DiffGrasp: Whole-Body Grasping Synthesis Guided by Object Motion Using a Diffusion Model

Yonghao Zhang[1,2*], **Qiang He**[1,2*], **Yanguang Wan**[1,2], **Yinda Zhang**[3], **Xiaoming Deng**[1,2†], **Cuixia Ma**[1,2†], **Hongan Wang**[1,2]

[1]Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Google
{zhangyonghao2022, heqiang2022, wanyanguang2021, xiaoming, cuixia, hongan}@iscas.ac.cn,
yindaz@google.com

Abstract

Generating high-quality whole-body human object interaction motion sequences is becoming increasingly important in various fields such as animation, VR/AR, and robotics. The main challenge of this task lies in determining the level of involvement of each hand given the complex shapes of objects in different sizes and their different motion trajectories, while ensuring strong grasping realism and guaranteeing the coordination of movement in all body parts. Contrasting with existing work, which either generates human interaction motion sequences without detailed hand grasping poses or only models a static grasping pose, we propose a simple yet effective framework that jointly models the relationship between the body, hands, and the given object motion sequences within a single diffusion model. To guide our network in perceiving the object's spatial position and learning more natural grasping poses, we introduce novel contact-aware losses and incorporate a data-driven, carefully designed guidance. Experimental results demonstrate that our approach outperforms the state-of-the-art method and generates plausible results.

Project Page — https://iscas3dv.github.io/DiffGrasp/

Introduction

Capturing, synthesizing, and controlling human motion plays a key role in many areas such as animation, VR/AR, and robotics. However, the movement of the human body is complex, especially the movement of hands. Hands are used in almost every scene of our daily life, from interacting with small objects like playing with toys or picking up and inspecting a phone, to handling larger objects like moving boxes or using the laptop with both hands. Manipulating different objects also necessitates different involvement of both hands, and controlling the motion of two hands is incredibly complex and nuanced. Realistically modeling such intricate interactions can greatly benefit downstream applications.

Existing work on human object interaction has made significant progress in modeling whole-body motion sequences, leveraging natural language (Diller and Dai 2024; Song et al. 2024; Ghosh et al. 2023a), object motion keypoints (Li, Wu, and Liu 2023; Li et al. 2023b; Taheri et al. 2022), and initial frames (Xu et al. 2023; Kulkarni et al. 2023) to synthesize whole-body motion sequences or collaborative motion sequences. However, these works are not capable of generating satisfactory grasp sequences, especially for small objects because of the more intricate shapes of the object and the more flexible manipulations involved. In the domain of fine-grained grasping, GRIP (Taheri et al. 2023) and COOP (Zheng et al. 2023b) can generate natural single frame whole-body grasping results based on the 3D positions of different small objects. However, these works are unable to generate realistic natural sequence results due to the lack of temporal smoothness modeling.

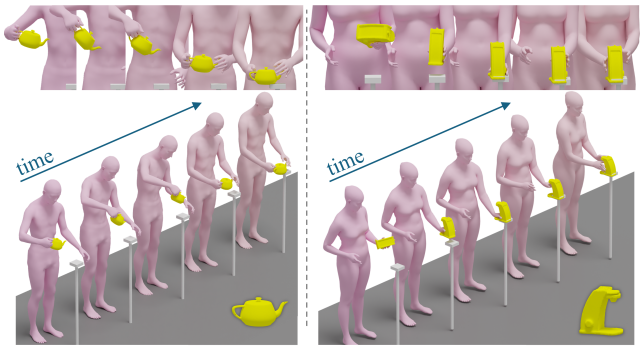

Figure 1: DiffGrasp generates whole-body human grasp sequence with realistic finger-object contact, conditioned on 3D object shape and object motion sequence.

For generating realistic whole-body human object grasp motion sequences, there are three key challenges: first, the range of motion of the body and the range of motion of the hands are at two different spatial scales in speed and fineness, resulting in great difficulty to accurately model both motions together within a single model;

second, when grasping objects of different sizes, shapes and motions, the strategies for using either a single hand or both hands can be distinct due to complex grasp pattern with twice degree-of-freedom of both hands; finally, once an object is grasped, the contacting hand must maintain stable

*These authors contributed equally.
†indicates corresponding author.

contact and avoid penetrating the object.

In this work, we propose DiffGrasp, a novel whole-body framework for generating fine-grained grasping sequences with both hands conditioned on an object motion sequence. A natural result of human interaction needs to consider both the shape of the object and the movement of the object to determine the involvement of both hands in grasping. Our method can generate natural, stable, and realistic whole-body grasping results for different object shapes while also maintaining coordinated movements of other body parts. To model the complex movements of the object, hands and body jointly, we employ a novel conditional diffusion model to learn the joint distribution of the human-object motion space. This also avoids the network degradation issue that we observed in our experiments with existing methods, which may occur due to stacking of multiple diffusion stages. To guide the network in learning complex motion patterns of the human body due to the high degrees-of-freedom in body with fingers, we propose two novel contact-aware losses, specifically for the hands, for training. The network can perceive the position of the object and generate realistic grasping results. Finally, to enhance the realism of grasping, we introduce a data-driven guidance term during the inference stage to maintain contact stabilization, while other terms can encourage contact and prevent penetration.

The main contribution of our method can be summarized as follows. 1) To the best of our knowledge, we propose the first diffusion-based framework to synthesize life-like whole-body human motion sequence with realistic finger-object contact, conditioned on 3D objects motion sequence. 2) We use a single diffusion model to learn the motion patterns between the hand and the object, while proposing two contact-aware losses that effectively guide the network to generate natural results that are also aware of the object's spatial position. 3) We propose a novel data-driven guidance strategy to prevent sliding between the grasping hand and the object, while other guidance strategies can achieve more stable contact and avoid penetration of the object surface. 4) Extensive experiments demonstrate that our proposed method outperforms the state-of-the-art method.

Related Work

Human Motion Generation. Generating human-like motions is a fundamental problem of artificial intelligence and has gained significant attention in recent years. Previous studies have demonstrated the effectiveness of the Variational Autoencoder (VAE) formulation in generating diverse human motions (Petrovich, Black, and Varol 2022; Guo et al. 2022; Lee, Moon, and Lee 2023; Lucas et al. 2022). More recent work (Liu et al. 2024; Zhang et al. 2024b; Chen et al. 2023; Ao, Zhang, and Liu 2023; Tseng, Castellon, and Liu 2023; Li et al. 2023c) has employed the diffusion model to generate motion sequences conditioned on various control signals, such as text, speech, music, etc. However, these control signals are generally ambiguous and the final motion sequences may not be well aligned with the user's real intention. In addition, several recent studies (Pinyoanuntapong et al. 2024; Li et al. 2023c; Lu et al. 2023) have explored the synthesis of whole-body human motion. These studies have typically adopted an independent modeling approach for hand parts and body parts, with the resulting outputs combined in a fusion step.

Hand Grasp Generation. There has been in-depth research on the problem of grasping with respect to given object poses, in the field of robotics (Sahbani, El-Khoury, and Bidaud 2012). With the emergence of human-object and hand-object interaction datasets (Zhang et al. 2021; Fan et al. 2023; Kwon et al. 2021; Taheri et al. 2020), numerous works have arisen to generate single-frame hand grasp poses with contact map priors of objects (Jiang et al. 2021; Li et al. 2023a; Liu et al. 2023). For hand grasp sequence generation, D-Grasp (Christen et al. 2022) synthesizes diverse dynamic sequences with the in-hand objects. Text2HOI (Cha et al. 2024) performs this task by first generating the contact map and then the sequence of hand-object interaction. ArtiGrasp (Zhang et al. 2024a) can generate physically plausible bimanual grasping. In terms of modeling the hand-object interaction for better modeling a reasonable hand grasping pose, some approaches (Zhou et al. 2022; Zheng et al. 2023a; Luo, Liu, and Yi 2024; Liu and Yi 2024) are proposed. These works achieve good results in hand grasp generation, but do not tackle whole-body grasp generation tasks. The generated hands positions also may not necessarily produce natural poses for the body.

Human Object Interaction Generation. Existing works on generating interactions between human body and objects can be categorized based on the given object. Interacting with scene objects (Cen et al. 2024; Kulkarni et al. 2023; Huang et al. 2023; Zhao et al. 2023) refers to interact with objects in the environment that cannot be moved by a person. In the realm of interactions with larger objects (Song et al. 2024; Diller and Dai 2024; Li et al. 2024; Li, Wu, and Liu 2023; Li et al. 2023b; Xu et al. 2023; Peng et al. 2023), most of these related works focus on reconstructing well-defined torso motion sequences with text guidance often provided as a prompt (Peng et al. 2023; Li et al. 2023b), while overlooking hand poses. For tasks involving interactions with smaller objects, existing works usually generate fine-grained single-frame grasping poses (Taheri et al. 2022; Wu et al. 2022; Tendulkar, Surís, and Vondrick 2023; Zheng et al. 2023b), while sequence generation tends to adopt methods by which hand motion guides object motion to achieve perceptually plausible results (Ghosh et al. 2023a; Xu et al. 2023; Li et al. 2024). In summary, previous work has not been able to generate realistic whole-body human motion sequences while maintaining a fine hand-object grasping posture.

Method

DiffGrasp generates whole-body human motion sequence with realistic finger-object contact, mainly conditioned on 3D objects motion sequence and human identity. The overview of DiffGrasp is shown in Fig. 2. We will introduce our model in three parts: data representation, diffusion model and contact-aware loss, and guidance.

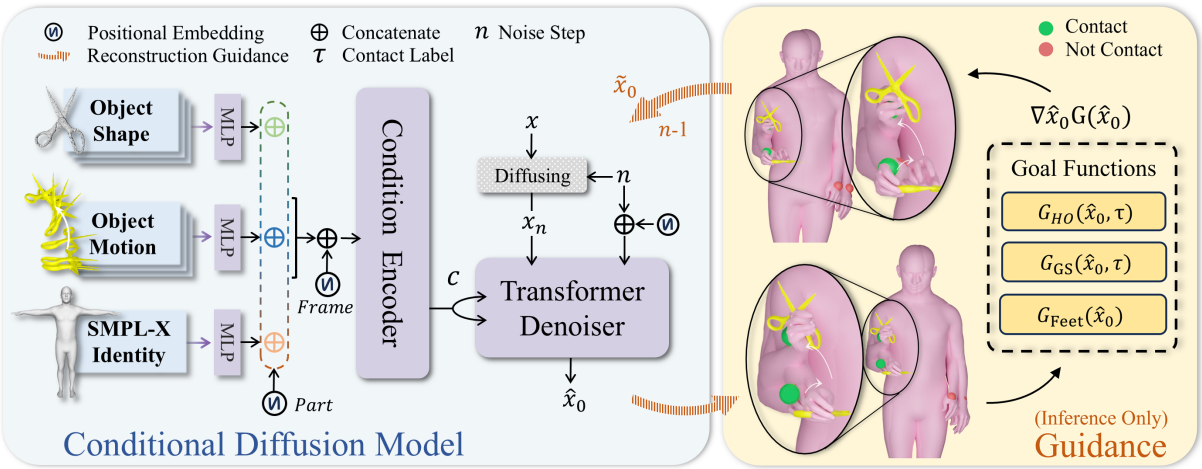

Figure 2: *Overview of DiffGrasp Framework.* In our conditional diffusion model, we use the given sequence of object motion, object shape and the SMPL-X identity as conditions. After specially designed positional encodings, these embedded conditions are inputted into a transformer-encoder-based condition encoder. Then, a transformer decoder as denoising network predicts a sequence of clean whole-body pose of SMPL-X as well as the wrist joints translations relative to the object centroid. During the inference stage, we reconstruct the SMPL-X pose sequence into a human mesh sequence. Based on carefully designed guidance functions, we control and optimize our predicted results for more stable hand grasping ($\mathcal{G}_{GS}$), less penetration ($\mathcal{G}_{HO}$) and better foot-floor contact ($\mathcal{G}_{Feet}$) through reconstruction guidance strategy.

Data Representation

Human Motion Representation. We denote the generated human whole-body pose as $H \in \mathbb{R}^{T\times D}$, where T and D represent the sequence length and dimension of the human pose. In frame t, the human pose H_t consists of global translation and global orientation of the root joint, and the body pose using the 6D continuous representation(Zhou et al. 2020). We use SMPL-X(Pavlakos et al. 2019) to represent the human body and use its shape β and gender G to present *Human Identity* $S_{id} = \{\beta, G\}$.

Object Sequence Representation. The object motion, which is one of our conditions, is represented by two components: the global 3D motion sequences for *Object Motion* and the geometry information for *Object Shape*. The motion of the object in each frame is represented by the centroid of the object and the global rotation and translation of the object, denoted as $W_t \in \mathbb{R}^{12}$. For the shape of the object, following previous work (Taheri et al. 2022; Zheng et al. 2023b; Li, Wu, and Liu 2023), we represent the geometry of the object using the Basis Point Set (BPS) representation (Prokudin, Lassner, and Romero 2019). In each frame t, we calculate the 3D mesh BPS representation of the object, denoted as $V_t \in \mathbb{R}^{1024\times 3}$.

Condition Input. At each frame t, we first use three MLPs to map each object shape V_t, object motion W_t and human identity S_{id} to three 256 dim features $\mathcal{V}_t, \mathcal{W}_t$ and $\mathcal{S}_{id}$. Inspired by (Cha et al. 2024), in addition to the general frame-wise positional encoding PE_t^f, we introduce the part-wise positional encoding PE^p to provide a more detailed differentiation of the frames and the condition part. The final sequence of input conditions can be formulated as $c_t^{raw} = PE_t^f(PE_V^p(\mathcal{V}_t) \oplus PE_W^p(\mathcal{W}_t) \oplus PE_s^p(\mathcal{S}_{id})) \in \mathbb{R}^{256\times 3}$.

More details of our positional encoding strategy can be found in the supplementary materials.

Conditional Diffusion Model

Model Architecture. Fig. 2 shows our model architecture. We adopted a basic full Transformer (Vaswani et al. 2017) architecture to fulfill our sequence-to-sequence generation task. Our network includes a transformer encoder as a condition encoder to encode the condition sequence c^{raw} into condition features c, and a transformer decoder as a denoiser, which maps the noised motion sequence data x_n and the noise step n to predict the clean data x_0 conditioned on c.

Denoiser Outputs. The conditional diffusion model aims to learn the latent correspondence between input conditions c and the generated whole-body human motion sequence H. To get more accurate 3D Euclidean space awareness for our guidance stage, our model also predicts translations of both hands wrist joint $\kappa \in \mathbb{R}^{T\times 6}$ relative to the condition object center O_m. The generated result is denoted as $X = \{H, \kappa\}$.

Conditional Diffusion Loss. The diffusion model comprises both a forward (noising) process and a reverse (denoising) process. The forward process involves gradually introducing noise to the initial data representation x_0 over N steps, implemented through a Markov chain formulation,

$$q(x_n|x_{n-1}) = \mathcal{N}(x_n; \sqrt{1-\beta_n}x_{n-1}, \beta_n I) \quad (1)$$

$$q(x_{1:N}|x_0) := \sum_{n=1}^{N} q(x_n|x_{n-1}) \quad (2)$$

where β_n is a fixed variance schedule and I is an identity matrix. For our method, the goal is to learn a conditional

diffusion model f_θ to reverse the noising diffusion process,

$$f_\theta(x_{n-1}|x_n, c) := \mathcal{N}(x_{n-1}; \mu_\theta(x_n, n, c), \Sigma_n) \quad (3)$$

where μ_θ denotes the predicted mean and Σ_n is a fixed variance. The process of learning the mean μ_θ can be optimized by reconstructing clean data x_0 following existing motion generation methods (Tevet et al. 2023; Shafir et al. 2023). The model f_θ is optimized by the objective:

$$\mathcal{L}_{diff} = \mathbb{E}_{n\sim[1,N]} \| f_\theta(x_n, n, c) - x_0 \|_2^2 \quad (4)$$

Contact Label. To independently represent the contact relationship of each hand with the object, we design a binary hand-object mask $\tau \in \{0, 1\}^{T\times 2}$, named *contact label*. The two components of τ_t are represented by 1 or 0, respectively, indicating whether the left hand or the right hand are in contact with the object at frame t. Contact label is calculated by thresholding the minimum distance between each hands and the mesh vertices of a given object geometry. During inference, the contact label is obtained by calculating whether the relative distance κ between the generated wrist and object is less than a given threshold.

Contact-aware Losses. Since the SMPL-X pose representation is on SE(3), and modeling human interacting with objects requires explicit representation in the 3D Euclidean space, we additionally propose two contact-aware loss functions for training based on contact label τ to provide more clues in the 3D Euclidean space for our network.

Contact-aware Reconstruction Loss enforces the generated hand joints and wrist joints to be close to ground truth. After the network predicts $\hat{X} = \{\hat{H}, \hat{\kappa}\}$, we use the pose parameters $\hat{H}$ to reconstruct all joints of the left hand and right hand $\hat{J}$ using the forward process of SMPL-X, and we denote left hand joints and right hand joints as $\hat{J}_l$, and $\hat{J}_r$, respectively. Furthermore, by adding the 3D position of the object centroid with the translation of the predicted wrist, we reconstruct the positions of the hand wrist joint, denoted as $\hat{v}_l$ and $\hat{v}_r$ by $\hat{v} = O_m + \hat{\kappa}$, where O_m is the centroid of the object shape. Then, the contact-aware reconstruction loss can be formulated as:

$$\begin{aligned}\mathcal{L}_{recon} = \tau_0(\|J_l - \hat{J}_l\|_2 + \lambda_{wrist}\|v_l - \hat{v}_l\|_2) \\ + \tau_1(\|J_r - \hat{J}_r\|_2 + \lambda_{wrist}\|v_r - \hat{v}_r\|_2)\end{aligned} \quad (5)$$

where τ^0 and τ^1 are the contact label of left and right hands, respectively, and λ_{wrist} is the balance weight of wrist terms.

Contact-aware Interaction Loss can ensure an accurate spatial position of the hands with respect to the object. In order to make the model more sensitive to the proximity of both hands to the object, we introduce exponentially decaying distance-aware interaction weights w_k for each hand joints k, inspired by (Ghosh et al. 2023b):

$$w_k = \tau \exp\left(-\alpha \cdot d(J_k, O_m)\right) \quad (6)$$

where $d(\cdot, \cdot)$ represents the per-joint Euclidean distance, and α is a scalar weight. Thus, the contact-aware interaction loss is defined as:

$$\mathcal{L}_{inter} = \sum_{k=1}^{K} w_k \|d(\hat{J}_k, O_m) - d(J_k, O_m))\|_2 \quad (7)$$

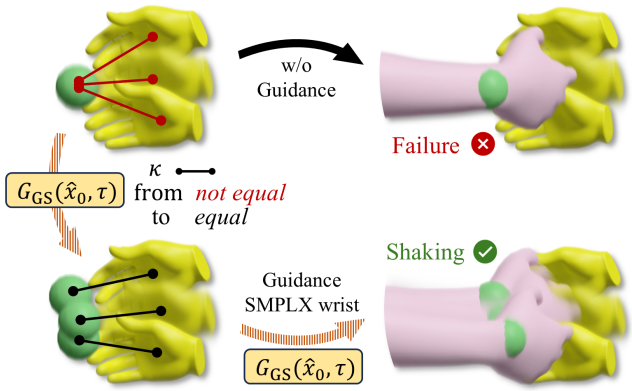

Figure 3: *Illustration of Grasp Stabilization Guidance* $\mathcal{G}_{GS}$. 'Handshaking' object movement example: Initially, the generated hand-object relative distance κ and the reconstructed hand do not follow the object's (yellow hand) shaking well. We stabilized the hand-object relative distance according to Eq. (10) to obtain the wrist position that follows the object's shaking, and then guided the reconstructed wrist position to successfully achieve the handshaking effect.

In total, we train our DiffGrasp to minimize a weighted sum of three loss terms: the diffusion loss, the contact-aware reconstruction loss and the contact-aware interaction loss:

$$\mathcal{L} = \lambda_{diff}\mathcal{L}_{diff} + \lambda_{recon}\mathcal{L}_{recon} + \lambda_{inter}\mathcal{L}_{inter} \quad (8)$$

where λ_{diff} and $\lambda_{contact}$ are assigned scalar weights to balance the individual losses.

Guidance

In human object interaction field, direct constraints between the human body mesh and the object mesh are crucial for generating realistic results. However, during the training stage, we do not explicitly impose loss constraints between the human body and the object meshes. This is because we found that, first, DiffGrasp, constrained by the aforementioned losses, can already generate reliable results; second, introducing explicit penetration and contact losses during training would significantly increase computational costs and training time. However, these factors do not significantly improve model generalization and visual realism. Therefore, we only explicitly optimize the interaction during the inference sampling process, because it would better balance the constraints and enhance visual realism by fitting the observed object motions.

Reconstruction Guidance. To align the model's inference sampling results more with a specific condition, the diffusion model often uses an optimization strategy called classifier guidance (Ho, Jain, and Abbeel 2020). This strategy can be formulated as $\tilde{\mu} = \mu - \eta\Sigma_n\nabla_\mu\mathcal{G}(\mu)$, where μ is the result of the denoised step n defined by Eq. (3), $\mathcal{G}$ is a goal function that determines a gradient direction that aligns better with the condition, and η is a scalar weight. Similarly, we optimize the reconstructed results, known as reconstruction guidance (Ho et al. 2022), which can be formulated as,

$$\tilde{x}_0 = \hat{x}_0 - \eta\Sigma_n\nabla_{\hat{x}_n}\mathcal{G}(\hat{x}_0) \quad (9)$$

where x_0 represents the clean data predicted by the network, in our work it corresponds to the SMPL-X human parameters $\hat{H}$ and the relative object translation of two wrists $\hat{J}$. We employ reconstruction guidance to gradually guide the optimization of the predicted results by the network. Through reconstructing $\hat{H}$, we obtain the human body mesh surface. We will detail our goal function $\mathcal{G}$ in the following.

Grasp Stabilization Guidance. Instead of hand-crafted constraints, we use data-driven constraints generated from our model. We introduce the Grasp Stabilization Guidance, which is motivated by a simple observation: *when a hand successfully grasps an object, it should not exhibit any relative sliding motion with respect to the object.* This guidance is especially helpful in optimizing the generated results for high-frequency, rapid object movements. Without loss of generality, we consider the generated relative distance $\hat{\kappa}_l$ between the left wrist and the object as an example.

Given an sequence of left wrist relative distance $\hat{\kappa}_l^0, \hat{\kappa}_l^1, \ldots, \hat{\kappa}_l^n$, its sequence of contact label $\tau_0^0, \tau_0^1, \ldots, \tau_0^n$ obtained by calculating whether the relative distance $\hat{\kappa}_l$ is less than a specific threshold, assuming the temporal segment of contact is from i to j, where $0 \leq i < j \leq n$, our aim is to ensure that the position of the wrist relative to the object in the i-th frame does not change in the next $j - i$ frame. For any frame k in the next $j - i$ frame, the corrected wrist position in the world coordinate system is calculated by the following formula:

$$\acute{v}_l^k = (\hat{\kappa}_l^i + O_m - W_T^i) {W_R^i}^{\mathrm{T}} W_R^k + W_T^k \tag{10}$$

where W_R^k and W_T^k is the object rotation and translation at frame k. We demonstrate this correction in Fig. 3. Next, we use $\hat{H}$ to reconstruct human wrist joints $\hat{v}$, and then constrain them to optimized wrist joints. Grasp stabilization guidance can be defined as

$$\mathcal{G}_{GS} = \tau \|\hat{v} - \acute{v}\|_2 \tag{11}$$

$$\tilde{x}_0^{up} = \hat{x}_0^{up} - \eta \Sigma_n \nabla_{\hat{x}_n^{up}} \mathcal{G}_{\mathcal{GS}}(\hat{x}_0, \tau) \tag{12}$$

Using contact labels τ, only the contact hand needs to be optimized. We optimize only the upper body parameters $\tilde{x}_0^{up}$ in experiments.

Hand-Object Contact Guidance. To reduce the penetration between the reconstructed human mesh and the object mesh, and to encourage contact between the hands and the object, we propose hand-object contact guidance. To expedite the computation process, we followed the work (Hasson et al. 2019) to sample only $|v_s|$ vertices on hands with the highest contact rates for penetration and contact distance calculations. Penetration distance D_{pene} and contact distance D_{cont} can be calculated as:

$$D_{pene} = \sum\nolimits_{h=1,\ldots,\hat{v}_s} -\min\{\mathrm{sdf}(\hat{v}_s[h]), 0\} \tag{13}$$

$$D_{cont} = \sum\nolimits_{h=1,\ldots,\hat{v}_s} \tau \cdot \mathrm{abs}(\mathrm{sdf}(\hat{v}_s[h])) \tag{14}$$

We compute the signed distance between the sampling points v_s on the hand and the object represented with signed distance fields (SDF), and simultaneously calculate the penetration distance and the contact distance. Then, we can define the hand-object contact guidance as follows:

$$\mathcal{G}_{HO} = \lambda_{ho} D_{pene} + (1 - \lambda_{ho}) D_{cont} \tag{15}$$

$$\tilde{x}_0^{hand} = \hat{x}_0^{hand} - \eta \Sigma_n \nabla_{\hat{x}_n^{hand}} \mathcal{G}_{HO}(\hat{x}_0, \tau) \tag{16}$$

During optimization, only the contact hand can be optimized. We only optimize the parameters of the hands $\tilde{x}_0^{hand}$.

Feet Penetration Guidance. We use feet penetration guidance that encourages human feet-floor contact:

$$\mathcal{G}_{Feet} = \sum\nolimits_{\hat{v}[z]<0} \mathrm{abs}(\hat{v}[z]) \tag{17}$$

$$\tilde{x}_0 = \hat{x}_0 - \eta \Sigma_n \nabla_{\hat{x}_n} \mathcal{G}_{Feet}(\hat{x}_0) \tag{18}$$

We use these three goal functions to explicitly guide the diffusion in generating results that better match the object motion condition. The goal function can be defined as:

$$\mathcal{G} = \mathcal{G}_{Feet} + \mathcal{G}_{GS} + \mathcal{G}_{HO} \tag{19}$$

Experiments

In this section, we first describe the dataset and evaluation metrics used in our experiment. Next, we present comparison experiments with the baseline method. Finally, we conduct ablation studies to evaluate the effectiveness of our method.

Datasets and Evaluation Metrics

Datasets. We use GRAB (Taheri et al. 2020) and ARCTIC (Fan et al. 2023) to conduct our experiment, which collects full-body hand-object interaction mesh sequences. Each dataset consist of 10 subjects, each grabbing and manipulating a number of different objects. We follow the conventional approach to divide the training and validation sets by 8 subjects used for training and 2 subjects for testing. To further evaluate our model's generalization ability to unseen objects, we exclude five objects from the GRAB training dataset for testing. In ARCTIC, we use the ground truth of the articulation angles of objects.

Evaluation Metrics. We use a set of evaluation metrics to this novel task, building upon established evaluation methods in the existing literature (Li, Wu, and Liu 2023; Taheri et al. 2022; Li et al. 2023b; He et al. 2022).

Motion Quality Metrics. *HandJPE* and *MPJPE* represent mean hand joint position errors (cm) and mean per-joint position errors (cm), respectively. *MPVPE* represents mean per-vertex errors (cm), and *FS* represents foot sliding metric.

Hand Collision Metrics. *Collision Percentage* and *Collision Depth* are used to evaluate the extent to which the vertices of the hands penetrate the surface of an object. At frame t, we employ the signed distance field of object to judge whether a hand vertex is located within the mesh and to calculate the distance between the vertex and the mesh surface. If the vertex is inside the mesh and the distance is below a specified threshold (5mm), we increase the collision count and record the collision depth. By iterating through the sequence, we can compute the collision percentage and the mean collision depth.

Dataset	Method	Hands JPE↓	MPJPE↓	MPVPE↓	FS↓	Coll. %↓	C depth ↓	F1↑	Cont dist↓
GRAB	OMOMO	31.28	17.57	13.80	1.05	0.0014	0.0007	0.0090	0.19
	OMOMO-V2	34.91	19.47	15.31	2.63	0.0007	0.0009	0.0641	0.25
	OMOMO-V3	32.72	17.45	13.75	**1.03**	0.0004	0.0001	0.1028	0.15
	DiffGrasp	**20.99**	**12.24**	**10.09**	2.22	0.0023	0.0001	**0.7840**	**0.04**
ARCTIC	OMOMO	25.95	11.68	8.53	**0.64**	0.0001	0.0001	0.0775	0.11
	OMOMO-V2	26.91	11.77	8.50	0.93	**0.0000**	**0.0000**	0.2771	0.12
	OMOMO-V3	26.57	11.91	8.74	0.65	**0.0000**	0.0001	0.3338	0.10
	DiffGrasp	**19.96**	**11.56**	**8.00**	1.25	0.0030	0.0001	**0.8067**	**0.04**

Table 1: *Comparative experimental results* on GRAB (Taheri et al. 2020) dataset and ARCTIC (Fan et al. 2023) dataset.

Hand Contact Metrics. In order to evaluate the effectiveness of hand contact, we employ the *F1 score* metric commonly used in object detection tasks. To obtain a more accurate measure of grasping, we calculate the mean *Contact distance* (cm) between the positions of the fingers and the object meshes. We empirically define a contact threshold of 5mm and use it to determine the contact labels for each frame. The same calculation is performed for the ground truth hand positions. Subsequently, we tally the true/false positive/negative cases to compute the F1 score.

Evaluations

Baselines. We compare our DiffGrasp with OMOMO (Li, Wu, and Liu 2023), a two-stage framework that generates whole-body motion sequences without the finger pose, conditioned on the object motion sequence. It employs two conditional diffusion models: the first generates hand positions conditioned on object geometry features, and the second generates whole-body poses based on the predicted hand positions. To facilitate a comprehensive comparison, we trained three versions of OMOMO based on its basic network architecture, namely OMOMO, OMOMO-V2, and OMOMO-V3. OMOMO follows the original settings, using the object motion trajectory and the BPS representation as the first-stage network inputs, which generates hand positions. These positions then condition the prediction of body pose in the second stage. OMOMO-V2 extends OMOMO by including the parameters of hand pose as additional outputs in its second model. OMOMO-V3 is built on the architecture of OMOMO, with an additional network that predicts hand pose parameters conditioned on the object motion trajectory, the BPS representation, and the hand positions. The additional network can be treated as predicting the hand pose parameters in its first hand positions generation model. Comparisons and results with more related work can be found in our supplementary materials.

Results. Table 1 shows that our approach outperforms all versions of the baseline. By comparing the *F1 scores* and *Hand Collision Percentage* between DiffGrasp and OMOMO, we can observe that our work can model and ensure more continuous and precise grasping. By comparing with OMOMO-V2, we see that jointly modeling the hand pose and body pose based on the baseline yields less dynamic mean poses. And qualitative results in our supplementary materials show that compared with modeling hand pose failure simultaneously leads to a degradation in its ability to model body pose. Although OMOMO-V3 achieves better results in terms of fine and continuous grasping compared to OMOMO and OMOMO-V2 , it still performs worse than DiffGrasp. OMOMO and OMOMO-V3 have smaller FS because the generated character body pose tends to be static. As shown on the right side of Fig. 4, DiffGrasp achieves a stooping motion while the baseline remains upright.

Ablation Study

We conduct ablation studies on contact-aware losses, frame-wise positional encoding PE^P and our guidance strategy. The experimental results are shown in Table 2 and Fig. 5.

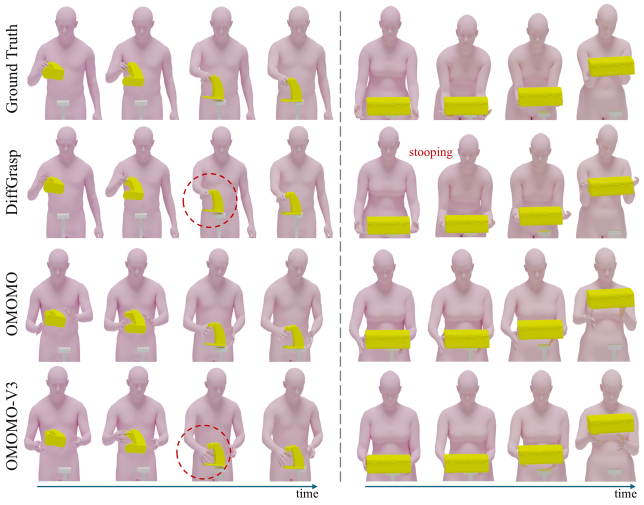

Figure 4: *Qualitative Results of Comparison Experiments.* Our model (DiffGrasp) generates more realistic results, with more hand-object contact and less penetration.

Effect of Contact-aware Losses. Ablation studies on *the contact-aware reconstruction loss* show that starting from a version with only the diffusion loss (Full loss w/o Inter and Recon), progressively adding the simple reconstruction loss (Full loss w/o Inter w/ Simp Recon), which does not

Method	Hands JPE↓	MPJPE↓	MPVPE↓	FS↓	Coll. %↓	C depth ↓	F1↑	Cont dist↓
Full loss w/o Recon	25.49	13.69	11.07	2.89	**0.0001**	0.0001	0.5543	0.07
Full loss w/o Inter and Recon	25.80	14.13	11.42	3.43	0.0022	0.0003	0.3861	0.10
Full loss w/o Inter w/ Simp Recon	23.42	13.32	10.92	**2.17**	0.0004	0.0001	0.3961	0.07
Full loss w/o Inter	22.35	13.12	10.86	2.20	0.0002	0.0001	0.5319	0.06
Full loss w/o PE^P	21.97	12.66	10.40	2.51	0.0007	0.0001	0.6448	0.05
Full loss	21.42	12.48	10.26	2.59	0.0033	0.0002	0.6982	0.05
DiffGrasp (Full loss w/ Guidance)	**20.99**	**12.24**	**10.09**	2.22	0.0023	**0.0001**	**0.7840**	**0.04**

Table 2: *Ablation study results* on GRAB (Taheri et al. 2020) dataset.

use the contact label constraint, and then using our contact-aware reconstruction loss (Full loss w/o Inter) results in significant improvements across all metrics. In the ablation of *the contact-aware interaction loss*, the version with interaction loss term (Full loss w/o Recon) outperforms all other loss ablations in the F1 metric, demonstrating that the interaction loss encourages contact between the hands and the object. Qualitative results in Fig. 5 shows that the interaction loss (Full w/o R.) brings the generated hand closer to the object but results in less natural poses, while the reconstruction loss (Full w/o I.) generates more natural poses but with weaker perception of the position of object.

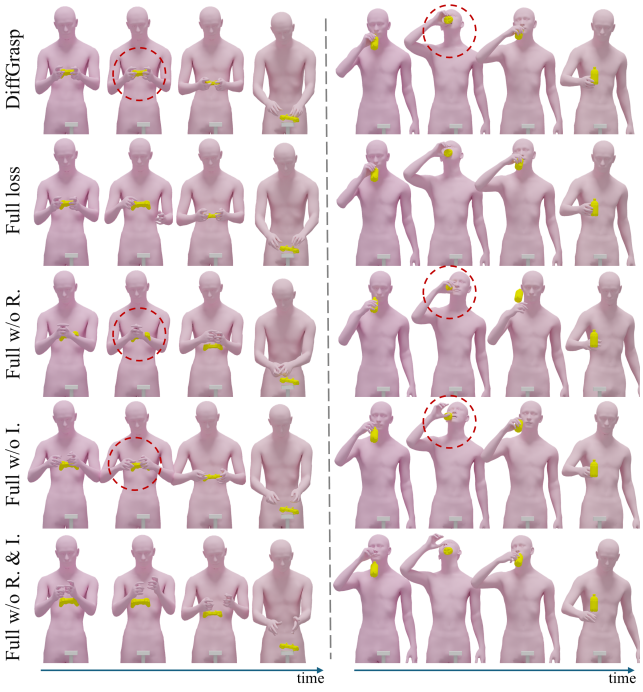

Figure 5: *Qualitative Results of Ablation Study.* In this figure, Full is the abbreviation for Full loss, R. is the abbreviation for Recon, and I. is the abbreviation for Inter.

Effect of Frame-wise Positional Encoding. Removing the frame-wise positional encoding PE^P (Full loss w/o PE^P) from the full loss version (Full loss) results in a slight decrease in almost all metrics, demonstrating that PE^P help the network understand the input conditions better.

Effect of Guidance. Comparing our model without guidance (complete loss) with our full model incorporating guidance (DiffGrasp), we can see slight improvements in accuracy, and user experiment results in the supplementary demonstrate that guidance generated more realistic results.

Limitations

Although promising results have been achieved, our method has several limitations, as demonstrated in Fig. 6. We find that due to the lack of constraints on the self-penetration of human body meshes, DiffGrasp may generate results with self-penetration (Fig. 6(a)). Moreover, because we do not add physical constraints between the hand and object, the generated hand results may exhibit unrealistic grasp poses (Fig. 6(b)). Furthermore, due to limitations in the dataset used, we have not yet achieved walking during grasping.

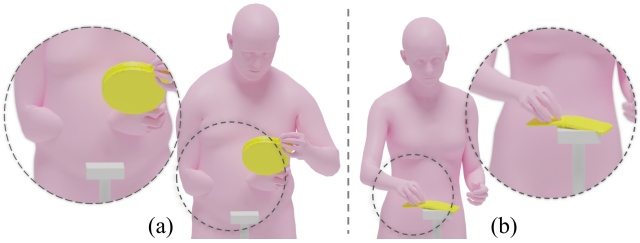

Figure 6: *Limitations.* Our method may generate self-penetration (a) or unrealistic poses (b) in some cases.

Conclusion

In this paper, we propose the first framework to generate realistic whole-body human motion sequence with fine finger-object grasp, conditioned on a 3D object motion sequence with different object sizes and shapes. Furthermore, by using reliable contact-aware losses, we leverage a single conditional diffusion model generate natural and reliable results. Finally, we utilize an innovative data-driven guidance strategy, along with others, to achieve a stable, non-penetrating, and non-sliding grasp. Extensive experiments show that our method achieves state-of-the-art results.

Acknowledgments

This work was supported in part by National Science and Technology Major Project (2022ZD0117904), National Natural Science Foundation of China (62473356,62373061), Beijing Natural Science Foundation (L232028), CAS Major Project (RCJJ-145-24-14), and Beijing Hospitals Authority Clinical Medicine Development of Special Funding Support No. ZLRK202330.

References

Ao, T.; Zhang, Z.; and Liu, L. 2023. GestureDiffuCLIP: Gesture diffusion model with CLIP latents. *arXiv preprint arXiv:2303.14613*.

Cen, Z.; Pi, H.; Peng, S.; Shen, Z.; Yang, M.; Shuai, Z.; Bao, H.; and Zhou, X. 2024. Generating Human Motion in 3D Scenes from Text Descriptions. In *CVPR*.

Cha, J.; Kim, J.; Yoon, J. S.; and Baek, S. 2024. Text2HOI: Text-guided 3D Motion Generation for Hand-Object Interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1577–1585.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18000–18010.

Christen, S.; Kocabas, M.; Aksan, E.; Hwangbo, J.; Song, J.; and Hilliges, O. 2022. D-Grasp: Physically Plausible Dynamic Grasp Synthesis for Hand-Object Interactions. arXiv:2112.03028.

Diller, C.; and Dai, A. 2024. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19888–19901.

Fan, Z.; Taheri, O.; Tzionas, D.; Kocabas, M.; Kaufmann, M.; Black, M. J.; and Hilliges, O. 2023. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ghosh, A.; Dabral, R.; Golyanik, V.; Theobalt, C.; and Slusallek, P. 2023a. IMoS: Intent-Driven Full-Body Motion Synthesis for Human-Object Interactions. In *Eurographics*.

Ghosh, A.; Dabral, R.; Golyanik, V.; Theobalt, C.; and Slusallek, P. 2023b. ReMoS: Reactive 3D Motion Synthesis for Two-Person Interactions. arXiv:2311.17057.

Guo, C.; Zuo, X.; Wang, S.; and Cheng, L. 2022. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, 580–597. Springer.

Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning joint reconstruction of hands and manipulated objects. arXiv:1904.05767.

He, C.; Saito, J.; Zachary, J.; Rushmeier, H.; and Zhou, Y. 2022. Nemf: Neural motion fields for kinematic animation. *Advances in Neural Information Processing Systems*, 35: 4244–4256.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *arXiv:2204.03458*.

Huang, S.; Wang, Z.; Li, P.; Jia, B.; Liu, T.; Zhu, Y.; Liang, W.; and Zhu, S.-C. 2023. Diffusion-based Generation, Optimization, and Planning in 3D Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiang, H.; Liu, S.; Wang, J.; and Wang, X. 2021. Hand-Object Contact Consistency Reasoning for Human Grasps Generation. arXiv:2104.03304.

Kulkarni, N.; Rempe, D.; Genova, K.; Kundu, A.; Johnson, J.; Fouhey, D.; and Guibas, L. 2023. NIFTY: Neural Object Interaction Fields for Guided Human Motion Synthesis. arXiv:2307.07511.

Kwon, T.; Tekin, B.; Stühmer, J.; Bogo, F.; and Pollefeys, M. 2021. H2O: Two Hands Manipulating Objects for First Person Interaction Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10138–10148.

Lee, T.; Moon, G.; and Lee, K. M. 2023. MultiAct: Long-term 3D human motion generation from multiple action labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37-1, 1231–1239.

Li, H.; Lin, X.; Zhou, Y.; Li, X.; Huo, Y.; Chen, J.; and Ye, Q. 2023a. Contact2Grasp: 3D Grasp Synthesis via Hand-Object Contact Constraint. arXiv:2210.09245.

Li, J.; Clegg, A.; Mottaghi, R.; Wu, J.; Puig, X.; and Liu, C. K. 2023b. Controllable human-object interaction synthesis. *arXiv preprint arXiv:2312.03913*.

Li, J.; Wu, J.; and Liu, C. K. 2023. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6): 1–11.

Li, Q.; Wang, J.; Loy, C. C.; and Dai, B. 2024. Task-Oriented Human-Object Interactions Generation With Implicit Neural Representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3035–3044.

Li, R.; Zhao, J.; Zhang, Y.; Su, M.; Ren, Z.; Zhang, H.; Tang, Y.; and Li, X. 2023c. FineDance: A Fine-grained Choreography Dataset for 3D Full Body Dance Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10234–10243.

Liu, S.; Zhou, Y.; Yang, J.; Gupta, S.; and Wang, S. 2023. ContactGen: Generative Contact Modeling for Grasp Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Liu, X.; and Yi, L. 2024. GeneOH Diffusion: Towards Generalizable Hand-Object Interaction Denoising via Denoising Diffusion. *arXiv preprint arXiv:2402.14810*.

Liu, Y.; Cao, Q.; Wen, Y.; Jiang, H.; and Ding, C. 2024. Towards Variable and Coordinated Holistic Co-Speech Motion Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1566–1576.

Lu, S.; Chen, L.-H.; Zeng, A.; Lin, J.; Zhang, R.; Zhang, L.; and Shum, H.-Y. 2023. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*.

Lucas, T.; Baradel, F.; Weinzaepfel, P.; and Rogez, G. 2022. Posegpt: Quantization-based 3d human motion generation and forecasting. In *European Conference on Computer Vision*, 417–435. Springer.

Luo, H.; Liu, Y.; and Yi, L. 2024. Physics-aware Hand-object Interaction Denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2341–2350.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Peng, X.; Xie, Y.; Wu, Z.; Jampani, V.; Sun, D.; and Jiang, H. 2023. HOI-Diff: Text-Driven Synthesis of 3D Human-Object Interactions using Diffusion Models. *arXiv preprint arXiv:2312.06553*.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, 480–497. Springer.

Pinyoanuntapong, E.; Wang, P.; Lee, M.; and Chen, C. 2024. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1546–1555.

Prokudin, S.; Lassner, C.; and Romero, J. 2019. Efficient Learning on Point Clouds With Basis Point Sets. In *Proceedings of the IEEE International Conference on Computer Vision*, 4332–4341.

Sahbani, A.; El-Khoury, S.; and Bidaud, P. 2012. An overview of 3D object grasp synthesis algorithms. *Robotics and Autonomous Systems*, 60(3): 326–336.

Shafir, Y.; Tevet, G.; Kapon, R.; and Bermano, A. H. 2023. Human Motion Diffusion as a Generative Prior. arXiv:2303.01418.

Song, W.; Zhang, X.; Li, S.; Gao, Y.; Hao, A.; Hou, X.; Chen, C.; Li, N.; and Qin, H. 2024. HOIAnimator: Generating Text-prompt Human-object Animations using Novel Perceptive Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 811–820.

Taheri, O.; Choutas, V.; Black, M. J.; and Tzionas, D. 2022. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13263–13273.

Taheri, O.; Ghorbani, N.; Black, M. J.; and Tzionas, D. 2020. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In *European Conference on Computer Vision (ECCV)*.

Taheri, O.; Zhou, Y.; Tzionas, D.; Zhou, Y.; Ceylan, D.; Pirk, S.; and Black, M. J. 2023. Grip: Generating interaction poses using latent consistency and spatial cues. *arXiv preprint arXiv:2308.11617*.

Tendulkar, P.; Surís, D.; and Vondrick, C. 2023. Flex: Full-body grasping without full-body grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21179–21189.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*.

Tseng, J.; Castellon, R.; and Liu, K. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 448–458.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wu, Y.; Wang, J.; Zhang, Y.; Zhang, S.; Hilliges, O.; Yu, F.; and Tang, S. 2022. SAGA: Stochastic Whole-Body Grasping with Contact. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Xu, S.; Li, Z.; Wang, Y.-X.; and Gui, L.-Y. 2023. InterDiff: Generating 3D Human-Object Interactions with Physics-Informed Diffusion. In *ICCV*.

Zhang, H.; Christen, S.; Fan, Z.; Zheng, L.; Hwangbo, J.; Song, J.; and Hilliges, O. 2024a. ArtiGrasp: Physically Plausible Synthesis of Bi-Manual Dexterous Grasping and Articulation. In *International Conference on 3D Vision (3DV)*.

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

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024b. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhao, K.; Zhang, Y.; Wang, S.; Beeler, T.; and Tang, S. 2023. Synthesizing diverse human motions in 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14738–14749.

Zheng, J.; Zheng, Q.; Fang, L.; Liu, Y.; and Yi, L. 2023a. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 585–594.

Zheng, Y.; Shi, Y.; Cui, Y.; Zhao, Z.; Luo, Z.; and Zhou, W. 2023b. COOP: Decoupling and Coupling of Whole-Body Grasping Pose Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2163–2173.

Zhou, K.; Bhatnagar, B. L.; Lenssen, J. E.; and Pons-Moll, G. 2022. TOCH: Spatio-Temporal Object Correspondence to Hand for Motion Refinement. In *European Conference on Computer Vision (ECCV)*. Springer.

Zhou, Y.; Barnes, C.; Lu, J.; Yang, J.; and Li, H. 2020. On the Continuity of Rotation Representations in Neural Networks. arXiv:1812.07035.