DOI: 10.1111/cgf.14739
EUROGRAPHICS 2023 / K. Myszkowski and M. Nießner
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 42 (2023), Number 2

IMoS: Intent-Driven Full-Body Motion Synthesis for Human-Object Interactions

Anindita Ghosh[1,2,3] Rishabh Dabral[2,3] Vladislav Golyanik[2,3] Christian Theobalt[2,3] Philipp Slusallek[1,3]

[1] German Research Center for Artificial Intelligence (DFKI)
[2]Max-Planck Institute for Informatics (MPII)
[3]Saarland Informatics Campus
https://vcai.mpi-inf.mpg.de/projects/IMoS

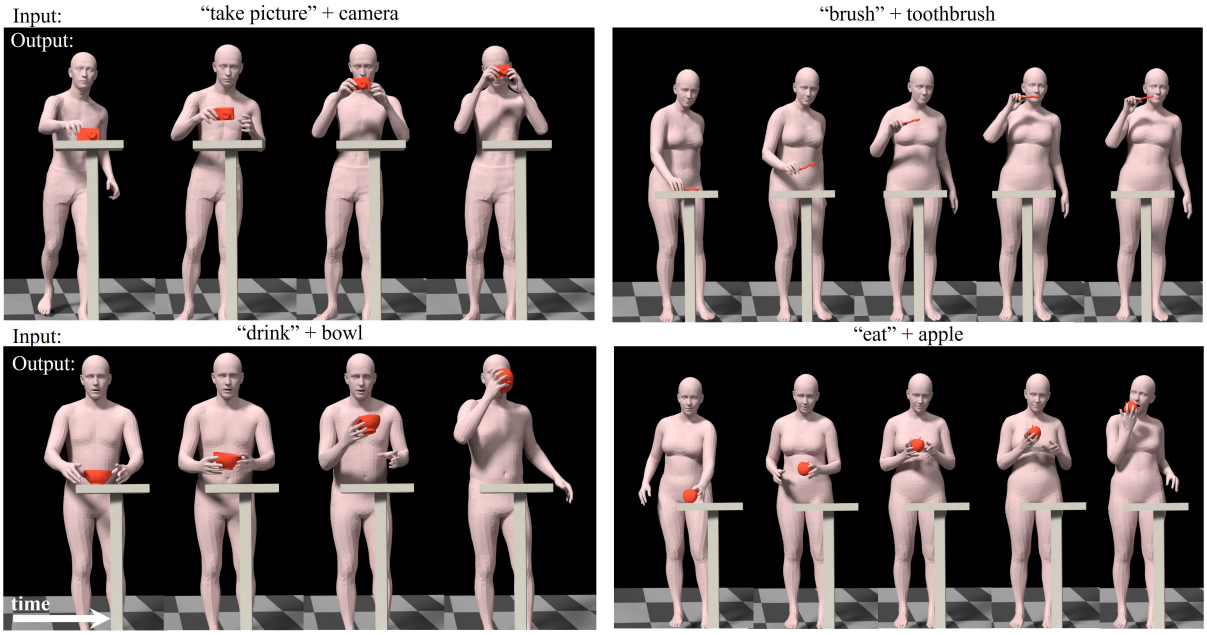

Figure 1: ***Visualizations of Motion Sequences of Virtual Characters Performing Various Intended Actions with Different Objects, as Generated by Our Method.*** *We synthesize the full-body pose sequences and the 3D object positions from text-based instruction labels. Our method can synthesize both single-handed and two-handed interactions depending on the intent and the type of object used.*

Abstract

Can we make virtual characters in a scene interact with their surrounding objects through simple instructions? Is it possible to synthesize such motion plausibly with a diverse set of objects and instructions? Inspired by these questions, we present the first framework to synthesize the full-body motion of virtual human characters performing specified actions with 3D objects placed within their reach. Our system takes textual instructions specifying the objects and the associated 'intentions' of the virtual characters as input and outputs diverse sequences of full-body motions. This contrasts existing works, where full-body action synthesis methods generally do not consider object interactions, and human-object interaction methods focus mainly on synthesizing hand or finger movements for grasping objects. We accomplish our objective by designing an intent-driven full-body motion generator, which uses a pair of decoupled conditional variational auto-regressors to learn the motion of the body parts in an autoregressive manner. We also optimize the 6-DoF pose of the objects such that they plausibly fit within the hands of the synthesized characters. We compare our proposed method with the existing methods of motion synthesis and establish a new and stronger state-of-the-art for the task of intent-driven motion synthesis.

1. Introduction

Humans regularly use and interact with objects in numerous ways in the real world. Interactions like eating a fruit or brushing the teeth, as shown in Fig. 1, are part of our daily routines. Being able to synthesize such interactions in a *virtual* 3D environment through textual instructions has widespread applications in several areas, including computer graphics and robotics [ALNM20; HTBT22; WLK*22], movie script visualization [HMLC09] and game design [SSR07]. For instance, in a digitally created movie scene or a virtual role-playing game, it is natural for the character to interact with the scene objects based on a set of instructions, such as yielding tools, using objects, or eating various items. Manually modeling such 3D character-object interactions or *intentions* is time-consuming and laborious, when we desire to synthesize a variety of possible motions with the same intention and object.

In this context, many recent methods automatically synthesize motions for virtual characters by encoding control signals such as music [LYL*19; LYC*20; LRX*21], speech [BCRM21; HXM*21; HES*22] or text, either as sentences [GZW*20; BRB*21; GCO*21; PBV22] or as high-level action descriptions [AHC*18; LWC*18; AM19]. Methods synthesizing full-body pose sequences typically follow an autoregressive approach to maintain continuity in the synthesized motions [LZCvdP20; RBH*21; GZZ*22]. These autoregressive motion synthesis frameworks predict short-term future sequences from a short history. There are also several methods for hand-object interactions [KYZ*20; TGBT20; JLWW21; ZYSK21; CKA*22], which focus on generating only the wrist and finger movements for grasping various objects. However, modeling hand motion alone is insufficient to create a plausible motion sequence for an intent-driven virtual character. Instead, *we believe it is crucial to operate in the space of full-body motion synthesis.* There are two prime reasons for this. *Firstly*, synthesizing full-body movements allows for a broader range of interactions (Fig. 1). For several intents, such as eating, drinking, inspecting, passing, or exchanging objects between hands, the head, the arms, and the torso are also part of the complete action sequence [TGBT20]. *Secondly*, trivially attaching the synthesized hand motion to the remaining body [PRB*18] leads to an uncanny and physically implausible motion generation (see suppl. video). Further, recent works [TCBT22; WWZ*22] have demonstrated the ability to generate whole-body grasping motion starting from a T-Pose *till* the moment of the grasp. However, synthesizing a plausible motion sequence *after* the first grasp moment, based on an intent guiding the human-object interaction, remains unaddressed.

To address these limitations, we propose IMoS, a novel framework to synthesize diverse, full-body motion sequences of human-object interactions. Crucially, we synthesize the motions based on the input textual instructions consisting of actions (intentions) and objects (Fig. 2). We learn generalizable intent encodings from the input intent-object pairs using a CLIP encoder [RKH*21], which is a large-scale language model trained on a large corpus of text-image pairs. Given the initial body poses and the 3D object positions, we design an intent-driven full-body motion generator model to autoregressively generate full-body motions (Sec.3). We follow a decoupling approach and model the arms and the body motions using separate Conditional Variational Autoregressors to make our output arm and body movements more precise. Since these autoregressors are variational in nature, they allow us to sample diverse motions from the latent space at inference time. We also observe that regressing the motion from a larger past context is crucial in modeling long-term temporal dependence between the joints. We use a position-encoded self-attention mapping to model correlations between the different joints to allow a broader range of interactions. Lastly, we perform an optimization routine to estimate the corresponding 6-DoF object positions relative to the hand position in each frame (Sec 3.2.4). We use the recovered object positions to condition future motion synthesis.

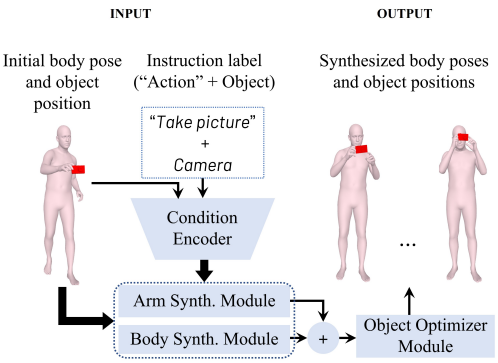

Figure 2: ***Overview of Our Intent-Driven Full-Body Motion Generator.*** *Our model takes in the initial 3D body poses and object positions (upper-left) and instruction labels (upper-middle) describing the object types and the intended actions. We design a pair of decoupled conditional variational auto-regressors, the Arms Synthesis Module and the Body Synthesis Module (lower-middle), to separately synthesize the arms and the rest of the body. We also design a Condition Encoder (middle) to condition our decoupled autoregressors based on the input instruction labels and the body shape parameters. We concatenate our synthesized arm and body motions and use our Object Optimizer Module (lower-right) to optimize the 6-DoF parameters of the object while satisfying the grasping constraints. Our model outputs the synthesized full-body motion sequence together with the object positions (upper-right).*

We train and evaluate our method on the recent GRAB dataset [TGBT20] (Sec. 5.1), consisting of ∼1.3K sequences of human-object interactions exhibiting multiple intents. We quantitatively evaluate our synthesized sequences on established metrics, such as the mean per-joint position error, the average variance error, the Fréchet Inception distance, recognition accuracy, diversity, and multimodality, to test the effectiveness of the model. Further, we conduct a visual perceptual study for subjective evaluation of our synthesized motions compared to recent conditional motion synthesis methods (Sec. 5.6).

In summary, our primary **technical contributions** are threefold:

- A new framework for generating diverse motion sequences of virtual human characters interacting with objects of known shapes placed within their reach, according to text-based instruction labels. In contrast to previous works on character-object in-

Method	Motion Synthesis			
	Full Body	Intent-Driven	Only Till Grasp	Object Manipulation
GRABNet [TGBT20]	✗	✗	✗	✗
D-Grasp [CKA*22]	✗	✗	✗	✓
A2M [PBV21]	✓	✓	✗	✗
ACTOR [PBV21]	✓	✓	✗	✗
GOAL [TCBT22]	✓	✗	✓	✗
SAGA [WWZ*22]	✓	✗	✓	✗
IMoS (ours)	✓	✓	✗	✓

Table 1: ***Overview of the Problem Definitions of Existing Methods.*** *Our method is the only one combining three important characteristics and the first one to synthesize intent-driven full-body pose sequences for motions with object manipulation.*

teractions, our proposed method also optimizes the 6-DoF object positions in 3D.
- Synthesizing interactions involving *both* hands, including sequences where the character exchanges an object between the hands ("offhand") – a previously unexplored setting.
- Learning separate variational latent embeddings for the arms from the rest of the body to enable diversity in the synthesized motions and accurate synthesis of both-handed interactions.

2. Related Work

Our work aligns with past works on modeling 3D human-object interactions. We study these works from four vantage points: human pose forecasting and synthesis, human-object 3D interaction modeling, hand-object grasp synthesis, and full-body grasp synthesis.

Human Pose Forecasting and Synthesis. Human pose forecasting methods predict future motions from a sequence of past poses as joint positions [MBR17] or joint rotations [PGA18; RBH*21]. Recent works on 3D human pose forecasting are stochastic methods [YK20; LLW*21] that use VAEs [KW14] or GANs [GPM*20] to introduce variability in the output motion sequences. HuMoR [RBH*21] proposes a CVAE architecture that learns a distribution of pose transitions in the latent space and ensures physical plausibility through post-processing optimization. MotionVAE [LZCvdP20] learns to drive a character based on a goal position by decoding from a variational latent space. Characteristic 3D pose [DFD22] stochastically predicts future 3D characteristic poses given short sequences of observations. Other human motion synthesis methods are trained to synthesize a motion sequence conditioned on semantic action labels [GZW*20; PBV21; DFD22], or text sentences [GZZ*22; PBV22]. Action2Motion [GZW*20] inputs an action label to generate the human pose in an autoregressive manner using a VAE-GRU. Differently, ACTOR [PBV21] employs a VAE-Transformer to generate the full sequence in one shot. TEMOS [PBV22] uses the VAE-Transformer concept on a multimodal setting to generate motions from text sentences. Our work extends full-body motion synthesis conditioned on semantic labels by additionally incorporating object interactions.

Human-Object 3D Interaction Modeling. With the availability of several human-object 3D datasets like [SZKS19], BEHAVE [BXP*22], PROX [HCTB19], D3D-HOI [XJMS21], H2O [KTS*21], GraviCap [DSJ*21], joint human-object motion modeling has been an actively researched topic. Among more recent methods, PHOSA [ZPJ*20] reconstructs the human and the object in the scene by jointly optimizing for the reprojection error of the object's silhouette and the human. Neural State Machines [SZKS19] synthesize human motion while interacting with objects like chairs or a wall in the scene. Likewise, SAMP [HCV*21] incorporates a path planning module to improve the character's motion in the scene. COUCH [ZBS*22] synthesizes sitting interactions with couches by generating contact points and then using them to constrain the sitting motion. All the previous methods perform interactions with large objects (chairs and couches) and are driven by low-level character control. In contrast, we synthesize fine-grained motions with *handheld* objects using *instruction labels* as input.

Hand-Object Grasp Synthesis. Grasp synthesis has been extensively studied in computer graphics [ES03; LFP07; KP15; KYZ*20; ZYSK21] and robotics [BI05; HL06; DKB*10; AGK18; LPX*19]. Analytical approaches have formulated grasp synthesis as a constrained optimization problem satisfying the grasp properties [KDCI10; SKK12]. Data-driven approaches [RA15; PG16] focus on learning the representations for synthesizing grasps through machine learning methods. More recent approaches [BHHF19; KYZ*20; TGBT20; JLWW21] predict the hand parameters of the MANO hand model [RTB17] for synthesizing a grasp using neural networks. Many image datasets [HVT*19; ZLM*19; BTT*20; LWM21; ZHT*21] featuring hand-object interaction with contact maps are also currently available. Taheri et al. [TGBT20] further introduce the GRAB dataset, which captures the contact map from hands and the full-body motions before and during the grasp. They also propose GrabNet, a network that estimates MANO parameters at the moment of grasp for unseen objects in a coarse-to-fine manner. [KYZ*20] proposes Grasping Field, a method that learns an implicit representation of the hand-object interaction using a generative model. Grady et al. [GTT*21] derive physically plausible hand pose estimation by optimizing estimated hand meshes with contact prediction. We differ from all these methods as our work focuses on synthesizing *full-body* sequences. While modeling hand-object interaction is a well-researched problem, it is inherently limited in its ability to model several types of human-object interactions that require the full human body (*e.g.*, tilting back the head when drinking from a glass).

Full-Body Grasp Synthesis. This is a relatively recent line of work following the success of hand-object grasp synthesis. GOAL [TCBT22] synthesizes full-body motion for grasping a given object by first estimating the whole-body grasping pose for the object and then treating this pose as their goal for a motion-infilling module that interpolates the motion between a T-Pose and the goal pose. SAGA [WWZ*22] also follows a similar strategy of motion infilling but uses markers to represent the body pose while also learning a contact map for the grasp for additional supervision. Both these methods synthesize full-body motions *until* the point of grasping. In contrast, we synthesize the motion taking place *after*

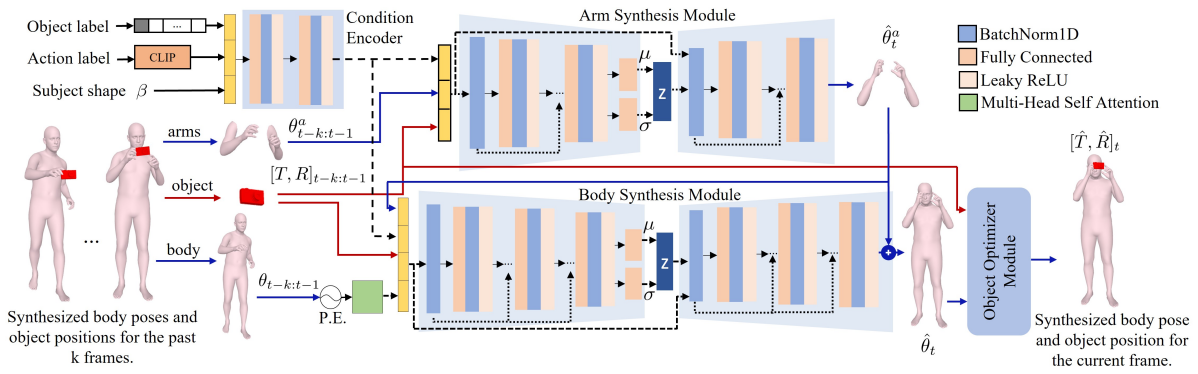

Figure 3: ***Architecture of Our Intent-Driven Full-Body Motion Generator Model.*** *Given previous k frames of body poses and object positions, we train the arms and the rest of the body separately using our Arm Synthesis (upper-middle) and the Body Synthesis (lower-middle) Modules, respectively. We jointly synthesize the entire motion sequences autoregressively, conditioned on the input intent, the object, and the body shape, all encoded through our Condition Encoder (upper-left). We use position-encoded self-attention on the past k frames for the body joints before passing them through our Body Synthesis Module. After generating the body pose, our Object Optimizer Module (lower-right) optimizes for the 6-DoF pose of the given object such that it plausibly fits within the hands of the synthesized character.*

the object is grasped (see Table 1). This is a non-trivial and more challenging setup. Conditioning human and object motions based on the intended actions while also ensuring diversity in the generated motion sequences requires additionally learning their intent-based mutual interactions in an efficient and generalizable manner.

3. Intent-Driven Full-Body Motion Generator

We show the architecture of our intent-driven full-body motion generator model in Fig. 3. Given a human character's shape and initial 3D body pose, a rigid 3D object placed within their reach, and an intended action to perform with that object, our goal is to synthesize a full-body motion sequence of the character performing the intended action with the object. We pose this problem as synthesizing the full-body motion sequence conditioned on the given object and a textual instruction label indicating the intent. We solve this problem through four modules. First, we encode the input instruction labels consisting of the type of the object and the associated action using our Condition Encoder. We also input the subject's body shape parameters into our Condition Encoder. We use this encoding as a conditioning signal for all the modules. A key characteristic of our problem is that the arms are the primary movers during human-object interactions. Therefore, we use a pair of decoupled conditional variational autoregressor networks to synthesize the arm movements and the rest of the body movements separately, using an Arm Synthesis Module and a Body Synthesis Module, respectively. Lastly, we use an Object Optimizer Module to optimize the 6-DoF pose of the given object such that it fits plausibly within the hands of the synthesized character.

3.1. 3D Human Body and Object Representation

We represent the human mesh using the SMPL-X [PCG*19] parametric body model. SMPL-X parametrizes the full human body along with the hands and the face as a differentiable function $SMPLX(\beta, \mathbf{r}, \Psi, \mathbf{t})$, consisting of body shape parameters $\beta \in \mathbb{R}^{10}$, the root translation $\mathbf{t} \in \mathbb{R}^3$, the axis-angle rotations for the body joints $\mathbf{r} \in \mathbb{R}^{J\times 3}$ $(J = 55)$, and the face expression parameters $\Psi \in \mathbb{R}^{10}$. It maps the parameters to a body mesh with $10,475$ vertices. To improve the stability and the convergence characteristics of our model, we use the 6D continuous representations [ZBL*19] $\theta \in \mathbb{R}^{J\times 6}$ to represent body joint rotations. We downsample all the objects in the dataset to 300 vertices for faster optimization. The object's 6-DOF pose is represented using a rotation matrix $\mathbf{R} \in \mathbb{R}^9$ and a translation vector $\mathbf{T} \in \mathbb{R}^3$.

3.2. Model Design

We now discuss each of our modules in detail. Our synthesis pipeline assumes that the character interacts with only one object at a time. Interactions can be either one-handed or both-handed, depending on the type of action and the object.

3.2.1. Condition Encoder

We input the object's category label using a one-hot vector $\mathbf{w_o} \in \mathbb{R}^{51}$. To represent the intended action information, we pass the intended action label, given as an English word, through the pre-trained CLIP [RKH*21] model and use the embeddings $\mathbf{w_a} \in \mathbb{R}^{512}$ that it outputs. The idea behind encoding the action labels with a pre-trained text encoder is the general relevance between the action semantics and the corresponding body movements. For example, actions such as "drink" and "pour" typically invoke similar arm movements and are semantically close. In contrast, other actions, such as "inspect" and "pass", invoke different body movements and are semantically different. Therefore, their embeddings, given by a large-scale language model such as CLIP, provide a regularized, semantics-based distribution of the intended actions and stabilizes further processing. Refer to the Appendix for more details.

We concatenate $\mathbf{w_o}$ and $\mathbf{w_a}$ with the body shape parameters ($\beta \in \mathbb{R}^{10}$) and pass them into our Condition Encoder $\mathbf{q_c}$. Our Condition

Encoder uses a series of MLPs to encode these input signals and projects them onto an encoded feature vector $\phi \in \mathbb{R}^{400}$ as

$$\phi = \mathbf{q_c}(\mathbf{w_o}, \mathbf{w_a}, \beta). \tag{1}$$

3.2.2. Arm Synthesis Module

Our Arm Synthesis Module is a conditional variational autoregressor that synthesizes the arm movements, conditioned on our condition encoder output ϕ and the previous k frames of synthesized arm poses along with the 3D object positions. The encoder of this module, $\mathbf{q_a}$, takes in the tuple $\mathbf{q_a}^{in} = \{\phi, \theta^{\mathbf{a}}_{\mathbf{t-k:t-1}}, \mathbf{T_{t-k:t-1}}, \mathbf{R_{t-k:t-1}}\}$, where $\theta^{\mathbf{a}}_{\mathbf{t-k:t-1}}$ are the rotations for the arm joints synthesized by the past k frames, and $\mathbf{T_{t-k:t-1}}$, $\mathbf{R_{t-k:t-1}}$ are the translation and rotation parameters of the object for the past k frames. During training, $\mathbf{q_a}$ uses a series of MLPs on the input and maps them to the parameters of a latent normal distribution, $\mu_{\mathbf{a}}, \sigma_{\mathbf{a}} \in \mathbf{R}^{32}$. The decoder, $\hat{\mathbf{q}}_{\mathbf{a}}$, samples $\mathbf{z_a} \in \mathbb{R}^{32}$ from the latent distribution and uses the previous pose information ($\mathbf{q^{in}_a}$) to synthesize the arm pose for the current frame ($\hat{\theta}^{\mathbf{a}}_{\mathbf{t}}$) through a series of MLPs with skip connections as

$$\hat{\theta}^{\mathbf{a}}_{\mathbf{t}} = \hat{\mathbf{q}}_{\mathbf{a}}(\mathbf{z_a}, \mathbf{q^{in}_a}). \tag{2}$$

3.2.3. Body Synthesis Module

Similar to the Arm Synthesis Module, the Body Synthesis Module is a variational autoregressor. We use the term 'body' to denote the rest of the body parts apart from the arms. It includes the head, the torso, the hips, and the legs. We also note that the movements of all these parts are correlated when performing a full-body action. For example, to drink from a cup, one has to tilt their head back when bringing the cup to their mouth. To model such fine-grained correlations, we first compute a self-attention mapping between all the joints in each pose as

$$\theta^{\mathbf{pe}}_{\mathbf{k}} = [Attn(\mathbf{Q}, \mathbf{K}, \mathbf{V})]_{\mathbf{k}}, \tag{3}$$

where the query $\mathbf{Q}$ is a joint position and the key-value pair $(\mathbf{K}, \mathbf{V})$ are information of all other joints provided as J sinusoidal positional encodings for each of the k frames. The encoder of the module, $\mathbf{q_b}$, takes in the tuple $\mathbf{q_b}^{in} = \left\{\phi, \hat{\theta}^{\mathbf{a}}_{\mathbf{t-k:t-1}}, \theta^{\mathbf{pe}}_{\mathbf{t-k:t-1}}, \mathbf{T_{t-k:t-1}}, \mathbf{R_{t-k:t-1}}\right\}$. The structure of $\mathbf{q_b}$ is similar to that of the Arm Synthesis Module encoder $\mathbf{q_a}$, and it maps the input $\mathbf{q^{in}_b}$ to the parameters of a latent normal distribution, $\mu_{\mathbf{b}}, \sigma_{\mathbf{b}} \in \mathbf{R}^{100}$. The decoder, $\hat{\mathbf{q}}_{\mathbf{b}}$, samples $\mathbf{z_b} \in \mathbb{R}^{100}$ from the latent distribution and outputs the rest of the body poses as

$$\hat{\theta}^{\mathbf{b}}_{\mathbf{t}} = \hat{\mathbf{q}}_{\mathbf{b}}(\mathbf{z_b}, \mathbf{q^{in}_b}). \tag{4}$$

We then concatenate $\hat{\theta}^{\mathbf{a}}_{\mathbf{t}}$ and $\hat{\theta}^{\mathbf{b}}_{\mathbf{t}}$ to obtain the full-body pose $\hat{\theta}_{\mathbf{t}}$ at time t. We pass $\hat{\theta}_{\mathbf{t}}$ to our Object Optimizer Module, along with the last predicted object position, to generate the object position for the current frame.

3.2.4. Object Optimizer Module

We have so far focused only on synthesizing the body poses for a given instruction. For a complete synthesis, we also need to estimate the corresponding 6-DoF positions of the object. Although fine-grained object synthesis is not the primary goal of our work,

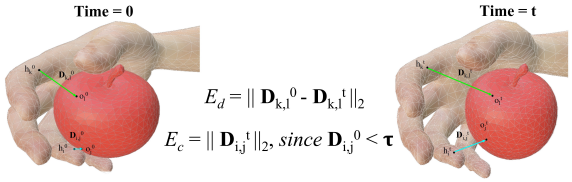

Figure 4: ***Our Hand-Object Setup.*** *We design the energy term E_d to enforce that the distances between the hand and the object vertices remain constant throughout the synthesis. Through the hand-object contact term E_c, we also enforce that the points in contact in the first frame remain in contact during the synthesis.*

we aim to produce plausible object trajectories faithful to the synthesized full-body motion. To this end, our core assumptions are that *(a)* at the moment of grasping in the initial frame, the object is at rest in an upright position and *(b)* inter-vertex distances between the vertices of the object and the hand remain constant throughout our intent-driven motion synthesis.

With these assumptions, we optimize the object's rotation $\mathbf{R}$, translation $\mathbf{T}$, as well as the pose parameters of the hand, $\mathbf{P}^h$, in the SMPL-X parameter space.

We first compute the matrix of Euclidean distances $\mathbf{D} \in \mathbb{R}^{N \times M}$ between the vertices on the hand, $\mathbf{V}^h \in \mathbb{R}^N$ and those on the surface of the object, $\mathbf{V}^o \in \mathbb{R}^M$ for the initial frame. We can retrieve the hand vertices using the SMPL-X parameterization,

$$\mathbf{V}^h = SMPLX(\mathbf{P}^h). \tag{5}$$

For each subsequent frame, we then minimize the objective:

$$\mathbf{R}^*, \mathbf{T}^*, \mathbf{P}^{h*} = \min_{\mathbf{R},\mathbf{T},\mathbf{P^h}} (\lambda_d E_d + \lambda_c E_c + \lambda_r E_r) \tag{6}$$

We use an energy term, E_d, to enforce the same inter-vertex distances between the hand and the object vertices in all the subsequent frames as in the first frame, as

$$E_d(\mathbf{R}, \mathbf{T}, \mathbf{P^h}) = \left\| \text{dist}(\mathbf{V^h}, \mathbf{RV^o} + \mathbf{T}) - \mathbf{D} \right\|_2, \tag{7}$$

However, this term alone does not guarantee that the object is in contact with the hand in subsequent frames because, in practice, the hand joints do not converge to plausible poses using E_d. We address this issue by introducing the contact term E_c, which forces the distance between the in-contact vertex pairs of the first frame to be zero, as

$$E_c(\mathbf{P^h}) = \left\| \delta . \text{dist}(\mathbf{V^h}, \mathbf{RV^o} + \mathbf{T}) \right\|_2. \tag{8}$$

Here, $\delta(\cdot, \cdot)$ is a contact indicator function for the elements of the distance matrix for which the distance is less than a threshold: $\delta(i, j)) = 1$, if $\mathbf{D}_{i,j} < \tau$ and 0 otherwise, as we show in Fig. 4.

Finally, E_r consists of L2 regularizers to ensure that the object and hand poses do not deviate significantly from the previous frame and thus enforce temporal consistency, as

$$E_r(\mathbf{R}, \mathbf{T}, \mathbf{P^h}) = \left\| \Delta\mathbf{R} + \Delta\mathbf{T} + \Delta\mathbf{P^h} \right\|_2, \tag{9}$$

where Δ signifies the difference in values between the current and

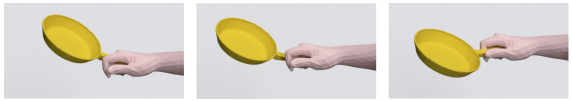

Figure 5: ***Object Position Optimization****. We optimize the 6-DoF pose of the object such that it fits plausibly within the hands of the virtual character. We show three snapshots of such fitting after the* 0^{th}*,* 500^{th} *and the* 1200^{th} *iteration of our optimization.*

the previous frame. We initialize the hand poses using a state-of-the-art grasp estimator proposed in [TCBT22]. The optimization routine iteratively corrects the initial estimates of the finger movements while placing the object within the person's hands. Fig. 5 illustrates the optimization routine.

4. Implementation

This section describes our training and inference routines and the implementation details for our generator network.

Training and Inference Routines. To maintain a fixed number of input frames for computational stability, to reduce the parameter load and associated training overheads, and to avoid overfitting to redundant frames, we represent our ground-truth motion sequences using $T = 15$ frames, taken at a sampling rate of 8-10 fps. These 15 frames act as the key frames determining the motion sequence.

The encoders and the decoders inside our four modules use fully-connected layers with skip connections, LeakyReLU activations, and batch normalization [Aga18; BGSW18]. We use $k = 4$ past frames (optimized through experiments) to synthesize the subsequent time steps. We train our autoregressor based Arm Synthesis and Body Synthesis Modules to minimize the KL divergence loss:

$$\begin{aligned}\mathcal{L}_{KL} =& D_{KL}\big(\mathbf{q}_a\left(\mathbf{z_a}|\mathbf{x_{t-k:t-1}},\phi\right)||\mathcal{N}\left(0,I\right)\big)\\ &+ D_{KL}\big(\mathbf{q}_b\left(\mathbf{z_b}|\mathbf{x_{t-k:t-1}},\phi\right)||\mathcal{N}\left(0,I\right)\big).\end{aligned} \tag{10}$$

We compute the pose and the velocity reconstruction loss between the ground-truth rotations θ and the predicted rotations $\hat{\theta}$ as

$$\mathcal{L}_{rec} = \left\|\theta - \hat{\theta}\right\|_1 + \left\|\Delta\theta - \Delta\hat{\theta}\right\|_1. \tag{11}$$

We train our model on the following weighted sum of these losses:

$$\mathcal{L} = \lambda_{KL}\mathcal{L}_{KL} + \lambda_p\mathcal{L}_{rec}, \tag{12}$$

where λ_{KL} and λ_p are the weight parameters. We can then use the regressed body motion parameters $\hat{\mathbf{p}}$ to optimize the 6-DoF object positions at every time step.

During inference, we synthesize motions for novel intent-object pairs and novel body shape parameters. We input an initial body pose, a 3D object placed within reach of the character and an intended action to be performed with the object, and autoregressively synthesize the intent-based full-body motion sequence.

Implementation Details. We train our model for 1,600 epochs using the Adam Optimizer [KB14] with a base learning rate of 5×10^{-4}, and a batch size of 64, which takes roughly four hours on an NVIDIA A100-PCIE-40GB GPU. We decay the learning rate (LR) using a Reduce-on-plateau LR scheduler with a patience of 3 epochs and a decay rate of 0.999. We set $\lambda_{KL} = 0.001$, $\lambda_p = \lambda_d = 1.0$ and $\lambda_c = \lambda_r = 0.005$. During inference, synthesizing the full-body poses and the corresponding object positions for a motion sequence of 15 frames take approximately 1-1.5 minutes. Finally, we perform a linear interpolation on our generated frames to up-sample the motion to 30 frames per sequence for cleaner visualization. We have implemented our network, training, and inference using the PyTorch framework [PGC*17].

5. Experiments and Results

This section reports the results of our experimental evaluation, including the dataset, the evaluation metrics we use, and our ablation studies. Since there are no existing methods for generating full-body human-object interactions, we use existing methods that generate full-body poses based only on action labels as our baselines.

5.1. Dataset

We use the GRAB dataset [TGBT20] consisting of whole-body grasping sequences performed by ten different subjects. The subjects interact with 51 different objects via four basic intents, "use", "pass", "lift", and "offhand". "Use" further has a sub-category of 26 different actions depicting plausible intent-object interactions such as drinking or pouring from a cup to taking a picture with or browsing a camera. Following the split of [DFD22], we take subject 'S1' for validation, 'S10' for testing, and the remaining subjects 'S2' through 'S9' for training. This data split ensures that we test on novel subjects with different body shapes and our inference contains novel (intent-object) pairs such as offhanding a water bottle, which is not present in our training set. We discard the "lift" intention from our setting as the motions depicting lifting an object were inconsistent in the dataset. Thus our train, validation, and test splits respectively consist of 789, 157, and 115 sequences.

5.2. Baselines

We compare our results with ACTOR [PBV21], Action2Motion [GZW*20] and TEMOS [PBV22]. Since these methods were not originally trained on the GRAB dataset, we re-train them for our setting. We re-train ACTOR and the Action2Motion methods for 1,600 epochs (the same number of epochs we train our model for, see Sec. 4) conditioned only on the action labels with no object information. For comparison with TEMOS, we create sentences of the form "A person ⟨action⟩ the ⟨object⟩" (*e.g.*, "a person eats the apple") to use as input sentences, and re-train the TEMOS model for 1,600 epochs as well. We apply our Object Optimizer Module for all three motion synthesis methods to generate the object positions for visual comparison.

5.3. Evaluation Metrics

We evaluate our method using the Mean Per-Joint Positional Error (MPJPE), which measures the mean joint error over all time steps, and the Average Variance Error (AVE) [GCO*21], which measures the variance error between the joint positions.

Method	MPJPE (↓)	AVE (↓)	FID (↓)	Accuracy (↑)	Diversity (→)	Multimodality (→)
Real Motions (GT)	-	-	-	0.97 ± 0.001	1.15 ± 0.015	0.30 ± 0.010
ACTOR	0.09 ± 0.005	8.05 ± 0.002	0.67 ± 0.002	0.78 ± 0.010	1.06 ± 0.015	0.19 ± 0.010
Action2Motion	0.11 ± 0.003	8.26 ± 0.002	1.08 ± 0.002	0.69 ± 0.011	1.10 ± 0.010	0.22 ± 0.010
TEMOS	0.10 ± 0.005	9.98 ± 0.001	1.21 ± 0.004	0.23 ± 0.010	0.83 ± 0.010	0.09 ± 0.010
Ablation 1	0.05 ± 0.002	4.41 ± 0.002	0.39 ± 0.002	0.78 ± 0.012	1.06 ± 0.015	0.21 ± 0.010
Ablation 2	0.04 ± 0.005	4.77 ± 0.002	0.38 ± 0.002	0.82 ± 0.010	1.10 ± 0.020	0.24 ± 0.020
Ablation 3	0.05 ± 0.005	5.41 ± 0.002	0.42 ± 0.002	0.82 ± 0.010	1.08 ± 0.010	0.25 ± 0.010
Ours	$\mathbf{0.03 \pm 0.005}$	$\mathbf{3.82 \pm 0.004}$	$\mathbf{0.27 \pm 0.002}$	$\mathbf{0.87 \pm 0.011}$	$\mathbf{1.11 \pm 0.015}$	$\mathbf{0.28 \pm 0.015}$

Table 2: ***Quantitative Evaluation.*** *We compare with other motion synthesis methods, namely ACTOR [PBV21], Action2Motion [GZW*20] and TEMOS [PBV22], and three ablated versions of our model (Sec. 5.4). We evaluate the methods on the MPJPE, AVE, FID, recognition accuracy, diversity, and multimodality metrics. "↓" denotes lower values are better, "↑" denotes higher values are better, and "→" denotes values closer to the ground-truth are better.*

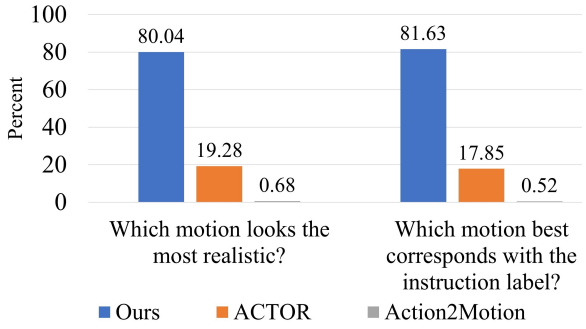

Figure 6: ***Perceptual Study Evaluation.*** *We conduct a user study where participants answer two questions: "Which animation looks more realistic?" and "which animation best corresponds with the input instruction label?". We show them* 30 *randomly sampled motion sequences synthesized by our method and the two baselines, ACTOR [PBV21] and Action2Motion [GZW*20]. We see our method is chosen more than* 80% *times.*

We further evaluate the naturalness and the overall diversity of our generated motions using the Fréchet Inception Distance (FID) [HRU*17], recognition accuracy, diversity, and multimodality. Following ACTOR [PBV21] and Action2Motion [GZW*20], we train a standard RNN action recognition classifier on the GRAB dataset and use the final layer of the classifier as the motion feature extractor for calculating FID, diversity, and multimodality. Please refer to our Appendix for further details.

5.4. Ablation Studies

We compare the performance of our model with the following ablated versions:

- **Ablation 1: Randomly initializing the input action labels with 512-d vectors:** To study how the CLIP model influences the conditioning of the synthesized motion, we conduct an ablation where we train our Condition Encoder with a 512 dimensional randomly initialized vector for the input action labels instead of taking the CLIP embeddings.
- **Ablation 2: Training the Body Synthesis module without using the self-attention mapping.** In this ablation, we exclude our position-encoded multi-head self-attention from the input of the Body Synthesis module of our framework to see how it influences the quality of our motion.
- **Ablation 3: Training the full body instead of decoupling to the Arm Synthesis and the Body Synthesis Modules.** We train the whole body movements in one module instead of separately synthesizing the arms and the rest of the body.

5.5. Quantitative Evaluation

Table 2 shows the MPJPE, AVE, FID, recognition accuracy, diversity, and multimodality on our test set compared to the three state-of-the-art methods of ACTOR [PBV21], Action2Motion [GZW*20], and TEMOS [PBV22]. We also include the ablated versions of our methods (Sec. 5.4) in our evaluation. We repeat each experiment 20 times as done in ACTOR [PBV21], and report a statistical interval with 95% confidence. Our method shows significant improvements in all the metrics compared to the existing methods and the ablated versions.

5.6. Perceptual Study

To evaluate the visual quality of our motions, we conduct a perceptual study where we compare our results with ACTOR [PBV21] and Action2Motion [GZW*20]. Except for TEMOS [PBV22], which would quickly settle on the mean pose, the other two methods generated plausible full-body motions after re-training. We, therefore, exclude TEMOS from the user study. We conduct our perceptual study in the following two sections.

Comparison with Motion Synthesis Methods. In the first section, we displayed our results and the results from ACTOR and Action2Motion side-by-side in random order, along with the input instruction label. We asked the participants to answer these questions for each sequence: *"Which motion looks the most realistic?"* and *"Which motion best corresponds with the input instruction label?"*. We collected answers for 30 such sequences from 75 participants. Fig. 6 illustrates the results of the study. In 80% responses, participants marked our method as the most realistic compared to

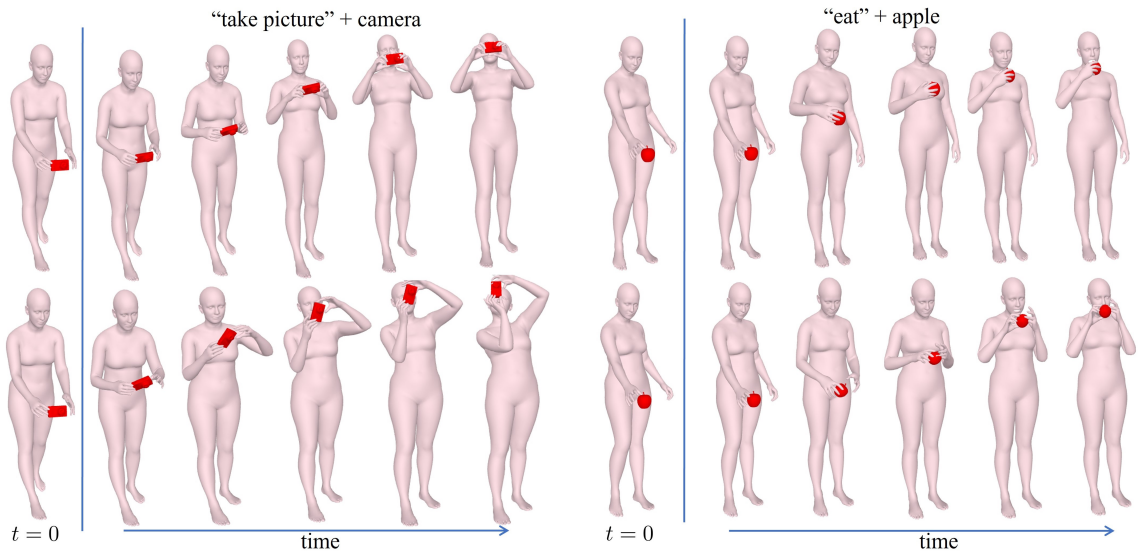

Figure 7: ***Qualitative Results Showing Diversity in the Synthesized Motions.*** *The two rows depict two diverse motion sequences generated by our model. Our method can generate variations for the same instructions using either or both hands, along with plausible coordination of the head and the body. Please refer to the supplementary video for more results.*

ACTOR and Action2Motion. Likewise, 81.6% participants chose our method to have the best semantic fidelity with the instruction label. Upon examining the cases for which the participants preferred ACTOR instead of us, we found that it performed better for a few actions, *e.g.*, "screwing" the light bulb and "toasting" with the wineglass, where the motion does not need to have hand-to-eye or hand-to-mouth coordination. These actions do not include significant variations within the dataset and are, therefore, easy to overfit.

Comparison with Ground-Truth. While ACTOR and Action2Motion are the closest methods for our paradigm, they were not originally designed to synthesize motions conditioned on intents. Therefore, to get an additional perspective on the performance of our method, we asked the participants to compare our *best* synthesis results with the ground-truth motions in the second section. To establish an upper bound on our performance, we chose the 10 best samples from various intent-object pairings to compare with the corresponding ground-truth motions. Again, we displayed our synthesized and ground-truth motions side-by-side in random order. This time, we kept an extra option: *"cannot distinguish"*. While our synthesized motions are, expectedly, less preferred than the ground-truth motions (15.6% vs. 36.9%), 47.5% of the responses rate our best syntheses as *indistinguishable* in terms of realism. We also note that participants rated our synthesized motions as more realistic than the ground-truth motions when it involves actions such as "eating" an apple with one hand, which has abundant training samples. On the other hand, our method encounters difficulties when synthesizing intents involving high-frequency wrist or finger movements such as "shaking" or "squeezing". This is because our ℓ_1 loss function (Eqn. (11)) tends to smooth out the high-frequency components from the motion sequence, and the GRAB dataset does not have sufficient samples of these actions to train them separately.

5.7. Qualitative Evaluation

We show full qualitative results in our supplementary video. When qualitatively compared with the ablated versions (Sec. 5.4), we find that Ablation 1 (using random initialized vectors instead of CLIP) and Ablation 3 (training one module for the whole-body) fail to synthesize precise hand-mouth or hand-eye coordination for actions such as "drinking" and "eating". Ablation 2 (without using self-attention mapping) lacks subtle body movements, such as tilting back the head or bending the knee to pick up an object, which improve the motion plausibility. We further analyze our generated motions under the following headings:

Diversity Analysis. As we noted earlier (Sec. 1), generating diverse motion sequences for the same input instruction label is crucial for an immersive user experience. Fig 7 shows our result for two different sequences (left and right). Sampling from the variational latent space allows us to synthesize diverse motion sequences. In Fig. 7, we show two different sequences: "taking picture" with a camera (left) and "eating" an apple (right). We show two variations of the same motions (upper and lower rows). We note that the variations are diverse w.r.t how the head, arm, and torso are angled to use the object. Our method benefits from operating in the full-body space and produces more natural results compared to naïvely performing a fixed mapping from the global hand pose parameters to the end effectors of the remaining body.

Synthesis of Both-Handed Interactions. Our method is the first to plausibly synthesize full-body motions for both-handed interactions. We achieve this by decoupling the arm synthesis from the full-body synthesis in our generator design (Sec. 3.2). The wrist and the elbow joints play a crucial role in tasks such as picking up an object with both hands or holding the object precisely. Learning

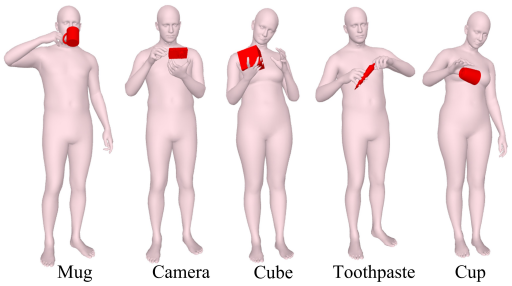

Figure 8: ***Examples of Imprecise Contacts in the GRAB Dataset [TGBT20].*** *We show five (ground-truth) frames where the body and the object are in contact. However, these contacts are not precise. The fingers do not touch the object when grasping the mug, the camera, or the cup. On the other hand, we see inter-penetration between the hand and the object for the cube and the toothpaste.*

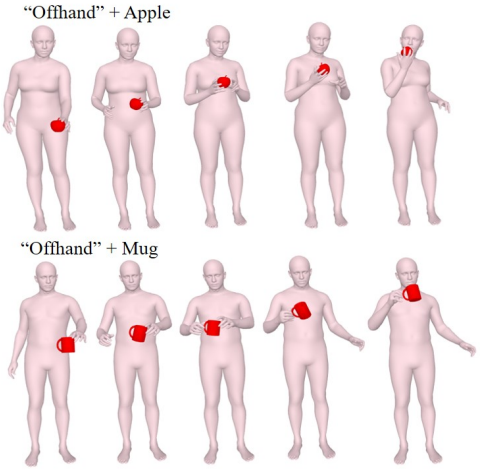

Figure 9: ***Off-Handing.*** *We show two interactions of "offhanding" where the character passes the object from one hand to the other. Such interactions pose a unique optimization challenge when the object is switching hands.*

the arm motions in a separate latent space helps our generator focus more on such precise synthesis.

Object Position Predictions for Off-Handing Interactions. In addition to both-handed interactions, we encounter sequences in the GRAB dataset where the character passes an object from one hand to the other. It is non-trivial to optimize the accurate object positions when the object switches hands. Here, we first compute the most likely frame at which the switching occurs and then transfer the optimized hand parameters to the other hand. Fig: 9 shows two such off-handing interactions with two objects.

Plausibility of Head Motions. Similar to the motion of the fingers and the arms, the coordinated movement of the head and the hands also determines the synthesis quality. While recent works like GOAL [TCBT22] explicitly account for the head direction vector during network training and optimization, we observe that our model learns visually plausible head orientations and hand-head coordination without any explicit supervision. This raises the question of whether explicit supervision is indeed necessary.

6. Discussion and Limitations

Through quantitative evaluations and a perceptual study, we establish that our method synthesizes plausible motions of virtual characters performing intended actions with given objects. While we can synthesize motions for various intents and objects, we observe failure cases for certain rare intents with high-frequency wrist motions, *e.g.*, "squeeze", "shake" (see supplementary video). Additionally, our Object Optimizer Module (Sec. 3.2.4) optimizes the fingers and the object positions based on an initial distance between them. This assumption works well with most of the intents in the GRAB dataset, which involve static grasps. However, dynamic grasping, which involves hand slipping and relative motion between the object and the hands (such as "rotating" a cube and "stretching" an elastic band), is limited in our setting. We also note that the contacts between the body and the objects are not fully precise for all samples in the GRAB dataset, possibly due to the sparse marker-based motion capture. In some sequences, the fingers do not touch the object while grasping, while in others, there are inter-penetrations between the hand and the object (Fig. 8). Lastly, we do not address long-term motion synthesis (in the order of minutes) involving a sequence of actions performed with an object.

Ethical Considerations. Our method does not support texture and fine appearance details and cannot be used to produce deceptive content. Our results are not photo-realistic by design and cannot be confused with real scenes. However, combining our technique with a method supporting more realistic texture might raise ethical concerns in the future.

7. Conclusion and Future Work

We presented the first full-body motion synthesis method for character-object interactions. Such a motion synthesis pipeline can become a useful, practical tool in applications requiring large-scale character animations. We demonstrate that a decoupling approach that separately models the arms and the body motions using conditional variational autoregression leads to measurable perceptual improvements and advances the state-of-the-art on multiple quantitative evaluations. We also synthesize interactions involving *both* hands, including sequences where the object exchanges hands.

In the future, we intend to extend our model to synthesize dynamic grasps and full-body poses such that the virtual character can change the grasp within a sequence. We also plan to explore descriptive sentence embeddings for the interactions (*e.g.*, "a person passes the bowl using the right hand") to generate more precise and controllable motions.

Acknowledgements. This research was supported by the BMBF grant XAINES (01|W20005), the BMBF and ITEA grant AIToC (01|S20073B), the EU Horizon 2020 grant Carousel+ (101017779), and the ERC Consolidator Grant 4DRepLy (770784). Open Access funding is enabled and organized by Projekt DEAL.

References

[Aga18] AGARAP, ABIEN FRED. "Deep learning using rectified linear units (relu)". *arXiv preprint arXiv:1803.08375* (2018) 6.

[AGK18] ANTOTSIOU, DAFNI, GARCIA-HERNANDO, GUILLERMO, and KIM, TAE-KYUN. "Task-oriented hand motion retargeting for dexterous manipulation imitation". *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018 3.

[AHC*18] AHN, H., HA, T., CHOI, Y., et al. "Text2Action: Generative Adversarial Synthesis from Language to Action". *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 2018 2.

[ALNM20] AHUJA, CHAITANYA, LEE, DONG WON, NAKANO, YUKIKO I, and MORENCY, LOUIS-PHILIPPE. "Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach". *European Conference on Computer Vision*. 2020 2.

[AM19] AHUJA, C. and MORENCY, L. "Language2Pose: Natural Language Grounded Pose Forecasting". *2019 International Conference on 3D Vision (3DV)*. 2019 2.

[BCRM21] BHATTACHARYA, UTTARAN, CHILDS, ELIZABETH, REWKOWSKI, NICHOLAS, and MANOCHA, DINESH. "Speech2AffectiveGestures: Synthesizing Co-Speech Gestures with Generative Adversarial Affective Expression Learning". *Proceedings of the 29th ACM International Conference on Multimedia*. MM '21. 2021 2.

[BGSW18] BJORCK, NILS, GOMES, CARLA P, SELMAN, BART, and WEINBERGER, KILIAN Q. "Understanding Batch Normalization". *Advances in Neural Information Processing Systems*. Ed. by BENGIO, S., WALLACH, H., LAROCHELLE, H., et al. 2018 6.

[BHHF19] BRAHMBHATT, SAMARTH, HANDA, ANKUR, HAYS, JAMES, and FOX, DIETER. "Contactgrasp: Functional multi-finger grasp synthesis from contact". *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2019 3.

[BI05] BORST, CHRISTOPH W and INDUGULA, ARUN P. "Realistic virtual grasping". *IEEE Proceedings. VR 2005. Virtual Reality, 2005.* 2005 3.

[BRB*21] BHATTACHARYA, UTTARAN, REWKOWSKI, NICHOLAS, BANERJEE, ABHISHEK, et al. "Text2Gestures: A Transformer-Based Network for Generating Emotive Body Gestures for Virtual Agents". *2021 IEEE Conference on Virtual Reality and 3D User Interfaces (IEEE VR)*. 2021 2.

[BTT*20] BRAHMBHATT, SAMARTH, TANG, CHENGCHENG, TWIGG, CHRISTOPHER D., et al. "ContactPose: A Dataset of Grasps with Object Contact and Hand Pose". *The European Conference on Computer Vision (ECCV)*. 2020 3.

[BXP*22] BHATNAGAR, BHARAT LAL, XIE, XIANGHUI, PETROV, ILYA, et al. "BEHAVE: Dataset and Method for Tracking Human Object Interactions". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 3.

[CKA*22] CHRISTEN, SAMMY, KOCABAS, MUHAMMED, AKSAN, EMRE, et al. "D-Grasp: Physically Plausible Dynamic Grasp Synthesis for Hand-Object Interactions". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 2, 3.

[DFD22] DILLER, CHRISTIAN, FUNKHOUSER, THOMAS, and DAI, ANGELA. "Forecasting Characteristic 3D Poses of Human Actions". (2022) 3, 6.

[DKB*10] DETRY, RENAUD, KRAFT, DIRK, BUCH, ANDERS GLENT, et al. "Refining grasp affordance models by experience". *2010 IEEE International Conference on Robotics and Automation*. 2010 3.

[DSJ*21] DABRAL, RISHABH, SHIMADA, SOSHI, JAIN, ARJUN, et al. "Gravity-Aware Monocular 3D Human-Object Reconstruction". *International Conference on Computer Vision (ICCV)*. 2021 3.

[ES03] ELKOURA, GEORGE and SINGH, KARAN. "Handrix: animating the human hand". *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*. 2003 3.

[GCO*21] GHOSH, ANINDITA, CHEEMA, NOSHABA, OGUZ, CENNET, et al. "Synthesis of Compositional Animations From Textual Descriptions". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021 2, 6.

[GPM*20] GOODFELLOW, IAN, POUGET-ABADIE, JEAN, MIRZA, MEHDI, et al. "Generative adversarial networks". *Communications of the ACM* (2020) 3.

[GTT*21] GRADY, PATRICK, TANG, CHENGCHENG, TWIGG, CHRISTOPHER D., et al. "ContactOpt: Optimizing Contact To Improve Grasps". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021 3.

[GZW*20] GUO, CHUAN, ZUO, XINXIN, WANG, SEN, et al. "Action2motion: Conditioned generation of 3d human motions". *Proceedings of the 28th ACM International Conference on Multimedia*. 2020 2, 3, 6, 7.

[GZZ*22] GUO, CHUAN, ZOU, SHIHAO, ZUO, XINXIN, et al. "Generating Diverse and Natural 3D Human Motions From Text". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 2, 3.

[HCTB19] HASSAN, MOHAMED, CHOUTAS, VASILEIOS, TZIONAS, DIMITRIOS, and BLACK, MICHAEL J. "Resolving 3D Human Pose Ambiguities with 3D Scene Constraints". *International Conference on Computer Vision*. 2019 3.

[HCV*21] HASSAN, MOHAMED, CEYLAN, DUYGU, VILLEGAS, RUBEN, et al. "Stochastic Scene-Aware Motion Prediction". *Proceedings of the International Conference on Computer Vision 2021*. 2021 3.

[HES*22] HABIBIE, IKHSANUL, ELGHARIB, MOHAMED, SARKAR, KRIPASHINDU, et al. "A Motion Matching-based Framework for Controllable Gesture Synthesis from Speech". *SIGGRAPH '22 Conference Proceedings*. 2022 2.

[HL06] HSIAO, KAIJEN and LOZANO-PEREZ, TOMAS. "Imitation learning of whole-body grasps". *2006 IEEE/RSJ international conference on intelligent robots and systems*. 2006 3.

[HMLC09] HANSER, EVA, MC KEVITT, PAUL, LUNNEY, TOM, and CONDELL, JOAN. "Scenemaker: Intelligent multimodal visualisation of natural language scripts". *Irish Conference on Artificial Intelligence and Cognitive Science*. 2009 2.

[HRU*17] HEUSEL, MARTIN, RAMSAUER, HUBERT, UNTERTHINER, THOMAS, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". *Advances in neural information processing systems* (2017) 7.

[HTBT22] HUANG, YINGHAO, TAHERI, OMID, BLACK, MICHAEL J., and TZIONAS, DIMITRIOS. "InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction". *German Conference on Pattern Recognition (DAGM)*. 2022 2.

[HVT*19] HASSON, YANA, VAROL, GÜL, TZIONAS, DIMITRIOS, et al. "Learning joint reconstruction of hands and manipulated objects". *CVPR*. 2019 3.

[HXM*21] HABIBIE, IKHSANUL, XU, WEIPENG, MEHTA, DUSHYANT, et al. "Learning Speech-driven 3D Conversational Gestures from Video". *IVA*. 2021 2.

[JLWW21] JIANG, HANWEN, LIU, SHAOWEI, WANG, JIASHUN, and WANG, XIAOLONG. "Hand-object contact consistency reasoning for human grasps generation". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021 2, 3.

[KB14] KINGMA, DIEDERIK P and BA, JIMMY. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980* (2014) 6.

[KDCI10] KRUG, ROBERT, DIMITROV, DIMITAR, CHARUSTA, KRZYSZTOF, and ILIEV, BOYKO. "On the efficient computation of independent contact regions for force closure grasps". *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2010 3.

[KP15] KIM, JUN-SIK and PARK, JUNG-MIN. "Physics-based hand interaction with virtual objects". *2015 IEEE International Conference on Robotics and Automation (ICRA)*. 2015 3.

[KTS*21] KWON, TAEIN, TEKIN, BUGRA, STÜHMER, JAN, et al. "H2O: Two Hands Manipulating Objects for First Person Interaction Recognition". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021 3.

[KW14] KINGMA, DIEDERIK and WELLING, MAX. "Auto-Encoding Variational Bayes". 2014 3.

[KYZ*20] KARUNRATANAKUL, KORRAWE, YANG, JINLONG, ZHANG, YAN, et al. "Grasping Field: Learning Implicit Representations for Human Grasps". *8th International Conference on 3D Vision*. 2020 2, 3.

[LFP07] LI, YING, FU, JIAXIN L, and POLLARD, NANCY S. "Data-driven grasp synthesis using shape matching and task-based pruning". *IEEE Transactions on visualization and computer graphics* (2007) 3.

[LLW*21] LIU, ZHENGUANG, LYU, KEDI, WU, SHUANG, et al. "Aggregated multi-gans for controlled 3d human motion prediction". *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021 3.

[LPX*19] LIU, MIN, PAN, ZHERONG, XU, KAI, et al. "Generating grasp poses for a high-dof gripper using neural networks". *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2019 3.

[LRX*21] LI, WEIPENG, REN, BOYUAN, XU, HAOYUE, et al. "AutoDance: Music Driven Dance Generation". 2021 2.

[LWC*18] LIN, ANGELA S, WU, LEMENG, CORONA, RODOLFO, et al. "Generating animated videos of human activities from natural language descriptions". (2018) 2.

[LWM21] LIN, FANQING, WILHELM, CONNOR, and MARTINEZ, TONY. "Two-Hand Global 3D Pose Estimation Using Monocular RGB". *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2021 3.

[LYC*20] LI, JIAMAN, YIN, YIHANG, CHU, HANG, et al. "Learning to generate diverse dance motions with transformer". *arXiv preprint arXiv:2008.08171* (2020) 2.

[LYL*19] LEE, HSIN-YING, YANG, XIAODONG, LIU, MING-YU, et al. "Dancing to Music". *Advances in Neural Information Processing Systems*. Ed. by WALLACH, H., LAROCHELLE, H., BEYGELZIMER, A., et al. 2019 2.

[LZCvdP20] LING, HUNG YU, ZINNO, FABIO, CHENG, GEORGE, and van de PANNE, MICHIEL. "Character Controllers Using Motion VAEs". *ACM Trans. Graph.* (2020) 2, 3.

[MBR17] MARTINEZ, JULIETA, BLACK, MICHAEL J, and ROMERO, JAVIER. "On human motion prediction using recurrent neural networks". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017 3.

[PBV21] PETROVICH, MATHIS, BLACK, MICHAEL J., and VAROL, GÜL. "Action-Conditioned 3D Human Motion Synthesis with Transformer VAE". *International Conference on Computer Vision (ICCV)*. 2021 3, 6, 7.

[PBV22] PETROVICH, MATHIS, BLACK, MICHAEL J., and VAROL, GÜL. "TEMOS: Generating diverse human motions from textual descriptions". *European Conference on Computer Vision (ECCV)*. 2022 2, 3, 6, 7.

[PCG*19] PAVLAKOS, GEORGIOS, CHOUTAS, VASILEIOS, GHORBANI, NIMA, et al. "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image". *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019 4.

[PG16] PINTO, LERREL and GUPTA, ABHINAV. "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours". *2016 IEEE international conference on robotics and automation (ICRA)*. 2016 3.

[PGA18] PAVLLO, DARIO, GRANGIER, DAVID, and AULI, MICHAEL. "Quaternet: A quaternion-based recurrent model for human motion". *arXiv preprint arXiv:1805.06485* (2018) 3.

[PGC*17] PASZKE, ADAM, GROSS, SAM, CHINTALA, SOUMITH, et al. "Automatic Differentiation in PyTorch". *NeurIPS 2017 Workshop on Autodiff*. 2017 6.

[PRB*18] PUIG, XAVIER, RA, KEVIN, BOBEN, MARKO, et al. "VirtualHome: Simulating Household Activities via Programs". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018 2.

[RA15] REDMON, JOSEPH and ANGELOVA, ANELIA. "Real-time grasp detection using convolutional neural networks". *2015 IEEE international conference on robotics and automation (ICRA)*. 2015 3.

[RBH*21] REMPE, DAVIS, BIRDAL, TOLGA, HERTZMANN, AARON, et al. "HuMoR: 3D Human Motion Model for Robust Pose Estimation". *International Conference on Computer Vision (ICCV)*. 2021 2, 3.

[RKH*21] RADFORD, ALEC, KIM, JONG WOOK, HALLACY, CHRIS, et al. "Learning transferable visual models from natural language supervision". *International Conference on Machine Learning*. 2021 2, 4.

[RTB17] ROMERO, JAVIER, TZIONAS, DIMITRIOS, and BLACK, MICHAEL J. "Embodied Hands: Modeling and Capturing Hands and Bodies Together". *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* (2017) 3.

[SKK12] SEO, JUNGWON, KIM, SOONKYUM, and KUMAR, VIJAY. "Planar, bimanual, whole-arm grasping". *2012 IEEE International Conference on Robotics and Automation*. 2012 3.

[SSR07] SUNG, KELVIN, SHIRLEY, PETER, and ROSENBERG, BECKY. "Experiencing aspects of games programming in an introductory computer graphics class". *Proceedings of the 38th SIGCSE technical symposium on Computer science education*. 2007 2.

[SZKS19] STARKE, SEBASTIAN, ZHANG, HE, KOMURA, TAKU, and SAITO, JUN. "Neural State Machine for Character-Scene Interactions". *ACM Trans. Graph.* (2019) 3.

[TCBT22] TAHERI, OMID, CHOUTAS, VASSILEIOS, BLACK, MICHAEL J., and TZIONAS, DIMITRIOS. "GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping". *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 2, 3, 6, 9.

[TGBT20] TAHERI, OMID, GHORBANI, NIMA, BLACK, MICHAEL J., and TZIONAS, DIMITRIOS. "GRAB: A Dataset of Whole-Body Human Grasping of Objects". *European Conference on Computer Vision (ECCV)*. 2020 2, 3, 6, 9.

[WLK*22] WANG, XI, LI, GEN, KUO, YEN-LING, et al. "Reconstructing Action-Conditioned Human-Object Interactions Using Commonsense Knowledge Priors". *International Conference on 3D Vision (3DV)*. 2022 2.

[WWZ*22] WU, YAN, WANG, JIAHAO, ZHANG, YAN, et al. "SAGA: Stochastic Whole-Body Grasping with Contact". *Proceedings of the European Conference on Computer Vision (ECCV)*. 2022 2, 3.

[XJMS21] XU, XIANG, JOO, HANBYUL, MORI, GREG, and SAVVA, MANOLIS. "D3D-HOI: Dynamic 3D Human-Object Interactions from Videos". *arXiv preprint arXiv:2108.08420* (2021) 3.

[YK20] YUAN, YE and KITANI, KRIS. "Dlow: Diversifying latent flows for diverse human motion prediction". *European Conference on Computer Vision*. 2020 3.

[ZBL*19] ZHOU, YI, BARNES, CONNELLY, LU, JINGWAN, et al. "On the Continuity of Rotation Representations in Neural Networks". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 4.

[ZBS*22] ZHANG, XIAOHAN, BHATNAGAR, BHARAT LAL, STARKE, SEBASTIAN, et al. "COUCH: Towards Controllable Human-Chair Interactions". (2022) 3.

[ZHT*21] ZHANG, XIONG, HUANG, HONGSHENG, TAN, JIANCHAO, et al. "Hand Image Understanding via Deep Multi-Task Learning". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021 3.

[ZLM*19] ZHANG, XIONG, LI, QIANG, MO, HONG, et al. "End-to-end hand mesh recovery from a monocular rgb image". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019 3.

[ZPJ*20] ZHANG, JASON Y., PEPOSE, SAM, JOO, HANBYUL, et al. "Perceiving 3D Human-Object Spatial Arrangements from a Single Image in the Wild". *European Conference on Computer Vision (ECCV)*. 2020 3.

[ZYSK21] ZHANG, HE, YE, YUTING, SHIRATORI, TAKAAKI, and KOMURA, TAKU. "ManipNet: Neural Manipulation Synthesis with a Hand-Object Spatial Representation". *ACM Trans. Graph.* (2021) 2, 3.