COOP: Decoupling and Coupling of Whole-Body Grasping Pose Generation

Yanzhao Zheng Yunzhou Shi Yuhao Cui Zhongzhou Zhao Zhiling Luo† Wei Zhou
Alibaba DAMO Academy
{zhengyanzhao.zyz,yunzhou.syz,cyh262498,zhongzhou.zhaozz,godot.lzl,fayi.zw}@alibaba-inc.com

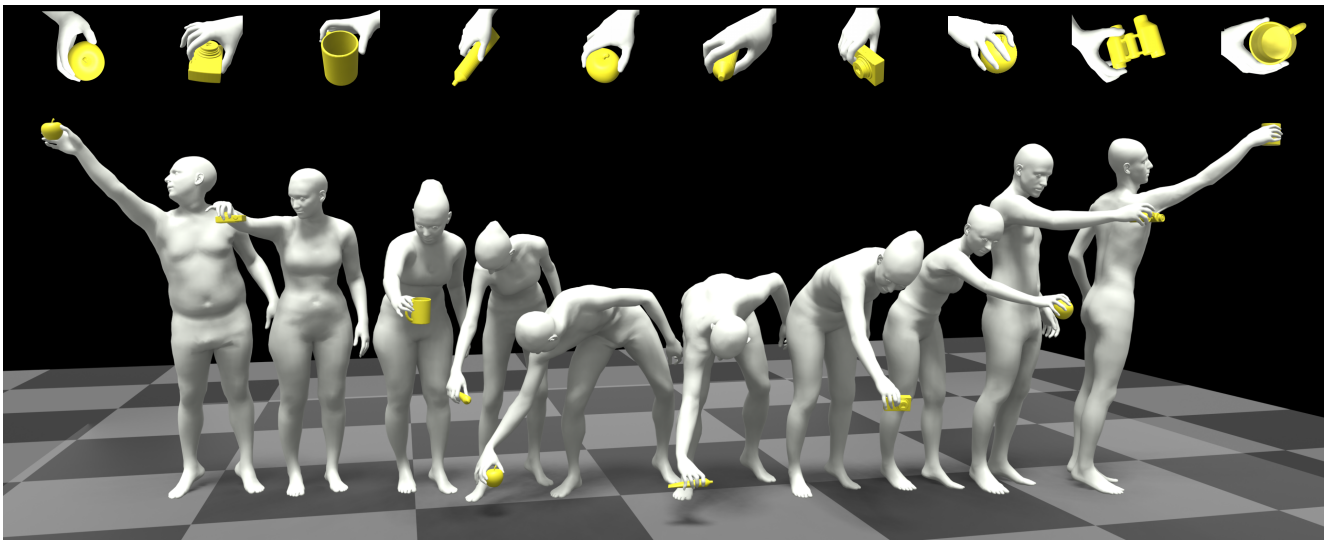

Figure 1. COOP can consistently generate realistic whole-body grasping poses as the position of the object changes. The figure shows generated poses for 10 people grasping novel objects in the vicinity of them.

Abstract

Generating life-like whole-body human grasping has garnered significant attention in the field of computer graphics. Existing works have demonstrated the effectiveness of keyframe-guided motion generation framework, witch focus on modeling the grasping motions of humans in temporal sequence when the target objects are placed in front of them. However, the generated grasping poses of the human body in the key-frames are limited, failing to capture the full range of grasping poses that humans are capable of.

*To address this issue, we propose a novel framework called **COOP** (De**CO**upling and C**O**upling of Whole-Body Gras**P**ing Pose Generation) to synthesize life-like whole-body poses that cover the widest range of human grasping capabilities. In this framework, we first decouple the whole-body pose into body pose and hand pose and model them separately, which allows us to pre-train the body model with out-of-domain data easily. Then, we couple these two generated body parts through a unified optimization algorithm.*

Furthermore, we design a simple evaluation method to evaluate the generalization ability of models in generating grasping poses for objects placed at different positions. The experimental results demonstrate the efficacy and superiority of our method. And COOP holds great potential as a plug-and-play component for other domains in whole-body pose generation. Our models and code are available at https://github.com/zhengyanzhao1997/COOP.

1. Introduction

Grasping is a common activity in human daily life. To make virtual humans act realistically in a 3D scene, an avatar needs to be able to perform diverse grasping poses, similar to those of real humans, adapting to the targets from different positions.

In this paper, we focus on the task of synthesizing life-like whole-body grasping poses with objects located at var-

†Corresponding author.

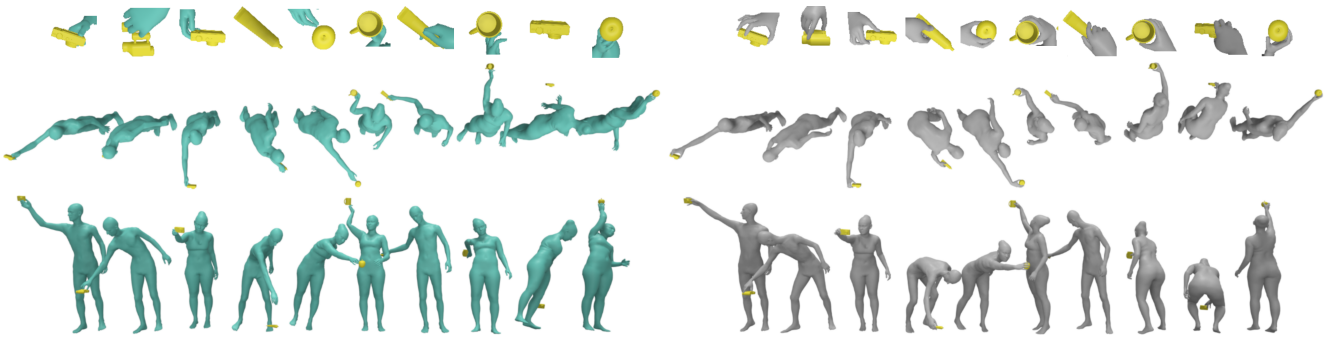

Figure 2. Whole-body grasping poses generated by GNet [36] (Left) and COOP (Right) for 10 testing samples with 2 genders.

ious heights in the vicinity of a human. It is worth emphasizing that this task does not require modeling any temporal aspects, such as moving to the object. Although existing works [36, 40] can generate life-like grasping poses of humans, the objects in the generated key-frames are often concentrated close to the human body due to limitations in training data and generation methods. As shown on the left side of Fig. 2, [36] generates the body and hand pose as a whole. When the position of the given object is far away from the human or beyond the scope of the training data, either the body pose is distorted and unbalanced or the grasping hand is far away from the target object in the generated grasping key-frame. Although IK (Inverse Kinematic) [30] can force the hand to touch the given object by adjusting the generated body pose, it can also result in some side effects such as unrealistic body pose or floating feet.

Since whole-body grasping involves multiple body parts such as the body, feet, and hands, the realism of a grasping pose needs to be evaluated from multiple aspects. Firstly, the generated hand should fit the given 3D object with high contact and low penetration. Secondly, the generated body pose should resemble that of a real human while avoiding ground penetration or floating. Finally, the body poses need to be balanced.

To address these challenges, we propose a novel framework called COOP (for an overview, see Fig. 3). Due to the decoupling and subsequent coupling of body parts, COOP can simultaneously consider the precision of hand grasping pose and the fidelity of the body pose. In the whole-body pose generation stage, we design two networks (HNet and BNet) to generate hand poses and body poses separately. In HNet, we introduce a fine-grained representation of hand-object contact, which we call Point2Finger Contact Map (PF-map), as an important intermediate information for guiding the generation of the grasping hand pose. In BNet, we propose the Body Graph Transformer (BGT) to generate the body pose conditioned on the target position of the right wrist. The Body Graph Transformer mainly consists of the Pose Graph Layer we proposed, which can explicitly encode the hierarchy of body joints. Based on the decoupling, we can easily pre-train the BGT with out-of-domain data, witch significantly improves the performance. In the unified optimization stage, we design the Stitch Loss to couple these two generated body parts. We further utilize the sampled PF-map as a contact constraint to improve the hand grasping and introduce the Body Balance Loss to ensure the generated body pose is balanced.

Contributions. Our work makes the following contributions:

1. COOP, a novel generation framework that can synthesize life-like whole-body grasping pose, given different 3D objects in various positions;
2. Body Graph Transformer, a body pose generation model that explicitly encodes the hierarchy of body joints, along with a unique pre-training method that can use a large amount of out-of-domain data for pre-training;
3. Our framework and pre-training method can be easily applied to other domains in whole-body pose generation or downstream tasks as a plug-and-play component.
4. We propose an evaluation method that can test the generalization ability of different whole-body grasping pose generation methods on object positions. The experimental results demonstrate the efficacy and superiority of our method.

2. Related work

Hand Grasp Synthesis. Hand Grasp Synthesis is a challenging task and has been widely studied in both robotics and computer graphics [4–6, 8, 9, 13, 22, 48]. In the field of robotic grasping [6, 22, 31], many works have proposed methods to plan collision-avoiding trajectories for robotic grippers to grasp objects. Most previous research in computer graphics [14, 19, 25, 26] has focused on synthesizing realistic hand grasping in 3D virtual environments based on physics simulations. In recent years, deep learning has been widely applied to hand grasp synthesis [5, 13, 15, 37, 42]. [15] proposed Grasping Field to represent the interaction

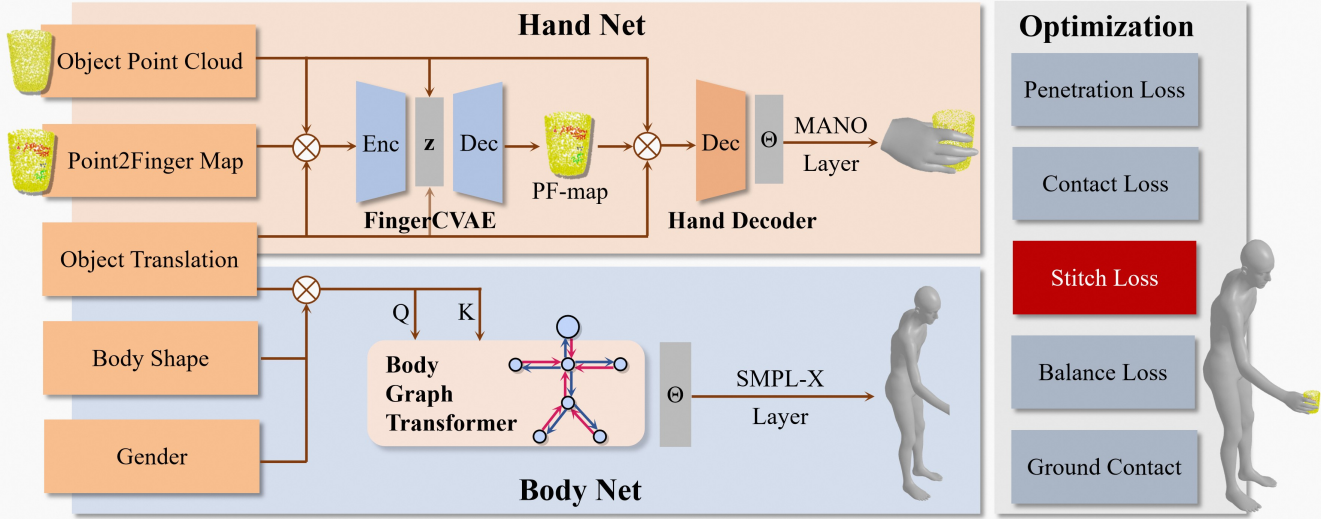

Figure 3. Overview of COOP. Given the 3D shape and global translation of the target object, Hand Net first samples a PF-map and generates a right grasping hand. The Body Net generates the SMPL-X parameters of the body, excluding the right hand, conditioned on the object translation and body shape. Finally, the unified optimization module couples these two body parts and optimizes them.

between hand and object. The hand mesh is generated by fitting MANO model [27] to the hand point cloud reconstructed from the representation. However, the quality of generated hand poses largely depends on the generated representation, and the generated hand may not fit the given object well. [37] adopted a coarse-to-fine pipeline for 3D grasp predictions. Given a 3D object, it first generates a reasonable grasping pose by CoarseNet and then refine the distance between hand and object by RefineNet. [13] argues that it is crucial to model the consistency between the hand contact points and object contact regions, and introduces hand-object contact consistency to further update the model parameters. In contrast, [3, 41] found a strong association between contacted binoculars points with fingers and used this information to optimize hand gripping poses further.

While some works have focused on static interactions, others have focused on dynamic grasp synthesis [4, 5, 42, 44, 46]. [5] formulated the dynamic grasp synthesis task as a reinforcement learning problem and proposed a policy learning approach that leverages a physics simulation. [42] proposed a neural network-based motion synthesis system that can generate detailed finger motions for one-/two-hand dexterous object manipulation. [4] presented a simple model-free framework that can learn to reorient objects with both the hand facing upwards and downwards.

Whole-Body Grasp Synthesis. In contrast to the hand grasp synthesis task, which only considers the hand, whole-body grasp synthesis considers the body, feet, and hands jointly. [35] proposed a body pose taxonomy for locomanipulation tasks. [12] introduced imitation learning to teach a robot to grasp objects using unrealistic humanoid models and simple objects. To generate a realistic whole-body grasp with complex 3D objects, [2, 37] captured a dataset named GRAB. Based on this dataset, two existing works [36, 40] generate whole-body grasp motion using a pipeline containing a grasping pose generation module and a motion filling module. The grasping pose generation module first generates a SMPL-X [24] whole-body grasping pose with the right hand in contact with the target object, and then the motion filling module fill the motions between the given start pose and the generated key-frame pose. [36] proposes GNet, a two-stage grasping pose generation module in which a trained CVAE network first generates a coarse whole-body grasping pose with a certain body shape, and then an optimization module optimizes the generated SMPL-X parameters to improve the fidelity based on several designed constraints. [40] also built a similar two-stage pipeline to generate the whole-body grasping pose. However, unlike [36] which generates a whole-body mesh directly, [40] first generates several critical markers on the body mesh based on the given object, and then recovers the body shape and SMPL-X parameters using a contact-aware pose optimization algorithm, which means that the shape, gender, and translation of the generated body pose are uncontrollable. Compared to the task of synthesizing missing middle frames between two key-frame poses [7, 10, 16], how to generate key-frame poses of whole-body grasping that match the grasping ability of real humans has not been fully studied. Therefore, we focus on building a generative method that can synthesize realistic key-frame poses to

grasp objects within a wide range.

3. Method

Problem definition. It is worth emphasizing again that our focus is to generate key-frame poses of human grasping objects at different positions. To encourage the model to generate a various poses, such as bending, squatting, tiptoeing, turning, etc., we do not require modeling any temporal aspects. Therefore, we fix the human's horizontal coordinates at the origin of the 3D global coordinate system. Given the target objects, represented by point cloud $\mathcal{P}$, in the vicinity of the human, we need to generate life-like whole-body grasping poses Θ, represented by SMPL-X[24], with body shape β and gender G as conditions.

As shown in Figure 3, the framework of COOP consists of two separate components trained during the training phase: Hand Net (HNet), which generates diverse hand grasping poses based on the object shape, and Body Net (BNet), which generates a stable body pose with the right wrist reaching the target position. During inference, the unified optimization module couples these two body parts and optimizes them together simultaneously.

3.1. Hand Net (HNet)

Based on the existing concept of fine-grained contact map, we are the first to introduce the **Point2Finger Map** as an important condition to control the generation of hand poses. It significantly outperforms traditional binary contact map and further enhances precision when combined with the **Finger Contact Loss** we designed.

Point2Finger Map (PF-map). [3, 41] have found a strong association between contacted binoculars points with fingers. Based on this observation, we introduce a more fine-grained representation of hand-object contact similar to that of [3, 41], which we named Point2Finger Map (PF-map). As shown in the HNet in Fig. 3, we pre-define 5 prior contact regions of the right hand motivated by [13]. For each grasping sample, we calculated the nearest finger joint for each contact point in the object point cloud and label the contact category for it.

HNet. As shown at the top of Fig. 3, the HNet we build contains FingerCVAE and a hand decoder. FingerCVAE is a Conditional Variational Auto-Encoder [34] based generative network, which is trained to sample diverse PF-maps instead of the binary contact map. And the hand decoder generates the corresponding MANO [27] parameters of the right hand based on the PF-map and hand shape $\beta^h \in \mathbb{R}^{10}$. We introduce the translations of objects as an additional condition to control the rotation of the wrist. To better capture the distribution information of the contact points, we apply Point Cloud Transformer [45] as the encoder and decoder of FingerCVAE and the hand decoder.

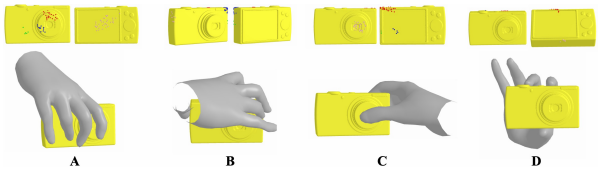

Figure 4. The PF-maps generated by FingerCVAE are displayed at the top, with the corresponding generated grasping hands presented below. The heights of the object positions are A: 0.4m, B: 0.8m, C: 1.2m, D: 1.6m.

Training. During the training stage, the input features consist of the centered point cloud feature of the object $\mathcal{P}^o \in \mathbb{R}^{(N\times 6)}$ (where N is the number of points, and 6 represents the 3D locations and 3 normal features), and the PF-map $\mathcal{C} \in \mathbb{R}^{(N\times E^f)}$ (where E^f is the dimension of the contact categories embedding). To ensure that the generated hand poses match the body pose better, we introduce an additional normalization step for the translation of the object $t^o \in \mathbb{R}^3$. The normalized translation is used as a condition to control the direction of the generated grasping. All of these features are concatenated point-wise and input to the encoder of FingerCVAE. The encoder outputs the mean $\mu \in \mathbb{R}^{16}$ and variance $\sigma^2 \in \mathbb{R}^{16}$ of the posterior Gaussian distribution $\mathcal{Q}(z|\mu, \sigma^2)$ [18]. The decoder of FingerCVAE takes the concatenation of the latent code z, the point cloud feature, and the object translation as input to reconstruct the PF-map. The training objective of FingerCVAE is given by:

$$\mathcal{L}_{FingerCVAE} = \mathcal{L}_{ce} + \lambda_{KL}\mathcal{L}_{KL} \tag{1}$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss of reconstructed contact categories of the PF-map and $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence, and λ_{KL} is the weight.

The hand decoder learns to predict the MANO parameters Θ^h of the grasping hand, which is conditioned on the PF-map and the object translation. Here, $\Theta^h = \{\theta^h, t^h\}$ represents the hand pose $\theta^h \in \mathbb{R}^{15\times 6}$ and the relative translation $t^h \in \mathbb{R}^3$ of the hand with respect to the object. The differentiable MANO layer takes the parameters Θ^h and outputs the mesh vertices $\mathcal{V}^h$ and faces $\mathcal{F}^h$ of the hand. The training loss is given by:

$$\mathcal{L}_{hand} = \|\Theta^h - \widehat{\Theta^h}\|_2 + \|V^h - \widehat{V^h}\|_1 \tag{2}$$

where $\widehat{\Theta^h}$ and $\widehat{\mathcal{V}^h}$ are the parameters and vertices of the ground true hand pose/mesh.

Inference. During the inference time, given the object point cloud, the decoder of FingerCVAE samples a PF-map conditioned on the object translation and the latent code z randomly sampled from a Gaussian distribution. The hand decoder uses the sampled PF-map to form the grasping hand mesh. Fig. 4 shows the high correlation between the sampled PF-maps and hand poses. Given a novel object placed

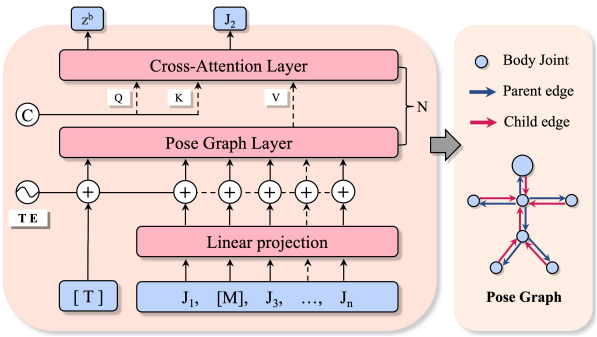

Figure 5. Architecture of Body Pose Graph Transformer. The rotation of body joints is input as token embedding after linear projection. $[T]$ is a special token used to predict the z-coordinate z^b of the body. TE represents the token type embedding which will be added to the input tokens to distinguish different joints. Each block of BGT includes one Pose Graph Layer and one Cross-Attention Layer. Pose Graph Layer encodes the hierarchy of body joints by applying modified relative self-attention, and the cross-attention layer induces the condition information C such as the target position of the right wrist, human gender, and body shape.

at different positions, the sampled PF-maps adapt to the object translation and determine the direction of grasp. To further improve the accuracy of the hand grasping, we utilize the sampled PF-map to refine the hand pose in the optimization module, as detailed in Sec. 3.3.

3.2. Body Net (BNet)

We propose **Body Graph Transformer**, a novel body model which can explicitly models the body joint hierarchy. Along with the unsupervised pre-training method, we are free to pre-train our model with large amounts of out-of-domain data.

Task definition. We adopt the SMPL-X model to represent the human body. Due to the decouping of the body parts, BNet doesn't need to pay attention to the hand pose, and its objective is to generate the SMPL-X parameters Θ^b of the body with the right wrist at the target position. $\Theta^b = \{\theta^b, t^b\}$, $\theta^b = J \in \mathbb{R}^{n\times 6}$ represents the articulated rotations of joints [49], where n is the number of joints to be predicted. In this work $n = 40$, which equals to the total number of joints of SMPL-X minus the number of right-hand joints. And $t^b \in \mathbb{R}^3$ represents the body's translation. Since the horizontal position of the human body is fixed at the origin, BNet only predicts the vertical coordinate, represented by the z-axis coordinate $z^b \in \mathbb{R}^1$, of the human body.

Body Pose Graph. To explicitly encode the SMPL-X joint hierarchy, which is defined by a kinematic tree that keeps the parent relation for each joint, we convert the hierarchy into a directed graph $\mathcal{G} = \langle \mathcal{V}_j, \mathcal{E} \rangle$, where the vertices $\mathcal{V}_j$ represent the body joints and the directed edges $\mathcal{E}$ represent the relations between joints. The edges are categorized as parent, child, or none.

Body Graph Transformer (BGT). The architecture of the proposed Body Graph Transformer is illustrated in Fig. 5. Unlike prior work [36, 37], which directly flattens the joint rotation parameters and feeds them into the network, we modify the Transformer architecture [39] to encode the body pose graph. In this model, each joint $J_i \in \mathbb{R}^6$, treated as a vertex in the body graph, is input as an independent token. It is linearly projected to the high-dimensional embedding $E_i^J \in \mathbb{R}^H$, where H is the hidden size of BGT. To enable the model to distinguish different joints, we add a learnable token type embedding $TE \in \mathbb{R}^{(n+1)\times H}$ to the joint embedding.

Based on the body pose graph, the relationships between different joints are diverse, and many may be less relevant. Previous work [1] directly applied full-type self-attention [39], which may introduce excessive noise or misguidance. Instead, we propose **Pose Graph Layer (PGL)**, which employs a modified Relative Self-Attention Mechanism [32, 47] to explicitly encode the directed edges in the pose graph whose vertices are at the token level. We redefine the calculation of the self-attention module as follows:

$$e_{ij} = \frac{x_i W^Q (x_j W^K + r_{ij}^K)^T}{\sqrt{d_z}}, z_i = \sum_{j=1}^{n} \alpha_{ij} (x_j W^V + r_{ij}^V) \tag{3}$$

where $\alpha_{ij} = \underset{j}{softmax}\{e_{ij}\}$.

Specifically, we initialize learnable embedding $R^V, R^K \in \mathbb{R}^{3\times H}$ for each type of edge defined above. For all the input samples, we build a relation matrix $\mathcal{R} \subseteq (n \times n)$. $\mathcal{R}^{(i,j)}$ represents the index of relation type of token J_i for J_j. While computing the relative attention, we set the $r_{ij}^K = R^K_{\mathcal{R}^{(i,j)}} \in \mathbb{R}^H$ and $r_{ij}^V = R^V_{\mathcal{R}^{(i,j)}} \in \mathbb{R}^H$.

Condition encoding. To incorporate conditions to control the generated body poses, we stack a cross-attention layer on the Pose Graph Layer in each block. In this attention layer, the $\mathcal{K}$ and $\mathcal{V}$ are projected from the concatenated condition feature $\mathcal{F}_c = [t^w, \beta^b, G]$, where $t^w \in \mathbb{R}^3$ represents the target right wrist position, $\beta^b \in \mathbb{R}^{10}$ represents the body shape, and $G \in \mathbb{R}^g$ represents the gender embedding with g being the embedding size for each gender.

Translation decoding. We initialize a special token $[T] \in \mathbb{R}^H$ and concatenate it to the joint embedding, and set up a sub-task for the model to predict the z-coordinate z^b of the body from this special token.

Unsupervised pre-training. By utilizing the Transformer-like model structure and special objectives that generate body poses where the right wrist reaches a designated target position, BGT can effectively learn the prior knowledge of the hierarchy of body joints through pre-training with out-of-domain data.

We pre-train BGT using AMASS [23]. We randomly mask 50% joints embedding to encourage the model to recover them conditioned on the right wrist position. To prevent the model from being confused by masking joints that are weakly correlated with the right wrist position, we mask only 20 critical joints. The training objective is as follows:

$$\begin{aligned}\mathcal{L}_{pre} = &\sum_{k=1} \|J_k^{mask} - \widehat{J_k^{mask}}\|_2 + \|z^b - \hat{z^b}\|_1 \\ &+ \|V^b - \widehat{V^b}\|_1 + \|t_w^b - t^w\|_1\end{aligned} \tag{4}$$

where the distance between the right wrist position t_w^b and the target position t^w is minimized.

It's worth emphasizing that the entire pre-training process is unsupervised, which is an advantage of our framework that previous methods cannot achieve. Our framework, when combined with this pre-training method, can be easily applied to other generation domains as a plug-and-play component. By merely switching the target joint from the wrist to other joints like legs, head, and hips.

Fine-tuning. During the fine-tuning stage, we mask all joint embeddings directly, which encourages BGT to generate the ground truth grasping body pose from scratch based solely on the target position of the right wrist.

Inference. At the inference stage, since the accurate position of the right wrist is unavailable, BGT takes the translation of the target object t^o as the rough condition and generates a stable grasping pose, with the right wrist positioned in close proximity to the object.

3.3. Unified Optimization (U-Opt)

Given the MANO parameters Θ^h and SMPL-X parameters Θ^b initialized by the generation from HNet and BNet, we propose a unified optimization algorithm (U-Opt) to couple and optimize these two body parts with gradient descent using Adam [17]. U-Opt leverages the PF-map sampled by FingerCVAE and the constraint algorithm we developed for body rationalization to simultaneously enhance the fidelity of the body pose and grasping-hand pose.

The optimization objectives of U-Opt are:

$$\begin{aligned}E(\Theta) &= \lambda_s E_{stitch}(\Theta^b) + E_h(\Theta^h) + E_b(\Theta^b) + \lambda\|\Theta\|_2, \\ E_h(\Theta^h) &= \lambda_c E_{contact} + \lambda_p E_{penet}, \\ E_b(\Theta^b) &= \lambda_b E_{balance} + \lambda_g E_{ground} + \lambda_h E_{head}\end{aligned} \tag{5}$$

Finger Contact Loss. We further improve the stability and accuracy of the hand grasping by minimizing the consistency loss between the hand vertices $\mathcal{V}^h$ and the predicted PF-map $\mathcal{C}$:

$$E_{contact}(\Theta^h) = \sum_{f=1}^{5} \sum_{o \in C_f} d(o, \mathcal{V}_f^h), \tag{6}$$

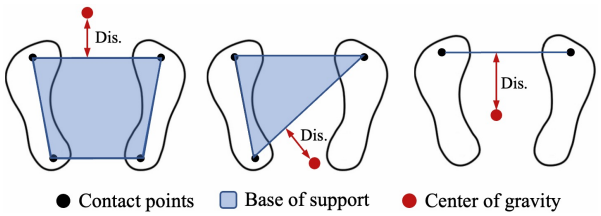

Figure 6. Three cases of minimizing Body Balance Loss. The base of support is a closed polygon calculated by the pre-defined points in contacted with the ground on the feet. If there are only two contact points, the base of support turns into a line segment. The goal of Balance Loss is to minimize the distance between the projection of the center of gravity and the supporting polygon.

where C_f are the points corresponding to the finger f in the contact point cloud, $\mathcal{V}_f^h$ are the vertices in the prior contact regions of finger f. And $d(a, B) = \min_{b \in B} \|a - b\|_2$ is the minimum distance from point a to point cloud B.

Penetration Loss E_{penet} is implemented to penalize the penetration between hand and object following [13].

Stitch Loss. To couple the generated body pose and hand pose, Stitch Loss is proposed to minimise the distance between the right wrist of the body from BNet and that from HNet. Note that Stitch Loss only adjusts the body pose parameters Θ^b, while ensuring that the accuracy of the hand grasp is not affected:

$$E_{stitch}(\Theta^b) = \|t_w^b - t_w^h\|_2 \tag{7}$$

where t_w^b is the right wrist position of the body from BNet and t_w^h is the wrist position of the hand from HNet.

Body Balance Loss. Previous studies [20, 28, 29] have established that the balance of a human is only sufficient when the gravity line is aligned with the base of support. Building on this insight and inspired by the successful implementation of [33], we are the first to introduce **Body Balance Loss** in SMPL-Body pose generation and carefully designed a novel calculation method for this loss.

Specifically, we encourage that the vertical projection of the body's center of gravity (approximated by the hip joint) to pass through the base of support formed by four pre-defined support points located at the forefoot and heel, as illustrated in Fig. 6. Body Balance Loss is defined as

$$E_{balance}(\Theta^b) = D(J_{hip-xy}^b, S) \tag{8}$$

where $D(p, P)$ is the distance between point p to the polygon P, it will be zero if p is in P. J_{hip-xy}^b is the vertical projection of the hip to the ground, and S represents the base of support formed by the ground contact points of feet.

Ground contact Loss E_{ground} is implemented to keep the feet contacting the ground, where we encourage the z-coordinate of the lowest vertices of two feet to be zero.

Head orientation Loss E_{head} is implemented to encourage the head of the human body to face the object.

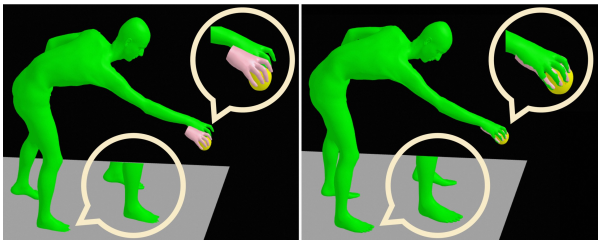

Figure 7. The body pose (green) and hand pose (pink) generated by COOP before optimization (left) and after optimization (right).

Model	FingerCVAE		Hand Decoder
	Encoder	Decoder	
Nblocks	4	4	4
Nneighbor	16	16	16
Transformer dim	256	256	512
Contact dim	4	-	4
Npoints	2048	2048	2048
Latent dim	8	8	-

Table 1. The main hyperparameter settings of HNet. Contact dim represents the embedding size of each contact category.

Layers	Attention head	Hidden size	Dropout	Gender
8	8	256	0.2	16

Table 2. The main hyperparameter settings of the Body Graph Transformer. Gender represents the embedding size of gender.

3.4. Whole-Body Pose Synthesis

After finishing U-Opt, we compose the hand pose parameters $\theta^h \in \mathbb{R}^{15\times6}$ and the body pose parameters $\theta^b \in \mathbb{R}^{40\times6}$ into the complete SMPL-X parameters $\Theta = \{\theta \in \mathbb{R}^{55\times6}, t^b\}$ as the final output. Fig. 7 displays a case of the unified optimization.

4. Experiment

4.1. Model Architecture

HNet. We utilize the Point Cloud Transformer [45] as the baseline model for FingerCVAE and the hand decoder in HNet. We replicated most of the hyperparameter settings from [45] and adjusted the model structure to suit the task of HNet. In particular, we only retain the transition-down layers for the Encoder of FingerCVAE and added two fully connected layers to map the global feature to the parameters of a normal distribution, $\mu, \sigma \in \mathbb{R}^8$. The hand decoder, like the FingerCVAE Encoder, maps the hand-object contact information (PF-map concatenated with the object point cloud and object translation) to the MANO [27] parameters Θ^h of the grasping right hand. Meanwhile, for the FingerCVAE Decoder, we adopted the settings from the "Part Segmentation" task, keeping the transition-down and transition-up layers to categorize the points in the object point cloud into 6 different categories. Table 1 presents the primary hyperparameter settings for HNet.

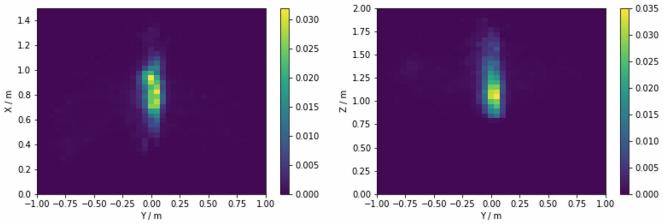

Figure 8. The distribution of translated objects in the training set is expressed as a percentage. The 3D global coordinate system is used, with X and Y as the horizontal axes and Z as the vertical axis.

Body Graph Transformer (BGT). Table 2 presents the primary hyperparameter settings for our proposed Body Graph Transformer (BGT). To input the 6D rotation of each body joint $J_i \in \mathbb{R}^6$ [49] and the condition feature $\mathcal{F}_c \in \mathbb{R}^{(3+10+16)}$ into the model, we separately map them into hidden size embeddings using linear projection.

4.2. Datasets

We utilize **GRAB** [2, 37], a dataset capturing the whole-body 3D SMPL-X human grasping sequences, to train and evaluate HNet and BNet. We down-sample the motion sequences of GRAB from 120fps to 30fps, then collect the frames in which the object is only in contact with the human's right hand and leaves the table. Moreover, we filter out frames in which the object is too close to the human, causing deviations from grasping poses (e.g. eyes not looking at the object). Subsequently, for each frame, we normalize the translation of human to the origin of the horizontal plane. To ensure no overlap between objects in the training, validation, and testing set, we divide the dataset according to the objects as done in [36]. The distribution of object translation in the training set we created is shown in Fig. 8.

AMASS [23] is a vast dataset of human motion that comprises multiple action types, such as sports, dance, and more. Although these action types differ significantly from grasping, which we refer to as out-of-domain data, we can still leverage them to pre-train the BGT due to the decoupling method used in COOP. We down-sampled the motion sequences from AMASS to 30fps and filtered out frames where at least one foot of the human is in contact with the ground. For each frame, we normalize the translation of human to the origin of the horizontal plane.

4.3. Testing set

To evaluate the generalization of different methods for grasping objects from different positions, we build a test set with object positions distributed far beyond the range

Method	Contact Ratio ↑	Valid Error (e^{-2}) ↓	Ground Contact ↑	Penetration Volume*(cm^3) ↓	Displacement* (cm) ↓
GrabNet [37]-SMPLX	0.015	**0.114**	0.294	16.102	7.023
GNet [36]	0.676	0.312	0.748	8.172	4.365
WholeGrasp [40]	0.948	0.258	0.495	4.335	4.871
COOP (ours)	**0.987**	0.223	**0.996**	**4.098**	**3.592**
GT-GRAB [37]	1.000	0.131	1.000	7.017	3.553

Table 3. Results of the comparison experiment on the test data we build. ***: Notably**, in calculating the metrics of grasping quality, i.e., penetration and displacement, we only consider the samples where the human body and object were in contact. This was done because when the body mesh is not in contact with the object, the penetration value is 0, and the displacement of the object becomes a very large value, rendering the result meaningless. However, even with this evaluation logic, COOP still achieves significant advantages.

Metric	GNet [36]	WholeGrasp [40]	COOP (ours)
Overall Pose ↑	2.55 ±1.07	2.29 ±0.90	**3.89** ±0.80
Body Balance ↑	2.66 ±1.21	2.45 ±0.98	**3.97** ±0.82
Ground Contact ↑	2.27 ±1.13	1.78 ±1.04	**3.86** ±0.72
Hand Grasp ↑	2.43 ±1.16	2.35 ±1.07	**3.70** ±0.92
Average ↑	2.48 ±1.15	2.22 ±1.03	**3.85** ±0.82

Table 4. The average scores and standard deviation of the results obtained from the perceptual study.

of the training set. For each testing group, consisting of one unseen object and one specific body shape, we place a novel object at 128 different 3D positions surrounding the human in global coordinates and evaluate the generated whole-body grasping poses. The x-coordinate (the human's T-pose is facing) of the positions ranged from -0.8 to 0.8m, the y-coordinate ranged from -0.6 to 0.6m, and the z-coordinate, which represents the vertical position in our paper, ranged from 0.2 to 1.8m. The object orientation remained fixed within each testing group but varied across groups, resulting in different relative rotations of the object with respect to the human at different test positions. The test set comprised a total of 6400 samples, with 5 novel objects and 10 different human body shapes.

4.4. Evaluation Metrics

(1) Contact Ratio [43]. To evaluate the validity of the grasping pose, we measure the ratio of the generated samples with body-object contact. **(2) Valid Error.** We utilize VPoser [24], a learning-based variational human pose prior, to evaluate the authenticity of the generated body pose. Specifically, we calculate the L2 loss of vertex reconstruction as Valid Error. Although VPoser tends to output lower error for more common poses, high valid error can represent the great likelihood of an impossible body pose to a certain extent. **(3) Ground Contact** represents the percentage of physically plausible human-to-ground contact in the generated samples. A sample is considered "good" if the distances between both feet's lowest vertex and the ground are less than 3cm. **(4) Penetration** is measured by penetration volume between the object and generated body mesh following [11, 13]. **(5) Grasp Displacement.** To measure the stability of the grasping, we utilize physics simulation to compute the displacement of the object subjected to gravity following [11, 38]. It should be noted that the calculated frictional force applied to the object has a positive correlation with the interpenetration. So the grasping stability must be analyzed in conjunction with the penetration.

4.5. Quantitative Evaluation

We compare our method with three others that we retrained on the training set we build. **"GrabNet-SMPL-X"** is a variant of GrabNet [37] that adapted to SMPL-X parameters. **GNet-GOAL** [36] and **WholeGrasp-SAGA** [40] are the current SOTA works for whole-body grasping poses generation, including the optimization stage. Although the shape, gender, and translation of the generated body pose are uncontrollable in WholeGrasp, we replicated it for our task by fixing the body parameter and horizontal translation during the optimization stage.

The evaluation results on the test set we build are presented in Table 3. While the object positions in the ground truth by GRAB are not entirely consistent with those in our testing samples, we display the evaluation results as a reference for actual grasping poses. The low Valid Loss achieved by GrabNet-SMPL-X is due to the majority of generated poses being "frozen" into invalid regular grasping poses. Our approach outperforms other methods by demonstrating a strong ability to generalize across different object positions while ensuring body poses rationality. Although WholeGrasp [40] achieves a high contact ratio by setting a strong contact constraint in the optimization, it can easily lead to unnatural body poses such as ground penetration or floating. As for the grasping quality, our method achieves not only smaller penetration but also stronger stability.

4.6. Perceptual Evaluation

We conduct a perceptual study to further evaluate the generated grasping poses in visual perception. For each generated sample, 10 users are asked rate it on four met-

	Contact. ↑	Valid Error ↓	MAE-w(cm) ↓
Ours	**1.000**	**0.215**	**12.672**
w/o PGL	0.998	0.316	14.492
w/o fine-tuning	0.985	1.119	25.650
w/o pre-training	0.991	1.396	27.004
w/o pre. + PGL	0.942	2.165	39.318

Table 5. Results from the ablation study on model structure and training methods. **MAE-w** represents the distance error between the target position and the right wrist of the generated body without optimization. **w/o PGL**: we replace Pose Graph Layer with the normal self-attention layer in BGT.

	Contact. ↑	Pen Vol.* ↓	Disp.* ↓	V2V(cm) ↓
Ours	**1.000**	4.048	**3.190**	**1.950**
w/o PF-map	0.902	**2.311**	6.656	2.560

Table 6. Results from the ablation study on PF-map. The PF-map in HNet is replaced by the binary contact map, making HNet degrade to be similar to the current state-of-the-art hand-grasp poses generation model [21]. **V2V** is the vertex-to-vertex error for the hand mesh generated by the hand decoder on the GRAB test set.

	Contact. ↑	Pen. Vol.* ↓	Disp.* ↓
w/o U-Opt	0.377	26.091	4.873
E_b + E_{stitch}	0.978	6.500	3.274
E_b + E_{stitch} + $E_{contact}$	**1.000**	10.301	**2.035**
E_b + E_{stitch} + $E_{contact}$ + E_{penet}	**1.000**	**4.048**	3.190

Table 7. Results from the ablation study on the losses in U-Opt.

Method	Overall Pose ↑	Body Balance ↑
Ours	**0.61** ±0.49	**0.63** ±0.48
w/o $E_{balance}$	0.39 ±0.49	0.37 ±0.48

Table 8. Result of perceptual ablation study on $E_{balance}$. Every two generated poses are compared by the users, with a score of 1 assigned to the preferred pose and 0 to the non-preferred pose.

rics using a 1-5 point scale. Table 4 shows that COOP outperforms others in all the metrics related to body and hand grasp. Additionally, 90.6% of the samples COOP synthesized are considered to be life-like in terms of Overall Pose, with an average score greater than 3.5 .

4.7. Ablation Study

To verify the effectiveness of the model structure, training methods, and optimization constraints we designed, we perform the ablation studies on the built test data with a randomly selected body shape.

Model design. Table 5 presents the results of the ablation study that verifies the effectiveness of different model structures and training methods. The results demonstrate that our proposed Pose Graph Layer enhances the accuracy of body poses generated by BNet. Moreover, both pre-training and fine-tuning further improve the fidelity of the grasping poses. The results suggests that when BNet is unable to generate a proper body pose, the pose may be significantly distorted after being optimized by U-opt. To demonstrate that PF-map, as a contact representation, brings more valid information than the binary contact map. We replace the PF-map with the ordinary binary contact map in HNet. Table 6 shows that the PF-map improves the fit of hand-object grasping and reduces confusion of the hand decoder.

Optimization Loss. Table 7 demonstrates the significance of U-Opt as a coupling module in COOP. It reveals that using $E_{contact}$ alone leads to improvement in grasping stability, but also results in increased penetration. By incorporating E_{penet}, a balance between penetration and grasping stability can be achieved. Additionally, we conduct a additional perceptual study to evaluate the effectiveness of $E_{balance}$. Table 8 indicates that $E_{balance}$ effectively enhances the visual balance of the generated grasping poses.

5. Limitations and Conclusion

Limitations. Considering the limited distribution of training data, we constructed our Body Net as a non-VAE structure, resulting in a lack of diversity in the generated poses for the same set of conditions. During the experiment, we found that when an object is positioned beyond the graspable range of the human, the generated pose either cannot make contact with the object or suffer from serious penetration.

Conclusion. In this paper, we introduce COOP, a decoupling and coupling framework for generating whole-body grasping poses. COOP can synthesize life-like whole-body grasping poses for target objects over a wide range of positions that resemble those of humans in the real world. In future research, we plan to study how to utilize these unseen generated grasping poses to synthesize dynamic motions. Additionally, COOP may be easily adapted for other human daily activities such as kicking and sitting. We will continue explore it in future work.

References

[1] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *3DV*, pages 565–574, 2021. 5

[2] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *CVPR*, 2019. 3, 7

[3] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *ECCV*, pages 361–378. Springer, 2020. 3, 4

[4] Tao Chen, Jie Xu, and Pulkit Agrawal. A system for general in-hand object re-orientation. In *Conference on Robot Learning*, volume 164, pages 297–307, 2021. 2, 3

[5] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *CVPR*, pages 20577–20586, 2022. 2, 3

[6] Renaud Detry, Dirk Kraft, Anders Glent Buch, Norbert Krüger, and Justus H. Piater. Refining grasp affordance models by experience. In *ICRA*, pages 2287–2293, 2010. 2

[7] Yinglin Duan, Yue Lin, Zhengxia Zou, Yi Yuan, Zhehui Qian, and Bohan Zhang. A unified framework for real time motion completion. *AAAI*, pages 4459–4467, 2022. 3

[8] Oliver Glauser, Shihao Wu, Daniele Panozzo, Otmar Hilliges, and Olga Sorkine-Hornung. Interactive hand pose estimation using a stretch-sensing soft glove. *TOG*, 38(4):41:1–41:15, 2019. 2

[9] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *CVPR*, pages 1471–1481, 2021. 2

[10] Félix G. Harvey, Mike Yurick, Derek Nowrouzezahrai, and Chris Pal. Robust motion in-betweening. *TOG*, 39(4):60, 2020. 3

[11] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, 2019. 8

[12] Kaijen Hsiao and Tomás Lozano-Pérez. Imitation learning of whole-body grasps. In *IROS*, pages 5657–5662, 2006. 3

[13] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, pages 11087–11096, 2021. 2, 3, 4, 6, 8

[14] Maciej Kalisiak and Michiel van de Panne. A grasp-based motion planning algorithm for character animation. *Comput. Animat. Virtual Worlds*, 12(3):117–129, 2001. 2

[15] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *3DV*, pages 333–344, 2020. 2

[16] Jihoon Kim, Taehyun Byun, Seungyoun Shin, Jungdam Won, and Sungjoon Choi. Conditional motion in-betweening. volume 132, page 108894, 2022. 3

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015. 6

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. 2014. 4

[19] Paul G. Kry and Dinesh K. Pai. Interaction capture and synthesis. *TOG*, 25(3):872–880, 2006. 2

[20] Jean-Charles Le Huec, R Saddiki, J Franke, J Rigal, and S Aunoble. Equilibrium of the human body and the gravity line: the basics. *Spine*, 20(5):558–563, 2011. 6

[21] Haoming Li, Xinzhuo Lin, Yang Zhou, Xiang Li, Yuchi Huo, Jiming Chen, and Qi Ye. Contact2grasp: 3d grasp synthesis via hand-object contact constraint. 2022. 9

[22] Min Liu, Zherong Pan, Kai Xu, Kanishka Ganguly, and Dinesh Manocha. Generating grasp poses for a high-dof gripper using neural networks. In *IROS*, pages 1518–1525, 2019. 2

[23] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. *ICCV*, pages 5441–5450, 2019. 6, 7

[24] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 3, 4, 8

[25] Nancy S. Pollard and Victor B. Zordan. Physically based grasping control from example. In *SIGGRAPH*, pages 311–318, 2005. 2

[26] Hans Rijpkema and Michael Girard. Computer animation of knowledge-based human grasping. In *SIGGRAPH*, pages 339–348, 1991. 2

[27] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *TOG*, 36(6), 2017. 3, 4, 7

[28] Pierre Roussouly, Sohrab Gollogly, Olivier Noseda, Eric Berthonnaud, and Johanes Dimnet. The vertical projection of the sum of the ground reactive forces of a standing patient is not the same as the c7 plumb line: a radiographic study of the sagittal alignment of 153 asymptomatic volunteers. *Spine*, 31(11):E320–E325, 2006. 6

[29] Frank Schwab, Virginie Lafage, Reid Boyce, Wafa Skalli, and Jean-Pierre Farcy. Gravity line analysis in adult volunteers: age-related correlation with spinal parameters, pelvic parameters, and foot position. *Spine*, 31(25):E959–E967, 2006. 6

[30] Lorenzo Sciavicco and Bruno Siciliano. A solution algorithm to the inverse kinematic problem for redundant manipulators. *IEEE J. Robotics Autom.*, 4(4):403–410, 1988. 2

[31] Jungwon Seo, Soonkyum Kim, and Vijay Kumar. Planar, bimanual, whole-arm grasping. In *ICRA*, pages 3271–3277, 2012. 2

[32] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *NAACL-HLT*, volume 2, pages 464–468, 2018. 5

[33] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ToG*, 39(6):1–16, 2020. 6

[34] Kihyuk Sohn, Xinchen Yan, and Honglak Lee. Learning structured output representation using deep conditional generative models. *NIPS*, pages 3483–3491, 2015. 4

[35] Júlia Borràs Sol and Tamim Asfour. A whole-body pose taxonomy for loco-manipulation tasks. In *IROS*, pages 1578–1585, 2015. 3

[36] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *CVPR*, pages 13253–13263, 2022. 2, 3, 5, 7, 8

[37] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, pages 581–600, 2020. 2, 3, 5, 7, 8

[38] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 118(2):172–193, 2016. 8

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, pages 5998–

6008, 2017. 5

[40] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *ECCV*, 2022. 2, 3, 8

[41] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *ICCV*, pages 11097–11106, 2021. 3, 4

[42] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *TOG*, 40(4):121:1–121:14, 2021. 2, 3

[43] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3d people in scenes without people. In *CVPR*, pages 6193–6203, 2020. 8

[44] Yunbo Zhang, Wenhao Yu, C. Karen Liu, Charlie C. Kemp, and Greg Turk. Learning to manipulate amorphous materials. *TOG*, 39(6):189:1–189:11, 2020. 3

[45] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. *ICCV*, pages 16239–16248, 2020. 4, 7

[46] Wenping Zhao, Jianjie Zhang, Jianyuan Min, and Jinxiang Chai. Robust realtime physics-based motion control for human grasping. *TOG*, 32(6):207:1–207:12, 2013. 3

[47] Yanzhao Zheng, Haibin Wang, Baohua Dong, Xingjun Wang, and Changshan Li. Hie-sql: History information enhanced network for context-dependent text-to-sql semantic parsing. *arXiv*, 2022. 5

[48] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object correspondence to hand for motion refinement. In *ECCV*, 2022. 2

[49] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019. 5, 7