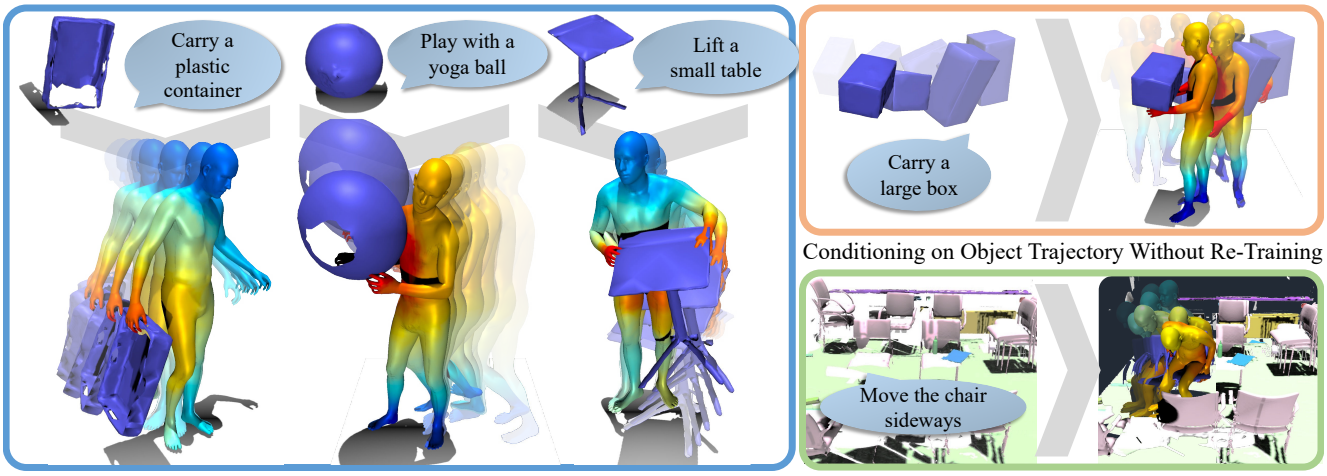
CG-HOI: Contact-Guided 3D Human-Object Interaction Generation

Christian Diller
Technical University of Munich
christian.diller@tum.de

Angela Dai
Technical University of Munich
angela.dai@tum.de



Contact-Guided 3D Human-Object Interaction Synthesis from Text

Conditioning on Object Trajectory Without Re-Training

Application to Objects in Static 3D Scene Scans

Figure 1. We present an approach to generate realistic 3D human-object interactions (HOIs), from a text description and given static object geometry to be interacted with (left). Our main insight is to explicitly model contact (visualized as colors on the body mesh, closer contact in red), in tandem with human and object sequences, in a joint diffusion process. In addition to synthesizing HOIs from text, we can also synthesize human motions conditioned on given object trajectories (top right), and generate interactions in static scene scans (bottom right).

Abstract

We propose CG-HOI, the first method to address the task of generating dynamic 3D human-object interactions (HOIs) from text. We model the motion of both human and object in an interdependent fashion, as semantically rich human motion rarely happens in isolation without any interactions. Our key insight is that explicitly modeling contact between the human body surface and object geometry can be used as strong proxy guidance, both during training and inference. Using this guidance to bridge human and object motion enables generating more realistic and physically plausible interaction sequences, where the human body and corresponding object move in a coherent manner. Our method first learns to model human motion, object motion, and contact in a joint diffusion process, inter-correlated through cross-attention. We then leverage this learned contact for guidance during inference synthesis of realistic, coherent HOIs. Extensive evaluation shows that our joint contact-based human-object interaction approach generates realistic and physically plausible sequences, and we show two applications highlighting the capabilities of our method. Conditioned on a given object trajectory, we can generate the corresponding human motion without re-training, demonstrating strong human-object interdependency learning. Our approach is also flexible, and can be applied to static real-world 3D scene scans.

1. Introduction

Generating human motion sequences in 3D is important for many real-world applications, e.g. efficient realistic character animation, assistive robotic systems, room layout planning, or human behavior simulation. Crucially, human interaction is interdependent with the object(s) being interacted with; the object structure of a chair or ball, for instance, constrains the possible human motions with the object (e.g., sitting, lifting), and the human action often impacts the object motion (e.g., sitting on a swivel chair, carrying a backpack).

Existing works typically focus solely on generating dynamic humans, and thereby disregarding their surroundings

[13, 16, 57, 61, 102, 105], or grounding such motion generations in a static environment that remains unchanged throughout the entire sequence [31, 36, 77, 79, 83, 99, 103, 104, 107]. However, real-world human interactions affect the environment. For instance, even when simply sitting down on a chair, the chair is typically moved: to adjust it to the needs of the interacting human, or to move it away from other objects such as a table. Thus, for realistic modeling of human-object interactions, we must consider the interdependency of object and human motions.

We present CG-HOI, the first approach to address the task of generating realistic 3D human-object interactions from text descriptions, by jointly predicting a sequence of 3D human body motion along with the object motion. Key to our approach is to not only model human and object motion, but also to explicitly model contact as a bridge between human and object. In particular, we model contact by predicting contact distances from the human body surface to the closest point on the surface of the object being interacted with. This explicit modeling of contact helps to encourage human and object motion to be semantically coherent, as well as provide a constraint indicating physical plausibility (e.g., discouraging objects to float without support).

CG-HOI jointly models human, object, and contact together in a denoising diffusion process. Our joint diffusion model is designed to encourage information exchange between all three modalities through cross-attention blocks. Additionally, we employ a contact weighting scheme, based on the insight that object motion, when being manipulated by a human, is most defined by the motion of the body part in closest contact (Fig. 3). We make use of this by generating separate object motion hypotheses for multiple parts of the human body and aggregating them based on that part's predicted contact. During inference, we leverage the predicted contact distances to refine synthesized sequences through our contact-based diffusion guidance, which penalizes synthesizing sequences with human-object contact far from the predicted contact distances.

Our method is able to generate realistic and physically plausible human-object interactions, and we evaluate our approach on two widely-used interaction datasets, BEHAVE [8] and CHAIRS [35]. In addition, we also demonstrate the usefulness of our model with two related applications: First, generating human motion given a specific object trajectory without any retraining, which demonstrates our learned human-object motion interdependencies. Second, populating a static 3D scene scan with human-object interactions of segmented object instances, showing the applicability of our method to general real-world 3D scans.

In summary, our contributions are three-fold:

- We propose an approach to generate realistic, diverse, and physically plausible human-object interaction sequences by jointly modeling human motion, object motion, and contact through cross-attention in a diffusion process.
- We formulate a holistic contact representation: Object motion hypotheses are generated for multiple pre-defined points on the surface of the human body and aggregated based on predicted contact distances, enabling comprehensive body influence on contact while focusing on the body parts in closer contact to the object.
- We propose a contact-based guidance during synthesis of human-object interactions, leveraging predicted contacts to refine generated interactions, leading to more physically plausible results.

2. Related Work

3D Human Motion Generation. Generating sequences of 3D humans in motion is a task which evolved noticeably over the last few years. Traditionally, many methods used recurrent approaches [2, 14, 20, 22, 32, 50] and, improving both fidelity and predicted sequence length, graph- and attention-based frameworks [45, 46, 68]. Notably, generation can either happen deterministically, predicting one likely future human pose sequence [18, 20, 45, 46, 50], or stochastically, thereby also modelling the uncertainty inherent to future human motion [4, 7, 9, 17, 47, 85, 86, 91].

Recently, denoising diffusion models [64, 65] showed impressive results in 2D image generation, producing high fidelity and diverse images [30, 65]. Diffusion models allow for guidance during inference, with classifier-free guidance [29, 52] widely used to trade off between generation quality and diversity. Inspired by these advances, various methods have been proposed to model 3D human motion through diffusion, using U-Nets [13, 16, 57, 61, 102, 105], transformers [1, 57, 63, 66, 71, 72, 78, 80, 81, 87, 88, 94], or custom architectures [3, 6, 12, 15, 95]. Custom diffusion guidance has also been shown to aid controllability [33, 38, 60] and physical plausibility [92].

In addition to unconditional motion generation, conditioning on text descriptions allows for more control over the generation result [61, 71, 78, 81, 94, 105]. In fact, generating plausible and corresponding motion from textual descriptions has been an area of interest well before the popularity of diffusion models [5, 13, 25, 37, 39, 55, 93].

These methods show strong potential for 3D human motion generation, but focus on a skeleton representation of the human body, and only consider human motion in isolation, without naturally occurring interactions. To generate realistic human-object interactions, we must consider the surface of the human body and its motion with respect to object motion, which we characterize as contact.

3D Human Motion in Scenes. As human motion typically occurs not in isolation but in the context of an object or surrounding environment, various methods have explored learning plausible placement of humans into scenes,

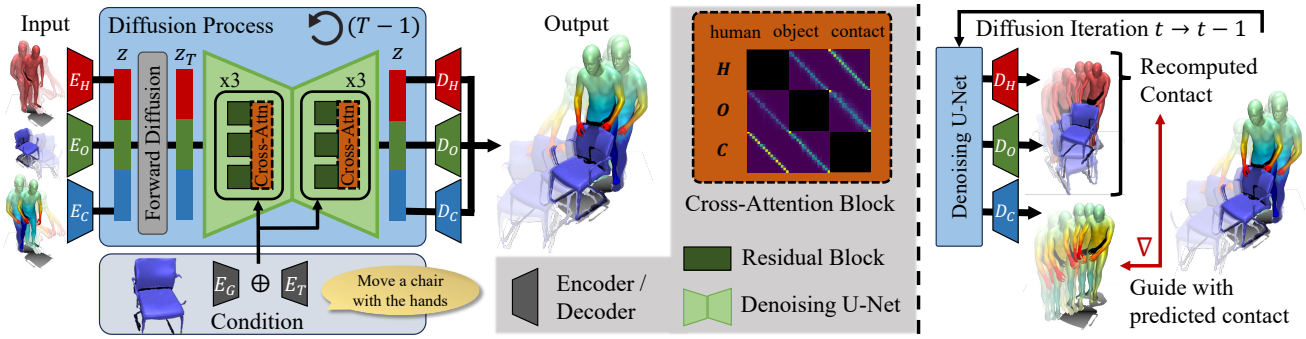


Figure 2. Method Overview. Given a brief text description an an object geometry, CG-HOI produces a human-object interaction sequence where both human and object motion are modeled. To produce realistic human-object interactions, we additionally model contact to bridge the interdependent motions. Our method jointly generates all three during training (left), using a U-Net-based diffusion with cross-attention across human, object, and contact. During inference (right), we drive synthesis under guidance of estimated contact to sample more physically plausible interactions.

both physically and semantically, [26, 28, 96, 100], forecasting future motion given context [11, 48], or generating plausible walking and sitting animations [27, 31, 36, 75–77, 79, 83, 99, 103, 104, 107]. This enables more natural modeling of human reactions to their environment; however, the generated interactions remain limited due to the assumption of a static scene environment, resulting in a focus on walking or sitting movements.

Recent methods have also focused on more fine-trained interactions by generating human motion given a single static object [41, 42, 67, 69, 82, 97, 98]. While these methods only focus on human motion generation for a static object, InterDiff [84] jointly forecasts both human and object motion sequences given an initial sequence observation. Our approach also models both human and object motion, but we formulate a flexible text-conditioned generative model for dynamic human and object motion, modeling the interdependency between human, object, and contact to synthesize more realistic interactions under various application settings.

Contact Prediction for Human-Object Interactions. While there is a large corpus of related work for human motion prediction, only few works focus on object motion generation [19, 51, 59, 110]. Notably, these methods predict object movement in isolation, making interactions limited, as they typically involve interdependency with human motion.

Contact prediction has been most studied in recent years for the task of fine-grained hand-object interaction [10, 21, 40, 43, 89, 106, 108]. It is defined either as binary labels on the surface [10, 21, 40, 43, 89, 106] or as the signed distance to a corresponding geometry point [108]. In these works, predicting object and hand states without correct contact leads to noticeable artifacts. Contact prediction itself has also been the focus of several works [23, 34, 73, 82], either predicting contact areas or optimizing for them.

Applied to the task of generating whole-body human-object interactions, this requires access to the full surface geometry of both object and human. Only few recent motion generation works focus on generating full-body geometric representations of humans [49, 54, 55, 70, 85, 99, 101] instead of simplified skeletons which is a first step towards physically correct interaction generation. However, while several of these works acknowledge that contact modeling would be essential for more plausible interactions [54, 55, 99], they do not model full-body contact.

We approach the task of generating plausible human-object motion from only the object geometry and a textual description as a joint task and show that considering the joint behavior of full-body human, object, and contact between the two benefits output synthesis to generate realistic human-object interaction sequences.

3. Method Overview

CG-HOI jointly generates sequences of human body and object representations, alongside contact on the human body surface. Reasoning jointly about all three modalities in both training and inference enables generation of semantically meaningful human-object interaction sequences.

Fig. 2 shows a high-level overview of our approach: We consider as condition a brief text description T of the action to be performed, along with the static geometry G of the object to be interacted with, and generate a sequence of F frames $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_F]$ where each frame $\mathbf{x}_i$ consists of representations for the object transformation o_i, for the human body surface h_i, and for the contact c_i between human and object geometry. We denote as $H = \{h_i\}$ the human body representations, $O = \{o_i\}$ the object transformations, and $C = \{c_i\}$ the contact representations.

We first train a denoising diffusion process to generate H, O, and C, using a U-Net architecture with per-modality residual blocks and cross-attention modules. Using cross-attention between human, object motion, and contact allows for effectively learning interdependencies and and feature sharing (Sec. 4). We use the generated contact to guide

both training and inference: Instead of predicting one object motion hypothesis per sequence, we generate multiple, and aggregate them based on predicted contacts, such that body parts in closer contact with the object have a stronger correlation with the final object motion (Sec. 4.3). During inference, the trained model generates H, O, and C. For each step of the diffusion inference, we use predicted contact C to guide the generation of H and O, by encouraging closeness of recomputed contact and predicted contact, producing more refined and realistic interactions overall (Sec. 5).

4. Human-Object Interaction Diffusion

4.1. Probabilistic Denoising Diffusion

Our approach uses a diffusion process to jointly generate a sequence of human poses, object transformations, and contact distances in a motion sequence from isotropic Gaussian noise in an iterative process, removing more noise at each step. More specifically, during training we add noise depending on the time step ("forward process") and train a neural network to reverse this process, by directly predicting the clean sample from noisy input. Mathematically, the forward process follows a Markov chain with T steps, yielding a series of time-dependent distributions $q(\mathbf{z}_t|\mathbf{z_{t-1}})$ with noise being injected at each time step until the final distribution $\mathbf{z}_T$ is close to $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Formally,

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\sqrt{\beta_t}\mathbf{z}_{t-1} + (1-\beta_t)\mathbf{I}) \tag{1}$$

with the variance of the Gaussian noise at time t denoted as β_t, and $\beta_0 = 0$.

Since we adopt the Denoising Diffusion Probabilistic Model [30], we can sample $\mathbf{z}_t$ directly from $\mathbf{z}_0$ as

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\epsilon \tag{2}$$

with $\alpha_t = \prod_{t'=0}^{t}(1-\beta_t)$, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

For the reverse process, we follow [57, 71, 84], directly recovering the original signal $\tilde{\mathbf{z}}$ instead of the added noise.

Human-Object Interactions To model human-object interactions with diffusion, we employ our neural network formulation $\mathcal{G}$. $\mathcal{G}$ operates on the noised vector of concatenated human, object, and contact representations, together with the current time step t, and a condition consisting of object point cloud G, encoded by an encoder E_G, and text information T, encoded by encoder E_T. Formally,

$$\tilde{\mathbf{z}} = \mathcal{G}(\mathbf{z}_t, t, E_G(G) \oplus E_T(T)) \tag{3}$$

More specifically, in our scenario E_T extracts text features with a pre-trained CLIP [58] encoder. Encoder E_G processes object geometry G as a uniformly sampled point cloud in world coordinate space with a PointNet [56] pre-trained on object parts segmentation.

Object transformations o_i are represented as global translation and rotation using continuous 6D rotation representation [109]. In contrast to prior work [17, 41, 71, 78, 91, 94, 97] which focused on representing human motion in a simplified manner as a collection of J human joints, disregarding both identity-specific and pose-specific body shape, we model physically plausible human-object contacts between body surface and geometry. Thus, we represent the human body h_i in SMPL [44] parameters: $h_i = \{h_i^p, h_i^b, h_i^r, h_i^t\}$ with pose parameters $h_i^p \in \mathbb{R}^{63}$, shape parameters $h_i^b \in \mathbb{R}^{10}$, as well as global rotation $h_i^r \in \mathbb{R}^3$ and translation $h_i^t \in \mathbb{R}^3$. These body parameters can then be converted back into a valid human body surface mesh in a differentiable manner, using the SMPL [44] model. This allows us to reason about the contact between human body surface and object geometry. We represent contact c_i on the human body as the distance between a set of $M = 128$ uniformly distributed motion markers on the body surface to the closest point of the object geometry, for each marker. Specifically, we represent contact for frame x_i and j-th contact marker ($j \in \{0..M\}$) c_i^j as its distance from the human body surface to the closest point on the same frame's object surface.

4.2. Human-Object-Contact Cross-Attention

We jointly predict human body sequences $\{h_i\}$, object transformations $\{o_i\}$, and corresponding contact distances $\{c_i\}$ in our diffusion approach. We employ a U-Net backbone for diffusion across these outputs, with separate residual blocks for human, object, and contact representations, building modality-specific latent feature representations. As we aim to model the inter-dependency across human, object, and contact, we introduce custom human-object-contact cross-attention modules after every residual block where each modality attends to the other two.

We follow the formulation of Scaled Dot-Product Attention [74], computing the updated latent human body feature:

$$h_i = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{D}}\right)\boldsymbol{V}, \tag{4}$$

with query $\boldsymbol{Q} = h_i$, and key and value $\boldsymbol{K} = \boldsymbol{V} = o_i \odot c_i$ ($\odot$ denotes concatenation). Applying this similarly to o_i and c_i yields the final features after each cross-attention module.

4.3. Contact-Based Object Transform Weighting

As visualized in Fig. 3, object motion is naturally most influenced by parts of the human body in very close contact to the object (as they are often the cause of the motion), and less impacted (if at all) by body parts further away. For instance, if a person moves an object with their hands, the object follows the hands but not necessarily other body parts (e.g., body and feet may remain static or walk in a different direction). Thus, instead of directly generating

Figure 3. An object's trajectory is largely defined by the motion of the region of the body in close contact with the object, e.g. the hand(s) when carrying an object (left, middle) or the lower body when moving with an object while sitting (right). This informs our contact-based approach to generating object motion.

one object motion hypothesis o_i alongside the corresponding human motion h_i, we couple o_i to the M body contact points $j \in \{1..M\}$ and their predicted distances $\{c_i^j\}$ between human body surface and object geometry.

Formally, we predict object transformation hypotheses o_i^j for each contact point on the human body, and weigh them with the inverse of their predicted contact distance c_i^j:

$$o_i = \frac{1}{\sum_j c_i} \sum_{j=0}^{N} (\max(|c_i|) - |c_i^j|) o_i^j, \tag{5}$$

4.4. Loss Formulation

During training, the input is a noised vector $\mathbf{z}$, containing F frames $\{\mathbf{x}_i\}$, each a concatenation of human body representation h_i, object transformation o_i, and contact parameters c_i. As condition $\mathbf{C}$, we use additionally input encoded object geometry G and text description T. The training process is then supervised with the ground-truth sequence containing $\hat{h_i}, \hat{o_i}, \hat{c_i}$, minimizing a common objective:

$$\mathbf{L} = \lambda_h ||h_i - \hat{h_i}||_1 + \lambda_o ||o_i - \hat{o_i}||_1 + \lambda_c ||c_i - \hat{c_i}||_2, \tag{6}$$

with $\lambda_h = 1.0, \lambda_o = 0.9, \lambda_c = 0.9$. We use classifier-free guidance [29] for improved fidelity during inference, thus masking out the conditioning signal with 10% probability.

5. Interaction Generation

Using our trained network model, we can generate novel human-object interaction sequences for a given object geometry and a short text description using our weighting scheme for generating object transformations, and a custom guidance function on top of classifier-free guidance to generate physically plausible sequences.

Specifically, we use our trained model to reverse the forward diffusion process of Eq. 2: Starting with noised sample $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we iteratively use our trained network model $\mathcal{G}$ to estimate cleaned sample $\mathbf{z}_0$:

$$\mathbf{z}_{t-1} = \sqrt{\alpha_t}\tilde{\mathbf{z}} + \sqrt{1-\alpha_t}\epsilon. \tag{7}$$

5.1. Contact-Based Diffusion Guidance

While our joint human-object-contact training already leads to some plausible motions, generated sequences are not explicitly constrained to respect contact estimates during inference, which can lead to inconsistent contact between human and object motion (e.g., floating objects). Thus, we introduce a contact-based guidance during inference to refine predictions, using a cost function $G((x)_t)$ which takes as input the denoised human, object, and contact predictions $\mathbf{z}_t = [h_t, o_t, c_t]$ at diffusion step t. Based on the difference between predicted and actual contact distances for each contact point, we then calculate the gradient $\nabla_{\mathbf{z}_t} G(\mathbf{z}_t)$.

We use this gradient for diffusion guidance, following [38], by re-calculating the mean prediction μ_t at each time t:

$$\hat{\mu}_t = \mu_t + s \sum_t \nabla_{x_t} G(x_t), \tag{8}$$

for a scaling factor s. This guidance is indirect but dense in time, and is able to correct physical contact inconsistencies in the predicted sequences during inference time, without requiring any explicit post-processing steps.

5.2. Conditioning on Object Trajectory

While our model has been trained with text and static object geometry as condition, we can also apply the same trained model for conditional generation of a human sequence given an object sequence and text description. Note that this does not require any re-training, as our model has learned a strong correlation between human and object motion. Instead, we use a replacement-based approach, and inject the given object motion O' into the diffusion process during inference at every step. Following Eq. 7, we obtain:

$$\mathbf{z}_{t-1} = \sqrt{\alpha_t}\tilde{\mathbf{z}'_t} + \sqrt{1-\alpha_t}\epsilon, \tag{9}$$

with $\tilde{\mathbf{z}'} = [h_t, o'_t, c_t]$, concatenating human motion h_t, contact distances c_t, and injected given object motion o'_t.

6. Results

We evaluate our approach using two commonly used human-object interaction datasets BEHAVE [8] and CHAIRS [35] on a range of metrics, measuring motion fidelity and diversity. We show that our approach is able to generate realistic and diverse motion on both datasets, across a variety of objects and types of interactions.

6.1. Experimental Setup

Datasets We conduct our experiments on two datasets containing interactions between whole-body 3D humans and corresponding objects. CHAIRS [35] captures 46 subjects as their SMPL-X [53] bodies interacting with 81 different types of chairs and sofas. We extract sequences in which

		BEHAVE				CHAIRS			
Task	Approach	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →
	Real (human)	0.73	0.09	4.23	4.55	0.83	0.01	7.34	3.00
Text-Cond. Human Only	MDM [71]	0.52	4.54	5.44	5.12	0.72	5.99	6.83	3.45
	InterDiff [84]	0.49	5.36	3.98	3.98	0.63	6.76	5.24	2.44
	Ours	**0.60**	**4.26**	**4.92**	**4.10**	**0.78**	**5.24**	**7.90**	**3.22**
	Real	0.81	0.17	6.80	6.24	0.87	0.02	9.91	6.12
Motion-Cond. HOI	InterDiff [84]	0.68	3.86	5.62	5.90	0.67	4.83	7.49	4.87
	Ours	**0.71**	**3.52**	**6.89**	**6.43**	**0.79**	**4.01**	**8.42**	**6.29**
Text-Cond. HOI	MDM [71]	0.49	9.21	6.51	8.19	0.53	9.23	6.23	7.44
	InterDiff [84]	0.53	8.70	3.85	4.23	0.69	7.53	5.23	4.63
	Ours	**0.62**	**6.31**	**6.63**	**5.47**	**0.74**	**6.45**	**8.91**	**5.94**

Table 1. Quantitative comparison with state-of-the-art approaches, MDM [71] and InterDiff [84]. Human Only results are evaluated only on the human pose sequence, and motion-cond. denotes predictions additionally conditioned on past observations of both human and object behavior. For metrics with →, results closer to the real distribution are better. Our approach outperforms these baselines in all three settings, indicating a strong learned correlation between human and object motion.

both human and object are in motion, yielding ≈ 1300 HOI sequences, each labeled with a text description. We use a random 80/10/10 split along object classes, ensuring that test objects are not seen during training. BEHAVE [8] captures 8 participants as their SMPL-H [62] parameters alongside 20 different objects. This yields ≈ 520 sequences with corresponding text descriptions. We use their original train/test split. We sample both datasets at 20 frames per second, and generate 32 frames for CHAIRS and 64 for BEHAVE, leading to generated motion that lasts up to 3 seconds.

Implementation Details We train our model with batch size 64 for 600k steps (≈24 hours), after which we choose the checkpoint that minimized validation FID, following [84]. Our attention uses 4 heads and a latent dimension of 256. Input text is encoded using a frozen CLIP-ViT-B/32 model. For classifier-free guidance during inference time, we use a guidance scale of 2.5, which empirically provides a good trade-off between diversity and fidelity. For our inference-time contact-based guidance, we use scale $s = 100.0$.

6.2. Evaluation Metrics

We measure realism and diversity of combined human and object motion, alongside closeness to the text description, following established practices [24, 25, 71]. We first train a joint human-object motion feature extractor and separate text feature extractor using a contrastive loss to produce geometrically close feature vectors for matched text-motion pairs, and vice versa. These encoders are then used for the following metrics:

R-Precision measures the closeness of the text condition and generated HOI in latent feature space, and reports whether the correct match falls in the top 3 closest feature vectors.

Frechet Inception Distance (FID) is commonly used to evaluate the similarity between generated and ground-truth distribution in encoded feature space.

Diversity and MultiModality. Diversity measures the motion variance across all text descriptions and is defined as $\frac{1}{N}\sum_{i=1}^{N}||v_i - v_i'||_2$ between two randomly drawn subsets $\{v_i\}$ and $\{v_i'\}$. MultiModality (MModality) measures the average such variance intra-class, for each text description.

Perceptual User Study. The exact perceptual quality of human-object interactions is difficult to capture with any single metric; thus, we conducted a user study with 32 participants to evaluate our method in comparison to baseline approaches. Participants are shown side-by side views of sequences with the same geometry and text conditioning, and asked to choose 1) Which one follows the given text better and 2) Which one looks more realistic overall.

6.3. Comparison to Baselines

As our method is the first to enable generationg human and object motion from text, there are no baselines available for direct comparison. InterDiff [84] is closest to our approach, performing forecasting from observed human and object motion as input and predicting a plausible continuation. In Tab. 1, we compare to ours first in their setting, using observed motion as condition (motion-cond.), for a fair comparison. Additionally, we modify their approach by replacing observed motion encoders with our text encoder, allowing for a comparison in our setting (text-cond.). We also compare with MDM [71], a state-of-the-art method for human-only sequence generation from text, both in their original setting, only predicting human sequences, and extending theirs to also generate object sequences, by adding additional tokens and geometry conditioning to their transformer encoder formulation. For more details of baseline setup, we refer to the appendix. We evaluate the quality of generated human-object interactions as well as human-only generation, only evaluating the human sequence for our method, as compared to the generated sequences of MDM.

Both Tab. 1 and the user study in Tab. 5 show that our approach is able to generate more realistic and physically plausible human-object interaction sequences than baselines.

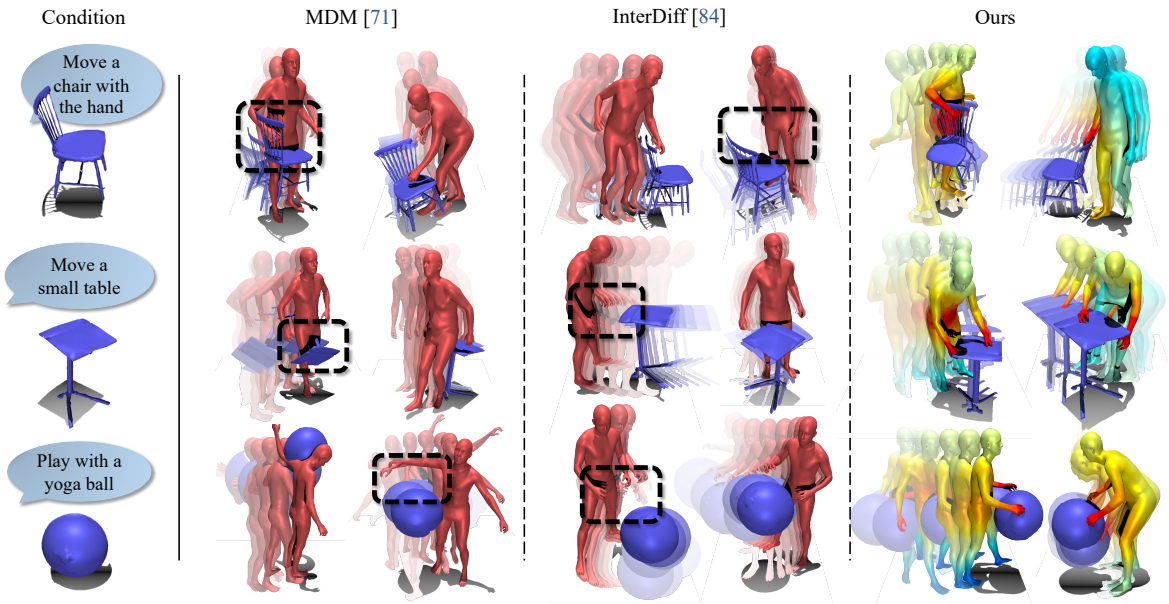


Figure 4. Qualitative comparison to state-of-the-art methods MDM [71] and InterDiff [84]. Our approach generates high-quality HOIs by jointly modeling contact (closer contact in red), reducing penetration and floating artifacts (black highlight boxes).

In Fig. 4, we see that our approach synthesizes more meaningful human-object interaction with respect to contact and mitigating independent object floating.

6.4. Ablation Studies

Cross-attention enables learning human-object interdependencies. Tab. 2 shows that our human-object-contact cross-attention (Sec. 4.2) significantly improves performance by effectively sharing information between human, contact, and object sequence modalities. In Fig. 6, we see this encourages realistic contact between human and object.

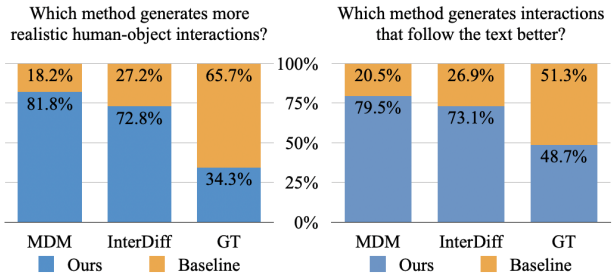


Figure 5. Perceptual User Study. Participants significantly favor our method over baselines, for overall realism and text coherence.

Contact prediction improves HOI generation performance. Predicting contact (Sec. 5) is crucial to generating more realistic human-object sequences, resulting in more realistic interactions between human and object (Fig. 6), and improved fidelity (Tab. 2). Notably, learning contact jointly with human and object motion improves overall quality, compared to a separately trained contact model used for inference guidance ("Separate contact pred.", Tab. 2).

Contact-based object transformation weighting improves generation performance. Weighting predicted object motion hypotheses with predicted contact (Sec. 4.3) improves HOI generation over naive object sequence prediction, both

	BEHAVE				CHAIRS			
Approach	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →
Real	0.81	0.17	6.80	6.24	0.87	0.02	9.91	6.12
No cross-attention	0.35	10.44	8.23	7.40	0.49	10.84	12.22	10.64
No contact prediction	0.41	9.64	10.10	6.89	0.41	8.53	11.56	9.15
Separate contact pred.	0.47	8.01	5.12	5.12	0.52	9.34	7.65	4.62
No contact weighting	0.55	8.54	6.52	5.29	0.64	7.55	8.56	5.45
No contact guidance	0.59	7.22	7.84	5.30	0.70	7.41	8.05	5.76
Ours	**0.62**	**6.31**	**6.63**	**5.47**	**0.74**	**6.74**	**8.91**	**5.94**

Table 2. Ablation on our design choices. Joint contact prediction with cross-attention encourages the generation of more natural HOIs, and our weighting scheme and inference-time contact guidance together enable the best generation performance.

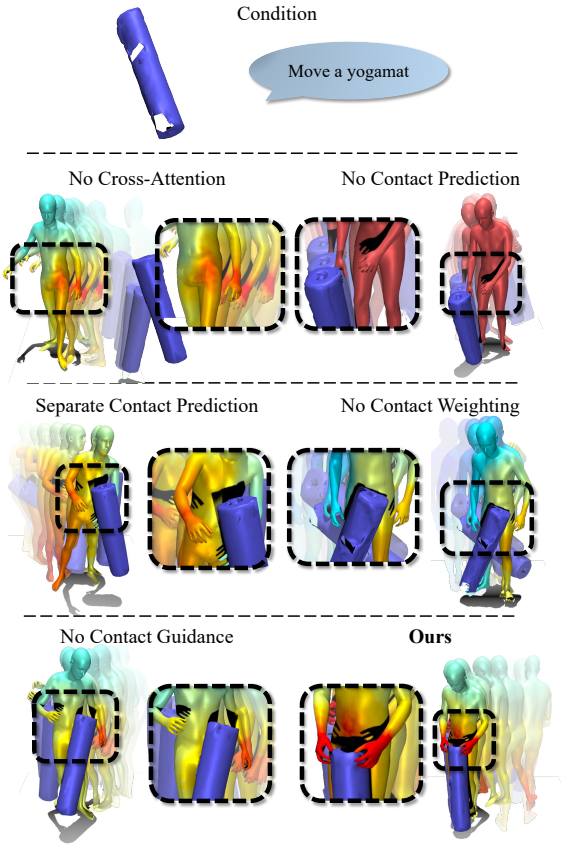


Figure 6. Visualization of ablation of our method design: Generation, weighting, and inference-time guidance work together to enable realistic interactions in our method, resolving artifacts such as object floating.

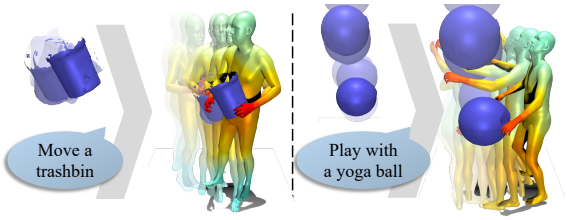


Figure 7. Given an object trajectory at inference time, our method can generate corresponding human motion without re-training.

quantitatively in Tab. 2 ("No contact weighting") and visually as realistic human-object interactions in Fig. 6.

Contact-based guidance during inference helps produce physically plausible interactions. As visualized in Fig. 6 and evaluated in Tab. 2, using custom guidance based on predicted contacts leads to a higher degree of fidelity and physical plausibility.

6.5. Applications

Human motion generation given object trajectory. Our approach can also be directly applied to conditionally generate human sequences given object sequences as condition, as shown in Fig. 7. As our model learns a strong correspondence between object and human motion, facilitated by contact distance predictions, we are able to condition without any additional training.

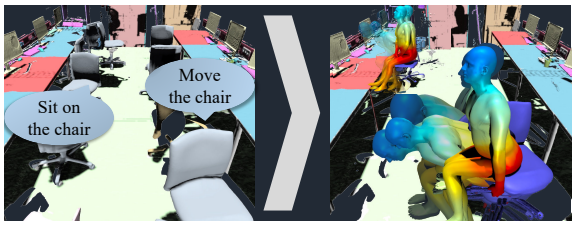


Figure 8. Application to static scene scans. Our method can generate HOIs from segmented objects in such environments.

Populating 3D scans. Fig. 8 shows that we can also apply our method to generate human-object interactions in static scene scans. Here, we use a scene from the ScanNet++ dataset [90], with their existing semantic object segmentation. This enables potential to generate realistic human motion sequences only given a static scene environment.

6.6. Limitations

While we have demonstrated the usefulness of joint contact prediction in 3D HOI generation, several limitations remain. For instance, our method focuses on realistic interactions with a single object. We show that this can be applied to objects in static 3D scans; however, we do not model multiple objects together, which could have the potential to model more complex long-term human behavior. Additionally, our method requires expensive 3D HOI captures for training; a weakly supervised approach leveraging further supervision from 2D action data might be able to represent more diverse scenarios.

7. Conclusion

We propose an approach to generating realistic, dynamic human-object interactions based on contact modeling. Our diffusion model effectively learns interdependencies between human, object, and contact through cross-attention along with our contact-based object transformation weighting. Our predicted contacts further facilitate refinement using custom diffusion guidance, generating diverse, realistic interactions based on text descriptions. Since our model learns a strong correlation between human and object sequences, we can use it to conditionally generate human motion sequences from given object sequences in a zero-shot manner. Extensive experimental evaluation confirms both fidelity and diversity of our generated sequences and shows improved performance compared to related state-of-the-art baselines.

Acknowledgements

This project is funded by the Bavarian State Ministry of Science and the Arts and coordinated by the Bavarian Research Institute for Digital Transformation (bidt), and the German Research Foundation (DFG) Grant "Learning How to Interact with Scenes through Part-Based Understanding".

References

[1] Hyemin Ahn, Esteve Valls Mascaro, and Dongheui Lee. Can we use diffusion probabilistic models for 3d motion prediction? pages 9837–9843, 2023. 2

[2] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7143–7152. IEEE, 2019. 2

[3] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph.*, 42(4):44:1–44:20, 2023. 2

[4] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5222–5231. Computer Vision Foundation / IEEE, 2020. 2

[5] Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. Make-an-animation: Large-scale text-conditional 3d human motion generation. *CoRR*, abs/2305.09662, 2023. 2

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

[7] Emad Barsoum, John R. Kender, and Zicheng Liu. HP-GAN: probabilistic 3d human motion prediction via GAN. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1418–1427. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[8] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: dataset and method for tracking human object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15914–15925. IEEE, 2022. 2, 5, 6, 16, 17

[9] Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. Accurate and diverse sampling of sequences based on a "best of many" sample objective. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8485–8493. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[10] Jona Braun, Sammy Joe Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. *CoRR*, abs/2309.07907, 2023. 3

[11] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. 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*, pages 387–404. Springer, 2020. 3

[12] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. *CoRR*, abs/2302.03665, 2023. 2

[13] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18000–18010. IEEE, 2023. 2

[14] Hsu-Kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1423–1432. IEEE, 2019. 2

[15] Jeongjun Choi, Dongseok Shim, and H. Jin Kim. Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. *CoRR*, abs/2212.02796, 2022. 2

[16] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 9760–9770. IEEE, 2023. 2

[17] Christian Diller, Thomas A. Funkhouser, and Angela Dai. Forecasting characteristic 3d poses of human actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15893–15902. IEEE, 2022. 2, 4

[18] Christian Diller, Thomas A. Funkhouser, and Angela Dai. Forecasting actions and characteristic 3d poses. *CoRR*, abs/2211.14309, 2022. 2

[19] Danny Driess, Zhiao Huang, Yunzhu Li, Russ Tedrake, and Marc Toussaint. Learning multi-object dynamics with compositional neural radiance fields. In *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, pages 1755–1768. PMLR, 2022. 3

[20] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4346–4354. IEEE Computer Society, 2015. 2

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

[22] Anand Gopalakrishnan, Ankur Arjun Mali, Dan Kifer, C. Lee Giles, and Alexander G. Ororbia II. A neural temporal model for human motion prediction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12116–12125. Computer Vision Foundation / IEEE, 2019. 2

[23] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. Contactopt: Optimizing contact to improve grasps. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1471–1481. Computer Vision Foundation / IEEE, 2021. 3

[24] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 2021–2029. ACM, 2020. 6, 14

[25] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5142–5151. IEEE, 2022. 2, 6, 14

[26] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2282–2292. IEEE, 2019. 3

[27] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J. Black. Stochastic scene-aware motion prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11354–11364. IEEE, 2021. 3

[28] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3d scenes by learning human-scene interaction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14708–14718. Computer Vision Foundation / IEEE, 2021. 3

[29] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. 2, 5

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020. 2, 4

[31] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 16750–16761. IEEE, 2023. 2, 3

[32] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5308–5317. IEEE Computer Society, 2016. 2

[33] Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. 162:9902–9915, 2022. 2

[34] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11087–11096. IEEE, 2021. 3

[35] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. CHAIRS: towards full-body articulated human-object interaction. *CoRR*, abs/2212.10621, 2022. 2, 5, 16

[36] James F. Mullen Jr., Divya Kothandaraman, Aniket Bera, and Dinesh Manocha. Placing human animations into 3d scenes by learning interaction- and geometry-driven keyframes. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 300–310. IEEE, 2023. 2, 3

[37] Sai Shashank Kalakonda, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. Action-gpt: Leveraging large-scale language models for improved and generalized zero shot action generation. *CoRR*, abs/2211.15603, 2022. 2

[38] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. GMD: controllable human motion synthesis via guided diffusion models. *CoRR*, abs/2305.12577, 2023. 2, 5

[39] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. FLAME: free-form language-based motion synthesis & editing. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 8255–8263. AAAI Press, 2023. 2

[40] Paul G. Kry and Dinesh K. Pai. Interaction capture and synthesis. *ACM Trans. Graph.*, 25(3):872–880, 2006. 3

[41] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas J. Guibas. NIFTY: neural object interaction fields for guided human motion synthesis. *CoRR*, abs/2307.07511, 2023. 3, 4

[42] Jiye Lee and Hanbyul Joo. Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. *CoRR*, abs/2301.02667, 2023. 3

[43] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. *CoRR*, abs/2303.13129, 2023. 3

[44] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. pages 248:1–248:16, 2015. 4

[45] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9488–9496. IEEE, 2019. 2

[46] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention.

In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, pages 474–489. Springer, 2020. 2

[47] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 13289–13298. IEEE, 2021. 2

[48] Wei Mao, Miaomiao Liu, Richard I. Hartley, and Mathieu Salzmann. Contact-aware human motion forecasting. 2022. 3

[49] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Weakly-supervised action transition learning for stochastic human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8141–8150. IEEE, 2022. 3

[50] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4674–4683. IEEE Computer Society, 2017. 2

[51] Damian Mrowca, Chengxu Zhuang, Elias Wang, Nick Haber, Li Fei-Fei, Josh Tenenbaum, and Daniel L. K. Yamins. Flexible neural representation for physics prediction. pages 8813–8824, 2018. 3

[52] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. 162:16784–16804, 2022. 2

[53] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. 5, 16

[54] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer VAE. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10965–10975. IEEE, 2021. 3

[55] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: generating diverse human motions from textual descriptions. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXII*, pages 480–497. Springer, 2022. 2, 3

[56] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 77–85. IEEE Computer Society, 2017. 4, 17

[57] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H. Bermano, and Daniel Cohen-Or. Single motion diffusion. *CoRR*, abs/2302.05905, 2023. 2, 4

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 4

[59] Davis Rempe, Srinath Sridhar, He Wang, and Leonidas J. Guibas. Predicting the physical dynamics of unseen 3d objects. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 2823–2832. IEEE, 2020. 3

[60] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 13756–13766. IEEE, 2023. 2

[61] Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE, 2023. 2

[62] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6):245:1–245:17, 2017. 6, 17

[63] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior. *CoRR*, abs/2303.01418, 2023. 2

[64] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2256–2265. JMLR.org, 2015. 2

[65] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. 2021. 2

[66] Jiarui Sun and Girish Chowdhary. Towards globally consistent stochastic human motion prediction via motion diffusion. *CoRR*, abs/2305.12554, 2023. 2

[67] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: generating 4d whole-body motion for hand-object grasping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13253–13263. IEEE, 2022. 3

[68] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamics. pages 935–941, 2018. 2

[69] Purva Tendulkar, Dídac Surís, and Carl Vondrick. FLEX: full-body grasping without full-body grasps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 21179–21189. IEEE, 2023. 3

[70] Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to CLIP space. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022. 3

[71] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 4, 6, 7, 14

[72] Sibo Tian, Minghui Zheng, and Xiao Liang. Transfusion: A practical and effective transformer-based diffusion model for 3d human motion prediction. *CoRR*, abs/2307.16106, 2023. 2

[73] Tze Ho Elden Tse, Zhongqun Zhang, Kwang In Kim, Ales Leonardis, Feng Zheng, and Hyung Jin Chang. S^2contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I*, pages 568–584. Springer, 2022. 3

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 4

[75] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9401–9411. Computer Vision Foundation / IEEE, 2021. 3

[76] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12206–12215. Computer Vision Foundation / IEEE, 2021.

[77] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20428–20437. IEEE, 2022. 2, 3

[78] Yin Wang, Zhiying Leng, Frederick W. B. Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. pages 22035–22044, 2023. 2, 4

[79] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: language-conditioned human motion generation in 3d scenes. 2022. 2, 3

[80] Dong Wei, Huaijiang Sun, Bin Li, Jianfeng Lu, Weiqing Li, Xiaoning Sun, and Shengxiang Hu. Human joint kinematics diffusion-refinement for stochastic motion prediction. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 6110–6118. AAAI Press, 2023. 2

[81] Dong Wei, Xiaoning Sun, Huaijiang Sun, Bin Li, Shengxiang Hu, Weiqing Li, and Jianfeng Lu. Understanding text-driven motion synthesis with keyframe collaboration via diffusion models. *CoRR*, abs/2305.13773, 2023. 2

[82] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. SAGA: stochastic whole-body grasping with contact. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, pages 257–274. Springer, 2022. 3

[83] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. *CoRR*, abs/2309.07918, 2023. 2, 3

[84] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14928–14940, 2023. 3, 4, 6, 7, 14

[85] Sirui Xu, Yu-Xiong Wang, and Liangyan Gui. Stochastic multi-person 3d motion forecasting. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 3

[86] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. MT-VAE: learning motion transformations to generate multimodal human dynamics. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, pages 276–293. Springer, 2018. 2

[87] Siqi Yang, Zejun Yang, and Zhisheng Wang. Longdanced-iff: Long-term dance generation with conditional diffusion model. *CoRR*, abs/2308.11945, 2023. 2

[88] Zhao Yang, Bing Su, and Ji-Rong Wen. Synthesizing long-term human motions with diffusion models via coherent sampling. pages 3954–3964, 2023. 2

[89] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22479–22489. IEEE, 2023. 3

[90] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 8

[91] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX*, pages 346–364. Springer, 2020. 2, 4

[92] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16010–16021, 2023. 2

[93] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: generating human motion from textual descriptions with discrete representations. *CoRR*, abs/2301.06052, 2023. 2

[94] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *CoRR*, abs/2208.15001, 2022. 2, 4

[95] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. *CoRR*, abs/2304.01116, 2023. 2

[96] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: proximity learning of articulation and contact in 3d environments. In *8th International Conference on 3D Vision, 3DV 2020, Virtual Event, Japan, November 25-28, 2020*, pages 642–651. IEEE, 2020. 3

[97] Wanyue Zhang, Rishabh Dabral, Thomas Leimkühler, Vladislav Golyanik, Marc Habermann, and Christian Theobalt. ROAM: robust and object-aware motion generation using neural pose descriptors. *CoRR*, abs/2308.12969, 2023. 3, 4

[98] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. COUCH: towards controllable human-chair interactions. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part V*, pages 518–535. Springer, 2022. 3

[99] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20449–20459. IEEE, 2022. 2, 3

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

[101] Yan Zhang, Michael J. Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3372–3382. Computer Vision Foundation / IEEE, 2021. 3

[102] Zihan Zhang, Richard Liu, Kfir Aberman, and Rana Hanocka. Tedi: Temporally-entangled diffusion for long-term motion synthesis. *CoRR*, abs/2307.15042, 2023. 2

[103] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, pages 311–327. Springer, 2022. 2, 3

[104] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. *CoRR*, abs/2305.12411, 2023. 2, 3

[105] Mengyi Zhao, Mengyuan Liu, Bin Ren, Shuling Dai, and Nicu Sebe. Modiff: Action-conditioned 3d motion generation with denoising diffusion probabilistic models. *CoRR*, abs/2301.03949, 2023. 2

[106] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. CAMS: canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 585–594. IEEE, 2023. 3

[107] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C. Karen Liu, and Leonidas J. Guibas. GIMO: gaze-informed human motion prediction in context. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIII*, pages 676–694. Springer, 2022. 2, 3

[108] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. TOCH: spatio-temporal object-to-hand correspondence for motion refinement. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part III*, pages 1–19. Springer, 2022. 3

[109] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5745–5753. Computer Vision Foundation / IEEE, 2019. 4

[110] Guangxiang Zhu, Zhiao Huang, and Chongjie Zhang. Object-oriented dynamics predictor. pages 9826–9837, 2018. 3

Appendix

We show in this appendix additional qualitative and quantitative results (Sec. A and Sec. B), detail our baseline evaluation protocol (Sec. C), elaborate on the metrics used in the main paper (Sec. D), show the architecture used in our approach (Sec. E), and provide additional details regarding the data (Sec. F).

A. Additional Qualitative Results

We show additional generated 3D human-object interactions of our method in Fig. 9, with object geometry and text condition on the left, and our generated sequence on the right.

B. Additional Quantitative Results

B.1. Penetration metric

The exact fidelity and diversity of our results is hard to capture with any single metric. Thus, we evaluate multiple such metrics in the main paper (R-Precision, FID, Diversity, MultiModality), and conduct a perceptual user study to verify the metrics' expressiveness.

Here, we provide an additional evaluation based on an intuitive physics-based metric: The ratio of frames with human-object inter-penetrations. Due to the imperfect nature of human-object interaction capture, a non-zero amount of penetrations is expected; however, a high amount of penetration indicates low quality interactions with independently floating and often intersecting objects.

We see in Tab. 3 that our approach leads to less overall penetrations, which confirms the higher quality of our sequences, compared to the baselines.

	BEHAVE	CHAIRS
InterDiff	0.92	0.81
MDM	0.67	0.78
Ours	0.31	0.36

Table 3. Ratio of frames with penetrations between human and object. Combined with the metrics in the main paper, a lower number corresponds to more realistic interactions. Our approach produces less such inter-penetrations, compared to baselines.

B.2. Novelty of Generated Interactions

We perform an additional interaction novelty analysis to verify that our method does not simply retrieve memorized train sequences but is indeed able to generate novel human-object interactions. To do so, we generate ≈ 500 sequences from both datasets and retrieve the top-3 most similar train sequences, as measured by the l_2 distance in human body and object transformation parameter space.

Fig. 10 shows the top-3 closest train sequences, along with a histogram of l_2 distances computed on our test set of ≈ 500 generated sequences. In red, we mark the intra-trainset distance between samples in the train set. We observe that the distance between our generated sequences and the closest train sequence is mostly larger than the intra-train distance. Thus, our method is able to produce samples that are novel and not simply retrieved train sequences.

C. Baseline Evaluation Setup

There is no previous approach to modeling 3D human-object interactions from text and object geometry for direct comparison. Thus, we compare to the two closest methods, and compare to them in multiple settings, for a fair comparison.

The most related approach is InterDiff [84]. Their setting is to generate a short sequence of human-object interactions, from an observed such sequence as condition, with geometry but no text input. Their goal is to generate one, the most likely, sequence continuing the observation. We use their full approach, including the main diffusion training together with the post-processing refinement step. We compare in two different settings: First, in their native setup, running their method unchanged and modifying ours to take in geometry and past sequence observation instead of text (Motion-Cond. HOI in Tab. 1 main). Then, we modify their approach to take in geometry and text, replacing their past motion encoder with our CLIP-based text encoder (Text-Cond. HOI in Tab. 1 main). We observe that our method is able to outperform InterDiff in both scenarios, for both datasets.

We additionally compare to MDM [71], a recent diffusion-based state-of-the-art human motion generation approach. Their approach is based on a transformer encoder formulation, using each human body as a token in the attention. We run their method on SMPL parameters and first compare in their native setting, only predicting human motion. We compare to the human motion generated by our method which is trained to generate full human-object interactions (Text-Cond. Human Only in Tab. 1 main). We also compare to human motion sequences generated by InterDiff in this setting. We see that our method is able to outperform both baselines even in this setting, demonstrating the added benefit of learning interdependencies of human and object motion. For the comparison in our setting, we modify MDM by adding additional tokens for the objects to the attention formulation. Our approach performs more realistic and diverse sequences in both settings which better follow the text condition.

D. Fidelity and Diversity Metrics

We base our fidelity and diversity metrics R-Precision, FID score, Diversity, and MultiModality on practices established for human motion generation [24, 25, 71], with minor modifications: We use the same networks used by these previous approaches, and adapt the input dimensions to fit our feature lengths, $F = 79$ when evaluating human body motion only,

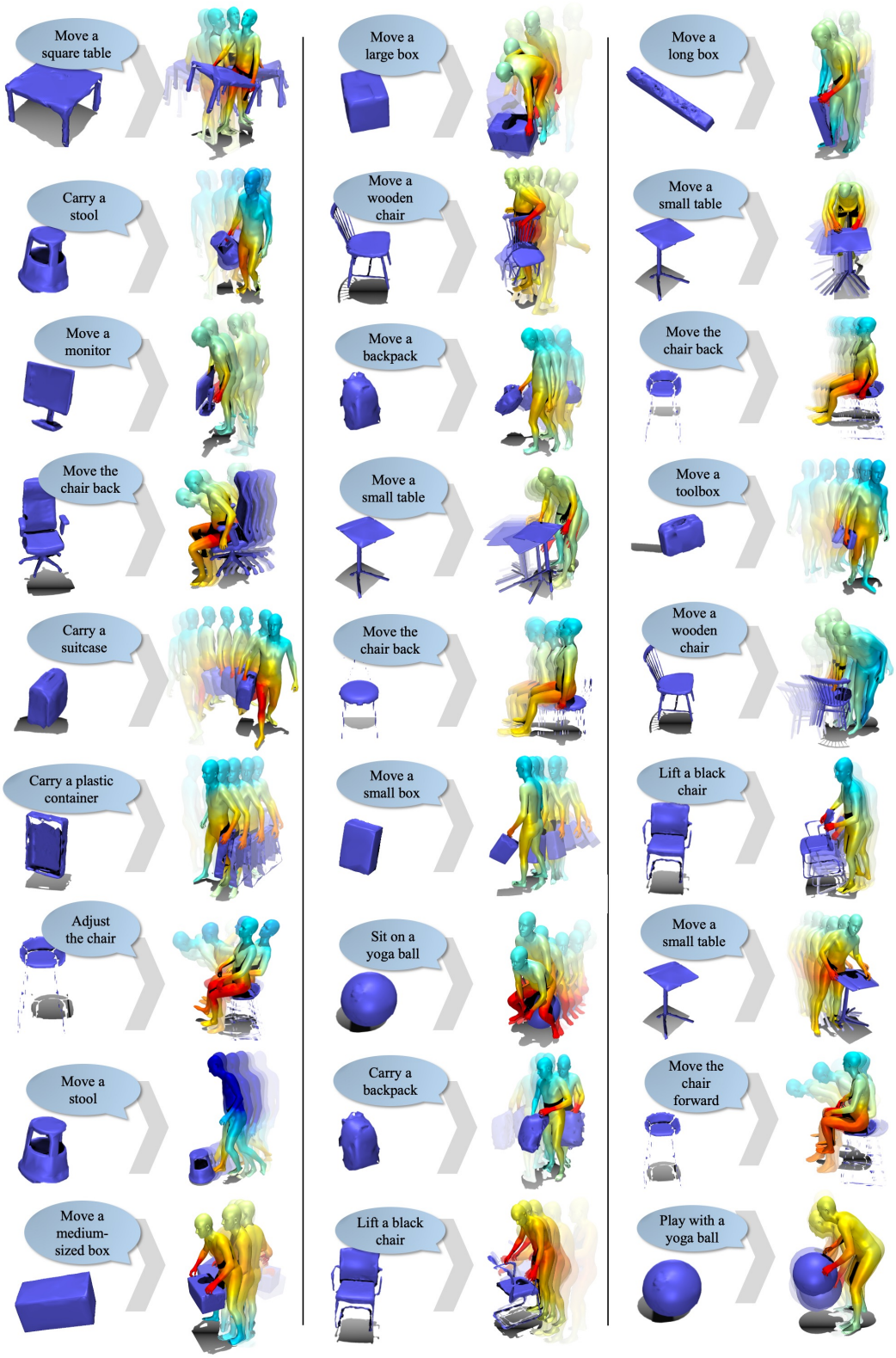


Figure 9. Additional qualitative evaluation. Our method produces diverse and realistic 3D human-object interaction sequences, given object geometry and short text description of the action. The sequences depict high-quality human-object interactions by modeling contact, mitigating floating and penetration artifacts.

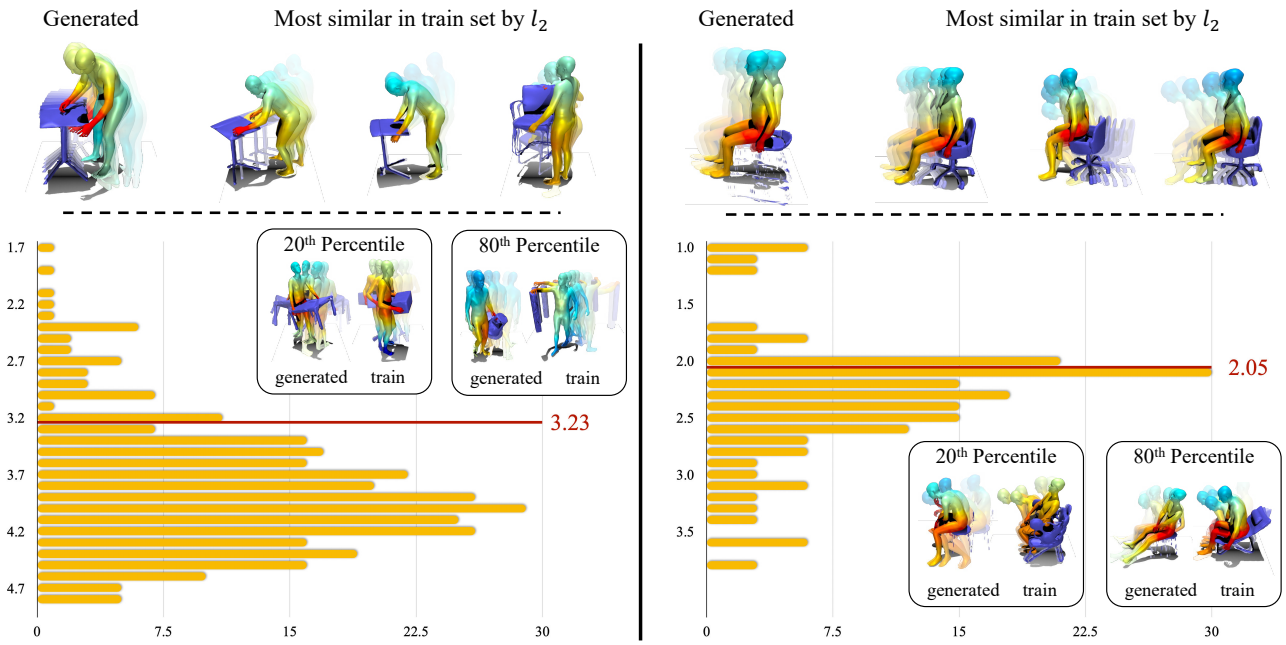


Figure 10. Human-Object Interaction Sequence Novelty Analysis. Performed on BEHAVE [8] (left) and CHAIRS [35] (right). We retrieve top-3 most similar sequences from the train set, and plot a histogram of distances to the closest train sample. While sequences at the 20th percentile still resemble the generated interactions, there is a large gap in the 80th percentile. We show the intra-trainset distance in red. Our approach generates novel shapes, not simply retrieving memorized train samples.

and $F = 79 + 128 + 9 = 216$ (SMPL parameters, contact distances, object transformations) for full evaluation in the human-object interaction scenario.

E. Architecture Details

Fig. 11 shows our detailed network architecture, including encoder, bottleneck, and decoder formulations.

F. Data Details

F.1. Datasets

CHAIRS [35] captures 46 subjects as their SMPL-X [53] parameters using a mocap suit, in various settings interacting with a total of 81 different types of chairs and sofas, from office chairs over simple wooden chairs to more complex models like suspended seating structures. Each captured sequence consists of 6 actions and a given script; the ex-

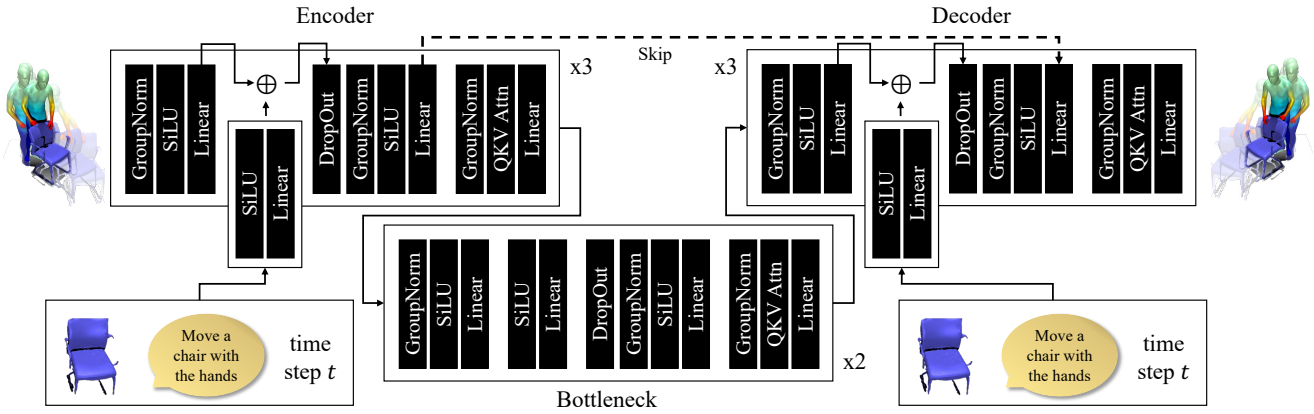


Figure 11. Network architecture specification.

act separation into corresponding textual descriptions was manually annotated by the authors of this paper. In total, this yields ≈ 1300 sequences of human and object motion, together with a textual description. Every object geometry is provided as their canonical mesh; we additionally generate ground-truth contact and distance labels based on posed human and object meshes per-frame for each sequence. We use a random 80/10/10 split along object types, making sure that test objects are not seen during training.

BEHAVE [8] captures 8 participants as their SMPL-H [62] parameters captured in a multi-Kinect setup, along with the per-frame transformations and canonical geometries of 20 different object with a wide range, including yoga mats and tables. This yields ≈ 130 longer sequences. We use their original train/test split.

F.2. Object Geometry Representation

We represent object geometry as a point cloud, to be processed by a PointNet [56] encoder. For this, we sample $N = 256$ points uniformly at random on the surface of an object mesh. Each object category is sampled once as a preprocessing step and kept same for training and inference.