Object Motion Guided Human Motion Synthesis

JIAMAN LI, Stanford University, USA
JIAJUN WU†, Stanford University, USA
C. KAREN LIU†, Stanford University, USA

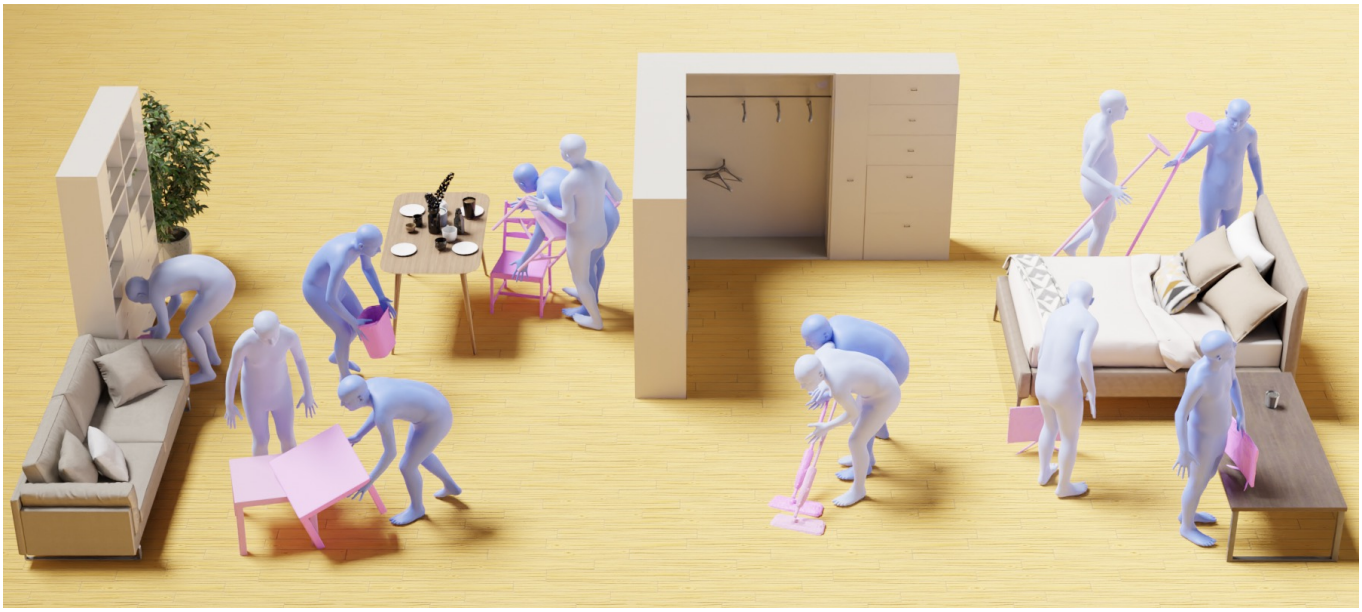


Fig. 1. OMOMO takes a sequence of object states as input and generates full-body human motion interacting with the given object.

Modeling human behaviors in contextual environments has a wide range of applications in character animation, embodied AI, VR/AR, and robotics. In real-world scenarios, humans frequently interact with the environment and manipulate various objects to complete daily tasks. In this work, we study the problem of full-body human motion synthesis for the manipulation of large-sized objects. We propose **O**bject **MO**tion guided human **MO**tion synthesis (OMOMO), a conditional diffusion framework that can generate full-body manipulation behaviors from only the object motion. Since naively applying diffusion models fails to precisely enforce contact constraints between the hands and the object, OMOMO learns two separate denoising processes to first predict hand positions from object motion and subsequently synthesize full-body poses based on the predicted hand positions. By employing the hand positions as an intermediate representation between the two denoising processes, we can explicitly enforce contact constraints, resulting in more physically plausible manipulation motions. With the learned model, we develop a novel system that captures full-body human manipulation motions by simply attaching a smartphone to the object being manipulated. Through extensive experiments, we demonstrate the effectiveness of our proposed pipeline and its ability to generalize to unseen objects. Additionally, as high-quality human-object interaction datasets are scarce, we collect a large-scale dataset consisting of 3D object geometry, object motion, and human motion. Our dataset contains human-object interaction motion for 15 objects, with a total duration of approximately 10 hours.

CCS Concepts: • **Computing methodologies** → **Animation**.

Additional Key Words and Phrases: Human-Object Interaction, Animation, Conditional Diffusion Model, Contact

ACM Reference Format:
Jiaman Li, Jiajun Wu†, and C. Karen Liu†. 2023. Object Motion Guided Human Motion Synthesis. *ACM Trans. Graph.* 42, 6, Article 202 (December 2023), 11 pages. https://doi.org/10.1145/3618333

1 INTRODUCTION

Capturing and synthesizing human movements in contextual environments is critical to progressing embodied AI, character animation, VR/AR, and robotics. The real world in which humans live is complex and highly dynamic. Humans routinely interact with dynamic objects to accomplish everyday tasks, demonstrating a diverse range of full-body manipulations. For example, humans pull and push a mop to tidy a floor, reposition a floor lamp to illuminate a specific area, drag a chair toward a desk, and place a monitor on a desk. Realistically simulating such complex manipulation behaviors is a fundamental problem in computer graphics with a lot of downstream applications.

†indicates equal contribution.

Authors' addresses: Jiaman Li, Stanford University, USA, jiamanli@stanford.edu; Jiajun Wu†, Stanford University, USA, jiajunwu@cs.stanford.edu; C. Karen Liu†, Stanford University, USA, karenliu@cs.stanford.edu.

0730-0301/2023/12-ART202
https://doi.org/10.1145/3618333

Prior works have made significant progress in addressing the contextual human motion synthesis problem for activities such as navigating through a 3D scene or sitting on a chair [Hassan et al. 2021; Mir et al. 2023; Wang et al. 2021a,b; Zhang et al. 2022a; Zhao et al. 2023]. They model interactions with static 3D scenes or static objects based on large-scale human motion datasets. In comparison, datasets containing full-body interaction with moving objects are scarce. Prior works rely on reinforcement learning to model such behaviors [Hassan et al. 2023; Merel et al. 2020; Xie et al. 2023], but the learned policies are often limited to manipulating specific types of geometry used for training.

We present a new approach to synthesizing the dynamic interactions between humans and large-sized objects, particularly in manipulation tasks requiring full-body movements and precise coordination between hands and objects. We aim to bridge the gap between current research and real-world manipulation behaviors by introducing a large-scale dataset and developing a robust approach to synthesize full-body motion from object motion.

We present a new framework – **O**bject **MO**tion guided human **MO**tion synthesis (OMOMO). We leverage a conditional diffusion formulation to predict plausible full-body poses with a sequence of object geometry as input. One key observation is that hand position is a deciding factor for full-body movement during manipulation. Thus, we devise a two-stage approach to generate hand positions conditioned on object geometry features and then synthesize full-body poses based on the predicted hand positions. The two-stage design enables us to apply contact constraints to our predicted hand joint positions, which significantly enhances the contact realism of the generated results. We demonstrate the effectiveness of our proposed method in our dataset and showcase its generalizability to unseen objects.

Moreover, we introduce an innovative application that generates full-body human poses based on object motion captured by an iPhone. In particular, we mount an iPhone on an unseen object, employ the iPhone ARKit to obtain camera poses and deduce the motion of the object. Subsequently, we apply these object poses to 3D geometry reconstructed using Luma [AI 2023]. Our pipeline takes the sequence of object geometry as input and generates the corresponding full-body human motion. This application demonstrates an affordable and user-friendly method for capturing human interaction motions during everyday tasks.

An additional contribution of this work is a new dataset with paired object motion and human motion to facilitate the learning of full-body human manipulation behaviors. We leverage an advanced 3D reconstruction technique to extract 3D object geometry from a monocular video. We then use motion capture devices to capture human and moving objects simultaneously. To capture motions that resemble real-world scenarios, we provide language descriptions to guide our volunteers to perform meaningful interactions with various objects. Our dataset can be used for different tasks to model full-body human manipulation behaviors.

To summarize, the contributions of this work include:

(1) A novel approach to full-body manipulation synthesis by generating full-body motion from object motion. We introduce an effective framework based on conditional diffusion to synthesize full-body movements from object motion.
(2) A novel application that employs an iPhone to capture object motion from the egocentric view of the object, enabling the synthesis of full-body movements by simply attaching an iPhone to various objects.
(3) A large-scale high-quality dataset consisting of 3D object geometry, object motion, and full-body motion.

2 RELATED WORK

Human Motion and Interaction Datasets. Human motion modeling has been extensively studied with motion capture datasets [Mahmood et al. 2019]. Recently, there has been a surge of interest in human scene interactions. PROX [Hassan et al. 2019] provides paired 3D scenes and human motions extracted from RGB videos. HPS [Guzov et al. 2021] contributes a dataset of paired scenes, egocentric video, and human motion captured with an IMU-based suit. EgoBody [Zhang et al. 2022c] collects a dataset consisting of 3D scenes, egocentric video, eye gaze, and human motions extracted from multi-view RGBD frames with a focus on social interactions. GIMO [Zheng et al. 2022] explores the problem of gaze-guided motion prediction using a similar data modality. Synthetic datasets [Li et al. 2023; Wang et al. 2022] combine scene datasets [Dai et al. 2017; Straub et al. 2019] with motion datasets [Mahmood et al. 2019] to produce paired human motions in 3D environments. CIRCLE [Araujo et al. 2023] integrates VR and MoCap techniques to collect high-quality motion within virtual scenes.

A couple of datasets focus on human-object interactions. For example, SAMP [Hassan et al. 2021] contains sitting and lying down motions while interacting with chairs and sofas. COUCH [Zhang et al. 2022a] is dedicated to data collection for sitting on different chairs. These datasets primarily contain motions interacting with static objects.

Moreover, some datasets collect both human motion and object motion [Bhatnagar et al. 2022; Fan et al. 2023; Guzov et al. 2023; Taheri et al. 2020]. GRAB [Taheri et al. 2020] focuses on the interaction between humans and small-sized objects, involving mostly hand motions. BEHAVE [Bhatnagar et al. 2022] records interactions with larger-sized objects, making it closely related to our dataset. However, it relies on multi-view RGBD input to extract human and object motion, which does not yield motion of sufficient quality for motion synthesis tasks. Furthermore, the limited data for each object impedes its capacity for training a motion generative model. In contrast, our work focuses on synthesizing dynamic human interactions with large-sized objects and we introduce a large-scale dataset consisting of high-quality human motion and object motion.

Contextual Human Motion Synthesis. Motion synthesis is a long-standing problem in computer graphics, and here we survey prior works centered on motion synthesis in 3D environments. Leveraging the dataset with paired scenes and human motions [Hassan et al. 2019], a couple of work [Wang et al. 2021a,b] learn separate modules to predict root trajectory first and generate full-body poses conditioned on both scene and the planned path. However, constrained by the scale and motion quality of the dataset, these methods struggle to synthesize realistic human motions. To improve the motion quality of generation results, SAMP [Hassan et al. 2021] collects a

high-quality dataset consisting of walking, sitting and lying down motions. And they present a pipeline that first produces a collision-free path based on A* algorithm, generates full-body motion following the path and then synthesizes interaction motions to sit on chairs and sofas. A recent work [Mir et al. 2023] introduces action keypoints as scene abstraction, enabling continual motion synthesis generation across various scenes. In order to produce physically plausible movements, several works employ reinforcement learning techniques to learn interaction policies through meticulously designed task rewards [Chao et al. 2021; Hassan et al. 2023; Lee and Joo 2023].

Another line of work focuses on reaching motion synthesis within contextual environments. GOAL [Taheri et al. 2022] and SAGA [Wu et al. 2022] generate full-body poses aimed at grasping a specific object. IMoS [Ghosh et al. 2023] further synthesizes human and object motions simultaneously after grasping an object. However, these works only consider the target object and do not involve navigation in cluttered scenes. Meanwhile, CIRCLE [Araujo et al. 2023] incorporates human-scene interaction features and formulates the problem with a scene-aware motion refinement model, enabling reaching synthesis in complex static scenes.

While most existing work that involves interaction with dynamic objects aims to synthesize dexterous hand motions [Li et al. 2007; Ye and Liu 2012; Zhang et al. 2021], our work diverges from this line of work. Instead, we focus on the synthesis of full-body movements for manipulation without synthesizing detailed hand movements.

Full-body human motion synthesis for manipulation has been explored in both kinematic-based [Starke et al. 2019] and physics-based methods [Hassan et al. 2023; Merel et al. 2020; Xie et al. 2023]. NSM [Starke et al. 2019] learns a gating network and a motion prediction network to synthesize interaction movements including sitting and carrying objects. As for physics-based character animation, reinforcement learning has been widely used to learn different skills [Liu and Hodgins 2018; Peng et al. 2018, 2021; Xie et al. 2022]. In terms of manipulation, Merel et al. [2020] devise a hierarchical reinforcement learning framework to synthesize box catching and carrying movements with egocentric observations. More recently, Hassan et al. [2023] propose to learn policies based on the Adversarial Motion Priors framework [Peng et al. 2021] for box manipulation task.

In summary, most prior research has not considered the dynamic interaction between humans and large-sized objects. A few works studied the problem of full-body manipulation but were constrained to interactions with boxes. In contrast, our work examines contextual environments with diverse dynamic objects. Leveraging our large-scale dataset, we develop an approach to synthesize manipulation movements for diverse objects. And inspired by the success of diffusion in motion modeling [Dabral et al. 2023; Huang et al. 2023; Li et al. 2023; Tevet et al. 2023; Tseng et al. 2023; Zhang et al. 2022b], we design our framework based on conditional diffusion.

3 METHOD

Our goal is to generate full-body poses $\boldsymbol{X} \in \mathbb{R}^{T\times D}$ from a sequence of object geometry $\boldsymbol{V} \in \mathbb{R}^{T\times K\times 3}$, where T denotes the time steps of the sequence, D, K represents the dimension of human pose state and the number of vertices on object mesh respectively. This problem presents two significant challenges. First, there is inherent uncertainty in predicting full-body poses from object motions, as humans can produce the same object motion with varying movements. Second, the generated human poses need to maintain correct contact with the given object when it is being manipulated. The first challenge can be addressed by using a generative model, such as a diffusion model [Ho et al. 2020]. However, naively applying diffusion models would not address the second challenge of precisely enforcing contact constraints between the hands and the object. We develop a two-staged method based on a diffusion framework with hand positions as an intermediate representation. The first stage predicts both right and left hand positions $\boldsymbol{H} \in \mathbb{R}^{T\times 6}$ from the object geometry. The second stage generates full-body poses $\boldsymbol{X} \in \mathbb{R}^{T\times D}$ conditioned on the predicted hand joint positions. Our pipeline is shown in Figure 2.

3.1 Data Representation

Human Pose Representation. Our pose state representation at time step t consists of global joint position $\boldsymbol{J}_t \in \mathbb{R}^{24\times 3}$ and global joint rotation $\boldsymbol{Q}_t \in \mathbb{R}^{22\times 6}$ represented using 6D continuous rotation [Zhou et al. 2019]. We adopt a widely used parametric human model, SMPL-X [Pavlakos et al. 2019], to reconstruct human mesh from pose and shape parameters.

Object Representation. Given a sequence of object geometry $\boldsymbol{V} \in \mathbb{R}^{T\times K\times 3}$, we adopt Basis Point Set (BPS) representation [Prokudin et al. 2019] to encode object geometry. We use the BPS representation for two reasons. First, it gives us a lightweight and compact representation using fixed-length vectors. Second, BPS does not rely on special model architecture, such as PointNet [Qi et al. 2017], to process and can be encoded with an MLP to learn downstream tasks effectively as demonstrated in the previous work [Prokudin et al. 2019]. We define a ball with a radius $r = 1$, a value chosen to encompass all objects in our dataset. The ball is centered at the centroid of the object, $(g_t^x, g_t^y, g_t^z) = \frac{1}{K}\sum_{i=1}^{K} \boldsymbol{V}_t^i$ at time step t. We sample 1024 points from the volume of the ball $\boldsymbol{B}_t \in \mathbb{R}^{1024\times 3}$. The BPS representation is computed by calculating the difference between each sampled point and its nearest neighbor vertex on the object mesh, and denoted as $d(\boldsymbol{B}_t, \boldsymbol{V}_t) \in \mathbb{R}^{1024\times 3}$. As the global position is not encoded in the BPS representation, we concatenate the 3D location of the object at time step t to yield object geometry features $[g_t^x, g_t^y, g_t^z, d(\boldsymbol{B}_t, \boldsymbol{V}_t)] \in \mathbb{R}^{3+1024\times 3}$. Then we employ a Multilayer Perceptron (MLP) to project the high-dimensional features onto a lower-dimensional space. The projected geometry features are denoted as $\boldsymbol{O}_t, \boldsymbol{O}_t \in \mathbb{R}^{256}$.

3.2 Conditional Diffusion Formulation

The diffusion model consists of a forward diffusion process and a reverse diffusion process. The forward diffusion process is gradually adding noise to the data representation $\boldsymbol{x}^0$ for N steps formulated using a Markov chain,

$$q(\boldsymbol{x}^{1:N}|\boldsymbol{x}^0) := \prod_{n=1}^{N} q(\boldsymbol{x}^n|\boldsymbol{x}^{n-1}). \tag{1}$$

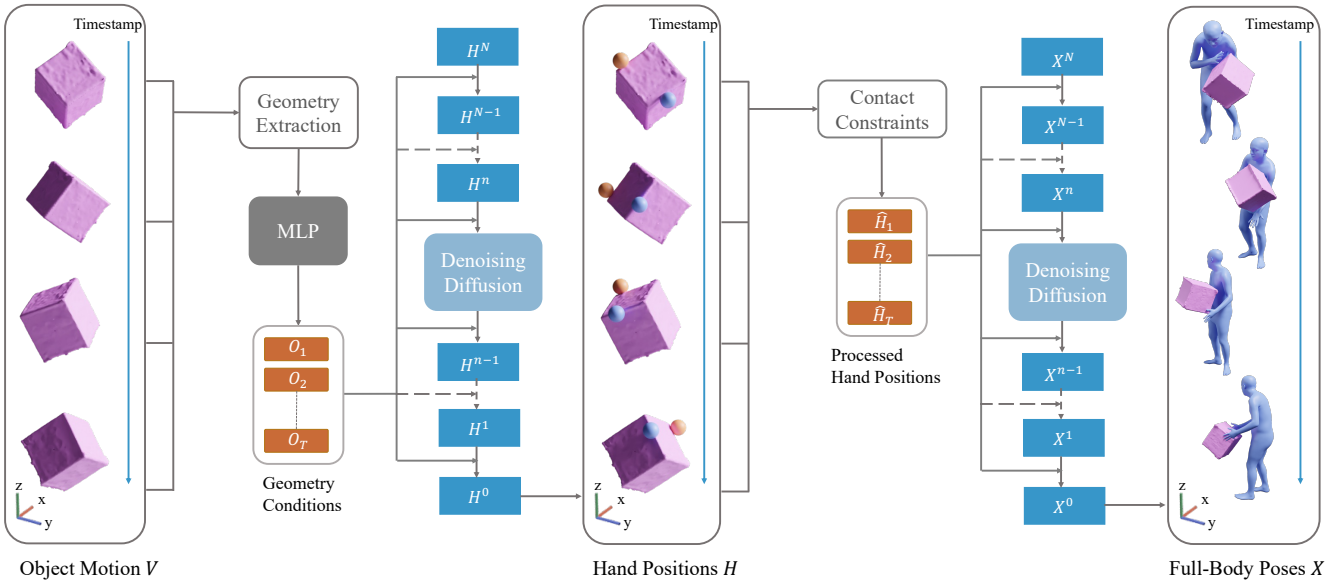


Fig. 2. Method Overview. Given a sequence of object geometry, we use BPS representation to encode geometry features and project the representation to a low dimensional vector at each time step using an MLP. We use conditional diffusion to synthesize hand joint positions and apply contact constraints. Then we feed the updated hand joint positions to our full-body synthesis module and produce human poses in contact with the given dynamic object.

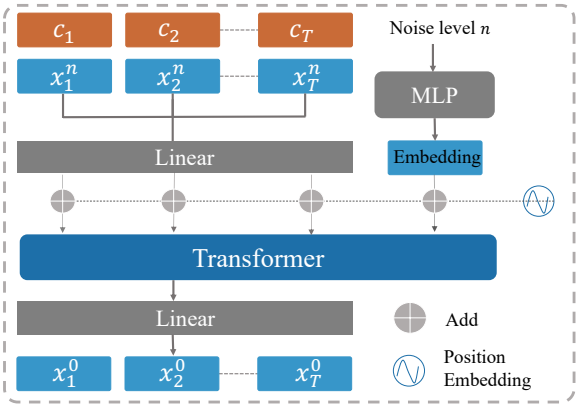


Fig. 3. Model architecture of denoising network. In stage 1, the conditions $\boldsymbol{c}$ are object geometry features $\boldsymbol{O}$, and $\boldsymbol{x}$ are hand joint positions $\boldsymbol{H}$. In stage 2, the conditions $\boldsymbol{c}$ are rectified hand joint positions $\hat{\boldsymbol{H}}$, and $\boldsymbol{x}$ are full-body human poses $\boldsymbol{X}$.

The transition of forward diffusion is modeled by a posterior distribution q. And each step is decided by a fixed variance schedule using β_n and is defined as

$$q(\boldsymbol{x}^n|\boldsymbol{x}^{n-1}) := \mathcal{N}(\boldsymbol{x}^n; \sqrt{1-\beta_n}\boldsymbol{x}^{n-1}, \beta^n \boldsymbol{I}), \tag{2}$$

where $\boldsymbol{I}$ represents identity matrix.

The reverse diffusion process is to generate desired data representation from random noise $\boldsymbol{x}^N \sim \mathcal{N}(0, \boldsymbol{I})$. This is achieved by learning a neural network p_θ to denoise recursively. Specifically, at noise level n, we use $\boldsymbol{c}$ to represent the conditions, and we have the reverse diffusion process represented as follows:

$$p_\theta(\boldsymbol{x}^{n-1}|\boldsymbol{x}^n, \boldsymbol{c}) := \mathcal{N}(\boldsymbol{x}^{n-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}^n, n, \boldsymbol{c}), \sigma_n^2 \boldsymbol{I}), \tag{3}$$

where $\boldsymbol{\mu}_\theta(\boldsymbol{x}^n, n, \boldsymbol{c})$ is the learned mean, σ_n is the fixed variance. $\boldsymbol{\mu}_\theta(\boldsymbol{x}^n, n, \boldsymbol{c})$ (we use $\boldsymbol{\mu}_\theta$ in the following equation for brevity) can be formulated as,

$$\boldsymbol{\mu}_\theta = \frac{\sqrt{\alpha_n}(1-\bar{\alpha}_{n-1})\boldsymbol{x}^n + \sqrt{\bar{\alpha}_{n-1}}(1-\alpha_n)\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}^n, n, \boldsymbol{c})}{1-\bar{\alpha}_n}, \tag{4}$$

where $\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}^n, n, \boldsymbol{c})$ is the prediction of x^0, α_n, $\bar{\alpha}_n$ are fixed parameters that satisfy $\bar{\alpha}_n = \prod_{i=1}^n \alpha_n$.

Learning the mean can be reparameterized as learning to predict the original data $\boldsymbol{x}^0$. We use reconstruction loss of $\boldsymbol{x}^0$ during training:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}^0, n}||\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}^n, n, \boldsymbol{c}) - \boldsymbol{x}^0||_1. \tag{5}$$

3.3 Our Pipeline

Generating Hand Positions from Object Geometry. In the first stage, we employ conditional diffusion to generate hand joint positions $\boldsymbol{H}_1, \boldsymbol{H}_2, ..., \boldsymbol{H}_T$ from object geometry features $\boldsymbol{O}_1, \boldsymbol{O}_2, ..., \boldsymbol{O}_T$. Here, the conditions $\boldsymbol{c}$ are represented by $\boldsymbol{O}_1, \boldsymbol{O}_2, ..., \boldsymbol{O}_T$. We adopt a transformer model architecture [Vaswani et al. 2017] as our denoising network which consists of four self-attention blocks. Each self-attention block contains a multi-head attention layer followed by a position-wise feedforward layer. As shown in Figure 3, we introduce an additional step to include noise level embedding as an input to our transformer model.

Apply Hand Contact Constraints. The hand joint positions generated in the initial stage may not always be precise. They may occasionally deviate from the object, resulting in perceived non-contact at certain time steps. To mitigate this, we propose a post-processing strategy based on the observation that human hands typically maintain consistent relative positions with respect to objects during contact.

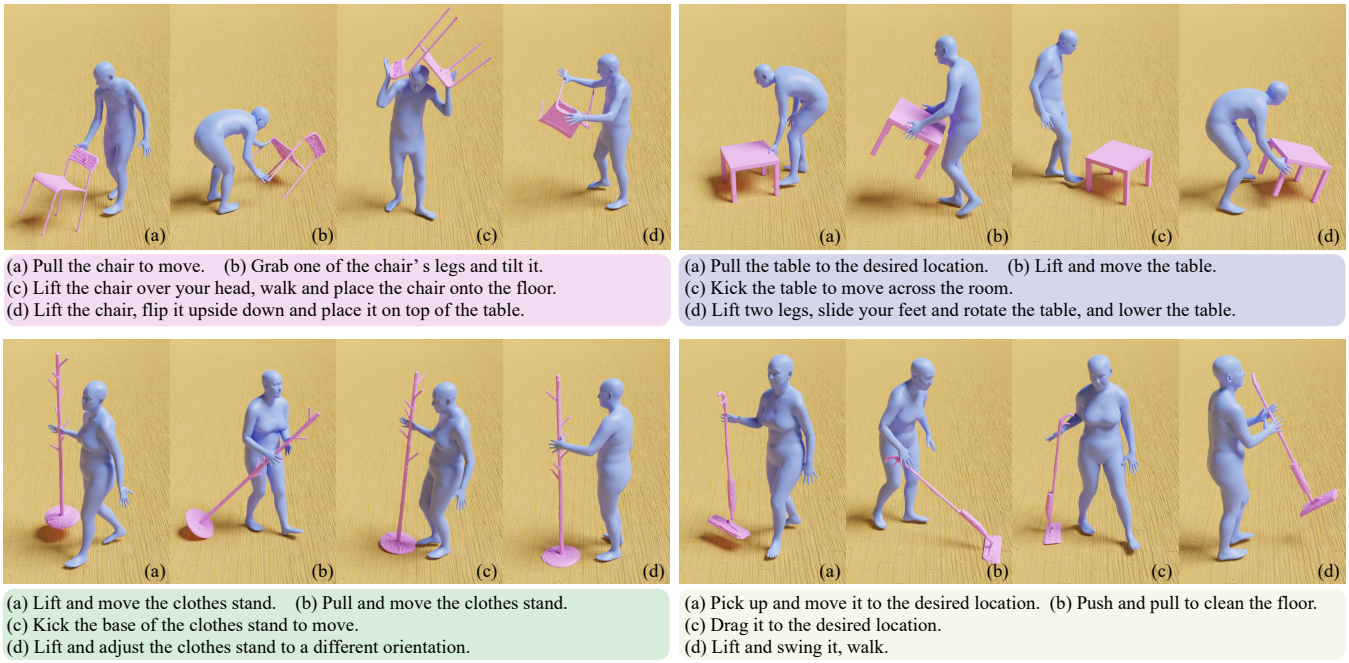


Fig. 4. Selected language descriptions used during our mocap sessions.

Given a sequence of hand joint positions $\boldsymbol{H}_1, \boldsymbol{H}_2, ..., \boldsymbol{H}_T$, we begin by computing the minimum distance from the hand joints to the corresponding object mesh $\boldsymbol{V}_1, \boldsymbol{V}_2, ..., \boldsymbol{V}_T$ at each time step, denoted as $d_1, d_2, ..., d_T$. We then traverse the sequence $d_1, d_2, ..., d_T$ starting from the first frame. We set an empirical contact threshold th = 0.03 and record a specific time step k where $d_k <$ th.

Next, we calculate the difference vector $\boldsymbol{p} = \boldsymbol{H}_k - \boldsymbol{V}_k^i$ at step k, where $\boldsymbol{V}_k^i$ is the nearest neighbor vertex of the hand joint on the object mesh. The difference vector $\boldsymbol{p}$, $\boldsymbol{p} \in \mathbb{R}^3$, is then used to compute updated hand joint positions in subsequent time steps. We denote the object rotation sequence as $\boldsymbol{R}_1, \boldsymbol{R}_2, ..., \boldsymbol{R}_T$, and for $t > k$, we compute the updated hand joint position as $\hat{\boldsymbol{H}}_t = \boldsymbol{V}_t^i + \boldsymbol{R}_t \boldsymbol{R}_k^{-1} \boldsymbol{p}$. This ensures the generated hand joint positions maintain a realistic, consistent contact with the object across the entire sequence. From the input object geometry, the first stage determines the joint positions for both the left and right hands. If the positions of both hands are in close proximity to the object, it results in a two-handed manipulation. If not, a single-handed manipulation is established. Specifically, close proximity is determined by computing the Euclidean distance between the hand position and its nearest neighbor points on the object mesh. If this distance is smaller than a predefined threshold (we empirically set the threshold to 0.03), it is inferred that there is contact.

Generating Full-body Poses from Hand Positions. In the second stage, we utilize the same denoising network architecture as in stage one to generate full-body poses from the hand joint positions. The conditions in this stage are the hand joint positions $(\hat{\boldsymbol{H}}_1, \hat{\boldsymbol{H}}_2, ..., \hat{\boldsymbol{H}}_T)$ that have been rectified using the contact constraints. The model is trained using human motion data only.

Table 1. Duration of 15 objects in our dataset.

Object	Large Table	Small Table	Monitor
Duration (min)	37	41	37
Object	Large Box	Small Box	Container
Duration (min)	40	37	39
Object	Wooden Chair	White Chair	Trashcan
Duration (min)	52	50	34
Object	Floor Lamp	Clothes Stand	Tripod
Duration (min)	35	38	42
Object	Mop	Vacuum	Suitcase
Duration (min)	42	43	40

By integrating these three components, we establish a complete pipeline to generate full-body poses from object motion. This pipeline models the one-to-many mapping from object motion to human poses and ensures that the generated poses maintain realistic contact with the object.

4 DATASET

We collected a large-scale high-quality dataset consisting of 3D object geometry, human and object motions. In this section, we elaborate on our object geometry acquisition, motion capture, and data processing.

Table 2. Quantitative evaluation on 15 objects.

Method	Hand JPE	MPJPE	MPVPE	T_{root}	O_{root}	Collision %	FS	C_{prec}	C_{rec}	F1 Score
GOAL	49.90	15.64	21.82	34.35	0.76	**0.12**	**0.18**	0.83	0.23	0.32
OMOMO-single-stage	26.60	**12.07**	**16.13**	**17.93**	**0.47**	0.19	0.38	0.78	0.42	0.51
OMOMO w/o constraints	24.79	12.55	16.66	18.62	0.51	0.21	0.36	**0.83**	0.58	0.64
OMOMO	**24.01**	12.42	16.67	18.44	0.50	0.22	0.38	0.82	**0.70**	**0.72**

Table 3. Quantitative evaluation on 5 unseen objects.

Method	Hand JPE	MPJPE	MPVPE	T_{root}	O_{root}	Collision %	FS	C_{prec}	C_{rec}	F1 Score
GOAL	53.97	15.27	20.92	40.53	0.78	**0.06**	**0.18**	**0.79**	0.21	0.29
OMOMO-single-stage	26.06	**12.33**	**16.67**	**19.49**	**0.51**	0.15	0.43	0.72	0.40	0.47
OMOMO w/o constraints	26.15	13.25	17.77	21.55	0.53	0.16	0.45	0.76	0.44	0.52
OMOMO	**25.12**	13.06	17.60	21.19	0.53	0.17	0.43	0.74	**0.58**	**0.61**

Object Geometry Capture. We selected 15 objects commonly used in everyday tasks, which include a vacuum, mop, floor lamp, clothes stand, tripod, suitcase, plastic container, wooden chair, white chair, large table, small table, large box, small box, trashcan, and monitor. For each object, we filmed a video circling the object and employed Luma [AI 2023] to reconstruct the 3D object geometry from this monocular video. We then utilized Meshlab to manually remove noisy points and downsample object meshes to contain a reasonable number of points for training.

Motion Capture. We utilized a Vicon system comprised of 12 cameras controlled by Vicon Shogun, which record at a rate of 120 FPS. For each object, we attached 5 markers and captured the object and human motion simultaneously. We invited 17 subjects (13 males, 4 females) to participate in our motion capture sessions. During each mocap session, the volunteer was provided with verbal instructions on how to interact with each object to avoid meaningless interactions. We show some examples of our language guidance in Figure 4. Each mocap session lasted approximately 1.5 to 2 hours. The total duration of captured motion for each object is shown in Table 1.

Data Processing. For the object geometry data, we employed a public python library [Kleineberg 2023] to compute the SDF for objects. In cases where objects contained noisy SDFs, we used SIREN [Sitzmann et al. 2020] to train neural networks and extract the SDF.

In terms of motion data processing, we used Mosh++ [Loper et al. 2014; Mahmood et al. 2019] to process our raw mocap files and extract SMPL-X model [Pavlakos et al. 2019] parameters for each sequence. In order to compute object transformations based on marker positions, we initially manually annotate the marker positions on the reconstructed object mesh. Subsequently, we utilize the analytical solution of the orthogonal Procrustes problem to compute the scale, rotation, and translation needed to align the annotated points with the marker positions. Furthermore, we visualize the object and human meshes, and conduct a manual verification on our collected dataset, discarding any sequences that fail to meet our high-quality standard.

Fig. 5. Training objects are annotated in blue, and testing objects are annotated in purple.

5 EXPERIMENT

We first introduce the dataset and evaluation metrics used for this task. Then we describe the chosen baselines and showcase comparisons against them. Additionally, we conduct an ablation study to investigate the effects of hand positions on overall performance. We encourage readers to watch our supplementary video for more qualitative evaluations.

5.1 Dataset and Evaluation Metrics

Dataset. We conduct all experiments using our collected dataset. This dataset consists of motion capture data from 17 subjects, with 15 subjects used for training and 2 subjects for testing. We adopt two data partitioning for evaluation. In the first setting, we use 15 objects for both training and testing. To further evaluate the model's generalization ability to new objects, we divide the 15 objects into 10 for training and 5 for testing as shown in Figure 5.

Evaluation Metrics. We evaluate the synthesized results from two perspectives. Firstly, we compare the generated poses against the ground truth motion data. Additionally, we assess the physical plausibility of the results, considering contact correctness, object penetration, and foot sliding. We detail our evaluation metrics as follows.

- **HandJPE**, **MPJPE** and **MPVPE** represent mean hand joint position errors, mean per-joint position errors, and mean per-vertex errors in centimeters (cm).
- T_{root} and O_{root} represent the root translation error computed using Euclidean distance in centimeters (cm) and orientation

(a) OMOMO-single-stage (b) OMOMO (c) GT

Fig. 6. Qualitative Results. We compare our single-stage model, our two-stage model with contact constraints, and ground truth motion. For more qualitative comparisons with GOAL, please watch our supplementary video.

error defined by the Frobenius norm of the difference between the 3 × 3 rotation matrix $||R_{pred}R_{gt}^{-1} - I||_2$.

- **FS** represents foot sliding metric and is computed following previous work [He et al. 2022].
- **Collision Percentage**. At time step t, for ith vertex on reconstructed human mesh, we query the object SDF and acquire a signed distance value d_t^i. We use a threshold (4cm) to compute collision. If there exists vertices that satisfy $d_t^i < 0, |d_t^i| > 4$, we increment the collision count. By traversing the sequence, we can compute the collision percentage.
- **Contact Metrics**. We adopt metrics precision (C_{prec}), recall (C_{rec}), and F1 score from the object detection task to evaluate contact performance. We first compute the distance between hand positions and object meshes. We empirically set a contact threshold (5cm) and use it to extract contact labels for each frame. We perform the same calculation for ground truth hand positions. Then we count true/false positive/negative cases to compute precision, recall, and F1 score.

5.2 Evaluations

Baselines. Since no existing work specifically addresses the task of object motion-guided human motion synthesis, we adapt a prior work GOAL [Taheri et al. 2022] on object-reaching motion synthesis as our baseline. GOAL proposed an autoregressive model that predicts future 10 frames conditioned on past 5 poses, hand distance between the current frame and the target goal frame, and the BPS representation which encodes hand-to-object distance at the target frame. In our problem setting, the input is a sequence of object geometry that guide the motion generation instead of a single target frame in GOAL. Thus, we make changes to the input features and use the next frame as the target frame. Specifically, the input features in our modified version consist of the past 5 poses, and the BPS representation that encodes the distance features between the current human mesh and the object mesh at the next frame. We use their default model setting which consists of four learning blocks.

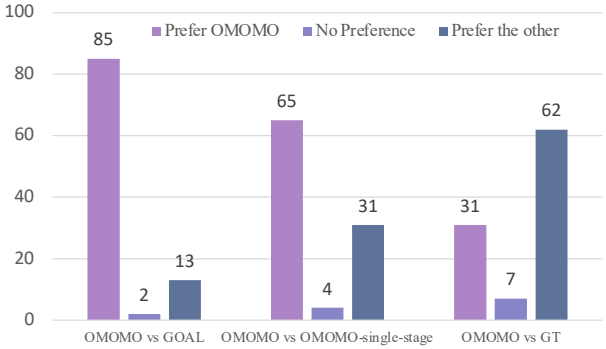


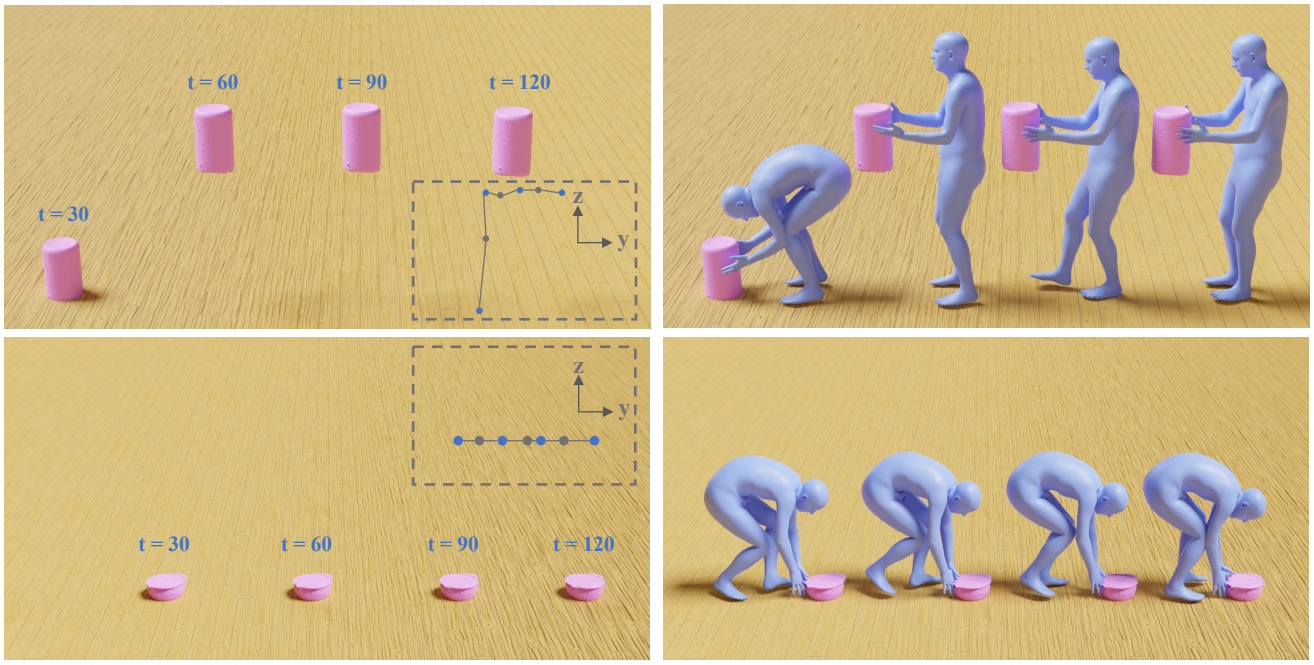
Fig. 7. Human Perceptual Study.

Each block contains a set of MLPs. The output dimension for each block is 2048, 1024, 1024, 2048 respectively.

Implementation Details. Our denoising network in OMOMO-single-stage, stage 1 model of OMOMO, and stage 2 model of OMOMO all consist of 4 self-attention blocks with 4 attention heads. The dimension of key, query, and value in the transformer architecture is 256. The output dimension of each layer is 512. Our implementation uses PyTorch [Paszke et al. 2019]. For training stage 1 and stage 2, we both use Adam optimizer [Kingma and Ba 2015] and start the training with a learning rate 0.0002. The training takes about 18 hours to converge for both stage 1 and stage 2 using a single NVIDIA Titan RTX GPU.

Results. Since our approach is based on conditional diffusion, there can be multiple plausible generation results given the same object motion. To make a quantitative comparison, we sample 20 times for the same object motion input and select the one with the smallest MPJPE. We show quantitative evaluations in Table 2 and Table 3 for two different data splits. One splits training and testing on all 15 objects. The other one uses 10 objects for training and the other 5 unseen objects for testing. For each configuration, there is only one random seed used to train our model, and the

Fig. 8. Examples of the generated motion sequences in human perceptual study.

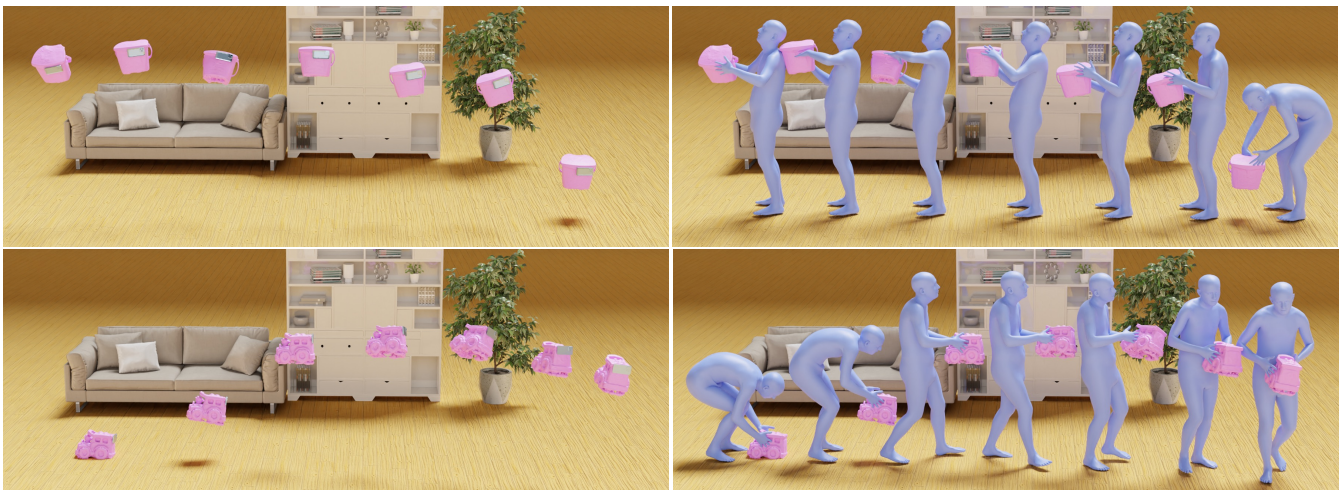


(a) Created Object Animation using Blender

(b) Synthesized Full-Body Human Motion

Fig. 9. We use Blender to create keyframes for objects with an interval of 15 frames and obtain a sequence of object geometry as input to our pipeline. We show the object geometry every 30 frames and the trajectory of keyframes on yz plane in (a). (b) shows the synthesized human motion.

statistics were computed using a single model. Note that our training is not sensitive to random seeds. We outperform baselines in both settings. In particular, OMOMO has superior results in terms of contact evaluations compared to the other two OMOMO variants, which demonstrate the effectiveness of our two-stage design and contact constraints. Note that the reason for the smaller collision percentage in GOAL is that the character in the baseline results often does not attempt to manipulate the object at all, hence the



(a) Object Motion Capture with iPhone (b) Synthesized Ful-Body Human Motion

Fig. 10. Application. We mount an iPhone on an object (shown in (a)) and use iPhone ARKit to capture object motion. (b) shows the synthesized human motion.

low collision percentage. As for smaller FS scores, we observed that the feet position in the baseline results usually drifts above the floor which will not be counted as foot sliding according to the foot sliding metric. In addition, it is worth mentioning that applying the contact constraints to GOAL is not straightforward. Since GOAL predicts all the joints' rotations, it requires inverse kinematics to rectify the human pose based on the corrected hand positions.

We also showcase qualitative results in Figure 6. OMOMO contains better contacts compared to the setting without hand joint positions as an intermediate representation, as evidenced by Figure 6 and higher contact F1 scores. For more qualitative comparisons, please watch our supplementary video.

Human Perceptual Study. We further conduct a human perceptual study to complement the evaluations. The goal is to evaluate the motion quality and contact realism. We random sample 100 generated sequences for each approach including OMOMO, OMOMO-single-stage, GOAL, and ground truth, covering all 15 objects. We show some generated results of each approach in Figure 8. We compare OMOMO and the other three settings and totally form 300 pairs for evaluation. For each question, we ask amazon mechanical turk workers which sequence looks more natural and interacts with objects more realistically. Each question is evaluated by 20 different workers (Figure 7).

We show that our OMOMO clearly outperforms the baseline GOAL and OMOMO-single-stage. And when compared with ground truth, 31% preferred our results (the upper bound would be 50%). It is worth noting that our results of OMOMO are produced via a single forward pass, without any optimization or post-processing for the full-body poses. Therefore, certain artifacts such as penetration may be produced in the generated motion, which results in ground truth motion is preferred in some sequences.

5.3 Ablation Study

To investigate the effects of hand positions on our overall performance, we compare the full-body human poses generation results that use the predicted hand joint position as input (OMOMO) and ground truth hand positions as input (OMOMO-GT). In Table 4, we show that the synthesis results can be further improved by feeding more accurate hand joint positions.

Table 4. Ablation Study. $*$ represents the setting that tests on unseen objects.

Method	MPJPE	T_{root}	C_{prec}	C_{rec}	F1 Score
OMOMO	12.42	18.44	0.82	0.70	0.72
OMOMO-GT	**7.01**	**10.08**	**0.89**	**0.77**	**0.79**
OMOMO*	13.06	21.19	0.74	0.58	0.61
OMOMO*-GT	**7.73**	**11.08**	**0.80**	**0.64**	**0.67**

5.4 Test on Manually Animated Object Trajectory

We further evaluated our pipeline using manually crafted animations of previously unseen objects. In this process, we began by reconstructing the 3D geometry of the object with the aid of Luma [AI 2023]. Once reconstructed, the 3D object was imported into Blender. Within Blender, we manually established keyframes at 15-frame intervals. Based on these keyframes, Blender then produced a complete object motion sequence. This sequence, exported from Blender, served as the input for our OMOMO. The resulting outputs are shown in Figure 9.

6 APPLICATION

We introduce our novel approach to capturing human motion interacting with objects using a single smartphone attached to the object. Specifically, we mount an iPhone XR on the target object and ask the subject to interact with the object while the iPhone camera is filming the environment. We leverage the API ARWorldTrackingConfiguration provided by iPhone ARKit to extract camera poses. This feature is based on visual-inertial odometry techniques that combine visual information and sensor information to estimate accurate camera pose in the world coordinate system. Since the

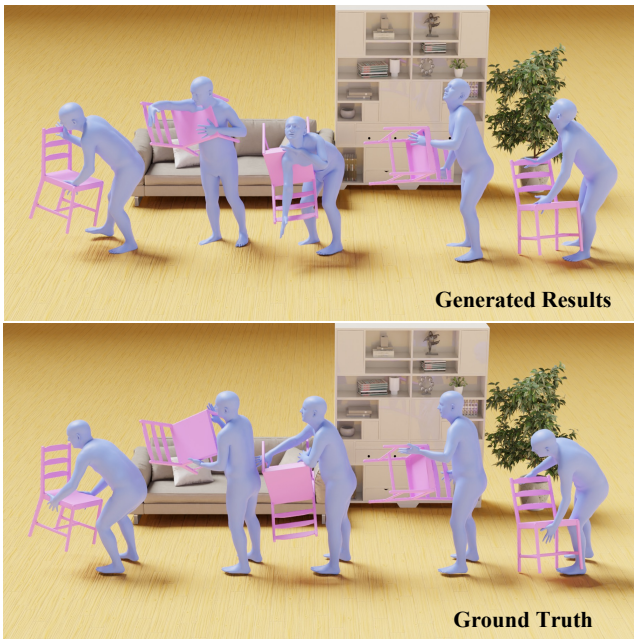


Fig. 11. Limitations. Our contact constraint cannot produce generations that involve intermittent contacts with the object. From top to bottom, we show the generated results and corresponding ground truth motion. In the generation results, the hand positions are processed to be fixed on the object, which introduces implausible human motions penetrating with objects.

camera is rigidly mounted on objects, we can derive object motion from camera poses. Similar to the data collection process, we film a video and use Luma [AI 2023] to reconstruct 3D geometry of the target object. From a sequence of object-moving geometries, we can generate full-body human poses with our proposed pipeline. We showcase some results in Figure 10. Note that these objects are not used during model training.

7 CONCLUSION

In summary, we presented a novel approach for synthesizing human motion guided by moving objects. Specifically, we proposed a framework based on a two-stage paradigm to enforce contact constraints, demonstrating its effectiveness in generating realistic human motions in interaction. Moreover, we introduced a novel application that enables capturing human interaction motion using a smartphone only. To facilitate the research on human-object interactions, we also introduced a large-scale dataset consisting of 3D object geometry, high-quality object motion, and human motion.

Limitations. Our current dataset falls short of accurately representing dexterous hand movements, which often result in implausible hand motions. A promising avenue for future research would be incorporating hand priors and optimization techniques, enhancing the realism of hand motions in our full-body pose generations. Furthermore, the contact constraints in our current framework cannot effectively address scenarios with intermittent contacts with the object as shown in Figure 11. This could be addressed by identifying and predicting contact states to enable the generation of more complex, long-term manipulation with the objects. Lastly, while our methodology is based on kinematics, future efforts could benefit from integrating physics-based components to mitigate the occurrence of artifacts.

ACKNOWLEDGMENTS

This work is in part supported by the Wu Tsai Human Performance Alliance at Stanford University, the Stanford Institute for Human-Centered AI (HAI), NSF CCRI 2120095, ONR MURI N00014-22-1-2740, the Toyota Research Institute (TRI), and Meta.

REFERENCES

Luma AI. 2023. *Capture 3D.* https://lumalabs.ai/

Joao Pedro Araujo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Deepak Gopinath, Jiajun Wu, Alexander Clegg, and C Karen Liu. 2023. CIRCLE: Capture In Rich Contextual Environments. In *Conference on Computer Vision and Pattern Recognition (CVPR).*

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. 2022. BEHAVE: Dataset and Method for Tracking Human Object Interactions. In *Conference on Computer Vision and Pattern Recognition (CVPR).* 15935–15946.

Yu-Wei Chao, Jimei Yang, Weifeng Chen, and Jia Deng. 2021. Learning to sit: Synthesizing human-chair interactions via hierarchical control. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. 2023. MoFusion: A Framework for Denoising-Diffusion-based Motion Synthesis. In *Computer Vision and Pattern Recognition (CVPR).*

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Computer Vision and Pattern Recognition (CVPR).* 5828–5839.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. 2023. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. In *Conference on Computer Vision and Pattern Recognition (CVPR).*

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. 2023. IMoS: Intent-Driven Full-Body Motion Synthesis for Human-Object Interactions. In *Eurographics.*

Vladimir Guzov, Julian Chibane, Riccardo Marin, Yannan He, Torsten Sattler, and Gerard Pons-Moll. 2023. Interaction Replica: Tracking human–object interaction and scene changes from human motion. In *arXiv.*

Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. 2021. Human POSEitioning System (HPS): 3D Human Pose Estimation and Self-localization in Large Scenes from Body-Mounted Sensors. In *Conference on Computer Vision and Pattern Recognition (CVPR).*

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. 2021. Stochastic Scene-Aware Motion Prediction. In *International Conference on Computer Vision (ICCV).* 11354–11364.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. 2019. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV).* 2282–2292.

Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. 2023. Synthesizing Physical Character-Scene Interactions. (2023), 1–9.

Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. 2022. Nemf: Neural motion fields for kinematic animation. *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), 6840–6851.

Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. 2023. Diffusion-based Generation, Optimization, and Planning in 3D Scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR).*

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR).*

Marian Kleineberg. 2023. *Mesh2SDF.* https://github.com/marian42/mesh_to_sdf

Jiye Lee and Hanbyul Joo. 2023. Locomotion-Action-Manipulation: Synthesizing Human-Scene Interactions in Complex 3D Environments. *arXiv preprint arXiv:2301.02667* (2023).

Jiaman Li, C Karen Liu, and Jiajun Wu. 2023. Ego-Body Pose Estimation via Ego-Head Pose Estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR).*

Ying Li, Jiaxin L Fu, and Nancy S Pollard. 2007. Data-driven grasp synthesis using shape matching and task-based pruning. *IEEE Transactions on Visualization and Computer Graphics* 13, 4 (2007), 732–747.

Libin Liu and Jessica Hodgins. 2018. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.

Matthew M. Loper, Naureen Mahmood, and Michael J. Black. 2014. MoSh: Motion and Shape Capture from Sparse Markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 33, 6 (Nov. 2014), 220:1–220:13. https://doi.org/10.1145/2661229.2661273

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*. 5442–5451.

Josh Merel, Saran Tunyasuvunakool, Arun Ahuja, Yuval Tassa, Leonard Hasenclever, Vu Pham, Tom Erez, Greg Wayne, and Nicolas Heess. 2020. Catch & carry: reusable neural controllers for vision-guided whole-body tasks. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 39–1.

Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. 2023. Generating Continual Human Motion in Diverse 3D Scenes. *arXiv preprint arXiv:2304.02061* (2023).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. 2018. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)* 37, 4 (2018), 1–14.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. 2021. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–20.

Sergey Prokudin, Christoph Lassner, and Javier Romero. 2019. Efficient learning on point clouds with basis point sets. In *International Conference on Computer Vision (ICCV)*. 4332–4341.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 652–660.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. 2020. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), 7462–7473.

Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. 2019. Neural state machine for character-scene interactions. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 209–1.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. 2019. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv:1906.05797* (2019).

Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. 2022. GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 13263–13273.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. 2020. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In *European Conference on Computer Vision (ECCV)*. https://grab.is.tue.mpg.de

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. 2023. Human Motion Diffusion Model. In *International Conference on Learning Representations (ICLR)*.

Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. 2023. EDGE: Editable Dance Generation From Music. In *Computer Vision and Pattern Recognition (CVPR)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 30.

Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. 2021a. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 9401–9411.

Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. 2021b. Scene-aware generative network for human motion synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 12206–12215.

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. 2022. HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. 2022. Saga: Stochastic Whole-Body Grasping with Contact. In *European Conference on Computer Vision (ECCV)*. 257–274.

Zhaoming Xie, Sebastian Starke, Hung Yu Ling, and Michiel van de Panne. 2022. Learning soccer juggling skills with layer-wise mixture-of-experts. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.

Zhaoming Xie, Jonathan Tseng, Sebastian Starke, Michiel van de Panne, and C Karen Liu. 2023. Hierarchical Planning and Control for Box Loco-Manipulation. (2023).

Yuting Ye and C Karen Liu. 2012. Synthesis of detailed hand manipulations using contact sampling. *ACM Transactions on Graphics (ToG)* 31, 4 (2012), 1–10.

He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. 2021. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)* 40, 4 (2021), 1–14.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022b. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *arXiv preprint arXiv:2208.15001* (2022).

Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. 2022c. EgoBody: Human Body Shape and Motion of Interacting People from Head-mounted Devices. In *European Conference on Computer Vision (ECCV)*. 180–200.

Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. 2022a. COUCH: Towards Controllable Human-Chair Interactions. In *European Conference on Computer Vision (ECCV)*. 518–535.

Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, , and Siyu Tang. 2023. Synthesizing Diverse Human Motions in 3D Indoor Scenes. In *International Conference on Computer Vision (ICCV)*.

Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, Karen Liu, and Leonidas Guibas. 2022. GIMO: Gaze-Informed Human Motion Prediction in Context. In *European Conference on Computer Vision (ECCV)*.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Computer Vision and Pattern Recognition (CVPR)*.