Controllable Human-Object Interaction Synthesis

Jiaman Li[1], Alexander Clegg[2], Roozbeh Mottaghi[2],
Jiajun Wu[1], Xavier Puig[2†], C. Karen Liu[1†]

[1]Stanford University, [2]FAIR, Meta

Abstract. Synthesizing semantic-aware, long-horizon, human-object interaction is critical to simulate realistic human behaviors. In this work, we address the challenging problem of generating synchronized object motion and human motion guided by language descriptions in 3D scenes. We propose **C**ontrollable **H**uman-**O**bject **I**nteraction **S**ynthesis (CHOIS), an approach that generates object motion and human motion simultaneously using a conditional diffusion model given a language description, initial object and human states, and sparse object waypoints. Here, language descriptions inform style and intent, and waypoints, which can be effectively extracted from high-level planning, ground the motion in the scene. Naively applying a diffusion model fails to predict object motion aligned with the input waypoints; it also cannot ensure the realism of interactions that require precise hand-object and human-floor contact. To overcome these problems, we introduce an object geometry loss as additional supervision to improve the matching between generated object motion and input object waypoints; we also design guidance terms to enforce contact constraints during the sampling process of the trained diffusion model. We demonstrate that our learned interaction module can synthesize realistic human-object interactions, adhering to provided textual descriptions and sparse waypoint conditions. Additionally, our module seamlessly integrates with a path planning module, enabling the generation of long-term interactions in 3D environments. Please refer to our project page for the qualitative results.

Keywords: motion synthesis · interaction synthesis · diffusion model

1 Introduction

Synthesizing human behaviors in 3D environments is critical for various applications in computer graphics, embodied AI, and robotics. Humans effortlessly navigate and engage within their surroundings, performing a plethora of tasks routinely. For example, drawing a chair closer to a desk to create a workspace, adjusting a floor lamp to cast the perfect glow, or neatly storing a suitcase. Each of these tasks requires precise coordination between the human, the object, and the surroundings. These tasks are also deeply rooted in purpose. Language serves as a powerful tool to articulate and convey these intentions. Synthesizing

† indicates equal contribution.

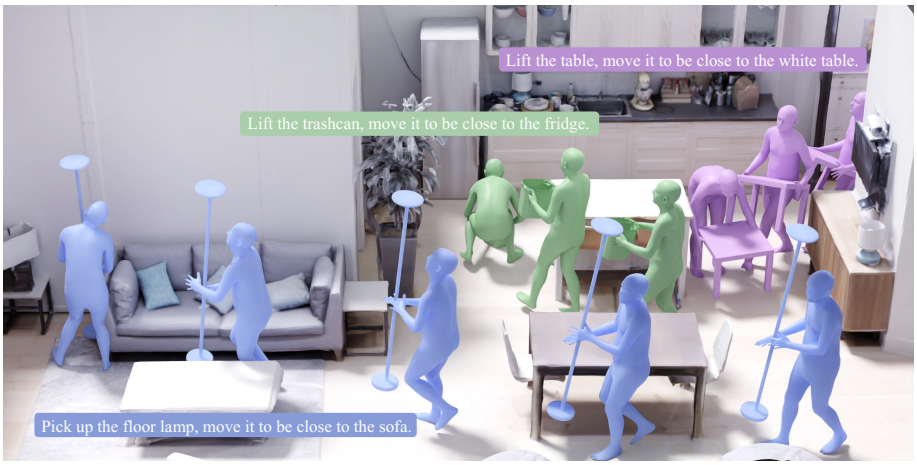

Fig. 1: Given an initial object and human state, a language description, and sparse object waypoints in a 3D scene, CHOIS generates synchronized object motion and human motion at the same time.

realistic human and object motion guided by language and scene context is the cornerstone of building advanced AI systems that simulate continuous human behaviors in diverse 3D environments.

Although some existing works study the problem of human-scene interaction [17], they are restricted to scenarios with static objects, such as sitting on a chair, neglecting the highly dynamic interactions that occur frequently in daily life. Recent advances have been made in modeling dynamic human-object interactions, yet these approaches focus solely on smaller objects [13, 29] or lack the ability to manipulate diverse objects [19, 60]. The most recent work on manipulation of larger, diverse objects relies on sequences of past interaction states or complete sequences of object motion [28, 54, 61], thus being incapable of synthesizing both object motion and human motion from initial states alone. Unlike these existing methods, we focus on synthesizing realistic human-object interactions for diverse objects in 3D environments from language and initial states. This problem is challenging primarily for two reasons. First, we need to generate realistic and synchronized motions for both objects and humans. Human hands should maintain appropriate contact with objects during interaction, and object motion should maintain a causal relationship to human actions. Second, 3D scenes are often cluttered with numerous objects, constraining the space of feasible motion trajectories. Thus, the method should accommodate environment clutter, rather than operating under the assumption of an empty scene.

To address these challenges, we leverage waypoints to guide the synthesis process. Starting with a language description outlining the desired human actions and an initial object and human state, we first extract a set of waypoints from the environment. Our goal is thus to generate motions for both humans and objects that align with the directives specified by language, while also conforming

to the environmental constraints defined by waypoint conditions derived from 3D scene geometry.

To achieve this, we employ a conditional diffusion model to generate synchronized object and human motion simultaneously, conditioned on language descriptions, initial states, and sparse object waypoints. However, naively applying a diffusion model fails to generate object motions that precisely adhere to the input object waypoints. Additionally, the generated interactions often exhibit issues such as unrealistic contact, foot floating, and objects penetrating the floor. To improve the accuracy of the predicted object motion, we incorporate an object geometry loss during training. Furthermore, we devise guidance terms applied during the sampling process to explicitly enforce contact constraints, thereby directly enhancing the realism of the generated interactions. We demonstrate the effectiveness of our learned interaction synthesis module within a system that produces continuous, realistic, and context-aware interactions given language descriptions and 3D scenes.

To summarize, our work makes the following contributions. First, we identify that the combination of language and object waypoints provides precise and expressive information for human-object interaction synthesis. We show that object waypoints do not need to be dense, which allows us to utilize existing path planning algorithms to generate sparse waypoints that represent long-horizon interactions in complex scenarios. Second, based on this finding, we devise a method that synthesizes human-object interaction guided by language and sparse waypoints of the object, using a conditional diffusion model. Third, we demonstrate that our approach synthesizes realistic interactions on the FullBodyManipulation dataset [28] and generalizes to novel objects from 3D-FUTURE [12]. We also integrate our method into a pipeline that synthesizes long-horizon environment-aware human-object interactions from 3D scenes and language input.

2 Related Work

Motion Synthesis from Language. With the development of large-scale high-quality motion capture datasets like AMASS [30], there has been a growing interest in generative human motion modeling. BABEL [40] and HumanML3D [14] further introduce action labels and language descriptions to enrich the mocap dataset, enabling the development of action-conditioned motion synthesis [37] and text-conditioned motion synthesis [14,38,50]. Prior work has shown that VAE formulation is effective in generating diverse human motion from text [14, 15]. Recently, with the success of the diffusion model in this domain [2, 6, 23, 27, 28, 41, 44, 45, 52, 62, 67], extensive work has explored generating motion from text using conditioning [8, 24, 51, 64]. In this work, we also take language descriptions as input to guide our generation. Instead of synthesizing human motion alone, we generate both object motion and human motion conditioned on the text.

Motion Synthesis in 3D Scenes. With the advent of paired scene-motion data [1, 16, 18, 57, 70] and paired object-motion data [17, 65], approaches [1, 17, 25, 55, 56, 65] have been developed to generate human interactions such as

sitting on a chair and reaching a target position in 3D scenes. To populate human-object interactions without training on paired scene-motion data, path planning algorithms have been deployed to generate collision-free paths which then guide the human motion generation [17, 33, 66, 68]. Another line of work leverages reinforcement learning frameworks to train scene-aware policies for synthesizing navigation and interaction motions in static 3D scenes [26,59]. In this work, instead of focusing on static scenes or objects, we synthesize interactions with dynamic objects. Also, inspired by approaches that decompose scene-aware motion generation into path planning and goal-guided generation phases, we design an interaction synthesis module conditioned on sparse object waypoints that can be effectively integrated into a scene-aware synthesis pipeline.

Interaction Synthesis. The field of modeling dynamic human-object interactions has largely focused on hand motion synthesis [7, 63, 69]. Recently, with the advent of full-body motion datasets with hand-object interactions [11, 49], models [48, 58] have been developed to synthesize full-body motions preceding object grasping. Some recent studies predict object motion based on human movements [36], and others [4,13,29] have taken this further by synthesizing both body and hand motion, subsequently applying optimization to predict object motion. However, these approaches focus on smaller objects where hand motion is the primary focus. In terms of manipulating larger objects, some methods train reinforcement learning policies to synthesize box lifting and moving behaviors [19, 32, 60], yet these models struggle to generalize to manipulation of diverse objects. Based on paired human-object motion data [3, 28, 54], recent works predict interactions from a sequence of past interaction states [54,61] or an object motion sequence [28], incapable of synthesizing interactions in 3D scenes solely from initial states. In this work, we generate synchronized object and human motion conditioned on sparse object waypoints, serving to ground the resulting trajectories in 3D scenes.

Concurrent Work. Our work is concurrent with CG-HOI [10] and HOI-Diff [35], which use the BEHAVE dataset [3] to synthesize both human motion and object motion from text. Our work also aims to synthesize human-object interactions but differs from the concurrent approaches by integrating textual conditions with sparse waypoint conditions. This unique combination enables our interaction synthesis module to be integrated with path planning modules, enabling long-term interactions in 3D environments. Additionally, we leverage the FullBodyManipulation dataset [28], specifically designed for interaction synthesis, offering superior data scale and motion quality compared to BEHAVE [3].

3 Method

Our goal is to generate synchronized object and human motion, conditioned on a language description, object geometry, initial object and human states, and sparse object waypoints. Two primary challenges arise in this context: first, modeling the complexity of synchronized object and human motion while also respecting the sparse condition signals; and second, ensuring the realism of contact between

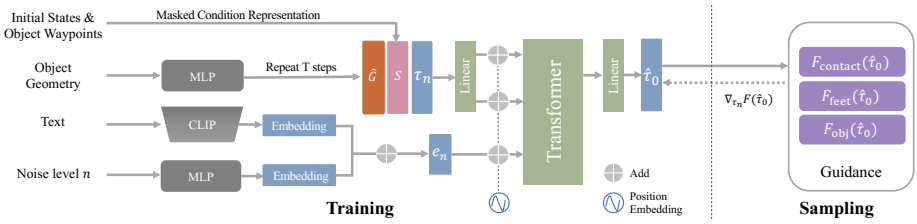

Fig. 2: Method Overview. Given an object geometry, we use the BPS representation to encode the geometry and an MLP to project the features into a low-dimensional vector. This feature vector is concatenated with masked pose states to form conditions for the denoising network. During sampling, we use analytical functions to compute gradients and perturb the generation to satisfy our defined constraints.

the human and object. To tackle the generation problem of complex interactions, we employ a conditional diffusion model to generate object motion and human motion at the same time. However, naively learning a conditional diffusion model to generate both object motion and human motion cannot ensure the precise contact between hand and object and the realism of the interaction. Thus, we incorporate several constraints as guidance during the sampling process of our trained diffusion model. We illustrate our approach in Figure 2.

3.1 Data Representation

Object and Human Motion Representation. We denote the human motion as $\boldsymbol{X} \in \mathbb{R}^{T \times D}$, where T and D represent the time steps and dimension of the human pose. $\boldsymbol{X}_t$, corresponding to the human pose at frame t, consists of global joint positions and 6D continuous rotations [71]. We adopt the widely used parametric human model, SMPL-X [34] to reconstruct the human mesh from the pose and shape parameters. To represent the object motion, we use two components: the global 3D position and the relative rotation. The global position is represented by the centroid of the object, while the relative rotation, denoted as $\boldsymbol{R}_{\text{rel}}$ at frame t, is expressed with respect to the input object's geometry $\boldsymbol{V}$ such that $\boldsymbol{V}_t = \boldsymbol{R}_{\text{rel}}\boldsymbol{V}$, where $\boldsymbol{V}_t$ represent the vertices of object at frame t. We denote the object motion by $\boldsymbol{O} \in \mathbb{R}^{T \times 12}$.

Object Geometry Representation. We represent the object geometry using the Basis Point Set (BPS) representation [39]. Following prior work [28], we begin by sampling a set of basis points from the volume of a ball with a 1-meter radius. Subsequently, for each sampled point, we calculate the minimum Euclidean distance to the nearest point on the object's mesh. Alongside this, we record the directional vectors from the basis points to their nearest neighbors. The resulting BPS representation is denoted as $\boldsymbol{G} \in \mathbb{R}^{1024 \times 3}$, representing 1024 sampled points each with a vector indicating their spatial relationship to the object's surface.

Input Condition Representation. We first use an MLP to project the object BPS representation $\boldsymbol{G}$ to a low-dimensional vector which is then broadcasted to each frame denoted as $\hat{\boldsymbol{G}} \in \mathbb{R}^{T \times 256}$ following [28]. We then adopt a masked motion data representation denoted as $\boldsymbol{S} \in \mathbb{R}^{T \times (12+D)}$ to represent the initial

states and waypoint conditions. The initial state contains the human pose and object pose at the first frame. The waypoint conditions consist of a series of 2D object positions for every 30 frames, and a 3D object position at the final frame. The remainder of $\boldsymbol{S}$ is padded with zeros. The encoded object geometry vector and the masked motion condition vector are then concatenated, serving as part of the input for our denoising network. For effectively integrating language conditions, we utilize CLIP [42] as a text encoder to extract language embeddings.

3.2 Interaction Synthesis Model

Conditional Diffusion Model. We utilize a conditional diffusion model [21] to generate synchronized object and human motion. To improve the realism of hand-object interaction, our model also predicts contact labels $\boldsymbol{H} \in \mathbb{R}^{T\times 2}$ for both the left and right hands. These predicted contact labels play a crucial role in guiding the sampling process, ensuring more accurate and realistic hand-object contacts in the generated motion sequence. The complete data representation in our model is denoted as $\boldsymbol{\tau} = \{\boldsymbol{X}, \boldsymbol{O}, \boldsymbol{H}\}$, encapsulating motion and contact data.

The conditional signals of our model, denoted as $\boldsymbol{c}$, include initial states, sparse object waypoints, the object BPS representation, and language descriptions. The diffusion model consists of a forward diffusion process that progressively adds noise to the clean data $\boldsymbol{\tau}_0$ and a reverse diffusion process which is trained to reverse this process. The forward diffusion process introduces noise for N steps formulated using a Markov chain,

$$q(\boldsymbol{\tau}_n|\boldsymbol{\tau}_{n-1}) := \mathcal{N}(\boldsymbol{\tau}_n; \sqrt{1-\beta_n}\boldsymbol{\tau}_{n-1}, \beta_n \boldsymbol{I}), \tag{1}$$

$$q(\boldsymbol{\tau}_{1:N}|\boldsymbol{\tau}_0) := \prod_{n=1}^{N} q(\boldsymbol{\tau}_n|\boldsymbol{\tau}_{n-1}), \tag{2}$$

where β_n represents a fixed variance schedule and $\boldsymbol{I}$ is an identity matrix. Our goal is to learn a model p_θ to reverse the forward diffusion process,

$$p_\theta(\boldsymbol{\tau}_{n-1}|\boldsymbol{\tau}_n, \boldsymbol{c}) := \mathcal{N}(\boldsymbol{\tau}_{n-1}; \boldsymbol{\mu}_\theta(\boldsymbol{\tau}_n, n, \boldsymbol{c}), \boldsymbol{\Sigma}_n), \tag{3}$$

where $\boldsymbol{\mu}_\theta$ denotes the predicted mean and $\boldsymbol{\Sigma}_n$ is a fixed variance. Learning the mean can be re-parameterized as learning to predict the clean data representation $\boldsymbol{\tau}_0$. The objective [21] is defined as

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{\tau}_0, n}||\hat{\boldsymbol{\tau}}_\theta(\boldsymbol{x}_n, n, \boldsymbol{c}) - \boldsymbol{\tau}_0||_1. \tag{4}$$

Model Architecture. We employ a transformer architecture [53] as our denoising network. Our input consists of object geometry $\hat{\boldsymbol{G}}$, masked motion conditions $\boldsymbol{S}$, and noisy data representation $\boldsymbol{\tau}_n$ at noise level n. The input is projected to a sequence of feature vectors using a linear layer. We employ an MLP to embed the noise level n. Then we combine the noise level embedding and the language embedding to form a single embedding vector denoted as $\boldsymbol{e}_n$. The embedding vector $\boldsymbol{e}_n$ has the same dimension as these feature vectors and is fed to the transformer along with these vectors. The final prediction $\hat{\boldsymbol{\tau}}_0$ is made by

projecting the updated feature vectors of the transformer excluding the time step corresponding to $\boldsymbol{e}_n$. The interaction synthesis model is illustrated in Figure 2.

Object Geometry Loss. During the training phase, we incorporate an additional loss to improve the object motion prediction. Utilizing the Basis Point Set (BPS) representation, we initially compute the nearest neighbor points on the object mesh in rest pose for each of the fixed set of points. From these, we sample 100 points out of the 1024 nearest neighbors to capture a rough outline of the object's shape. These selected points are defined as $\boldsymbol{K}_{\text{rest}} \in \mathbb{R}^{100\times 3}$, representing our selected object vertices at rest pose.

At each time step in our model, the predicted object rotation (converted to relative rotation with respect to the object geometry in rest pose) and position are employed to calculate the corresponding positions of these selected vertices. This is represented by the following equation, where $\hat{\boldsymbol{R}}_t$ and $\hat{\boldsymbol{d}}_t$ denote the predicted rotation and translation of the object, and $\boldsymbol{K}_t$ refers to the ground truth vertices at time step t. The object geometry loss is computed as

$$\mathcal{L}_{\text{obj}} = \sum_{t=1}^{T} ||\hat{\boldsymbol{R}}_t \boldsymbol{K}_{\text{rest}} + \hat{\boldsymbol{d}}_t - \boldsymbol{K}_t||_1. \tag{5}$$

This loss function plays a critical role in guiding the model to accurately predict the transformation of the object.

3.3 Guidance

During the training phase of our interaction synthesis model, there are no explicit contact constraints enforced in the losses. Incorporating loss terms such as hand-object contact loss, and object-floor penetration loss poses a challenge for training. First, these types of loss terms are computationally expensive and would slow down training significantly. Second, introducing more loss terms requires meticulously balancing different losses which usually necessitates re-training models with different settings. Instead, enforcing these constraints during test time is more flexible and makes it easier to select appropriate weights for different terms. Thus, to refine our generated interactions, we propose the application of guidance during the sampling process.

In the diffusion model framework, classifier guidance is commonly applied during test time to control the generation process in order to satisfy specific objectives or constraints. A typical approach to applying classifier guidance [9] is to perturb the noisy predicted mean at each denoising step. This is formulated as $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} - \alpha \boldsymbol{\Sigma}_n \nabla_{\boldsymbol{\mu}} F(\boldsymbol{\mu})$, where $\boldsymbol{\mu}$ denotes the predicted mean at denoising step n defined by Equation 3. F represents a learned or analytical function that determines how much the predicted mean should be penalized and α represents the strength of the perturbation. This guidance computes the gradient with respect to the noisy mean, requiring F to be trained on noisy data or a deterministic function designed for noisy data. Another approach is reconstruction guidance [22], which has proven to be effective for controlling the generation process in prior work [24, 25, 43]. Instead of perturbing the noisy mean, it perturbs the predicted

clean data representation $\hat{\boldsymbol{\tau}}_0$ using the gradient with respect to the noisy input data representation $\boldsymbol{\tau}_n$. The process is formally represented as

$$\tilde{\boldsymbol{\tau}}_0 = \hat{\boldsymbol{\tau}}_0 - \alpha \boldsymbol{\Sigma}_n \nabla_{\boldsymbol{\tau}_n} F(\hat{\boldsymbol{\tau}}_0). \tag{6}$$

In this work, we leverage reconstruction guidance [22] in the sampling process as we empirically found it to be more stable. We define multiple analytical functions as guidance terms which we will introduce in the following sections.

Hand-Object Contact Guidance. We have implemented a specialized contact guidance function to improve the hand-object contact accuracy for frames generated by our model. This function is specifically designed to address cases where a noticeable distance exists between the hands and the object, thereby improving the realism of the interaction. The contact guidance function is defined as follows:

$$F_{\text{contact}} = \|\boldsymbol{M}_l \odot |\boldsymbol{J}_l - \boldsymbol{V}_l|\|_1 + \|\boldsymbol{M}_r \odot |\boldsymbol{J}_r - \boldsymbol{V}_r|\|_1 . \tag{7}$$

$\boldsymbol{M}_l$ and $\boldsymbol{M}_r$ are binary masks for the left and right hand, respectively. These masks are derived from the predicted contact labels $\boldsymbol{H}$, with $\boldsymbol{M}_l, \boldsymbol{M}_r \in \mathbb{R}^{T\times 1}$ and are defined as $\boldsymbol{M}_l, \boldsymbol{M}_r = (\boldsymbol{H} > 0.95)$. This thresholding identifies frames where contact is likely to occur. The symbol $\odot$ represents the Hadamard product, applying these masks element-wise to the absolute differences between the hand positions ($\boldsymbol{J}_l$, $\boldsymbol{J}_r \in \mathbb{R}^{T\times 3}$) and nearest points on the object mesh ($\boldsymbol{V}_l$, $\boldsymbol{V}_r \in \mathbb{R}^{T\times 3}$).

Feet-Floor Contact Guidance. When generating joint positions and rotations, our model operates without awareness of the body's shape. Consequently, using the SMPL-X model [34] with predicted root positions, joint rotations, and a test subject's specific body shape parameters to reconstruct the human mesh can sometimes lead to scenarios where the feet do not touch the floor. To rectify this, we implement a guidance term that encourages realistic feet-floor contact.

The joint positions of the left and right toes are represented as $\boldsymbol{J}_l$ and $\boldsymbol{J}_r$, respectively. We identify the supporting foot in each frame by comparing the z components of these two joints at each frame. We also introduce a threshold height $h = 0.02$ meters, which is determined from the analysis of foot height in the ground truth motion. The guidance term is defined as follows:

$$F_{\text{feet}} = ||min(\boldsymbol{J}_l^z, \boldsymbol{J}_r^z) - h||_2. \tag{8}$$

This function computes the norm of the vertical difference between the lowest point of either toe and the threshold height h.

Object-Floor Penetration Guidance. To address the issue of generated object states potentially penetrating the floor, we integrate an additional guidance function into the sampling process. Given that our floor is positioned at the plane where $z = 0$, we define the guidance term as follows:

$$F_{\text{obj}} = ||min(\boldsymbol{V}^z, 0)||_1, \tag{9}$$

where $\boldsymbol{V}^z$ represents the z-coordinate of the object vertices.

During inference, we apply multiple guidance concurrently defined as follows,

$$F_{\text{all}} = \lambda_1 F_{\text{contact}} + \lambda_2 F_{\text{feet}} + \lambda_3 F_{\text{obj}}, \tag{10}$$

where λ_1, λ_2, λ_3 denote the loss weights. We apply the guidance in the last 10 denoising steps only since the prediction in the early steps is extremely noisy.

4 Experiments

We first introduce the datasets and evaluation metrics. Then we show comparisons of our proposed approach against the baselines. We further conduct a human perceptual study to complement our evaluation and ablation study to verify the effectiveness of our proposed guidance terms. Moreover, we demonstrate an application that generates long-term interactions conditioned on object waypoints extracted from 3D scenes.

4.1 Datasets

The FullBodyManipulation dataset [28] consists of 10 hours of high-quality, paired object and human motion, including interaction with 15 different objects. However, our study does not encompass the generation of motion for articulated objects, leading us to exclude sequences related to two such objects (vacuum and mop). We employ this dataset both for training our interaction model and for evaluating the generated results. The training set comprises 15 subjects, with an additional 2 subjects designated for testing, adhering to the dataset partitioning used in OMOMO [28]. We specifically chose the FullBodyManipulation dataset [28] over BEHAVE [3] due to several limitations in the latter. BEHAVE is not tailored for interaction synthesis, presenting challenges such as noticeable jittery motions, limited data scale and a lack of locomotion.
The 3D-FUTURE dataset [12] includes 3D models of various furniture items. From this dataset, we select 17 objects representing diverse types (such as chairs, tables, floor lamps, and boxes). This dataset serves to test our model's ability to generalize to objects it has not previously encountered. Given that the 3D-FUTURE dataset only includes 3D models, we integrate these objects with motion from the testing set of the FullBodyManipulation dataset [28] to generate input conditions for evaluation. In particular, given an object in 3D-FUTURE, we take a motion from FullBodyManipulation belonging to the same object category, and extract the 2D coordinates for the object positions every 30 frames to represent the input waypoints.

4.2 Evaluation Metrics

Condition Matching Metric: This metric calculates the Euclidean distance between the predicted and input object waypoints. It includes the start and end position errors (T_s, T_e), and waypoint errors (T_{xy}) measured in centimeters (cm).
Human Motion Quality Metric: This metric encompasses the foot sliding score (FS), foot height (H_{feet}), *Fréchet Inception Distance* (FID) and *R-precision* (R_{prec}). FS is the weighted average of accumulated translation in the xy plane, following prior work [20], measured in centimeters (cm). H_{feet} assesses the height

of the feet, also in centimeters. R_{prec} and FID are computed following the text-to-motion task [14]. R_{prec} (top-3) measures whether the generated motion is consistent with the text. FID assesses the motion quality by computing the discrepancy between the distributions of ground truth and generated motions.
Interaction Quality Metric: This metric assesses the accuracy of hand-object interactions, encompassing both contacts and penetrations. For contact accuracy, it employs precision (C_{prec}), recall (C_{rec}), and F1 score (C_{F_1}) metrics following prior work [28]. Additionally, it includes contact percentage ($C_{\%}$), determined by the proportion of frames where contact is detected. To compute the penetration score (P_{hand}), each vertex of the hand V_i is used to query the precomputed object's Signed Distance Field (SDF). This process yields a corresponding distance value d_i for each vertex. The penetration score is then derived by computing the average of the negative distance values (representing penetration), formalized as $\frac{1}{n}\sum_{i=1}^{n}|min(d_i, 0)|$, measured in centimeters (cm).
Ground Truth (GT) Difference Metric: This metric measures the deviation of generated results from the ground truth motion. It comprises the mean per-joint position error (MPJPE), translation error of the root joint (T_{root}), and object position error (T_{obj}), all computed using the Euclidean distance between the predicted and actual ground truth positions in centimeters (cm). Additionally, this metric includes the root joint orientation error (O_{root}) and the object orientation error (O_{obj}). These errors are calculated with the Frobenius norm of the rotational difference, formulated as $||\boldsymbol{R}_{pred}\boldsymbol{R}_{gt}^{-1} - \boldsymbol{I}||_2$ where $\boldsymbol{R}_{pred}$ and $\boldsymbol{R}_{gt}$ represent the predicted and ground truth rotation matrices respectively.

4.3 Results

Baselines. As there is no prior work presenting a solution for our task, we adapt related works such as InterDiff [61], MDM [51], and OMOMO [28] to fit our problem setting in order to establish baseline comparisons. InterDiff [61] focuses on anticipating human-object interactions using the previous 10 frames. MDM [51] generates human motion from language descriptions. OMOMO [28] synthesizes human motion based on provided object motion trajectories. We adapt InterDiff to accept additional input conditions including text and sparse waypoints. For MDM, we update the model to incorporate our object geometry representation and sparse waypoints. Additionally, we enhance MDM to include our object motion representation as an additional output. OMOMO requires a sequence of object states to generate full-body human poses; therefore, we implement a linear interpolation strategy for object positions based on the provided start and end positions, as well as predefined waypoints in the xy-plane, while maintaining consistent object rotation from the initial frame throughout the sequence. Furthermore, we introduce two variations, Pred-OMOMO and GT-OMOMO, as part of our ablation studies. Pred-OMOMO combines our text-conditioned object motion synthesis module with OMOMO. GT-OMOMO utilizes ground truth object motion as input for OMOMO. Additionally, we evaluate our approach CHOIS against two ablations: CHOIS w/o L_{obj} and CHOIS w/o F_{all}. CHOIS w/o L_{obj} is trained as a conditional diffusion model but does not

Table 1: Interation synthesis on the FullBodyManipulation dataset [28].

	Condition Matching			Human Motion				Interaction					GT Difference			
Method	$T_s \downarrow$	$T_e \downarrow$	$T_{xy} \downarrow$	$H_{feet} \downarrow$	FS$\downarrow$	$R_{prec} \uparrow$	$FID \downarrow$	$C_{prec} \uparrow$	$C_{rec} \uparrow$	$C_{F_1} \uparrow$	$C_{\%}$	$P_{hand} \downarrow$	MPJPE$\downarrow$	$T_{root} \downarrow$	$T_{obj} \downarrow$	$O_{obj} \downarrow$
Interdiff [61]	0.00	158.84	72.72	**0.90**	0.42	0.08	208.0	0.63	0.28	0.33	0.27	0.55	25.91	63.44	88.35	1.65
MDM [51]	5.18	33.07	19.42	6.72	0.48	0.51	6.16	0.72	0.47	0.53	0.43	0.66	17.86	34.16	24.46	1.85
Lin-OMOMO [28]	0.00	0.00	0.00	7.21	0.41	0.29	15.33	0.68	0.56	0.57	0.54	**0.51**	21.73	36.62	17.12	1.21
Pred-OMOMO [28]	2.39	8.03	4.15	7.08	0.40	0.54	4.19	0.73	0.66	0.66	0.62	0.58	18.66	28.39	16.36	1.05
GT-OMOMO [28]	0.00	0.00	0.00	7.10	0.41	0.48	5.69	0.77	**0.66**	0.67	0.59	0.55	15.82	24.75	0.00	0.00
CHOIS w/o L_{obj}	5.76	14.16	8.44	6.55	0.40	0.65	3.26	0.75	0.50	0.55	0.43	0.66	**14.34**	**21.97**	15.53	**0.98**
CHOIS w/o F_{all}	1.75	6.61	**2.69**	6.64	0.38	**0.65**	3.58	0.78	0.49	0.55	0.41	0.65	15.23	24.13	**11.51**	0.99
CHOIS (ours)	**1.71**	**6.31**	2.87	**4.20**	**0.35**	0.64	**0.69**	**0.80**	0.64	**0.67**	0.54	0.59	15.30	24.43	12.53	0.99

Table 2: Interaction synthesis on the 3D-FUTURE dataset [12].

	Condition Matching			Human Motion				Interaction	
	$T_s \downarrow$	$T_e \downarrow$	$T_{xy} \downarrow$	$H_{feet} \downarrow$	FS$\downarrow$	$R_{prec} \uparrow$	$FID \downarrow$	$C_{\%}$	$P_{hand} \downarrow$
InterDiff [61]	0	161.26	72.77	-0.26	0.42	0.09	207.3	0.24	**0.11**
MDM [51]	12.58	40.55	28.72	7.02	0.49	0.53	8.50	0.34	0.26
Lin-OMOMO [28]	0	0	0	6.32	0.42	0.23	23.17	0.44	0.11
Pred-OMOMO [28]	4.15	9.03	3.89	6.08	0.40	0.46	3.74	0.50	0.18
CHOIS w/o L_{obj}	6.70	13.73	7.99	5.68	0.41	**0.66**	3.26	0.36	0.30
CHOIS w/o F_{all}	5.75	7.96	**2.68**	5.84	0.39	0.62	4.78	0.33	0.26
CHOIS (ours)	**4.12**	**7.35**	2.92	**3.75**	**0.38**	0.62	**1.60**	0.48	0.15

include an additional object geometry loss. This variant allows us to understand the baseline performance of the diffusion model in a straightforward setup. In contrast, CHOIS w/o F_{all} incorporates the object geometry loss in its training process but operates without guidance during inference. This approach lets us explore the effectiveness of object geometry loss during training while assessing the model's capability in the absence of guidance.

Results on the FullBodyManipulation Dataset. We evaluate our approach using objects from the FullBodyManipulation dataset [28] as shown in Table 1. Introducing object geometry loss notably improves the condition matching metric. Furthermore, adding guidance during inference leads to better contact accuracy, reduced hand-object penetration, and less foot floating.

InterDiff cannot adequately adhere to the input waypoints and text since the input conditions are entangled. The condition embedding, which includes the past 10 frames, point cloud features, text, and sparse waypoints, is summed up to predict future frames. This approach leads to suboptimal performance in condition matching metrics and R_{prec}. Also, we observe feet-floor penetration issues in InterDiff's generated results, resulting in a lower foot height. MDM can synthesize plausible interactions but, as seen in the Interaction metrics, struggles with generating realistic contacts since it does not enforce any contact constraints.

Lin-OMOMO shows zero deviation from the input object trajectory as it only predicts human motion and does not alter the object motion input at the sparse input locations. Pred-OMOMO demonstrates improved contact metrics compared to the baselines but is still inferior to our CHOIS in terms of condition matching and human motion quality. Moreover, Pred-OMOMO requires three stages during inference, one from our object motion synthesis module and two from OMOMO, whereas CHOIS operates as a single-stage model. GT-OMOMO,

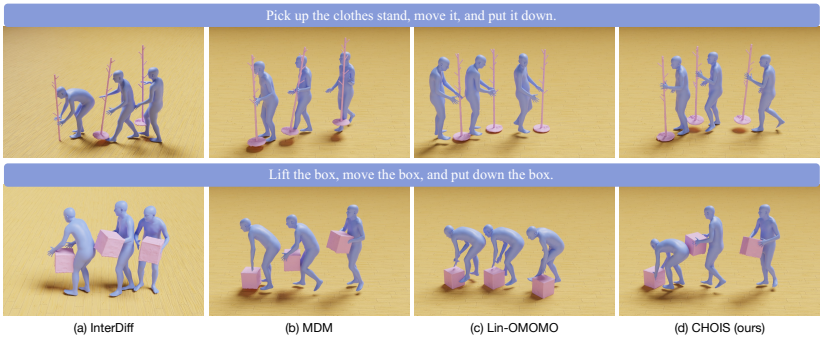

Fig. 3: Qualitative results of the FullBodyManipulation dataset [28].

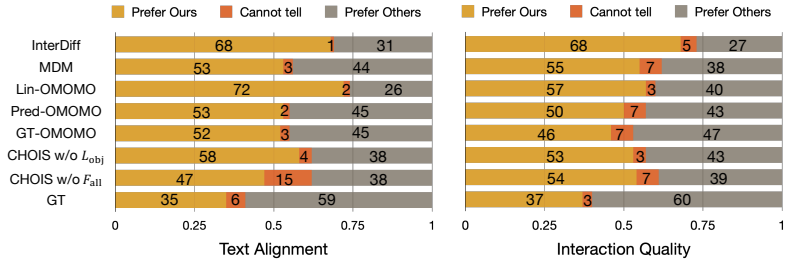

Fig. 4: Results of human perceptual studies. The numbers shown in the chart represent the percentage (%) over motion preferences.

Table 3: Ablation study on the FullBodyManipulation dataset [28]. We measure the effect of different guidance terms in the human and object motion generation.

	Condition Matching			Human Motion				Interaction					GT Difference			
Method	$T_s \downarrow$	$T_e \downarrow$	$T_{xy} \downarrow$	$H_{feet} \downarrow$	FS$\downarrow$	$R_{prec} \uparrow$	$FID \downarrow$	$C_{prec} \uparrow$	$C_{rec} \uparrow$	$C_{F_1} \uparrow$	$C_{\%}$	$P_{hand} \downarrow$	MPJPE$\downarrow$	$T_{root} \downarrow$	$T_{obj} \downarrow$	$O_{obj} \downarrow$
CHOIS w/o F_{contact}	**1.70**	6.42	2.70	**3.93**	**0.32**	**0.66**	0.74	0.78	0.49	0.55	0.41	0.65	15.41	**23.63**	**11.44**	0.99
CHOIS w/o F_{feet}	1.72	6.34	2.90	6.65	0.39	0.63	3.76	**0.81**	**0.64**	0.66	0.54	**0.58**	15.44	25.09	13.31	0.99
CHOIS w/o F_{all}	1.75	6.61	**2.69**	6.64	0.38	0.65	3.58	0.78	0.49	0.55	0.41	0.65	**15.23**	24.13	11.51	0.99
CHOIS (ours)	1.71	**6.31**	2.87	4.20	0.35	0.64	**0.69**	0.80	0.64	**0.67**	0.54	0.59	15.30	24.43	12.53	0.99

requiring a ground truth object motion sequence for input, shows comparable performance in interaction and GT difference metrics. However, the motions it generates suffer from foot floating issues, leading to a larger H_{feet} and FID. We also showcase qualitative comparisons against the baselines in Figure 3.

Results on the 3D-FUTURE Dataset. To test our model's ability to generalize to new objects, we conduct evaluations using the 3D-FUTURE dataset [12]. As shown in Table 2, our proposed method outperforms the baselines.

Human Perceptual Study. We conduct two human perceptual studies to further complement the evaluation of our approach. The first study assesses the consistency between the generated interactions and the text input. The second study evaluates the overall quality of these generated interactions. For each of these studies, we generate 100 sequences using each method, including baselines, OMOMO ablations, our CHOIS model, our own ablations, and the ground truth. This results in a set of 800 pairs. We employ Amazon Mechanical Turk (AMT)

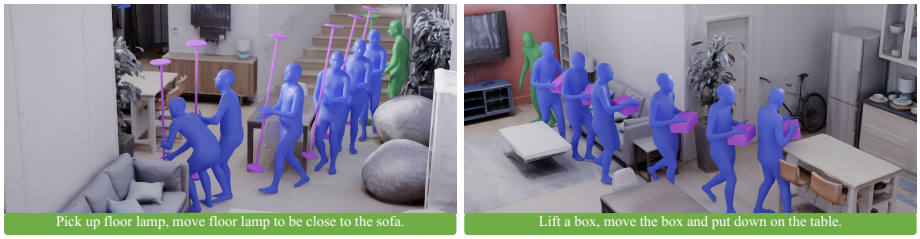

Fig. 5: Long-term interaction synthesis. Given language descriptions, a 3D scene with semantic labels, and initial human and object states, we synthesize long-term human-object interactions. The initial state is shown in green.

Table 4: Long-term interaction synthesis results on the FullBodyManipulation [28] and 3D-FUTURE datasets [12]. * represents the results on the 3D-FUTURE dataset.

	Condition Matching			Human Motion		Interaction	
	$T_s \downarrow$	$T_e \downarrow$	$T_{xy} \downarrow$	$H_{feet} \downarrow$	FS$\downarrow$	$C_{\%}$	$P_{hand} \downarrow$
CHOIS w/o F_{all}	**1.50**	**7.19**	**5.10**	6.01	**0.43**	0.49	0.70
CHOIS	2.22	9.94	5.73	**4.57**	0.46	**0.63**	**0.69**
CHOIS* w/o F_{all}	6.29	**9.39**	**5.26**	4.77	**0.39**	0.42	0.55
CHOIS*	**5.62**	12.08	5.95	**4.27**	0.41	**0.65**	**0.32**

for evaluation. Each sequence pair is reviewed by 10 different AMT workers. The results are illustrated in Figure 4.

4.4 Ablation Study

We conduct an ablation study to validate the effectiveness of our proposed guidance terms. As shown in Table 3, our hand-object contact guidance and feet-floor contact guidance are both critical. Without the hand-object contact guidance, the contact percentage degrades obviously. Without the feet-floor contact guidance, the height of the feet increases indicating there exists severe foot floating issues. We are not ablating object-floor penetration guidance as object-floor penetration issues are not common and this term is primarily designed for preventing penetration artifacts in qualitative results.

4.5 Application

This section presents a practical application of our method, enabling the synthesis of human-object interactions within 3D scenes, driven by language descriptions. We utilize 3D scenes from the Replica Dataset [46]. The process begins by composing language descriptions that specify the desired interactions, identifying both the objects involved and their intended positions. For example, the language description can be "pull the floor lamp to be close to a shelf". We also define a set of primitive functions used to sample target 3D positions from 3D scenes. This set includes functions like sampling points on an object's surface or near it. GPT-3 [5] is used to extract key information including the interaction object

Fig. 6: Results of interaction synthesis using the same text but different waypoints (a) and using the same waypoints but different text (b). The initial state is in green.

and target objects, and to select the appropriate primitive functions from our predefined function set. Combining the information with the semantic labels of the scene point cloud, we can determine the target 3D positions.

We leverage Habitat [31, 47] to generate collision-free paths within the scene given the start and target object positions. However, as Habitat provides waypoints without corresponding time steps, we need to adapt these to our learned module. We apply heuristics to create waypoints at fixed intervals of 30 frames, which serve as the input conditions for our model. The text input for our learned module excludes the directional component (e.g., "Pick up the floor lamp and move it"), focusing solely on the action and object. An example of this application is shown in Figure 5, demonstrating how our learned interaction synthesis model effectively synthesizes human-object motion following a description in a 3D scene. Table 4 includes a quantitative evaluation of the generated motion. In addition, we showcase the results using the same text input but different waypoints and the results using the same waypoints but different text in Figure 6, demonstrating the effectiveness of the control using object waypoints and text.

5 Conclusion

In conclusion, our work addresses the problem of human-object interaction synthesis conditioned on language descriptions and sparse object waypoints. By employing a conditional diffusion model, we successfully generate object and human motions that are not only synchronized but also resonate with given language descriptions. We incorporate object geometry loss during training which significantly improves the performance of object motion generation. We also propose effective guidance terms used during the sampling process which enhance the realism of the generated results. Moreover, we demonstrate that our learned interaction module can be integrated into a pipeline that synthesizes long-term interactions given language and 3D scenes.

Acknowledgments. This work is in part supported by the Wu Tsai Human Performance Alliance at Stanford University, the Stanford Institute for Human-Centered AI (HAI), NSF CCRI #2120095, ONR MURI N00014-22-1-2740, and Meta. Part of the research was done during Jiaman Li's internship at FAIR, Meta.

References

1. Araujo, J.P., Li, J., Vetrivel, K., Agarwal, R., Gopinath, D., Wu, J., Clegg, A., Liu, C.K.: CIRCLE: Capture in rich contextual environments. In: CVPR (2023)
2. Barquero, G., Escalera, S., Palmero, C.: Belfusion: Latent diffusion for behavior-driven human motion prediction. In: ICCV (2023)
3. Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: BEHAVE: Dataset and method for tracking human object interactions. In: CVPR (2022)
4. Braun, J., Christen, S., Kocabas, M., Aksan, E., Hilliges, O.: Physically plausible full-body hand-object interaction synthesis. In: 3DV (2024)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020)
6. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR (2023)
7. Christen, S., Kocabas, M., Aksan, E., Hwangbo, J., Song, J., Hilliges, O.: D-Grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In: CVPR (2022)
8. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: MoFusion: A framework for denoising-diffusion-based motion synthesis. In: CVPR (2023)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
10. Diller, C., Dai, A.: CG-HOI: Contact-guided 3D human-object interaction generation. In: CVPR (2024)
11. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: CVPR (2023)
12. Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., Tao, D.: 3D-FUTURE: 3D furniture shape with texture. International Journal of Computer Vision **129**, 3313–3337 (2021)
13. Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: IMoS: Intent-driven full-body motion synthesis for human-object interactions. In: Eurographics (2023)
14. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3D human motions from text. In: CVPR (2022)
15. Guo, C., Zuo, X., Wang, S., Cheng, L.: TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: ECCV (2022)
16. Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: Human POSEitioning System (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors. In: CVPR (2021)
17. Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.: Stochastic scene-aware motion prediction. In: ICCV (2021)

18. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: ICCV (2019)
19. Hassan, M., Guo, Y., Wang, T., Black, M., Fidler, S., Peng, X.B.: Synthesizing physical character-scene interactions. In: SIGGRAPH 2023 Conference Papers (2023)
20. He, C., Saito, J., Zachary, J., Rushmeier, H., Zhou, Y.: Nemf: Neural motion fields for kinematic animation. NeurIPS (2022)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
22. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022)
23. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3D scenes. In: CVPR (2023)
24. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: GMD: Controllable human motion synthesis via guided diffusion models. In: ICCV (2023)
25. Kulkarni, N., Rempe, D., Genova, K., Kundu, A., Johnson, J., Fouhey, D., Guibas, L.: Nifty: Neural object interaction fields for guided human motion synthesis. arXiv preprint arXiv:2307.07511 (2023)
26. Lee, J., Joo, H.: Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. arXiv preprint arXiv:2301.02667 (2023)
27. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: CVPR (2023)
28. Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Trans. Graph. **42**(6) (2023)
29. Li, Q., Wang, J., Loy, C.C., Dai, B.: Task-oriented human-object interactions generation with implicit neural representations. arXiv preprint arXiv:2303.13129 (2023)
30. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: ICCV (2019)
31. Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: ICCV (2019)
32. Merel, J., Tunyasuvunakool, S., Ahuja, A., Tassa, Y., Hasenclever, L., Pham, V., Erez, T., Wayne, G., Heess, N.: Catch & carry: reusable neural controllers for vision-guided whole-body tasks. ACM Transactions on Graphics (TOG) **39**(4), 39–1 (2020)
33. Mir, A., Puig, X., Kanazawa, A., Pons-Moll, G.: Generating continual human motion in diverse 3D scenes. In: 3DV (2024)
34. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR (2019)
35. Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., Jiang, H.: HOI-Diff: Text-driven synthesis of 3D human-object interactions using diffusion models. arXiv preprint arXiv:2312.06553 (2023)
36. Petrov, I.A., Marin, R., Chibane, J., Pons-Moll, G.: Object pop-up: Can we infer 3D objects and their poses from human interactions alone? In: CVPR (2023)
37. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: ICCV (2021)
38. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: ECCV (2022)

39. Prokudin, S., Lassner, C., Romero, J.: Efficient learning on point clouds with basis point sets. In: ICCV (2019)
40. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: BABEL: Bodies, action and behavior with english labels. In: CVPR (2021)
41. Raab, S., Leibovitch, I., Tevet, G., Arar, M., Bermano, A.H., Cohen-Or, D.: Single motion diffusion. In: ICLR (2024)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
43. Rempe, D., Luo, Z., Bin Peng, X., Yuan, Y., Kitani, K., Kreis, K., Fidler, S., Litany, O.: Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In: CVPR (2023)
44. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. In: ICLR (2023)
45. Shi, Y., Wang, J., Jiang, X., Dai, B.: Controllable motion diffusion model. arXiv preprint arXiv:2306.00416 (2023)
46. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The Replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
47. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D., Maksymets, O., Gokaslan, A., Vondrus, V., Dharur, S., Meier, F., Galuba, W., Chang, A., Kira, Z., Koltun, V., Malik, J., Savva, M., Batra, D.: Habitat 2.0: Training home assistants to rearrange their habitat. In: NeurIPS (2021)
48. Taheri, O., Choutas, V., Black, M.J., Tzionas, D.: GOAL: Generating 4d whole-body motion for hand-object grasping. In: CVPR (2022)
49. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: ECCV (2020)
50. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: ECCV (2022)
51. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Bermano, A.H., Cohen-Or, D.: Human motion diffusion model. In: ICLR (2023)
52. Tseng, J., Castellon, R., Liu, C.K.: EDGE: Editable dance generation from music. In: CVPR (2023)
53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NIPS) (2017)
54. Wan, W., Yang, L., Liu, L., Zhang, Z., Jia, R., Choi, Y.K., Pan, J., Theobalt, C., Komura, T., Wang, W.: Learn to predict how humans manipulate large-sized objects from interactive motions. IEEE Robotics and Automation Letters **7**(2), 4702–4709 (2022)
55. Wang, J., Xu, H., Xu, J., Liu, S., Wang, X.: Synthesizing long-term 3D human motion and interaction in 3D scenes. In: CVPR (2021)
56. Wang, J., Rong, Y., Liu, J., Yan, S., Lin, D., Dai, B.: Towards diverse and natural scene-aware 3D human motion synthesis. In: CVPR (2022)
57. Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., Huang, S.: HUMANISE: Language-conditioned human motion generation in 3d scenes. In: NeurIPS (2022)
58. Wu, Y., Wang, J., Zhang, Y., Zhang, S., Hilliges, O., Yu, F., Tang, S.: SAGA: Stochastic whole-body grasping with contact. In: ECCV (2022)

59. Xiao, Z., Wang, T., Wang, J., Cao, J., Zhang, W., Dai, B., Lin, D., Pang, J.: Unified human-scene interaction via prompted chain-of-contacts. arXiv preprint arXiv:2309.07918 (2023)
60. Xie, Z., Tseng, J., Starke, S., van de Panne, M., Liu, C.K.: Hierarchical planning and control for box loco-manipulation. Symposium on Computer Animation (SCA) (2023)
61. Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: InterDiff: Generating 3D human-object interactions with physics-informed diffusion. In: ICCV (2023)
62. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: ICCV (2023)
63. Zhang, H., Ye, Y., Shiratori, T., Komura, T.: Manipnet: neural manipulation synthesis with a hand-object spatial representation. ACM Transactions on Graphics (ToG) **40**(4), 1–14 (2021)
64. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022)
65. Zhang, X., Bhatnagar, B.L., Starke, S., Guzov, V., Pons-Moll, G.: COUCH: Towards controllable human-chair interactions. In: ECCV (2022)
66. Zhang, Y., Tang, S.: The wanderings of odysseus in 3D scenes. In: CVPR (2022)
67. Zhang, Z., Liu, R., Aberman, K., Hanocka, R.: Tedi: Temporally-entangled diffusion for long-term motion synthesis. arXiv preprint arXiv:2307.15042 (2023)
68. Zhao, K., Zhang, Y., Wang, S., Beeler, T., Tang, S.: Synthesizing diverse human motions in 3d indoor scenes. arXiv preprint arXiv:2305.12411 (2023)
69. Zheng, J., Zheng, Q., Fang, L., Liu, Y., Yi, L.: Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In: CVPR (2023)
70. Zheng, Y., Yang, Y., Mo, K., Li, J., Yu, T., Liu, Y., Liu, K., Guibas, L.: GIMO: Gaze-informed human motion prediction in context. In: ECCV (2022)
71. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR (2019)