal Distance metric. As shown in Tab. 3, using our evaluation metrics, our method still achieves higher scores in Physical Realism, Non-collision, and Multi-modality, validating the effectiveness of our affordance-based motion generation even on limited-action datasets.

5.3 Ablation Studies

We conduct ablation studies to evaluate both our method and dataset construction.

Ablations on Pipeline Components. Tab.4 shows results after disabling key modules in our pipeline: 1) Without the grounding module, the model struggles to locate target objects and interaction regions, leading to a sharp drop in Goal Distance. 2) Removing the hand-object joint affordance module significantly reduces interaction realism. This is because the hand-object joint affordance provides fine-grained spatiotemporal guidance for interactions and offers unique conditions for different types of objects. 3) Without local scene modeling, predicted motions often collide with the environment, showing that scene constraints are crucial for spatial consistency. These experiments demonstrate the effectiveness of the proposed modules.

Ablations on Collision-aware Loss. Tab.5 compares three collision-aware losses. The inclusion of both $\mathcal{L}_{cont}$ and $\mathcal{L}_{pene}$ as training-phase supervisory terms moderately reduced the Goal Distance, indicating their effectiveness in optimizing spatial positioning during human-object interactions. Constraint $\mathcal{L}_{cont}$ demonstrated greater efficacy in enhancing Physical Realism, while constraint $\mathcal{L}_{pene}$ more substantially improved the Non-collision Score, confirming their respective functional priorities: interaction assurance and collision prevention. As an inference-phase constraint, $\mathcal{L}_{ttp}$ achieved the most significant reduction in intersection artifacts through remarkable Non-collision Score improvement. All constraints exhibited measurable reductions in Multimodality, which we consider an essential trade-off between stringent safety requirements and behavioral diversity preservation.

Distractor Impact. Our dataset contains multiple interactable objects of the same category, requiring models to perform fine-grained spatial reasoning and accurate object grounding based on text. We study the impact of the number of same-category distractors within the scene on the model's final performance. The results are in Tab.6. The task is much harder when multiple distractors exist in the scene, demonstrating that our proposed dataset and task are non-trivial.

Dataset Construction. We evaluate motion quality to assess the effectiveness of our motion alignment techniques, as shown in Tab.7. Original Motion denotes unmodified HOI motions from GRAB and BEHAVE; Forced Alignment refers to forcing aligning these motions to the scene without refinement; Refined Motion is our proposed motion alignment method. Results show that Refined Motion significantly improves motion quality over Forced Alignment, demonstrating its ability to preserve naturalness and reduce artifacts, thus validating the rationale behind our dataset construction strategy.

Visualizations. We provide a visualization showing several frames of the interaction generated by our method in Fig.6. The person lifts the bowl, drinks and puts it back on the desk without collision and correctly uses both hands.

6 Conclusions

In this paper, we introduce a novel task of text-controlled human-object interaction generation in 3D scenes with movable objects, and build the InteractMove dataset to support it. Our proposed pipeline, integrating 3D visual grounding, joint affordance learning, and collision-aware motion generation, effectively handles object identification, diverse interaction prediction, and generation of physically realistic motion. Experiments show that our method outperforms existing approaches in generating physically plausible and text-compliant interactions.

Acknowledgements. This work was supported by the grants from the National Natural Science Foundation of China 62372014, Beijing Nova Program, Beijing Natural Science Foundation 4252040 and the State Key Laboratory of General Artificial Intelligence, BIGAI, Beijing, China.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 422–440.

[2] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. 2023. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21211–21221.

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

[4] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. 2020. Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 387–404.

[5] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. 2024. Text2HOI: Text-guided 3D Motion Generation for Hand-Object Interaction. arXiv:2404.00562 [cs.CV] https://arxiv.org/abs/2404.00562

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5828–5839.

[7] Christian Diller and Angela Dai. 2024. CG-HOI: Contact-Guided 3D Human-Object Interaction Generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 19888–19901. doi:10.1109/cvpr52733.2024.01880

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

[9] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. 2019. Resolving 3D human pose ambiguities with 3D scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2282–2292.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *arXiv preprint arxiv:2006.11239* (2020).

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

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

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

[14] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Ji, Yuxin Peng, and Yang Liu. 2023. Efficient Adaptive Human-Object Interaction Detection with Concept-guided Memory.

[15] Ting Lei, Shaofeng Yin, and Yang Liu. 2024. Exploring the Potential of Large Foundation Models for Open-Vocabulary HOI Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16657–16667.

[16] Jiaman Li, Jiajun Wu, and C. Karen Liu. 2023. Object Motion Guided Human Motion Synthesis. *ACM Transactions on Graphics* 42, 6 (Dec. 2023), 1–11. doi:10.1145/3618333

[17] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. 2019. Putting Humans in a Scene: Learning Affordance in 3D Indoor Environments. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[18] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. 2023. HOI-Diff: Text-Driven Synthesis of 3D Human-Object Interactions using Diffusion Models. arXiv:2312.06553 [cs.CV]

[19] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. 2016. Pigraphs: learning interaction snapshots from observations. *ACM Transactions On Graphics (TOG)* 35, 4 (2016), 1–12.

[20] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015).

[21] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. 2022. GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 13253–13263. doi:10.1109/cvpr52688.2022.01291

[22] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. 2020. GRAB: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 581–600.

[23] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=SJ1kSyO2jwu

[24] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. 2024. Move as You Say Interact as You Can: Language-guided Human Motion Generation with Scene Affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 433–444.

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

[26] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. 2023. InterDiff: Generating 3D Human-Object Interactions with Physics-Informed Diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 14882–14894. doi:10.1109/iccv51070.2023.01371

[27] Zhu Xu, Qingchao Chen, Yuxin Peng, and Yang Liu. 2024. Semantic-Aware Human Object Interaction Image Generation. In *Forty-first International Conference on Machine Learning*.

[28] Dejie Yang, Zhu Xu, Wentao Mo, Qingchao Chen, Siyuan Huang, and Yang Liu. 2024. 3D Vision and Language Pretraining with Large-Scale Synthetic Data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-24*. International Joint Conferences on Artificial Intelligence Organization.

[29] Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. 2024. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20623–20633.

[30] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. 2020. Generating person-scene interactions in 3d scenes. In *International Conference on 3D Vision (3DV)*, Vol. 2.

[31] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. 2020. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6194–6204.

[32] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. 2021. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 16259–16268.