Task-Oriented Human-Object Interactions Generation with Implicit Neural Representations

Quanzhou Li[1]   Jingbo Wang[2]   Chen Change Loy[1]   Bo Dai[2]

[1] S-Lab, Nanyang Technological University   [2] Shanghai AI Laboratory

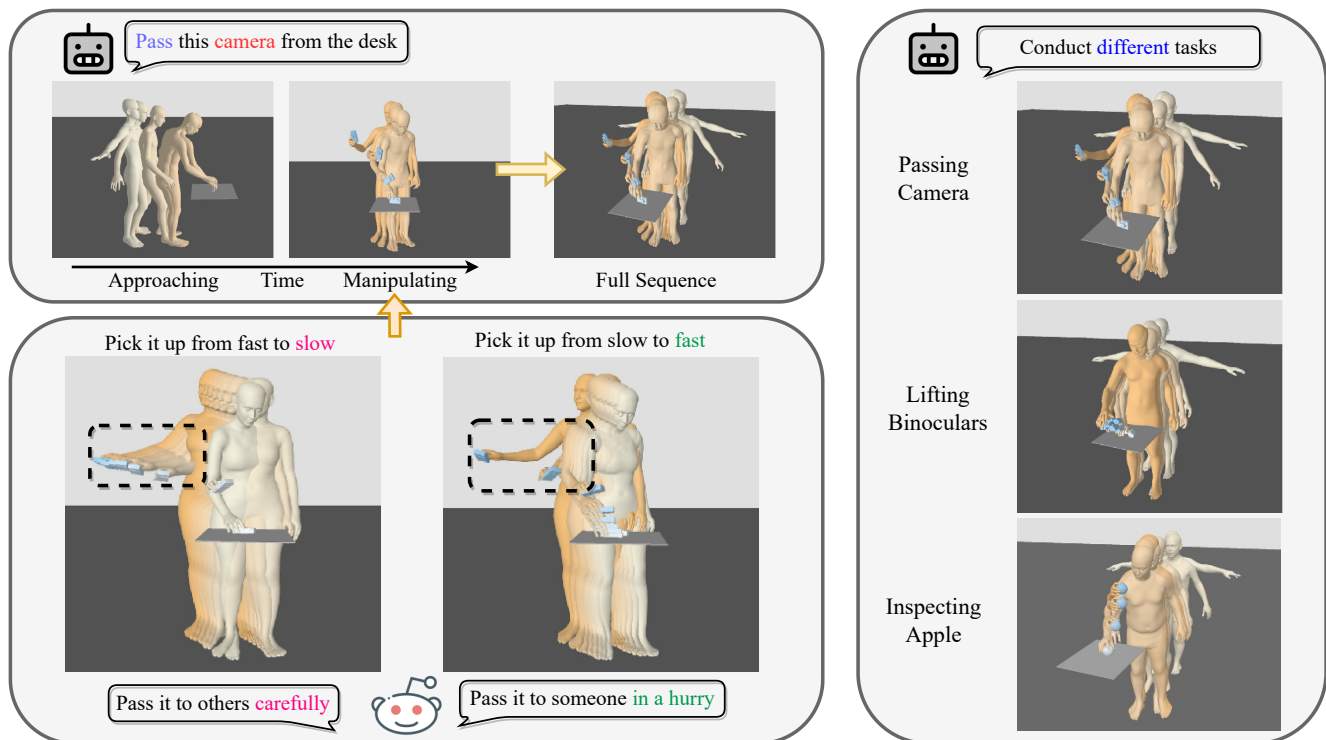{quanzhou001, ccloy}@ntu.edu.sg   {wangjingbo, daibo}@pjlab.org.cn

Figure 1. TOHO generates realistic and continuous task-oriented human-object interaction motions. To achieve a task, the human needs first to approach the object and then manipulate it. Unlike previous works which only synthesize the approaching motions or manipulation motions with known objects, TOHO generates complete task-oriented object manipulation sequences with unseen objects. In addition, as TOHO generates continuous sequences, we demonstrate that by inputting specific temporal coordinate vectors, TOHO can synthesize motions of various velocities and be upsampled to much higher framerates. On the right, we present three generated examples of digital humans with divergent shapes using novel objects to perform distinct tasks.

Abstract

Digital human motion synthesis is a vibrant research field with applications in movies, AR/VR, and video games. Whereas methods were proposed to generate natural and realistic human motions, most only focus on modeling humans and largely ignore object movements. Generating task-oriented human-object interaction motions in simula-tion is challenging. For different intents of using the objects, humans conduct various motions, which requires the human first to approach the objects and then make them move consistently with the human instead of staying still. Also, to deploy in downstream applications, the synthe-sized motions are desired to be flexible in length, provid-ing options to personalize the predicted motions for various purposes. To this end, we propose TOHO: Task-Oriented

Human-Object Interactions Generation with Implicit Neural Representations, which generates full human-object interaction motions to conduct specific tasks, given only the task type, the object, and a starting human status. TOHO generates human-object motions in four steps: 1) it first estimates the object's final position given the task intent; 2) it then generates keyframe poses grasping the objects; 3) after that, it infills the keyframes and generates continuous motions; 4) finally, it applies a compact closed-form object motion estimation to generate the object motion. Our method generates continuous motions that are parameterized only by the temporal coordinate, which allows for upsampling of the sequence to arbitrary frames and adjusting the motion speeds by designing the temporal coordinate vector. This work takes a step further toward general human-scene interaction simulation.

1. Introduction

Humans are in constant interactions with objects around them, and with different using intents, we conduct distinct motions with the objects. Generating such sequences in simulation is of great interest and value in various fields, from computer vision to robotics. Despite the tremendous progress in this topic in recent years, human-object interaction motion synthesis remains an under-studied problem. Previous works primarily focus on synthesizing human motions regardless of the objects or generating human interactions only with known objects already attached to the human hands. Moreover, existing works can only generate discrete frames and neglect actual motions' intrinsic diversity, making the synthesized motions hard to be manipulated in applications.

Synthesizing task-oriented human-object interaction motions is challenging due to various reasons. First, given an object with a task like lifting, the directions of moving the object are neither known nor unique and should be estimated by the given information. Second, a complete human-object interaction sequence should have the human approaching and grasping the object and making it move consistently with the human without floating around or staying still. Third, to deploy in applications that have various framerate requirements, generated motions should be flexible in length instead of having fixed framerates. Previous methods that are the most relevant to ours either only generate motions stopping at the grasping points, ignoring object motions and using intents [38,45], or only generate object manipulation sequences with objects already attached to the human hands [7], and neither can produce continuous results. Currently, synthesizing complete continuous task-oriented human-object interaction motions remains an unsolved problem.

Inspired by recent advances in human motion synthesis

and implicit neural representations, we propose TOHO, a novel task-oriented human-object interaction motion generation framework. A natural human-object motion requires the human first to approach the object, grasp it, and then use it to accomplish the task, sequentially. TOHO addresses each part of the motion synthesis process, taking only the task type, the object at its initial position, and the initial human status as inputs. It generates natural and realistic human motions conducting a task with an unseen object while allowing the generated results to be of arbitrary lengths.

We generate complete human-object motions with TOHO in four steps. 1) First, we design an object parameters sampler that estimates the object's translation and orientation offsets from its initial position with the task type and human shape, giving the final object position and orientation. 2) Similar to [38,45], we use a goal pose generation network to generate grasping poses given the object information. 3) Then, we propose an INR-based motion in-betweening network that generates continuous motions to infill the missing in-between part of two frames. 4) Finally, we present a fast and compact object motion estimation algorithm that outputs realistic and natural object motions based on the generated human motions. With the algorithm, we synthesize perceptually realistic human-object manipulations with objects moving consistently with the human.

With the four steps, our framework generates complete task-oriented human-object interaction motions. Unlike autoregressive and CNN-based methods, our framework generates motions that are continuous and parameterized only by the temporal coordinate. The continuity reflects what motions in the real world are like and allows for velocity adjustments and upsampling of sequences to arbitrary frames. Also, at inference time, as a frame is not conditioned on previous frames, all frames can be inferred parallelly with the same model. In Table 1, we compare our problem setting with previous methods.

We summarize our contributions as 1) we present a unified framework to generate complete human-object interaction motions with intents; 2) our framework generates continuous motions parameterized only by the temporal coordinates, exploring the intrinsic diversity of actual motions; 3) our generated results are flexible in length and speed, allowing upsampling and velocity adjustments in downstream applications.

2. Related Work

Hand Grasp Synthesis. Synthesizing realistic hand grasps is a challenging problem, and with the advancement of deep learning, many works have been proposed to approach the task, including [2,9,13,14,39,41]. Taheri et al. [39] propose a conditional variational autoencoder (cVAE) [36] based network to predict coarse MANO [32] parameters and then apply a refinement step to optimize them. Jiang et al. [13]

Table 1. Overview of our problem setting compared with previous methods. Our method is the only unified framework that generates complete and continuous intent-driven human-object manipulation motions with unseen objects.

	Motion Synthesis				
	Reaching Object	Object Manipulation	Unseen Object	Intent-Driven	Continuous Motion
GOAL [38]	✓	✗	✓	✗	✗
SAGA [45]	✓	✗	✓	✗	✗
IMoS [7]	✗	✓	✗	✓	✗
TOHO (ours)	✓	✓	✓	✓	✓

present a method to estimate the hand pose together with a contact map, which is then used to refine the grasp. Going beyond the synthesis of static object grasps, some previous studies [5, 48] also explore the generation of hand grasp motions. Christen et al. [5] suggest using reinforcement learning to synthesize physically plausible grasping motions. Given the object and wrists trajectories, Zhang et al. [48] propose an autoregressive model to synthesize the hand-object manipulation motions.

While these works have well-addressed hand grasp synthesis, they only focus on the hands in isolation from the body. Our work differs from the earlier works in that we propose to generate whole-body human-object interactions, which lie in a higher parameter dimension and require consistency of all body parts.

Human Motion Synthesis. Human motion synthesis has become a prevailing research topic in recent years and has drawn attention from both computer vision and computer graphics [3, 8, 15–17, 22–24, 28, 40]. Kaufmann et al. [15] propose considering motion sequences as images and utilizing convolutional neural networks to inpaint the missing parts. Some studies [3, 16] suggest the use of VAEs to generate stochastic human motions. Some more recent works [17, 22, 28, 40] make use of the attention [42] mechanism to model human motions. Even though these works have made significant progress in human motion synthesis, they only focus on modeling humans regardless of the scenes and objects around them.

There are some existing works proposed to tackle human motion synthesis involving the 3D scenes, including [4, 6, 10, 11, 19, 21, 31, 33, 37, 38, 43–45, 49]. Wang et al. [44] propose a hierarchical motion synthesis framework, first synthesizing several sub-goal poses in 3D scenes, then infilling the whole motion sequence, and finally refining the motion sequence with an optimization scheme. Hassan et al. [11] suggest first estimating the object's goal pose and contact to interact and a trajectory for the human to approach and then generating the human motion with an autoregressive module.

The works that are the most relevant to ours are [7, 38, 45]. GOAL [38] suggests first estimating a grasping pose

and then synthesizing the motion of the human from its initial position to the goal with an autoregressive model. Similarly, SAGA [45] first predicts a grasping pose and uses a CNN-based cVAE to infill the motion. While both works generate realistic motions, they only model the human approaching the objects and stopping at the 'touching' points, ignoring human-object interactions. Ghosh et al. [7] develop the IMoS method to generate human-object interactions assuming a grasp on a known object is already established. IMoS employs an autoregressive prediction to generate 15 frames for one motion clip and linearly interpolate them into 30 frames. Unlike the abovementioned methods, our framework synthesizes complete intent-driven human-object interactions with unseen objects, and the generated sequence can be upsampled to arbitrary frames.

Implicit Neural Representations. Implicit neural representations have gained considerable attention recently with the success of SIREN [34], and NeRF [25]. The critical insight of INR is that a complex signal can be represented by a function of spatial or temporal coordinates at its corresponding position, and high-frequency details can be well preserved through this mapping. INR has demonstrated its efficacy on multiple tasks, including image synthesis [1, 35], video generation [46], time-varying 3D geometries [26], and dynamic scenes [18, 30]. Recently, He et al. [12] propose a task-agnostic INR-based representation to interpret human motion as a function of time and conduct tasks through per-sequence optimization. Inspired by these works, we propose an INR-based generative model to infill the motion between two arbitrary human poses and positions.

3. Method

Preliminaries. 1) 3D human representation. We use the SMPL-X model [27], which models the human body with hand and face details. SMPL-X takes human shape, β, pose, θ, and body global translation, t, as inputs and generates a mesh with 10,475 vertices. In this work, we predict the 6D continuous pose rotation vector [51] $\theta \in \mathbb{R}^{55 \times 6}$ and the global translation $t \in \mathbb{R}^3$. The shape parameter $\beta \in \mathbb{R}^{10}$ is constant for a fixed body shape. **2) Object shape representation.** We use the Basis Point Set (BPS) [29] distances to represent the object shapes. Following [39], we randomly sample 1024 vertices from $[-0.15, 0.15]^3$ as the basis point set to calculate distances.

3.1. Overview

The overview of our method is shown in Fig. 2. Our goal is to synthesize continuous human-object motions conducting specific tasks with unseen objects. Given three inputs, namely: 1) a task type, 2) an object shape with its starting translation and orientation, and 3) a human pose with its shape at its initial position, our method generates a contin-
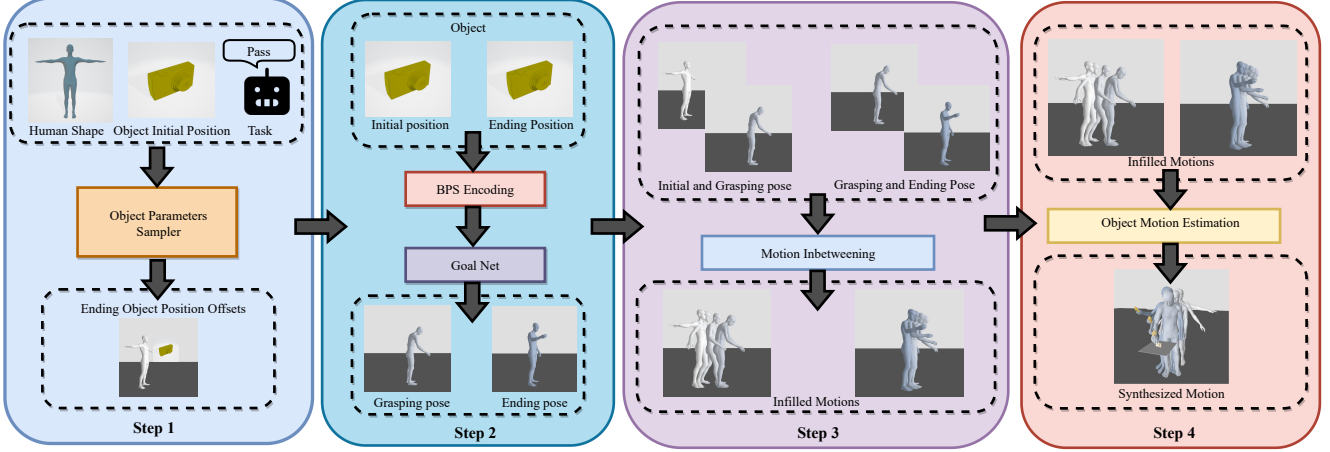
Figure 2. Overview of TOHO. We formulate the generation of object manipulation motions as a motion-infilling problem that consists of four steps. 1) With the shape parameters of the human and the task type, we estimate the object's final position using our object parameters sampler. 2) The goal net then generates human poses grasping the object at its initial and final positions. 3) Our motion inbetweening model subsequently generates continuous motions to infill the missing frames between the keyframe human poses. 4) Finally, our object motion estimation algorithm outputs a stable and consistent object motion based on the human motion in real-time.

uous motion sequence allowing the virtual human to grasp the object and conduct the task with it. The framework generates motions in four steps. We introduce each of the steps in the following sections.

3.2. Object Parameters Sampler

Previous works [38, 45] propose to use a pose prediction network to estimate the human pose at the grasping point based on the object's shape and position. However, the object's final position is not conveniently given in object manipulation. Also, with different types of tasks, the object could be taken to very diverse positions. [7] synthesizes object manipulation motions by predicting frames autoregressively. However, it assumes a grasp is already established and only synthesizes 15-frame motions into the future with known objects. To synthesize long-term human-object motion sequences with high fidelity, we first generate keyframes of a complete manipulation task and infill the in-between frames, which makes the whole sequence bounded by the keyframes and would not deviate. To this end, we formulate the task of generating complete human-object manipulation motions as a motion inbetweening problem. To generate the keyframes, we first propose a task-conditioned object position estimator. Our object parameters sampler is a cVAE, and in the training stage it takes

$$X_s = [a_{\text{one}}, \beta, t_o^{\text{init}}, r_o^{\text{init}}, t_o^{\text{off}}, t_o^{\text{off}}] \quad (1)$$

as inputs, where a_{one} is the a-th column of an identity matrix, i.e., a one-hot vector, $a \in \mathbb{R}$ is the task label, $\beta \in \mathbb{R}^{10}$ is the human shape, $t_o^{\text{init}} \in \mathbb{R}^3$ and $r_o^{\text{init}} \in \mathbb{R}^6$ are the initial translation and orientation of the object, and $t_o^{\text{off}} \in \mathbb{R}^3$

and $r_o^{\text{off}} \in \mathbb{R}^6$ are the translation and orientation offsets of the object's ending position from its initial position. The object position estimator encodes the input to a latent space with a dimensionality of 16 and is conditioned on a_{one}, β, t_o^{init}, and r_o^{init}. In inference, the model uses a sampled latent code $z_s \sim \mathcal{N}(\mu_s, \sigma_s^2)$, where $\mu_s \in \mathbb{R}^{16}$ and $\sigma_s \in \mathbb{R}^{16}$, and the conditioned parameters to decode the ending object translation offset $\hat{t}_o^{\text{off}}$ and orientation offset $\hat{r}_o^{\text{off}}$. The loss to training the sampler is defined as

$$\mathcal{L}_s = \lambda_t ||\hat{t}_o^f - t_o^f||_2 + \lambda_r ||\hat{r}_o^f - r_o^f||_2 + \lambda_{KL}\mathcal{L}_{KL}, \quad (2)$$

where $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence loss. We set the loss coefficients empirically to balance the different terms.

3.3. Goal Net

Given the object's position at the initial and final positions, we design a goal net that utilizes the object information to synthesize keyframe grasping poses. Our goal net is a cVAE conditioned on the object following [38, 45]. During training, the goal net inputs the whole-body grasp with the objects' shape and translation and reconstructs the poses and hand-object distances given the objects' shape and location. The input to the encoder is:

$$X = [\theta, t, \beta, v, d_{b \to o}, h, t_o, b_o, a], \quad (3)$$

where θ, t, and β are the human's pose, global translation, and shape parameters, respectively, $v \in \mathbb{R}^{400 \times 3}$ is the 3D coordinates of 400 sampled vertices on the human's body surface, $h \in \mathbb{R}^3$ denotes the head orientation, $t_o \in \mathbb{R}$ is the object translation, $b_o \in \mathbb{R}^{1024}$ is the BPS representation of the object shape, $a \in \mathbb{R}$ is the task label, and

$d_{b\rightarrow o} \in \mathbb{R}^{400\times3}$ is the offset vectors from the sampled body vertices to the closest object vertices. The decoder predicts the SMPL-X parameters $\hat{\theta}$, $\hat{t}$, the head orientation $\hat{h}$, and a right-hand offset vector $\hat{d}_{r\rightarrow o} \in \mathbb{R}^{99\times3}$ which is a subset of the 400 body vertices to object offsets, given a sampled latent code and conditioned on β, b_o, t_o, and a. We use the same loss in [38] for training and follow its post-optimization scheme to refine the generated poses.

3.4. Motion Inbetweening

Existing works of feed-forward motion inbetweening can only generate discrete motion clips. However, the framerate requirements could be diverse for real-world applications in AR/VR and gaming. Previous works need to either re-train the model or apply complex linear interpolations to synthesize a motion sequence of a different length [45], or have to run a per-sequence optimization which prohibits real-time deployment [12]. To this end, we propose a motion inbetweening method that generates continuous motions that are parameterized only by the temporal coordinates to infill two given frames. The continuity of the model allows the generated motions to be upsampled to arbitrary frames and sampled non-uniformly to give motions of diverse speeds, all by inputting different temporal coordinates to the same inference model.

Given two frames, our motion inbetweening module aims to predict the human poses and root translations in frames between them. Following [45], we find that first computing an interpolated root trajectory of the first and last frames and then predicting the translation offsets of each frame results in smoother results compared to directly predicting the translations themselves. To infill two frames, we view a motion clip as a motion image with the human pose parameters at frame i, $\theta_i \in \mathbb{R}^{55\times6}$, $i \in [1,...T]$, flattened and concatenated with its corresponding root translation $t_i \in \mathbb{R}^3$ as a column vector. For training, we define $T = 64$, while in inference, T can be any integer value. Thus, one motion clip is represented as a 333×64 motion image. We design our motion inbetweening model as $\mathrm{F}(\tau; \phi_{\mathrm{INR}}) = (\theta, t^{\mathrm{off}})$, where $\tau \in [0,1]$ is the temporal coordinate, θ and t^{off} are the pose and the translation offset from the interpolated trajectory at time τ, respectively. For training, we have each $\tau_i = i/64$ correspond to the i-th column of the motion image. The motion inbetweening net is a hypernetwork-based model: firstly, it takes poses of the two endpoint frames, θ_1 and θ_T, and their translation distance, $d_{1\rightarrow T} = t_T - t_1$, as input. Then, it generates the weights ϕ_{INR} for an INR block $\mathrm{F}_{\phi_{\mathrm{INR}}}$. The INR takes the temporal coordinate vector $\tau_{1:T}$, $\tau_i \in [0,1]$ as input and generates the complete motion image:

$$M = [\hat{\theta}_{1:T}, \hat{t}^{\mathrm{off}}_{1:T}], \tag{4}$$

where $\hat{\theta}_{1:T}$ and $t^{\hat{\mathrm{off}}}_{1:T}$ are reconstructed SMPL-X pose and translation offset sequences. We apply Factorized Multiplicative Modulation (FMM) [35] to parameterize the weights of our INR to reduce the number of model parameters.

With the inferenced SMPL-X parameters, we then calculate 99 surface marker locations $\hat{v}^M_{1:T} \in \mathbb{R}^{99\times3}$ and eight foot-ground contact labels $\hat{C}_{fg} \in \{0,1\}^8$. This aids the model in utilizing information from the 3D space and assists in circumventing the skating problem. The loss to train the motion inbetweening model is defined as:

$$\mathcal{L}_M = \lambda_\theta \mathcal{L}_\theta + \lambda_t \mathcal{L}_t + \lambda_v \mathcal{L}_v + \lambda_C \mathcal{L}_C, \tag{5}$$

where $\mathcal{L}_\theta = \sum_{n=1}^T ||\theta_n - \hat{\theta}_n||_2$, $\mathcal{L}_t = \sum_{n=1}^T ||t_n - \hat{t}_n||_1$, $\mathcal{L}_v = \sum_{n=1}^T ||v^M_n - \hat{v}^M_n||_1$, $\lambda_C \mathcal{L}_C = \mathcal{L}_{bce}(C_{fg}, \hat{C}_{fg})$. The hat denotes the reconstructed values.

In inference, we replace the first and last frames of the sequence with the input two frames. Unlike [45], which requires extensive post-optimization to get smooth motion sequences, we only add a lightweight post-processing step: to get a smoother result at the connection points, we interpolate the first and last five frames with the input frames. The motion inbetweening net is implemented with MLPs with skip connections. More details are given in the supplementary material.

3.5. Object Motion Estimation

Synthesizing human-object interactions without object motions is incomplete. Our method generates whole human-object interaction motions with objects moving consistently with humans. Instead of using a parametric model which is computationally expensive and slow, we propose a compact closed-form object motion estimation algorithm that is stable and yields realistic and natural object motion sequences based on human motions in real-time.

As most hand manipulation tasks only involve the right hand (for right-handed people), we focus on the right-hand-object interactions. When the human achieves a stable grasp, the algorithm computes an object motion based on the grasp to finish the task. We denote the frame of this stable grasp, i.e., the 'touching' pose generated by our pose predictor, as frame 1. Denote the locations of the five markers at five right-hand fingertips as v^{f_i}, $i \in \{1,2,3,4,5\}$ and the marker at the right palm as v^{f_0}, $v^{f_i} \in \mathbb{R}^3$. At each timestamp $n \in \{2,...,T\}$, we compute the offset vectors of the fingertip markers from the palm as $o^{f_i}_n = v^{f_i}_n - v^{f_0}_n$, $i \in \{1,2,3,4,5\}$, and find the rotation R_n that optimally aligns the set of vectors $o^{f_{1:5}}_n$ to $o^{f_{1:5}}_1$ with the Kabsch algorithm, i.e., find R_n s.t.,

$$R_n = \underset{R}{\mathrm{argmin}} \frac{1}{2} \sum_{i=1}^5 ||o^{f_i}_1 - Ro^{f_i}_n||_2 \tag{6}$$
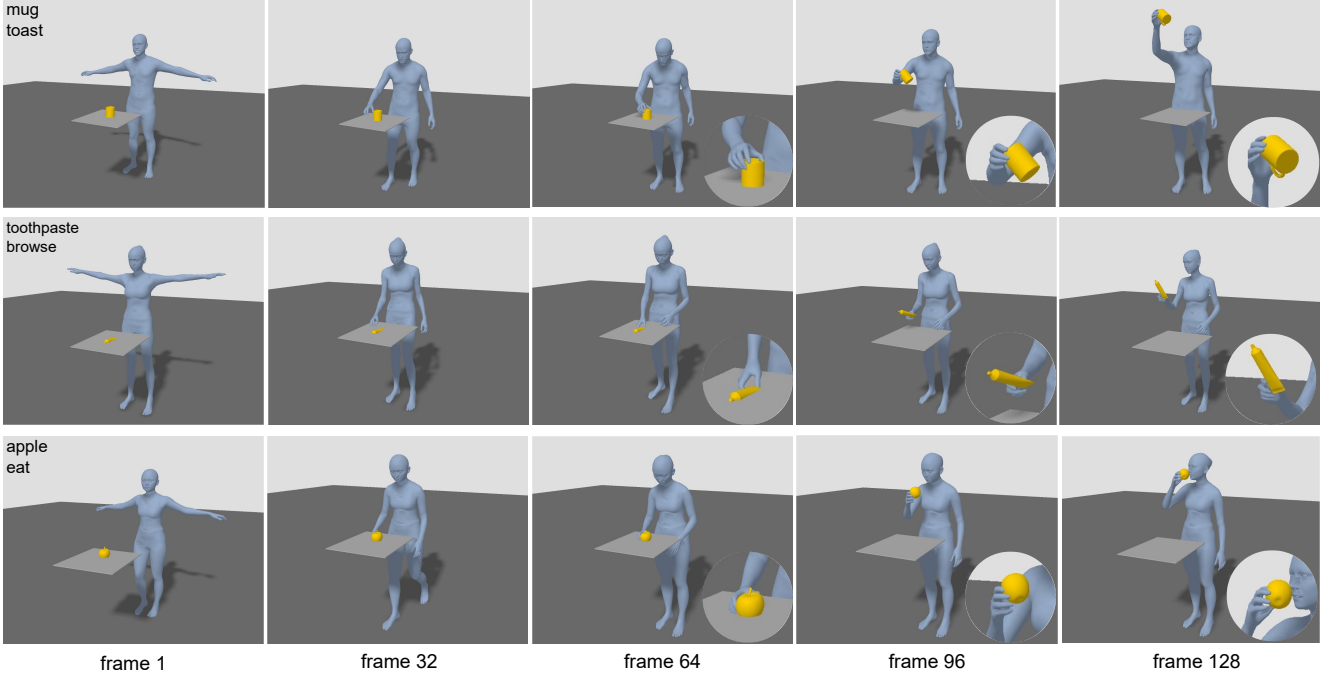
Figure 3. We present generated results of TOHO. Each row represents a human with a distinct shape conducting a task with a novel object unseen during training.

Thus, the object orientation at time n is given by:

$$rot_n = R_n^T rot_1, \tag{7}$$

where rot_i is the rotation matrix of the object orientation at frame n. The 6D orientation r_n can be computed from rot_n. Then the object translation at frame n is given by:

$$t_n^o = \frac{1}{6}\sum_{i=0}^{5} v_n^{f_i} + R_n^T(t_1^o - \frac{1}{6}\sum_{i=0}^{5} v_1^{f_i}) \tag{8}$$

The object motion parameters $(r_{1:T}, t_{1:T}^o)$ are then used to construct the object motion consistent with the right hand.

4. Experiments

We train our pipeline on the GRAB dataset [39], following the same split as [38]. The dataset consists of 51 objects and four manipulation intents: pass, lift, offhand, and use. For the use intent, there are 26 sub-tasks in total. However, some sub-tasks are specific to certain objects and have only one sequence for a subject in the dataset, and some sub-tasks are similar to each other though having different names. Thus, we merge some of the tasks into 6 tasks that have distinct behaviors to demonstrate the effectiveness of our method. We label the representative frames of each task in the dataset, which requires minimal labor work. For training our pose predictor, we additionally label

frames that the human first grasps the object from the table and the following 20 frames as a 'touch' task to indicate the grasping poses. For the motion inbetweening model, we downsample the framerate of GRAB from 120 fps to 30 fps, slide over each sequence with a skip frame of 16, and chop the sequences into 64 frames as unit training sequences.

4.1. Qualitative Results

Full pipeline. A natural human-object interaction motion requires the human to first walk towards the object, grasp it, and then conduct the task. Fig. 3 shows generated sequences of three humans of *different shapes* performing *distinct tasks* with *unseen objects*.

Upsampling As our motion inbetweening model generates continuous motions, it is theoretically guaranteed to allow upsampling to arbitrary frames. In the supplementary materials, we show our model generates smooth motion even at 512 frames although trained only with 64 frames sequences.

Velocity Modification Another advantage of our method is that by designing the temporal coordinate τ, users can generate motions with different velocities at specific parts of the sequence with the same model weights. This allows for affluent post-artworks to output motions reflecting divergent states of the human. Fig. 4 shows the motions generated with the same inferenced model weights but different temporal inputs. The four inputs include a uniform sampling of τ on $[0, 1]$, a uniform sampling that is two times sparser,
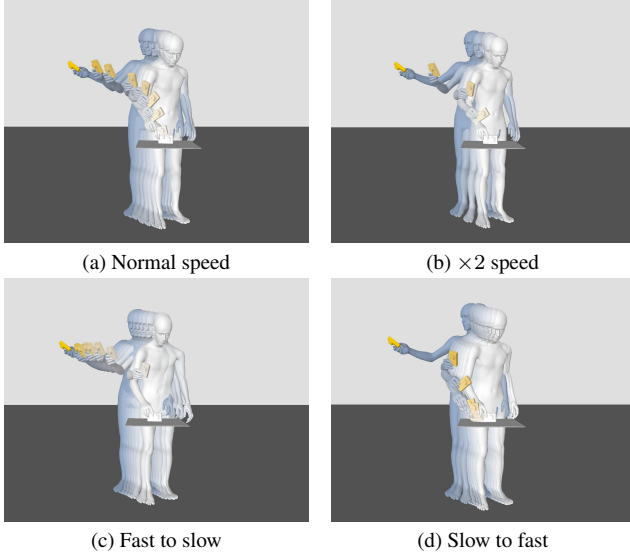
(a) Normal speed   (b) ×2 speed

(c) Fast to slow   (d) Slow to fast

Figure 4. Examples of motion velocity adjustment by modifying the temporal coordinate τ. a) A 64-frame generated result of normal speed. b) Speed up the sequence by uniformly sampling two times fewer values from $\tau \in [0, 1]$, which gives a result of doubled velocities. c) The human swiftly lifts the object and then slowly passes it, which is done by sampling sparsely near 0 and densely near 1. d) The human slowly lifts the object and swiftly passes it, which uses a reverse sampling scheme of c).

a non-uniform sampling that is denser at the end, and a sampling denser at the start. Here, we design the lengths of intervals of the non-uniform sampling as geometric sequences.

4.2. Quantitative Results

4.2.1 Motion Diversity

We calculate the Average L2 Pairwise Distance (APD) [47] of the object translation offsets predicted by our object position estimator, given the same objects and task types but different human shapes, to be 0.19 (m). This indicates our model generates diverse ending object positions for different human shapes within a reasonable range. We report the APD of our full generated motions on GRAB to be 0.34.

4.2.2 Motion Inbetweening

We conduct experiments to demonstrate the effectiveness of our motion inbetweening model. To our knowledge, our model is the first generalizable motion inbetweening model that generates continuous motions.
Baselines. SAGA [45] proposed a CNN-based motion-infilling network that predicts both the human poses and their root translations. Wang et al. [44] proposed an LSTM-based model, namely Route+PoseNet, to infill motion sequences. We take these two as our baselines as they are

Table 2. Comparisons of motion-infilling performance. Second row shows results of pose prediction using ground truth trajectories and third row shows results of prediction of both the poses and trajectories. Boldface represents best results.

	Methods	ADE (↓)	Skat (↓)	PSKL-J (↓)	
				(P, GT)	(GT, P)
GT Traj + local motion infilling	CNN-AE [15]	0.091	0.245	0.804	0.739
	LEMO [49]	0.083	0.152	0.507	0.447
	PoseNet [44]	0.090	0.236	0.611	0.668
	SAGA-Local [45]	**0.079**	**0.137**	**0.377**	**0.372**
Traj + local motion infilling	Route+PoseNet [44]	0.219	0.575	0.955	0.884
	SAGA [45]	**0.083**	0.394	0.772	0.609
	Ours	0.113	**0.247**	**0.232**	**0.218**
GRAB∗	IMoS [7]	0.087	0.413	0.411	0.339
	Ours	0.081	0.278	0.231	0.219

∗ Tested on the GRAB dataset with sequences starting from grasping poses, following the setting of IMoS.

the closest to our settings. We also compare our model with some existing works, including the convolution autoencoder network (CNN-AE) [15], LEMO [49], and PoseNet [44], that take the ground truth trajectory as inputs and infill the local motions. We use the same body markers as in [45] for fair comparisons.
Evaluation Metrics. We use the same evaluation metrics in [45]. 1) *3D marker accuracy.* We compute the Average L2 Distance (ADE) between the reconstructed marker sequences and the ground truth. 2) *Foot skating.* We follow [47] to decide on skating frames and report their ratio in the full sequences. 3) *Motion smoothness.* We measure the smoothness of the generated sequences by computing the Power Spectrum KL Divergence of their joints (PSKL-J) [49] to compare with the ground truth. We report the scores of PSKL-J in both directions as they are asymmetric. Note a lower PSKL-J score represents a closer generated distribution to the ground truth.
Results. In Table 2, we show the results of our method compared to the previous works mentioned above. We experiment on both GRAB [39] and AMASS [20] datasets following the settings of [45] to test our motion inbetweening model. We show the results of using the ground truth trajectory for [15, 49] as they only infill local motions. We also show the results of PoseNet and SAGA fed with ground truth trajectory to predict local poses. Then, we evaluate both the predicted pose and trajectory of our method against SAGA and Route+PoseNet. The results show that our model gives better scores of PSKL-J in both directions than previous works, which indicates that the continuity of our model facilitates the generation of smoother sequences. Our method also yields lower skating effects than SAGA and Route+PoseNet and achieves a comparable ADE score to SAGA. The results show that our method is on par with SOTA methods while providing the benefits of upsampling and velocity adjustment through its continuity property.

We report the results of IMoS-generated sequences to give a sense of how our method is compared with IMoS [7].

Table 3. Comparisons of our generated human-object interaction sequences to the ground truth. We report the results of the entire motion, and of the grasping poses generated by our pose predictor.

	Contact Ratio (↑)			Interp. Depth (m)		
	Max.	Min.	Avg.	Max.	Min.	Avg.
GT	1.00	0.99	1.00	0.007	0.005	0.006
Ours	0.93	0.77	0.85	0.012	0.006	0.007
Pose Predictor		0.91			0.008	

As IMoS is trained only on the GRAB dataset with sequences starting from the grasping poses, we also report our results tested in the same setting. We uniformly sample 60 frames from TOHO as motion sequences and interpolate the 15 generated frames of IMoS into 60-frame motion clips, such that all motion clips have the same length. Note this is not an exact comparison as IMoS is an autoregressive model while ours is a motion-infilling model. Sequences starting from the grasping poses tend to have small root translation movements, which leads to a relatively higher skating ratio.

4.2.3 Human-Object Motion

We conduct experiments to evaluate the realisticness of our human-object motions by computing the hand-object contact ratio [50] and the largest hand-object interpenetration depth along the whole motion sequence. We report the maximum, minimum, and average contact ratio and interpenetration depth along object movements. Our object motion estimation algorithm captures the stable grasp at the initial grasping frame and keeps this hand-object relationship across the sequence, so we also report the contact ratio and interpenetration depth of our pose predictor generated results. Table 3 shows that our method well preserves good grasps during the human-object interaction and yields results that are comparable to the ground truth.

4.2.4 Ablation Study

(1) Human shape for object parameters sampler. To evaluate the effects of human shapes in estimating final object positions, we conduct experiments by training an object parameters sampler which does not take the human shape as input. We report the average distance from the predicted object positions to the ground truth of this model as 0.073 (m), while ours taking the human shape information as input during training has a result of 0.048 (m). The results suggest that human shape is an impactful factor in estimating the final object position given a task type. **(2) Losses of the motion inbetweening model.** In Table 4, we show the results of our motion inbetweening model trained without the foot-ground contact loss/surface marker loss. All three models are trained and tested on the GRAB dataset. **(3) Object motion estimation.** To demonstrate the effectiveness of aligning the orientation of the object to the fingertip

Table 4. Ablation study of our motion inbetweening model. $\mathcal{L}_C$ is the foot-ground contact loss. $\mathcal{L}_v$ is the surface marker loss. Boldface represents best results.

	ADE (↓)	Skat (↓)	PSKL-J (↓)	
			(P, GT)	(GT, P)
Ours-w/o $\mathcal{L}_C$	0.088	0.229	0.242	0.235
Ours-w/o $\mathcal{L}_v$	0.096	0.201	0.264	0.256
Ours	**0.079**	**0.177**	**0.219**	**0.208**

Table 5. Ablation study of our object motion estimation algorithm. R_n is the optimal alignment rotation between the current and first frames of hand vectors.

	Contact Ratio (↑)			Interp. Depth (m)		
	Max.	Min.	Avg.	Max.	Min.	Avg.
Ours-w/o R_n	0.88	0.42	0.66	0.019	0.008	0.013
Ours	**0.98**	**0.89**	**0.93**	**0.009**	**0.004**	**0.007**

rotations, we compare our method with one without R_n, i.e., simply averaging the fingertip trajectories as the object trajectory and keeping the object orientation still. We show the results in Table 5. For both methods, we use the ground truth human motions to compute the object trajectories.

5. Conclusion

In this paper, we introduce TOHO, the first approach to synthesizing continuous task-oriented human-object interaction motions with unseen objects. TOHO generates complete human-object motions to conduct specific tasks with the task type, the object's initial information, and the starting human status as the only inputs. We address the synthesis process in four steps: 1) we estimate the ending object position given the task type; 2) use object positions and task labels to estimate keyframe poses; 3) the motion inbetweening model then generates continuous motions to infill the keyframes; 4) apply a novel closed-form object motion estimation algorithm to produce an object motion consistent to the human motion. The evaluation results show our framework can generate natural and realistic motions, and our generated motions are continuous, allowing for arbitrary upsampling and velocity adjustment. Future work may include generating interaction motions in scenes with occlusions and synthesizing physically corrected motions.

References

[1] I. Anokhin, K. Demochkin, T. Khakhulin, G. Sterkin, V. Lempitsky, and D. Korzhenkov. Image generators with conditionally-independent pixel synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[2] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox. Contact-Grasp: Functional Multi-finger Grasp Synthesis from Contact. In *International Conference on Intelligent Robots and Systems (IROS)*, 2019. 2

[3] Y. Cai, Y. Wang, Y. Zhu, T.J. Cham, J. Cai, J. Yuan, J. Liu, C. Zheng, S. Yan, H. Ding, and et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[4] Z. Cao, H. Gao, K. Mangalam, Q.Z. Cai, M. Vo, and J. Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[5] S. Christen, M. Kocabas, E. Aksan, J. Hwangbo, J. Song, and O. Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[6] E. Corona, A. Pumarola, G. Alenya, and F. Moreno-Noguer. Context-aware human motion prediction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[7] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Annual Conference of the European Association for Computer Graphics (EURAGRAPHICS)*, 2023. 2, 3, 4, 7

[8] A. Gopalakrishnan, A. Mali, D. Kifer, L. Giles, and A.G. Ororbia. A neural temporal model for human motion prediction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[9] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C.C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[10] F.G. Harvey, M. Yurick, D. Nowrouzezahrai, and C. Pal. Robust motion inbetweening. *ACM Transactions on Graphics (TOG)*, 2020. 3

[11] M. Hassan, D. Ceylan, R. Villegas, J. Saito, J. Yang, Y. Zhou, and M.J. Black. Stochastic scene-aware motion prediction. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[12] C. He, J. Saito, J. Zachary, H. Rushmeier, and Y. Zhou. NeMF: Neural motion fields for kinematic animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3, 5

[13] H. Jiang, S. Liu, J. Wang, and X. Wang. Hand-object contact consistency reasoning for human grasps generation. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[14] K. Karunratanakul, J. Yang, Y. Zhang, M.J. Black, K. Muandet, and S. Tang. Grasping field: Learning implicit representations for human grasps. In *International Conference on 3D Vision (3DV)*, 2020. 2

[15] M. Kaufmann, E. Aksan, J. Song, F. Pece, R. Ziegler, and O. Hilliges. Convolutional autoencoders for human motion infilling. In *International Conference on 3D Vision (3DV)*, 2020. 3, 7

[16] J. Li, R. Villegas, D. Ceylan, J. Yang, Z. Kuang, H. Li, and Y. Zhao. Task-generic hierarchical human motion prior using vaes. In *International Conference on 3D Vision (3DV)*, 2021. 3

[17] R. Li, S. Yang, D.A. Ross, and A. Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[18] T. Li, M. Slavcheva, M. Zollhöfer, S. Green, C. Lassner, C. Kim, T. Schmidt, S. Lovegrove, M. Goesele, R. Newcombe, and Z. Lv. Neural 3d video synthesis from multi-view video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[19] X. Li, S. Liu, K. Kim, X. Wang, M.H. Yang, and J. Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[20] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, 2019. 7

[21] O. Makansi, E. Ilg, O. Cicek, and T. Brox. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[22] W. Mao, M. Liu, and M. Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[23] J. Martinez, M.J. Black, and J. Romero. On human motion prediction using recurrent neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[24] J. Martinez, R. Hossain, J. Romero, and J.J. Little. A simple yet effective baseline for 3d human pose estimation. In *International Conference on Computer Vision (ICCV)*, 2017. 3

[25] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[26] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[27] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A.A.A. Osman, D. Tzionas, and M.J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[28] M. Petrovich, M.J. Black, and G. Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[29] S. Prokudin, C. Lassner, and J Romero. Efficient learning on point clouds with basis point sets. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[30] A. Pumarola, E. Corona, G. Pons-Moll, and F Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[31] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L.J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[32] J. Romero, D. Tzionas, and M.J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (TOG)*, 2017. 2

[33] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[34] V. Sitzmann, J.N.P. Martel, A.W. Bergman, D.B. Lindell, and G Wetzstein. Implicit neural representations with periodic activation functions. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 3

[35] I. Skorokhodov, S. Ignatyev, and M. Elhoseiny. Adversarial generation of continuous images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 5

[36] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2015. 2

[37] S. Starke, H. Zhang, T. Komura, and J. Saito. Neural state machine for characterscene interactions. *ACM Transactions on Graphics (TOG)*, 2019. 3

[38] O. Taheri, V. Choutas, M.J. Black, and D. Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 5, 6

[39] O. Taheri, N. Ghorbani, M.J. Black, and D. Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 6, 7

[40] Y. Tang, L. Ma, W. Liu, and W. Zheng. Longterm human motion prediction by modeling motion context and enhancing motion dynamics. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018. 3

[41] D. Turpin, L. Wang, E. Heiden, Y. Chen, M. Macklin, S. Tsogkas, S. Dickinson, and A. Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 3

[43] J. Wang, Y. Rong, J. Liu, S. Yan, D. Lin, and B. Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[44] J. Wang, H. Xu, J. Xu, S. Liu, and X. Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 7

[45] Y. Wu, J. Wang, Y. Zhang, S. Zhang, O. Hilliges, F. Yu, and S. Tang. SAGA: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3, 4, 5, 7

[46] S. Yu, J. Tack, S. Mo, H. Kim, J. Kim, J. Ha, and J. Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2022. 3

[47] Y. Yuan and K. Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision (ECCV)*, 2020. 7

[48] H. Zhang, Y. Ye, T. Shiratori, and T. Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (TOG)*, 2021. 3

[49] S. Zhang, Y. Zhang, F. Bogo, M. Pollefeys, and S. Tang. Learning motion priors for 4d human body capture in 3d scenes. In *International Conference on Computer Vision (ICCV)*, 2021. 3, 7

[50] Y. Zhang, M. Hassan, H. Neumann, M.J. Black, and S. Tang. Generating 3d people in scenes without people. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[51] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3