SAGA: Stochastic Whole-Body Grasping With Contact

Yan Wu*[1], Jiahao Wang*[2], Yan Zhang[1], Siwei Zhang[1], Otmar Hilliges[1], Fisher Yu[1], Siyu Tang[1]

[1] ETH Zürich, Switzerland
[2] Max Planck Institute for Informatics, Germany
yan.wu@vision.ee.ethz.ch, jiwang@mpi-inf.mpg.de,
{yan.zhang,siwei.zhang,otmar.hilliges,siyu.tang}@inf.ethz.ch, i@yf.io

Abstract. The synthesis of human grasping has numerous applications including AR/VR, video games and robotics. While methods have been proposed to generate realistic hand–object interaction for object grasping and manipulation, these typically only consider interacting hand alone. Our goal is to **synthesize whole-body grasping motions**. Starting from an arbitrary initial pose, we aim to generate diverse and natural whole-body human motions to approach and grasp a target object in 3D space. This task is challenging as it requires modeling both whole-body dynamics and dexterous finger movements. To this end, we propose **SAGA** (StochAstic whole-body Grasping with contAct), a framework which consists of two key components: (a) Static whole-body grasping pose generation. Specifically, we propose a multi-task generative model, to jointly learn static whole-body grasping poses and human-object contacts. (b) Grasping motion infilling. Given an initial pose and the generated whole-body grasping pose as the start and end of the motion respectively, we design a novel contact-aware generative motion infilling module to generate a diverse set of grasp-oriented motions. We demonstrate the effectiveness of our method, which is a novel generative framework to synthesize realistic and expressive whole-body motions that approach and grasp randomly placed unseen objects. Code and models are available at https://jiahaoplus.github.io/SAGA/saga.html.

Keywords: motion generation, whole-body grasping synthesis, human-object interaction.

1 Introduction

A fully automated system that synthesizes realistic 3D human bodies approaching and grasping a target object in 3D space will be valuable in various fields, from robotics and animation to computer vision. Although remarkable progress has been made towards synthesizing realistic hand–object interactions, most existing works only focus on hand pose synthesis without considering whole-body

* Equal contribution.

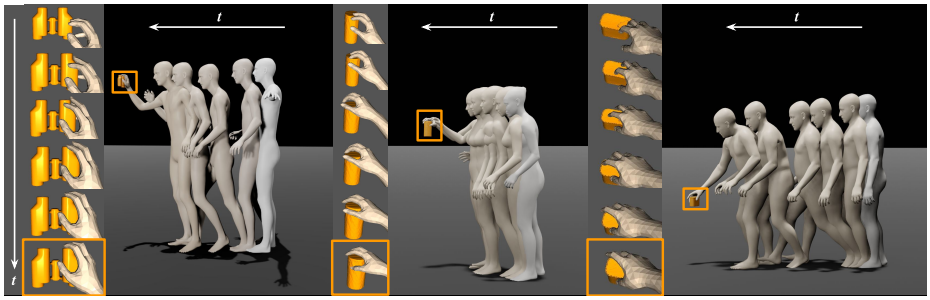

Fig. 1. Generated whole-body grasping motion sequences (in beige) starting from a given pose (in white) to approach and grasp randomly placed unseen objects. For each sample, we present hand motion details in the last few frames on the left column.

movements [22, 24, 50]. Meanwhile, whole-body motion synthesis [17, 19, 37, 60] largely ignores the presence of objects in the scene.

Modeling and synthesizing realistic whole-body grasping motions are challenging and remain unsolved due to a number of reasons. Firstly, whole-body grasping motions involve both full-body dynamics and dexterous finger movements [50], while the high dimensional degrees of freedom make the synthesis of grasping motions complicated. Secondly, a whole-body grasping sequence exhibits complex and frequent body–scene and hand–object contacts which are challenging to synthesize in a perceptually realistic way. For example, the hand's surface should conform naturally to the object and there should be no foot-skating artifacts in the whole-body motion. Thirdly, given only a starting pose and a target object in 3D space, there could be an infinite number of ways for the person to approach and grasp the object. The diversity of plausible grasping motions is further amplified by the large potential variation in object shape and pose. How to build an effective generative model that can capture this diversity and synthesize diverse realistic motions to grasp various 3D objects is an unsolved question.

To address these challenges, we propose **SAGA** (**S**toch**A**stic whole-body **G**rasping with cont**A**ct), a novel whole-body grasping generation framework that can synthesize stochastic motions of a 3D human body approaching and grasping 3D objects. Our solution consists of two components: (1) a novel 3D body generator that synthesizes diverse static whole-body grasping end poses, and (2) a novel human motion generator that creates diverse and plausible motions between given start and end poses. We present two key insights on both components. First, instead of directly using parametric body models (e.g. SMPL [38]) to represent 3D bodies, we employ the markers-based representation [60] which captures 3D human shape and pose information with a set of sparse markers on the human body surface. As demonstrated in [60], the markers-based representation is easier for neural networks to learn than the latent parameters of the parametric body models, yielding more realistic motion. We show that the

markers-based representation is especially advantageous to the latent body parameters for learning whole-body grasping, as the accumulation of errors along the kinematic chain has a significant impact on the physical plausibility of generated hand grasps, resulting in severe hand–object interpenetration. Second, *contact* plays a central role in our pipeline. As the human moves in 3D space and grasps a 3D object, physical contact is key for modeling realistic motions and interactions. For both components of our method, we learn contact representations from data and use them to guide interaction synthesis, greatly improving the realism of the generated motion.

For the static grasping pose generation, we built a multi-task conditional variational autoencoder (CVAE) to jointly learn whole-body marker positions and fine-grained marker–object contact labels. During inference, given a target object in 3D space, our model jointly generates a diverse set of consistent full-body marker locations and the contact labels on both the body markers and the object surface. A contact-aware pose optimization module further recovers a parametric body mesh from the predicted markers, while explicitly enforcing the hand–object contact by leveraging the predicted contact labels. Next, given the generated static whole-body grasping pose as the end pose, and an initial pose as a start pose, we propose a novel generative motion infilling network to capture motion uncertainty and generate diverse motions in between. We design a CVAE-based architecture to generate both the diverse in-between motion trajectories and the diverse in-between local pose articulations. In addition, contacts between feet and the ground are also predicted as a multi-task learning objective to enforce a better foot-ground interaction. Furthermore, leveraging the predicted human–object contacts, we design a contact-aware motion optimization module to produce realistic grasp-oriented whole-body motions from the generated marker sequences. By leveraging the GRAB [50] and AMASS [34] datasets to learn our generative models, our method can successfully generate realistic and diverse whole-body grasping motion sequences for approaching and grasping a variety of 3D objects.

Contributions. In summary, we provide (1) a novel generative framework to synthesize diverse and realistic whole-body motions approaching and grasping various unseen objects for 3D humans that exhibit various body shapes, (2) a novel multi-task learning model to jointly learn the static whole-body grasping poses and the body–object interactions, (3) a novel generative motion infilling model that can stochastically infill both the global trajectories and the local pose articulations, yielding diverse and realistic full-body motions between a start pose and end pose. We perform extensive experiments to validate technical contributions. Experimental results demonstrate both the efficacy of our full pipeline and the superiority of each component to existing solutions.

2 Related Work

Human Grasp Synthesis is a challenging task and has been studied in computer graphics [23, 27, 30, 39, 42, 58] and robotics [9, 20, 26, 30, 33, 46]. With the

advancement in deep learning, recent works also approach the realistic 3D human grasp synthesis task by leveraging large-scale datasets [3, 22, 24, 50, 58], however they only focus on hand grasp synthesis.

Grasping Field [24] proposes an implicit representation of hand–object interaction and builds a model to generate plausible human grasps. GrabNet [50] proposes a CVAE to directly sample the MANO [43] hand parameters, and additionally train a neural network to refine the hand pose for a more plausible hand–object contact. GraspTTA [22] suggests using consistent hand–object interactions to synthesize realistic hand grasping poses. It sequentially generates coarse hand grasping poses and estimates consistent object contact maps, and using the estimated object contact maps, the produced hand pose is further optimized for realistic hand–object interactions. Similarly, ContactOpt [11] proposes an object contact map estimation network and a contact-based hand pose optimization module to produce realistic hand–object interaction. Different from GraspTTA and ContactOpt, which predict consistent hand pose and hand–object contacts sequentially in two stages, we build a multi-task generative model that generates consistent whole-body pose and the mutual human-object contacts jointly to address a more complicated whole-body grasping pose learning problem. Going beyond the static grasp pose generation, provided the wrist trajectory and object trajectory, ManipNet [58] generates dexterous finger motions to manipulate objects using an autoregressive model. Nonetheless, to our best knowledge, none of the previous works studied 3D human whole-body grasp learning and synthesis.

3D Human Motion Synthesis. In recent years, human motion prediction has received a lot of attention in computer vision and computer graphics [4, 8, 10, 18, 21, 31, 32, 36, 48, 53, 60, 61]. Existing motion prediction models can also be split into two categories: deterministic [17, 19, 25, 37, 55] and stochastic [2, 5, 28, 56]. For deterministic motion prediction, [25] adopt convolutional models to provide spatial or temporal consistent motions, and [37] propose an RNN with residual connections and sampling-based loss to model human motion represented by joints. For stochastic motion prediction, recently, Li *et al.* [28] and Cai *et al.* [5] use VAE-based models to address general motion synthesis problems. While these methods make great contributions to human motion understanding, they do not study the interaction with the 3D environment.

There are several works predict human motion paths or poses in scene context [1,6,12,13,15,16,29,35,41,44,45,47,51,52,54,59]. Cao *et al.* [6] estimate goals, 3D human paths, and pose sequences given 2D pose histories and an image of the scene. However, the human is represented in skeletons, thus it is hard to accurately model body–scene contacts, which limits its application. Recently, Wang *et al.* [54] propose a pipeline to infill human motions in 3D scenes, which first synthesizes sub-goal bodies, then fills in the motion between these sub-goals, then refines the bodies. However, the generated motion appears unnatural especially in the foot–ground contact. [15] presents an RNN-based network with contact loss and adversarial losses to handle motion in-betweening problems. They use the humanoid skeleton as the body representation and require 10 start frames

and one end frame as input. [41] adopts a conditional variational autoencoder to correct the pose at each timestamp to address noise and occlusions. They also use a test-time optimization to get more plausible motions and human–ground contacts. [59] propose a contact-aware motion infiller to generate the motion of unobserved body parts. They predict motion with better foot–ground contact, but their deterministic model does not capture the nature of human motion diversity. Unlike the methods mentioned above, our generative motion infilling model, when given the first and the last frame, captures both the global trajectory diversity in between and the diversity of local body articulations.

Concurrent work. GOAL [49] builds a similar two-stage pipeline to approach the whole-body grasping motion generation, producing end pose first and then infilling the in-between motion. Unlike our work which captures both the diversity of grasping ending poses and in-between motions, however, GOAL builds a deterministic auto-regressive model to in-paint the in-between motion, which does not fully explore the uncertainty of grasping motions.

3 Method

Preliminaries. (a) 3D human body representation. (1) SMPL-X [38] is a parametric human body model which models body mesh with hand details. In this work, the SMPL-X body parameters $\boldsymbol{\Theta}$ include the shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, the body global translation $\boldsymbol{t} \in \mathbb{R}^3$, the 6D continuous representation [62] of the body rotation $\boldsymbol{R} \in \mathbb{R}^6$, and full-body pose parameters $\boldsymbol{\theta} = [\boldsymbol{\theta_b}, \boldsymbol{\theta_h}, \boldsymbol{\theta_e}]$, where $\boldsymbol{\theta_b} \in \mathbb{R}^{32}$, $\boldsymbol{\theta_h} \in \mathbb{R}^{48}$, $\boldsymbol{\theta_e} \in \mathbb{R}^6$ are the body pose in the Vposer latent space [38], the hands pose in the MANO [43] PCA space and the eyes pose, respectively; (2) Markers-based representation [60] captures the body shape and pose information with the 3D locations $\boldsymbol{M} \in \mathbb{R}^{N\times 3}$ of a set of sparse markers on the body surface, where N is the number of markers. We learn the markers representation in our neural networks, from which we further recover SMPL-X body mesh. **(b) 3D objects** are represented with centered point cloud data $\boldsymbol{O}$ and the objects height $t_{\boldsymbol{O}} \in \mathbb{R}^1$. We sample 2048 points on the object surface and each point has 6 features (3 XYZ positions + 3 normal features).

Notations. For clarity, in the following text, $\tilde{X}$ and $\hat{X}$ denote the CVAE reconstruction result of X, and random samples of X from CVAE, respectively.

3.1 Overview

Given an initial human pose and a 3D object randomly placed in front of the human within a reasonable range, our goal is to generate realistic and diverse whole-body motions, starting from the given initial pose and approaching to grasp the object. As presented in Fig. 2, we propose a two-stage grasping motion generation pipeline to approach this task.

Stage 1: Stochastic whole-body grasping ending pose generation (§ 3.2). We first build an object-conditioned multi-task CVAE which synthesizes whole-body grasping ending poses in markers and the explicit human–object contacts.

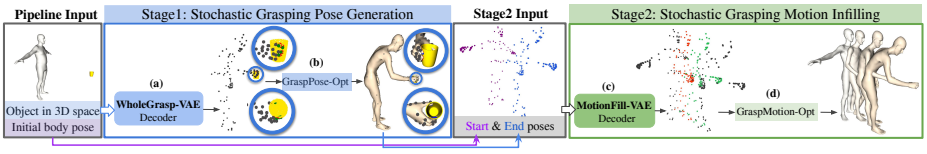

Fig. 2. Illustration of our two-stage pipeline. Given an object in 3D space and a human start pose, our method produces diverse human whole-body grasping motions. In stage 1, (a) taking the given 3D object information as inputs, our WholeGrasp-VAE (§ 3.2) decoder generates whole-body grasping poses represented by marker locations and mutual marker–object contact probabilities (green markers and red areas on the object surface indicate a high contact probability); (b) GraspPose-Opt (§ 3.2) further recovers body mesh from predicted markers. We use the generated grasping pose as the targeted end pose. Then in stage 2, (c) we feed in the start pose and the end pose into the MotionFill-VAE decoder (§ 3.3) to generate the in-between motions in markers representation, and (d) and GraspMotion-Opt (§ 3.4) further recovers smooth and realistic whole-body grasping motions.

We further perform contact-aware pose optimization to produce 3D body meshes with realistic interactions with objects by leveraging the contacts information.

Stage 2: Stochastic grasp-oriented motion infilling. We build a novel generative motion infilling model (§ 3.3) which takes the provided initial pose and the generated end pose in stage 1 as inputs, and outputs diverse intermediate motions. We further process the generated motions via a contact-aware optimization step (§ 3.4) to produce realistic human whole-body grasping motions.

3.2 Whole-Body Grasping Pose Generation

To synthesize diverse whole-body poses to grasp a given object, we propose a novel multi-task WholeGrasp-VAE to learn diverse yet consistent grasping poses and mutual contacts between human and object. The explicit human–object contacts provide fine-grained human–object interaction information which helps to produce realistic body meshes with high-fidelity interactions with the object.

Model Architecture. We visualize the multi-task WholeGrasp-VAE design in Fig. 3. The encoder takes the body markers' positions $\boldsymbol{M} \in \mathbb{R}^{N\times 3}$, body markers contacts $C_{\boldsymbol{M}} \in \{0,1\}^N$ and object contacts $C_{\boldsymbol{O}} \in \{0,1\}^{2048}$ as inputs, where N is the number of markers, and learns a joint Gaussian latent space $\mathbf{z_j}$. We use PointNet++ [40] to encode the object feature.

Training. The overall training objective is given by $\mathcal{L}_{train} = \mathcal{L}_{rec} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_c\mathcal{L}_c$, where λ_{KL}, λ_c are hyper-parameters.

Reconstruction loss includes the L1 reconstruction loss of body markers' postions and the binary cross-entropy (BCE) loss of contact probabilities:

$$\mathcal{L}_{rec} = |\boldsymbol{M} - \tilde{\boldsymbol{M}}| + \lambda_{\boldsymbol{M}}\mathcal{L}_{bce}(C_{\boldsymbol{M}}, \tilde{C}_{\boldsymbol{M}}) + \lambda_{\boldsymbol{O}}\mathcal{L}_{bce}(C_{\boldsymbol{O}}, \tilde{C}_{\boldsymbol{O}}). \quad (1)$$

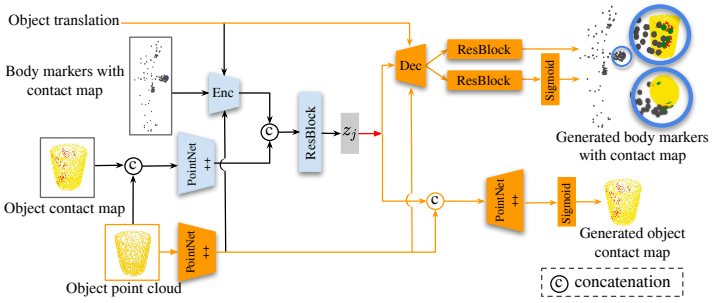

Fig. 3. The WholeGrasp-VAE design. WholeGrasp-VAE jointly learns the (1) body markers' locations; (2) body marker contacts (markers with high contact probability are shown in green); (3) Object contact map (the area with high contact probability is shown in red). The red arrow indicates sampling from the latent space. At inference time, activated modules are shown in orange.

KL-divergence loss. We employ the robust KL-divergence term [60] to avoid the VAE posterior collapse:

$$\mathcal{L}_{KL} = \Psi(D_{KL}(q(\mathbf{z_j}|\boldsymbol{M}, C_{\boldsymbol{M}}, C_{\boldsymbol{O}}, t_{\boldsymbol{O}}, \boldsymbol{O})||\mathcal{N}(\mathbf{0}, \mathbf{I}))), \tag{2}$$

where $\Psi(s) = \sqrt{s^2 + 1} - 1$ [7]. This function automatically penalizes the gradient to update the above KLD term, when the KL-divergence is small.

Consistency loss. We use a consistency loss to implicitly encourage consistent predictions of marker positions and mutual marker–object contacts:

$$\mathcal{L}_c = \sum_{m \in \boldsymbol{M}, \tilde{m} \in \tilde{\boldsymbol{M}}} \tilde{C}_m |d(\tilde{m}, \boldsymbol{O}) - d(m, \boldsymbol{O})| + \sum_{o \in \boldsymbol{O}} \tilde{C}_o |d(o, \tilde{\boldsymbol{M}}) - d(o, \boldsymbol{M})|, \tag{3}$$

$d(x, \mathcal{Y}) = \min_{y \in \mathcal{Y}} ||x - y||_2^2$ is the minimum distance from point x to point cloud $\mathcal{Y}$.

Inference. During inference, we feed the provided target object information into the WholeGrasp-VAE decoder to generate plausible body markers $\hat{\boldsymbol{M}}$ and marker–object contact labels $\hat{C}_{\boldsymbol{M}}$, $\hat{C}_{\boldsymbol{O}}$. We design a contact-aware pose optimization algorithm, GraspPose-Opt, to generate a realistic body mesh from markers and refine body pose for high-fidelity human–object interaction by leveraging the fine-grained human–object contacts. Specifically, by optimizing SMPL-X parameters $\boldsymbol{\Theta}$, the overall optimization objective is given by:

$$E_{opt}(\boldsymbol{\Theta}) = E_{fit} + E^o_{colli} + E^o_{cont} + E^g_{cont}. \tag{4}$$

Marker fitting loss. To project the predicted markers to a valid body mesh, we minimize the L1 distance between the sampled markers $\hat{\boldsymbol{M}}$ and the queried markers $\boldsymbol{M}(\boldsymbol{\Theta})$ on the SMPL-X body mesh:

$$E_{fit}(\boldsymbol{\Theta}) = |\hat{\boldsymbol{M}} - \boldsymbol{M}(\boldsymbol{\Theta})| + \alpha_{\boldsymbol{\theta}} |\boldsymbol{\theta}|^2, \tag{5}$$

where $\alpha_{\boldsymbol{\theta}}$ is the pose parameters regularization weight.

Object contact loss. By leveraging sampled contact maps, we propose a mutual contact loss to encourage body markers and object points with high

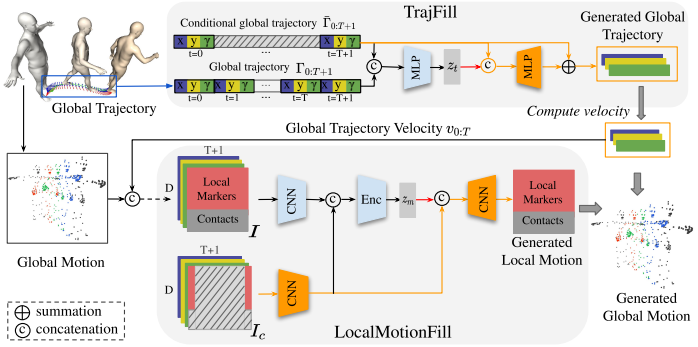

Fig. 4. MotionFill-VAE consists of two concatenated CVAEs: (1) TrajFill outputs the infilled global root trajectory when the start root and the end root are given; (2) LocalMotionFill takes the global trajectory information from TrajFill as one of the inputs, and it outputs the infilled local motion when the start pose, the end pose, and the global trajectory are given. We reconstruct the global motion from the generated global trajectory and the local motion. The red arrow indicates sampling from the latent space. The dash arrow indicates the input processing step for building the four-channel motion image (one local motion channel with contact states and three root velocity channels). At inference time, activated modules are shown in orange.

contact probabilities to contact the object surface and body surface, respectively.

$$E^{o}_{cont}(\boldsymbol{\Theta}) = \alpha^{o}_{cont} \sum_{o \in \boldsymbol{O}} \hat{C}_o d(o, \mathcal{V}_{\boldsymbol{B}}(\boldsymbol{\Theta})) + \alpha^{m}_{cont} \sum_{m \in \boldsymbol{M}(\boldsymbol{\Theta})} \hat{C}_m d(m, \boldsymbol{O}). \tag{6}$$

where $\mathcal{V}_{\boldsymbol{B}}(\boldsymbol{\Theta})$ denotes the SMPL-X body vertices.

Collision loss. We employ a signed-distance based collision loss to penalize the body–object interpenetration:

$$E_{colli}(\boldsymbol{\Theta}) = \alpha^{\boldsymbol{B}}_{colli} \sum_{b \in \mathcal{V}^{h}_{\boldsymbol{B}}(\boldsymbol{\Theta})} \max(-\mathcal{S}(b, \boldsymbol{O}), \sigma_b) + \alpha^{\boldsymbol{O}}_{colli} \sum_{o \in \boldsymbol{O}} \max(-\mathcal{S}(o, \mathcal{V}^{h}_{\boldsymbol{B}}(\boldsymbol{\Theta})), \sigma_o) \tag{7}$$

where $\mathcal{S}(x, \mathcal{Y})$ is the signed-distance from point x to point cloud $\mathcal{Y}$, $\mathcal{V}^{h}_{\boldsymbol{B}}(\boldsymbol{\Theta})$ denotes the hand vertices, and σ_b, σ_o are small interpenetration thresholds.

Ground contact loss is given by $E^{g}_{cont}(\boldsymbol{\Theta}) = \alpha_{cont} \sum_{v \in \mathcal{V}^{f}_{\boldsymbol{B}}} |h(v)|$, where we penalize the heights of feet vertices $\mathcal{V}^{f}_{\boldsymbol{B}}$ to enforce a plausible foot–ground contact.

3.3 Generative Motion Infilling

Given body markers on the start and end poses produced by GraspPose-Opt, *i.e.*, $\boldsymbol{M}_0$ and $\boldsymbol{M}_T$, many in-between motions are plausible. To model such uncertainty, we build a novel generative motion infilling model, namely MotionFill-VAE, to capture both the uncertainties of intermediate global root (pelvis joint) trajectories and intermediate root-related local body poses. Specifically, given motion

$\boldsymbol{M}_{0:T}$ represented in a sequence of markers positions, following [19, 25, 59], we represent the global motion $\boldsymbol{M}_{0:T}$ with a hierarchical combination of global root velocity $\boldsymbol{v}_{0:T}$ (where $\boldsymbol{v}_t = \boldsymbol{\Gamma}_{t+1} - \boldsymbol{\Gamma}_t, t \in [0, T]$, $\boldsymbol{\Gamma}$ and $\boldsymbol{v}$ denote the root trajectory and root velocity respectively) and the trajectory-conditioned local motion $\boldsymbol{M}^l_{0:T}$. Accordingly, we build the MotionFill-VAE to capture both the conditional global trajectory distribution $P(\boldsymbol{\Gamma}_{0:T+1}|\boldsymbol{\Gamma}_0, \boldsymbol{\Gamma}_T)$ and the conditional local motion distribution $P(\boldsymbol{M}^l_{0:T}|\boldsymbol{v}_{0:T}, \boldsymbol{M}^l_0, \boldsymbol{M}^l_T)$.

Model Architecture. As shown in Fig. 4, the MotionFill-VAE consists of two concatenated CVAEs: **(1) TrajFill** learns the conditional intermediate global root trajectory latent space $\boldsymbol{z_t}$. Taking the root states $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}_T$ as inputs, which are derived from the given start and end pose, our goal is to get the trajectory $\boldsymbol{\Gamma}_{0:T+1}$. Instead of directly learning $\boldsymbol{\Gamma}_{0:T+1}$, we build TrajFill to learn the trajectory deviation $\boldsymbol{\Delta\Gamma}_{0:T+1} = \boldsymbol{\Gamma}_{0:T+1} - \overline{\boldsymbol{\Gamma}}_{0:T+1}$, where $\overline{\boldsymbol{\Gamma}}_{0:T+1}$ is a straight trajectory which is a linear interpolation and one-step extrapolation of the given $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Gamma}_T$. We further compute the velocity $\boldsymbol{v}_{0:T}$ from the predicted trajectory $\boldsymbol{\Gamma}_{0:T+1}$. **(2) LocalMotionFill** learns the conditional intermediate local motion latent space $\boldsymbol{z_m}$. Taking the TrajFill output $\boldsymbol{v}_{0:T}$ and the given $\boldsymbol{M}_0, \boldsymbol{M}_T$ as inputs, LocalMotionFill generates the trajectory-conditioned local motion sequence. Specifically, following [59], we build a four-channel image $\boldsymbol{I}$, which is a concatenation of local motion information with foot–ground contact labels and root velocity, and we use it as the input to our CNN-based LocalMotionFill architecture. Similarly, we build the four-channel conditional image $\boldsymbol{I}_c$ with the unknown motion in between filled with all 0.

Training. The training loss is $\mathcal{L}_M = \mathcal{L}_{rec} + \lambda_{KL}\mathcal{L}_{KL}$, and λ_{KL} is hyper-parameter.

Reconstruction loss $\mathcal{L}_{rec}$ contains the global trajectory reconstruction, local motion reconstruction and foot–ground contact label reconstruction losses:

$$\begin{aligned}\mathcal{L}_{rec} = &\sum_{t=0}^{T+1} |\boldsymbol{\Gamma}_t - \tilde{\boldsymbol{\Gamma}}_t| + \lambda_1 \sum_{t=0}^{T} |\boldsymbol{v}_t^{\boldsymbol{\Gamma}} - \tilde{\boldsymbol{v}}_t^{\boldsymbol{\Gamma}}| + \lambda_2 \mathcal{L}_{bce}(C_F, \tilde{C}_F) \\ &+ \lambda_3 \sum_{t=0}^{T} |\boldsymbol{M}^l_t - \tilde{\boldsymbol{M}}^l_t| + \lambda_4 \sum_{t=0}^{T-1} |\boldsymbol{v}_t^{\boldsymbol{M^l}} - \tilde{\boldsymbol{v}}_t^{\boldsymbol{M^l}}|, \end{aligned} \tag{8}$$

where $\boldsymbol{v}_t^{(*)} = (*)_{t+1} - (*)_t$ denotes the velocity, and $\lambda_1 - \lambda_4$ are hyper-parameters.

KL-divergence loss. We use the robust KL-divergence loss for both TrajFill and LocalMotionFill:

$$\mathcal{L}_{KL} = \Psi(D_{KL}(q(\boldsymbol{z}_t|\boldsymbol{\Gamma}_{0:T+1}, \overline{\boldsymbol{\Gamma}}_{0:T+1})||\mathcal{N}(\mathbf{0}, \mathbf{I}))) + \Psi(D_{KL}(q(\boldsymbol{z}_m|\boldsymbol{I}, \boldsymbol{I}_c)||\mathcal{N}(\mathbf{0}, \mathbf{I}))). \tag{9}$$

Inference. At inference time, given the start and end body markers $\boldsymbol{M}_0, \boldsymbol{M}_T$ with known root states $\boldsymbol{\Gamma}_0, \boldsymbol{\Gamma}_T$, by first feeding the initial interpolated trajectory $\overline{\boldsymbol{\Gamma}}_{0:T+1}$ into the decoder of TrajFill, we generate stochastic in-between global motion trajectory $\hat{\boldsymbol{\Gamma}}_{0:T+1}$. Next, with the given $\boldsymbol{M}_0, \boldsymbol{M}_T$ and the generated $\hat{\boldsymbol{\Gamma}}_{0:T+1}$, we further build the condition input image $\boldsymbol{I}_c$ as the input to the LocalMotionFill decoder, from which we can generate infilled local motion sequences $\hat{\boldsymbol{M}}^l_{0:T}$ and also the foot–ground contact probabilities $\hat{C}_{F_{0:T}}$. Finally, we reconstruct the global motion sequences $\hat{\boldsymbol{M}}_{0:T}$ from the generated $\hat{\boldsymbol{\Gamma}}_{0:T}$ and $\hat{\boldsymbol{M}}^l_{0:T}$.

3.4 Contact-aware Grasping Motion Optimization

With the generated marker sequences $\hat{\boldsymbol{M}}_{0:T}$, foot–ground contacts $\hat{C}_F$ from MotionFill-VAE, and the human–object contacts $\hat{C}_{\boldsymbol{M}}, \hat{C}_{\boldsymbol{O}}$ from WholeGrasp-VAE, we design GraspMotion-Opt, a contact-aware motion optimization algorithm, to recover smooth motions $\boldsymbol{B}_{0:T}$ with natural interactions with the scene.

Similar to GraspPose-opt, we propose the contact-aware marker fitting loss:

$$E_{basic}(\boldsymbol{\Theta}_{0:T}) = \sum_{t=0}^{T}(E_{fit}(\boldsymbol{\Theta}_t) + E^o_{colli}(\boldsymbol{\Theta}_t)) + \sum_{t=T-4}^{T} E^o_{cont}(\boldsymbol{\Theta}_t), \tag{10}$$

where $E_{fit}, E^o_{cont}, E^o_{colli}$ are formulated in Eq. 5-7, and we only apply object contact loss E^o_{colli} on the last 5 frames.

We design the following loss to encourage a natural hand grasping motion by encouraging the palm to face the object's surface on approach.

$$E_g(\boldsymbol{\Theta}_{0:T}) = \sum_{t=0}^{T} \alpha_t \sum_{m \in \mathcal{V}^p_{\boldsymbol{B}}(\boldsymbol{\Theta}_t)} \mathbb{1}(d(m, \boldsymbol{O}) < \sigma)(\cos\gamma_m - 1)^2, \tag{11}$$

where $\alpha_t = 1 - (\frac{t}{T})^2$, $\mathcal{V}^p_{\boldsymbol{B}}(\boldsymbol{\Theta})$ denotes the selected vertices on palm, and γ_m is the angle between the palm normal vector and the vector from palm vertices to the closest object surface points. We only apply this constraint when palm vertices are close to the object's surface (within radius $\sigma = 1$cm).

Inspired by [59], we enforce smoothness on the motion latent space to yield smoother motion, and we also reduce the foot skating artifacts by leveraging the foot–ground contact labels $\hat{C}_F$. For more details, please refer to the Appendix.

4 Experiments

Datasets. (1) We use **GRAB** [50] dataset to train and evaluate our WholeGrasp-VAE and also finetune the MotionFill-VAE. For WholeGrasp-VAE training and evaluation, following [50], we take all frames with right-hand grasps and have the same train/valid/test set split. For MotionFill-VAE training, we downsample the motion sequences to 30fps and clip 62 frames per sequence, with last frames being in stable grasping poses. (2) We use the **AMASS** [34] dataset to pretrain our LocalMotionFill-CVAE. We down-sample the sequences to 30 fps and cut them into clips with 61 frames. (3) We take unseen objects from **HO3D** [14] dataset to test the generalization ability of our method.

We conduct extensive experiments to study the effectiveness of each stage in our pipeline. In Sec. 4.1 and Sec. 4.2, we study our static grasping pose generator and the stochastic motion infilling model respectively. In Sec. 4.3, we evaluate the entire pipeline performance for synthesizing stochastic grasping motions. We encourage readers to watch the video of generated grasping poses and motions.

Table 1. Comparisons with the extended GrabNet baseline and ablation study result on the multi-task WholeGrasp-VAE design. Numbers in **bold**/blue indicates the **best**/second-best respectively.

Method	APD (↑)	Contact Ratio (↑)	Inter. Vol. $[cm^3]$ (↓)	Inter. Depth $[cm]$ (↓)
GrabNet [50]-SMPLX	0.33	0.65	14.15	0.78
WholeGrasp-single w/o opt.*	**2.94**	0.90	11.44	0.78
WholeGrasp-single w/ heuristic opt.		0.81	**0.21**	**0.12**
WholeGrasp w/o opt.*	2.92	**0.96**	12.20	0.85
WholeGrasp w/ opt. (**Ours**)		0.94	0.48	0.16

* Body meshes are recovered from sampled markers with only E_{fit} in Eq. 5.

4.1 Stochastic Whole-body Grasp Pose Synthesis

We evaluate our proposed stochastic whole-body grasp pose generation module on GRAB dataset. We also conduct ablation studies to study the effectiveness of several proposed components, including the multi-task CVAE design and the contact-aware optimization module design.

Baseline. GrabNet [50] builds a CVAE to generate the MANO hand parameters for grasping a given object, and we extend GrabNet to whole-body grasp synthesis by learning the whole-body SMPL-X parameters. We compare our method against the extended GrabNet (named as GrabNet-SMPLX)*.

Evaluation Metrics. *(1) Contact Ratio.* To evaluate the grasp stability, we measure the ratio of body meshes being in minimal contact with object meshes. *(2) Interpenetration Volume and Depth.* We measure the interpenetration volumes and depths between the body and object mesh. Low interpenetration volume and depth with a high contact ratio are desirable for perceptually realistic body–object interactions. *(3) Diversity.* We follow [57] to employ the Average L2 Pairwise Distance (**APD**) to evaluate the diversity within random samples.

Results. In Table 1, we compare our method against the extended GrabNet baseline. Because the extended GrabNet baseline does not include an additional body mesh refinement step, we compare it to our results without GraspPose-Opt optimization (WholeGrasp w/o. opt. in Table 1). Our method w/o optimization outperforms the extended GrabNet baseline in the sample diversity (APD) and achieves higher contact ratio and smaller intersection. The extended GrabNet experiment demonstrates the challenges in learning the whole-body pose parameters for a plausible human–object interaction, with marker representation appearing to be more favorable for learning human grasping pose. Nevertheless, the derived body meshes from markers without pose optimization still have human–object interpenetration, and our contact-aware pose optimization (WholeGrasp w/ opt. in Table 1) drastically reduces the human–object collision issue while maintaining a high contact ratio.

* Please refer to the Appendix for experimental setup and implementation details.

Table 2. Ablation studies on different optimization losses ($E_{fit}, E_{colli}, E^o_{cont}$ in Eq. 5-Eq. 6). We fit ground truth markers (GT columns) and sampled markers (Samples columns), and numbers in **bold**/blue indicate the **best**/second-best respectively.

	Contact Ratio(↑)		Inter. Vol.(↓)		Inter. Depth(↓)	
	GT	Samples	GT	Samples	GT	Samples
GT Mesh	0.99	-	2.04	-	0.45	-
$E_{fit} + E^g_{cont}$	**0.99**	**0.96**	2.21	12.20	0.46	0.85
$E_{fit} + E^g_{cont} + E_{colli}$	0.25	0.24	**0.12**	**0.12**	**0.04**	**0.07**
$E_{fit} + E^g_{cont} + E_{colli} + E^o_{cont}$	0.94	0.94	0.52	0.48	0.17	0.16

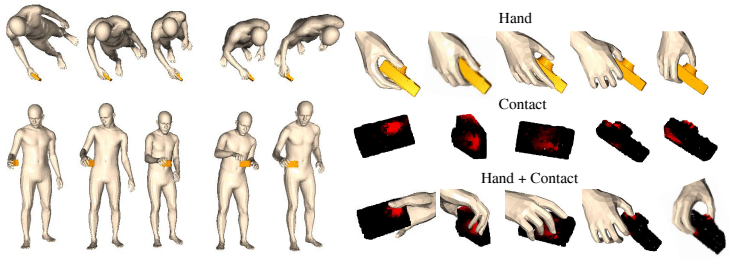

Fig. 5. Five **random** samples for an unseen object placed at the same positions. Left side: top view and front view of generated whole-body poses. Right side: hand grasping details and generated object contact maps (red areas indicate high contact probability).

Fig. 5 presents 5 **random** samples together with the generated object contact maps and hand grasping details for an unseen object. We can see that our models generate natural grasping poses with diverse body shapes and whole-body poses.

Ablation Study. *(1) Multi-task WholeGrasp design:* To study the effect of learning human–object contact labels, we build a single-task WholeGrasp-VAE architecture which only learns the markers' positions (WholeGrasp-single in Table 1). A similar pose optimization step as our GraspPose-opt further refines the grasping pose (WholeGrasp-single w/ heuristic opt. in Table 1), but we replace the mutual contact loss E_{cont} in Eq. 6 with a heuristic contact loss which is based on a pre-defined hand contact pattern. Both the single-task and multi-task WholeGrasp experiments demonstrate the benefit of using contact to refine the human–object interaction, and our multi-task WholeGrasp with explicit mutual contact learning outperforms the single-task setup with the pre-defined hand contact pattern. *(2) Study of GraspPose-Opt* (see Table 2): We evaluate recovered body meshes from both the ground truth markers and the randomly sampled markers, and also the ground truth body mesh. By fitting the body mesh to ground truth markers, our proposed GraspPose-Opt with only E_{fit} can recover an accurate body mesh with the human–object interaction metrics comparable to the ground truth mesh. The proposed E_{colli} and E_{cont} help to recover realistic human–object interaction significantly.

Table 3. Comparisons with motion infilling baselines. Best results are in boldface.

	Methods	ADE (↓)	Skat (↓)	PSKL-J (↓)	
				(P, GT)	(GT, P)
Local motion infilling*	CNN-AE [25]	0.091	0.245	0.804	0.739
	LEMO [59]	0.083	0.152	0.507	0.447
	PoseNet [54]	0.090	0.236	0.611	0.668
	Ours-Local†	**0.079**	**0.137**	**0.377**	**0.327**
Traj + local motion infilling	Route+PoseNet [54]	0.219	0.575	0.955	0.884
	Ours†	**0.083**	**0.394**	**0.772**	**0.609**

* Ground truth trajectories are used in the local motion infilling experiments.
† Generative models. And all the other methods are deterministic models.

4.2 Stochastic Motion Infilling

We evaluate our motion infilling model on AMASS and GRAB datasets. To our best knowledge, we are the first generative model to learn both the global trajectory and the local motion infilling given only one start pose and end pose. We compare our method with several representative motion infilling models.

Baselines. Wang *et al.* [54] proposed two sequential yet separate LSTM-based deterministic networks to first predict global trajectory (RouteNet) and then the local pose articulations (PoseNet) to approach the motion infilling task, and we take this sequential network (named as Route+PoseNet) as a baseline to our end-to-end generative global motion infilling model. There are some existing works which take the ground truth trajectory, start pose and end poses as inputs to predict the intermediate local poses, and following the same task setup, we also compare the generative local motion infilling component in our network against these baselines, including the convolution autoencoder network (CNN-AE) in [25], LEMO [59] and PoseNet [54]. We have chosen these baselines as they are the closest ones compared with our setting. For fair comparisons, we use the same body markers and the trajectory representation in all experiments*.

Evaluation Metrics. *(1) 3D marker accuracy.* For deterministic models, we measure the marker prediction accuracy by computing the Average L2 Distance Error (ADE) between the predicted markers and ground truth. For our generative model, we follow [57] to measure the sample accuracy by computing the minimal error between the ground truth and 10 random samples. *(2) Motion smoothness.* We follow [59] to use PSKL-J to measure the Power Spectrum KL divergence between the acceleration distribution of synthesized and ground truth joint motion sequences. PSKL-J being non-symmetric, we show the results of both direction, *i.e.*, (Predicted, Ground Truth) and (Ground Truth, Predicted). *(3) Foot skating.* Following [60], we measure the foot skating artifacts during motion and define skating as when the heel is within 5cm of the ground and the heel speed of both feet exceeds 75mm/s. *(4) foot–ground collision.* We also use a non-collision score, defined as the number of body mesh vertices above the ground divided by the total number of vertices.

Results. In Table 3, we compare our generative motion infilling model with the deterministic Route+PoseNet baseline [54], and both methods can infill the global trajectory and local pose motion. The results show that our generative model can yield much lower average 3D marker distance error (ADE). Also, our method has less foot skating and lower PSKL-J scores in both directions, which demonstrates that our method can generate more natural motions. We also compare our stochastic local motion infilling component (**Ours**-Local) against other deterministic local motion infilling baselines in Table 3. Our method outperforms all the other baselines in ADE, foot skating and PSKL-J, demonstrating that the our generative model can better capture human motion patterns and generate more natural motions. The motion sequences from the GRAB dataset and our generated motions have non-collision score of 0.9771 and 0.9743, respectively, showing that our method can effectively prevent foot–ground interpenetration.

4.3 Whole-body Grasp Motion Synthesis

Experiment setup. We test our grasping motion generation pipeline on 14 unseen objects from GRAB and HO3D dataset, and we generate 2s motions to grasp the object. Given different initial human poses, we place objects in front of the human at different heights (0.5m–1.7m) with various orientations (0–360° around the gravity axis) and different distances from start point to objects (5cm–1.1m). We conduct user studies on Amazon Mechanical Turk (AMT) for both ground truth grasping motion sequences from GRAB and our generated samples. On a five-point scale, three users are asked to rate the realism of presented motions, ranging from *strongly disagree* (score 0) to *strongly agree* (score 5)*.

Results. The perceptual scores for ground truth sequences and our synthesized sequences are 4.04 (around *agree*) and 3.15 (above *slightly agree*) respectively, showing that our proposed pipeline can synthesize high-fidelity grasping motions.

5 Conclusion and Discussion

In this work, we address an important task on how to synthesize realistic whole-body grasping motion. We propose a new approach consisting of two stages: (a) a WholeGrasp-VAE to generate static whole-body grasping poses; (b) a MotionFill-VAE to infill the grasp-oriented motion, given an initial pose and the predicted end pose. Our method, SAGA, is able to generate diverse motion sequences that have realistic interactions with the ground and random objects. We believe SAGA makes progress towards synthesizing human–object interaction, and provides a useful tool for computer graphics and robotics applications. However, in this work, we focus on the human motion synthesis task where a virtual human approaches to grasp an object without further hand–object manipulation. A future work is to synthesize the hand–object manipulation, while taking the object affordance, physics and the goal of the interaction into account.

Acknowledgement. This work was supported by the SNF grant 200021 204840 and Microsoft Mixed Reality & AI Zurich Lab PhD scholarship.

References

1. Alahi, A., Ramanathan, V., Fei-Fei, L.: Socially-aware large-scale crowd forecasting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2203–2210 (2014) 4
2. Barsoum, E., Kender, J., Liu, Z.: Hp-gan: Probabilistic 3d human motion prediction via gan. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1418–1427 (2018) 4
3. Brahmbhatt, S., Handa, A., Hays, J., Fox, D.: ContactGrasp: Functional Multi-finger Grasp Synthesis from Contact. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2019) 4
4. Cai, Y., Huang, L., Wang, Y., Cham, T.J., Cai, J., Yuan, J., Liu, J., Yang, X., Zhu, Y., Shen, X., et al.: Learning progressive joint propagation for human motion prediction. In: European Conference on Computer Vision. pp. 226–242. Springer (2020) 4
5. Cai, Y., Wang, Y., Zhu, Y., Cham, T.J., Cai, J., Yuan, J., Liu, J., Zheng, C., Yan, S., Ding, H., et al.: A unified 3d human motion synthesis model via conditional variational auto-encoder. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11645–11655 (2021) 4
6. Cao, Z., Gao, H., Mangalam, K., Cai, Q.Z., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: European Conference on Computer Vision. pp. 387–404. Springer (2020) 4
7. Charbonnier, P., Blanc-Feraud, L., Aubert, G., Barlaud, M.: Two deterministic half-quadratic regularization algorithms for computed imaging. In: Proceedings of 1st International Conference on Image Processing. vol. 2, pp. 168–172 vol.2 (1994) 7
8. Chiu, H.k., Adeli, E., Wang, B., Huang, D.A., Niebles, J.C.: Action-agnostic human pose forecasting. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1423–1432. IEEE (2019) 4
9. Detry, R., Kraft, D., Buch, A.G., Krüger, N., Piater, J.: Refining grasp affordance models by experience. In: 2010 IEEE International Conference on Robotics and Automation. pp. 2287–2293 (2010) 3
10. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4346–4354 (2015) 4
11. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmbhatt, S., Kemp, C.C.: ContactOpt: Optimizing contact to improve grasps. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4
12. Gupta, A., Satkin, S., Efros, A.A., Hebert, M.: From 3d scene geometry to human workspace. In: CVPR 2011. pp. 1961–1968. IEEE (2011) 4
13. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2255–2264 (2018) 4
14. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3196–3206 (2020) 10
15. Harvey, F.G., Yurick, M., Nowrouzezahrai, D., Pal, C.: Robust motion in-betweening. ACM Transactions on Graphics (TOG) **39**(4), 60–1 (2020) 4

16. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Physical review E **51**(5), 4282 (1995) 4
17. Hernandez, A., Gall, J., Moreno-Noguer, F.: Human motion prediction via spatio-temporal inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7134–7143 (2019) 2, 4
18. Holden, D., Komura, T., Saito, J.: Phase-functioned neural networks for character control. ACM Transactions on Graphics (TOG) **36**(4), 1–13 (2017) 4
19. Holden, D., Saito, J., Komura, T.: A deep learning framework for character motion synthesis and editing. ACM Transactions on Graphics (TOG) **35**(4), 1–11 (2016) 2, 4, 9
20. Hsiao, K., Lozano-Perez, T.: Imitation learning of whole-body grasps. In: 2006 IEEE/RSJ international conference on intelligent robots and systems. pp. 5657–5662. IEEE (2006) 3
21. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 5308–5317 (2016) 4
22. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: Proceedings of the International Conference on Computer Vision (2021) 2, 4
23. Kalisiak, M., Van de Panne, M.: A grasp-based motion planning algorithm for character animation. The Journal of Visualization and Computer Animation **12**(3), 117–129 (2001) 3
24. Karunratanakul, K., Yang, J., Zhang, Y., Black, M., Muandet, K., Tang, S.: Grasping field: Learning implicit representations for human grasps. In: 8th International Conference on 3D Vision. pp. 333–344. IEEE (Nov 2020) 2, 4
25. Kaufmann, M., Aksan, E., Song, J., Pece, F., Ziegler, R., Hilliges, O.: Convolutional autoencoders for human motion infilling. In: 2020 International Conference on 3D Vision (3DV). pp. 918–927. IEEE (2020) 4, 9, 13
26. Krug, R., Dimitrov, D., Charusta, K., Iliev, B.: On the efficient computation of independent contact regions for force closure grasps. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 586–591 (2010) 3
27. Kry, P.G., Pai, D.K.: Interaction capture and synthesis. ACM Trans. Graph. **25**(3), 872–880 (Jul 2006) 3
28. Li, J., Villegas, R., Ceylan, D., Yang, J., Kuang, Z., Li, H., Zhao, Y.: Task-generic hierarchical human motion prior using vaes. In: 2021 International Conference on 3D Vision (3DV). pp. 771–781. IEEE (2021) 4
29. Li, X., Liu, S., Kim, K., Wang, X., Yang, M.H., Kautz, J.: Putting humans in a scene: Learning affordance in 3d indoor environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12368–12376 (2019) 4
30. Li, Y., Fu, J.L., Pollard, N.S.: Data-driven grasp synthesis using shape matching and task-based pruning. IEEE Transactions on Visualization and Computer Graphics **13**(4), 732–747 (2007) 3
31. Ling, H.Y., Zinno, F., Cheng, G., Van De Panne, M.: Character controllers using motion vaes. ACM Transactions on Graphics (TOG) **39**(4), 40–1 (2020) 4
32. Liu, L., Hodgins, J.: Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. ACM Transactions on Graphics (TOG) **37**(4), 1–14 (2018) 4
33. Liu, M., Pan, Z., Xu, K., Ganguly, K., Manocha, D.: Generating grasp poses for a high-dof gripper using neural networks. In: 2019 IEEE/RSJ International Con-

ference on Intelligent Robots and Systems (IROS). pp. 1518–1525. IEEE (2019) 3
34. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019) 3, 10
35. Makansi, O., Ilg, E., Cicek, O., Brox, T.: Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7144–7153 (2019) 4
36. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9489–9497 (2019) 4
37. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2891–2900 (2017) 2, 4
38. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019) 2, 5
39. Pollard, N.S., Zordan, V.B.: Physically based grasping control from example. In: Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation. pp. 311–318 (2005) 3
40. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems **30** (2017) 6
41. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11488–11499 (2021) 4, 5
42. Rijpkema, H., Girard, M.: Computer animation of knowledge-based human grasping. ACM Siggraph Computer Graphics **25**(4), 339–348 (1991) 3
43. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017) 4, 5
44. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S.: Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1349–1358 (2019) 4
45. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: Pigraphs: learning interaction snapshots from observations. ACM Transactions on Graphics (TOG) **35**(4), 1–12 (2016) 4
46. Seo, J., Kim, S., Kumar, V.: Planar, bimanual, whole-arm grasping. In: 2012 IEEE International Conference on Robotics and Automation. pp. 3271–3277 (2012) 3
47. Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. ACM Trans. Graph. **38**(6), 209–1 (2019) 4
48. Starke, S., Zhao, Y., Komura, T., Zaman, K.: Local motion phases for learning multi-contact character movements. ACM Transactions on Graphics (TOG) **39**(4), 54–1 (2020) 4
49. Taheri, O., Choutas, V., Black, M.J., Tzionas, D.: Goal: Generating 4d whole-body motion for hand-object grasping. arXiv preprint arXiv:2112.11454 (2021) 5

50. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: European Conference on Computer Vision (ECCV) (2020) 2, 3, 4, 10, 11
51. Tai, L., Zhang, J., Liu, M., Burgard, W.: Socially compliant navigation through raw depth inputs with generative adversarial imitation learning. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 1111–1117. IEEE (2018) 4
52. Tan, F., Bernier, C., Cohen, B., Ordonez, V., Barnes, C.: Where and who? automatic semantic-aware person composition. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1519–1528. IEEE (2018) 4
53. Wang, B., Adeli, E., Chiu, H.k., Huang, D.A., Niebles, J.C.: Imitation learning for human pose prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7124–7133 (2019) 4
54. Wang, J., Xu, H., Xu, J., Liu, S., Wang, X.: Synthesizing long-term 3d human motion and interaction in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9401–9411 (2021) 4, 13, 14
55. Yan, S., Li, Z., Xiong, Y., Yan, H., Lin, D.: Convolutional sequence generation for skeleton-based action synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4394–4402 (2019) 4
56. Yan, X., Rastogi, A., Villegas, R., Sunkavalli, K., Shechtman, E., Hadap, S., Yumer, E., Lee, H.: Mt-vae: Learning motion transformations to generate multimodal human dynamics. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 265–281 (2018) 4
57. Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 11, 13
58. Zhang, H., Ye, Y., Shiratori, T., Komura, T.: Manipnet: neural manipulation synthesis with a hand-object spatial representation. ACM Trans. Graph. **40**, 121:1–121:14 (2021) 3, 4
59. Zhang, S., Zhang, Y., Bogo, F., Pollefeys, M., Tang, S.: Learning motion priors for 4d human body capture in 3d scenes. In: IEEE/CVF International Conference on Computer Vision (ICCV 2021) (2021) 4, 5, 9, 10, 13
60. Zhang, Y., Black, M.J., Tang, S.: We are more than our joints: Predicting how 3d bodies move. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3372–3382 (2021) 2, 4, 5, 7, 13
61. Zhang, Y., Yu, W., Liu, C.K., Kemp, C., Turk, G.: Learning to manipulate amorphous materials. ACM Transactions on Graphics (TOG) **39**(6), 1–11 (2020) 4
62. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 5